2010
2011

FACULTY OF SCIENCES

Master of Statistics: Biostatistics

Masterproef

Socio-Demographic determinants of anemia among children aged 6-59 months in mainland Tanzania

Promotor :
dr. Philippe HALDERMANS

Promotor :
Prof. ALDEGUNDA KOMBA

Stephano Cosmas

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Biostatistics

Maastricht University

universiteit hasselt
UNIVERSITEIT VAN DE TOEKOMST

2010
2011

FACULTY OF SCIENCES

Master of Statistics: Biostatistics

Masterproef

Socio-Demographic determinants of anemia among children aged 6-59 months in mainland Tanzania

Promotor :
dr. Philippe HALDERMANS

Promotor :
Prof. ALDEGUNDA KOMBA

Stephano Cosmas

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Biostatistics

Maastricht University

universiteit hasselt
UNIVERSITEIT VAN DE TOEKOMST

Acknowledgments

My sincere appreciation goes to my Internal Supervisor, Dr. Philippe Hardermans for all his plentiful inputs, guidance, and suggestions that leads to successful completion of this project. I shall not forget to express my profound gratitude to my External Supervisor, Aldegunda Komba of NBS- National Bureau of Statistics, Tanzania who continually guided me through this project.

I recognize the financial support from the Flemish Interuniversity Council (VLIR) which has enabled me to be among those that benefited from the scholarship grant to pursue this valuable Masters program. I would like to acknowledge my lecturers at Center for Statistics for imparting part of their statistical knowledge on me. I also thank my family members, friends as well as the whole community of Uhasselt, for their friendship, love, advice, encouragement, and support during the past two years.

I thank God for giving me strength and life to successfully complete this report.

Stephano G. Cosmas

University of Hasselt, Belgium, September, 2011

List of Abbreviations

ALR- Alternating Logistic Regression

GEE- Generalized Estimating Equations

GLMM- Generalized Linear Mixed Model

Hb- Haemoglobin level

MAR- Missing at Random

MCAR- Missing Completely at Random

MDG- Millennium Development Goal

MI- Multiple Imputation

MNAR- Missing Not at Random

MOHSW- Ministry of Health and Social Welfare

NBS- National Bureau of Statistics

POM- Proportional Odds Model

QIC- Quasi under Independence model Criterion

SES- Socio-Economic Status

THMIS- Tanzania HIV/AIDS and Malaria Indicator Survey

WHO- World Health Organization

Abstract

Anaemia is a disease which seriously affects young children and pregnant women. Knowledge of disease clustering is important because it may provide insights into the etiology of disease and risk factors operating within different levels of the clusters. In this study we identify Socio-demographic determinants of anaemia among children aged 6-59 months in the 21 regions Mainland Tanzania. Children with higher probability of occurrence of these determinant factors would be inferred to be most likely to experience anaemia. To answer the objective of the research question, models that handle the complexities of correlated data were employed. The models used were; Generalized Estimating Equations(GEE), Alternating Logistic Regression(ALR), Proportional Odds Model(POM) and Generalized Linear Mixed Model (GLMM). Statistical findings revealed that the risk of being anaemic reduced with age while boys showed higher risk. Children in rural were found to be less likely to be anaemic. Children in large households were found to be more having a higher risk of anaemia. Similarly, malaria occurrence was strongly correlated to anaemia.

Keywords: Anaemia; Clustered data analysis; GEE; ALR; POM; GLMM; Missingness.

Table of Contents

1. Introduction

1.1 Background

Anaemia is a disease which seriously affects young children and pregnant women. Anaemia is defined as a reduction of red blood cells or haemoglobin (Hb) concentration below the normal level. The etiologies of anaemia are often multi-factorial, with different causes interacting in a vicious cycle of nutritional deficiencies, infections and inherited red blood cell disorders. However, the relative contributions of these factors remain unclear (MOHSW, 2006). The clinical features of anaemia vary from mild to severe, according to the hemoglobin levels (Hb) present and possibly patient's state of immunity. Severe anaemia needs medical emergency. Delay in effective diagnosis and provision of appropriate medical emergency may lead to serious complications and even death. It is estimated that, more than 100 million African children are anaemic (Schellenberg *et al.*, 2003).

In September, 2000, 189 United Nations member States including Tanzania endorsed the eight Millennium Development Goals (MDGs) to be accomplished by 2015. The fourth Millennium Development Goal (MDG) targets to reduce child mortality rates by two-third between 1990 and 2015 (U.N, 2009). Reliable information about the rates and progress in child mortality is essential machinery tools for monitoring and assessment of trends towards Millennium Development Goals. If not effectively addressed, anaemia may be an impediment to realization of the Millennium Development Goal number four on child survival.

Although the immediate causes of anaemia are well documented, there is scanty information about determinants of anaemia across Regions, at household and individual levels in Tanzania. The identification of socio-demographic factors responsible for anaemia can be used to guide strategic planning and evaluation of programmes, and to complement and calibrate estimates obtained from other sources for addressing this problem in Mainland Tanzania.

Cross-sectional surveys can serve as tools to collect subset of possible risk factors which can be used to establish association with the response variable of interest. This association in turn enables the researcher to establish evidence based in policy planning and resource allocations and in evaluating progress in policy implementations. This can be achieved by applying proper statistical methods in measuring the outcome indicators as well as quantifying the impact of determinant factors, interventions coverage and other possible indicators are essential in the success of any policy.

1.2 Objectives of the Study

The objective of this project is to identify Socio-demographic determinants of anaemia among children aged 6-59 months in Mainland Tanzania.

This report is organized as follows; The study implementation, study design and data set used are introduced in section 2. Section 3 is dedicated on the statistical methods used for analyzing the data. In section 4, the results from the applied statistical methods are presented. The discussions and conclusions, and recommendations are presented in sections 5 and 6 respectively. The last part will have References and an Appendix.

2. Study Implementation

The 2007-08 Tanzania HIV/AIDS and Malaria Indicator Survey (THMIS) is the second population-based, comprehensive survey on HIV/AIDS to be carried out in Tanzania. In addition to the collection of information on household characteristics as well as health related issues during the survey interview, the THMIS also included anaemia and malaria testing for children aged 6-59 months and HIV testing for adults age 15-49.

The 2007-08 THMIS was implemented by the National Bureau of Statistics (NBS) in collaboration with the Office of the Chief Government Statistician—Zanzibar. The survey was commissioned by the Tanzania Commission for AIDS (TACAIDS) and the Zanzibar AIDS Commission (ZAC). Macro International, Inc provided technical assistance to the project through the MEASURE DHS, a USAID-funded project. Other agencies and organizations facilitated the successful implementation of the survey through technical or donor support include National AIDS Control Programme, National Malaria Control Programme, and Muhimbili University College of Health Sciences (MUCHS).

2.1 Study Design

The 2007-08 THMIS utilized the sampling frame developed by NBS after the 2002 Population and Housing Census (PHC). The sample excluded nomadic and institutional populations, such as persons staying in hotels, barracks, and prisons. The THMIS utilized a two-stage sample design. The first stage involved selecting sample points (clusters) consisting of enumeration areas demarcated for the 2002 PHC. A total of 475 clusters were selected. The sample was designed to allow estimates of key indicators for each of Tanzania's 26 regions. It is worth noting that Tanzania is formed out of the two sovereign states namely Mainland Tanzania (21 regions) and Zanzibar (5 regions).

On the Mainland Tanzania, 25 sample points were selected in Dar es Salaam and 18 in each of the other 20 regions. In Zanzibar, 18 sample points were selected in each of the five regions, for a total of 90 sample points. A household listing operation was undertaken in all the selected areas prior to the fieldwork. From these lists, households to be included in the survey were selected. The second stage of selection involved the systematic sampling of households from these lists. Approximately 16 households were selected from each sampling point in Dar es Salaam, and 18 households per sampling point were selected in other regions. In Zanzibar, approximately 18 households were selected from each sampling point in Unguja,

and 36 households were selected in Pemba to allow reliable estimates of HIV prevalence for each island group.

2.2 Questionnaires

The 2007-08 THMIS drawn in two questionnaires: a Household Questionnaire and an Individual Questionnaire. The Household Questionnaire was used to collect information on the characteristics of each person listed (usual members and visitors inclusive), including age, sex, education and relationship to the head of the household, questions on ownership and use of mosquito bednets. In addition, the Household Questionnaire was also used to collect non-income proxy indicators about the household's dwelling unit, ownership of various durable goods and land, and household food insecurity. The main purpose of the Household Questionnaire was to identify women and men who are eligible for the individual interview and children 6-59 months for anaemia and malaria testing. On the other hand, the Individual Questionnaire was used to collect information from women and men aged 15-49 years, covering different health related topics.

2.3 Platform Anaemia Testing

Haemoglobin measurement is a decisive method for individual anaemia screening. Haemoglobin measurement in population-based surveys like the THMIS provides a prospect to estimate the prevalence of anaemia and to examine the socio-economic, residential, and demographic differentials in anaemia levels in the population surveyed. Such information is useful in evaluating and developing health-intervention programs (such as iron fortification) to prevent iron-deficiency anaemia among young children.

In the THMIS, haemoglobin measurement for anaemia testing was performed in the field by a team member. Prior to anaemia testing, consent was obtained from the parent or guardian. For haemoglobin measurement, capillary blood was usually taken from a finger of the eligible children for whom consent had been obtained. A single-use, sterile lancet was used for this purpose. In cases where a child was very thin, a heel prick was used to obtain the blood sample. The concentration of haemoglobin (g/dl) in the blood was measured using the HemoCue system. This system consists of a battery-operated photometer and a disposable microcuvette, coated with a dried reagent that serves as the blood-collection device. The results of the anaemia were recorded on the household questionnaire. Moreover, these results were also reported to the parent or other responsible adult at the time of the testing, and

parents of children with low levels of haemoglobin were advised to take the child to health facilities for further evaluation and management.

WHO (2011) provides information about the use of haemoglobin concentration for diagnosing anaemia. Levels of anaemia for children under five years of age were classified as severe($Hb\ level < 7g/dl$), moderate($7g/dl \leq Hb\ level \leq 9.9g/dl$), mild ($10g/dl \leq Hb\ level \leq 10.9g/dl$) or non-anaemia ($Hb\ level \geq 11\ g/dl$) according to criteria developed by the World Health Organization.

2.4 Data Description

The proposed research will utilize the data from the 2007-08 Tanzania HIV/AIDS and Malaria Indicator Survey (2007/08 THMIS) in which Malaria and Anaemia modules among children aged 6-59 months were included as special modules. The recorded variables in the data included Region name, cluster identification number (CLID), household identification number (HhID), Child identification number(CASEID), age of child (in month), gender categorized as "1" to represent males and "2" to represent females, type of residence (1= urban, 2=rural), distance to the nearest health facility (km), household's size(Hh-size), duration of breast feeding (in month), mother's highest educational level (1=No education, 2=Primary Incomplete, 3=Primary Complete, 4=Secondary+), socio-economic status (1=Poorest, 2=Poorer, 3=Middle, 4=Richer, 5=Richest), ownership and use of bed net as well as health related issues including malaria and anaemia status of 4372 eligible children from the 375 sample points (clusters) in the 21 regions of Mainland Tanzania.

2.5 The Response Variable

Often in many epidemiologic, biomedical and related fields of studies, the outcome of interest is a binary variable such as anaemic versus non anaemic. In such circumstances, it is possible to employ plausible statistical tools for estimating the magnitude of the association between the response variable of interest as a function of independent predictor variables. The association provides information about the risk of developing an outcome. Unlike continuous outcome variables, binary or multi-categorical outcome variables are often prone to loss of information. However, one very practical advantage of using statistical methods for binary or multi-categorical response over statistical methods for continuous response variable in epidemiologic research is that parameter estimates of the possible risk factors can be directly converted to an odds ratio, which is interpretable. Additionally, the use of binary

outcome for defining anaemia and its severity at the population level, as well as the chronology of their founding allows the identification of populations at greatest risk of anaemia and priority areas for action, especially when resources are inadequate.

In view of the above, the haemoglobin level was first dichotomized based on the cut-off points as described in subsection 2.3 leading to the binary response;

$$Response = \begin{cases} 1 & if\ hb\ level < 11\ g/dl \\ 0 & if\ \ hb\ level \geq 11\ g/dl \end{cases} \quad (1)$$

The response was further categorized into the quartiles, leading to multinomial responses;

$$Response = \begin{cases} 1 & if \quad hb\ level < 7g/dl \\ 2 & if \quad 7g/dl\ \leq hb\ level \leq 9.9g/dl \\ 3 & if \quad 10g/dl\ \leq hb\ level \leq 10.9g/dl \\ 4 & if \quad hb\ level \geq 11\ g/dl \end{cases} \quad (2)$$

3. Statistical Methodology

3.1 Exploratory Data Analysis (EDA)

The exploratory data analysis was performed using simple descriptive statistics as well as graphical techniques to explore the association between the response variable and covariates of interest. This was intended to provide an insight about the data structure and the most plausible aspects or implications to be considered during statistical analysis.

3.1.2 Data Missingness

Missing data is an everywhere problem in the analysis of survey data. Sources of item-missing data in surveys may arise due to a number of reasons. For instance, respondents' participation in most surveys is a voluntary decision. The researcher has no influence to withhold a respondent who wants to withdraw from the course of the interview, mostly this happens when a respondent does not wish to answer sensitive or difficult questions. Item-missing data can also occur when a phase of a survey data collection activity, such as a blood draw from a respondent, may require special consent of the subject. Failure to obtain cooperation can lead to missing data for variables from that entire phase of the study. In the THMIS for example, about 5% of eligible children in the Mainland Tanzania approached for anaemia testing refused to take part in this exercise.

There are generally three recognized missing data mechanisms (Little and Rubin, 1987). A non-response process is said to be missing completely at random (MCAR) if the probability that a respondent does not report an item value is completely independent of the true underlying values of all of the observed and unobserved variables. Data are missing at random (MAR) if, conditional on the observed data, the missingness is independent of the unobserved values of the variables in the survey. Finally, data are not missing at random (MNAR) if missingness is neither MCAR nor MAR.

For categorical data the inferences with ALR are valid only under the strong assumption that the data are missing completely at random (MCAR) (Molenberghs and Verbeke, 2006). To relax on the MCAR assumption and allow a more flexible assumption by assuming that data is missing at random (MAR), one can use Multiple Imputation (Little and Rubin, 1987) to deal with missing data. With it, each missing value is replaced by two or more imputed values in order to represent the uncertainty about which value to impute, whose mean and variance can be determined to estimate the efficiency of the imputation procedure. When there is a combination of missing covariates and missing outcomes, as was the case for

THMIS, multiple imputation can be a useful tool to deal with such a case. In this report, multiple imputation was used to account for data missingness.

3.2 Marginal Models

In clustered data, observations are usually taken from the same unit, and thus this information forms a cluster of correlated observations. For instance, in the THMIS the dependent variable (anaemia status) was measured once for each eligible child (the unit of analysis) in the selected household, and the units of analysis are grouped into, or nested within households (cluster of units), which are in turn nested within clusters(cluster of clusters) which were selected from each region. Knowledge of disease clustering is important because it may provide insights into the etiology of disease and risk factors operating within different levels of clusters. Proper analysis of clustered data is required in modeling the association between the response variable and the given set of covariates.

Marginal models are among the most statistical models widely used to model clustered or repeated data. In marginal models, the primary scientific objective is to analyze the population-averaged effects of the given factors in the study on the binary response variable of interest. This means that the covariates are directly related to the marginal expectations.

The marginal models fitted in this report included the Generalized Estimating Equations (GEE), Proportional Odds Model (POM) and Alternating Logistic Regression (ALR).

3.2.1 Generalized Estimating Equations (GEE)

For binary data, one can use a GEE approach to account for the correlation between responses of interest for subjects from the same cluster (Diggle *et al.*, 1994). GEE is non-likelihood method that uses correlation to capture the association within the clusters or subjects in terms of marginal correlations (Molenberghs and Verbeke, 2006). For clustered and repeated data, Liang and Zeger (1986) proposed GEE which require only the correct specification of the univariate marginal distributions provided one is willing to adopt "working" assumptions about the correlation structure. The "working" assumptions as proposed by Liang and Zeger (1986) include independence, unstructured, exchangeable and auto-regressive AR(1). A detailed discussion of these assumptions can be found in Molenberghs and Verbeke (2006).

Let $Y_i = (y_{i1}, \ldots\ldots, y_{im})'$ be the response values on cluster $\boldsymbol{i}$ and let $X_i = (x_{i1}' \ldots, .., x_{im}')'$contains covariates associated with the response values. To model the relation

between the response and covariates, one can use a regression model similar to the generalized linear models given by

$$E[Y_i] = \mu_i, \;\; \eta(\mu_i) = \boldsymbol{X_i\beta}, \qquad i = 1,..,n \qquad (3)$$

where η is the specific link function and $\boldsymbol{\beta} = (\beta_1, \dots, .., \beta_m)'$ is a vector of unknown regression coefficients to be estimated. The GEE approach estimates β by solving estimating equations which consist of the working covariance matrix of Y_i (Liang and Zeger, 1986)

The score equation that is used to estimate the marginal regression parameters while accounting for the correlation structure is given by

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu_i}}{\partial \boldsymbol{\beta'}} \left(\boldsymbol{A_i^{\frac{1}{2}} R_i A_i^{\frac{1}{2}}} \right)^{-1} (\boldsymbol{y_i - \mu_i}) = \mathbf{0}, \qquad (4)$$

Where $\boldsymbol{R_i}$ is the working correlation matrix, and the covariance matrix $\boldsymbol{V}_i$ has been decomposed into $\boldsymbol{A_i^{\frac{1}{2}} R_i A_i^{\frac{1}{2}}}$ with $\boldsymbol{A}_i$ the matrix with the marginal variances on the main diagonal and zeros elsewhere, $\boldsymbol{y_i}$ multivariate vector of normal response variables with mean vector $\boldsymbol{\mu_i}$ ie $Y_i \sim N(X_i\beta, V_i)$. An advantage of the GEE approach is that it yields a consistent estimator of $\boldsymbol{\beta}$, even when the working correlation matrix R_i is misspecified. However, severe misspecification may seriously affect the efficiency of the GEE estimators (Pan, 2001).

3.2.2 Alternating Logistic Regression (ALR)

This method is very similar to the GEE, in that they are both quasi-likelihood based and they account for dependency in the data. But unlike GEE which measures the association among the observed data through the correlation structure, Alternating logistic regression (ALR) measures this association using the odds ratio, which is interpretable and more applicable for binary data. ALR extends beyond classical GEE in the sense that precision estimates follow for both the regression parameters $\boldsymbol{\beta}$ and the association parameters ($\boldsymbol{\alpha}$). Moreover with ALR inferences can be made, not only about marginal parameters but about pairwise associations between subjects as well (Molenberghs and Verbeke, 2006). The odds ratio modeled in the ALR can be expressed as:

$$OR(Y_{ij}, Y_{ik}) = \frac{P(Y_{ij} = 1, Y_{ik} = 1)P(Y_{ij} = 0, Y_{ik} = 0)}{P(Y_{ij} = 1, Y_{ik} = 0)P(Y_{ij} = 0, Y_{ik} = 1)} \qquad (5)$$

$$i = 1,..,n; \qquad j,k = 1,2, \dots \dots \dots \dots \dots, m$$

where Y_{ij} and Y_{ik} represent the response values for subjects j and k respectively from the same cluster $\boldsymbol{i}$.

In this report, ALR models were used to account for clustering of eligible children within households as well as clustering of households within sample points (clusters) which were selected from regions of Mainland Tanzania. These fitted models are called two level and three level models respectively.

3.2.3 Proportional Odds Model (POM)

Proportional Odds Models (POM) are part of the broad family of Multi-categorical Response models. Multi-categorical Response models are used when the response variable has more than two categories. Since the anaemia level was further categorized into four levels as described in subsection 2.1.4 and presented by equation (2), we found this type of modeling appropriate. For multinomial data, that is, multicategory responses, a multinomial observation with K categories is a vector of $K-1$ indicators of which the k^{th} is 1 when the observation falls in category *k* and 0 otherwise (Agresti, 2002).

POM considers ordinality of multi-level category response variable into account. The usual approach for modeling such type of response data is to use logits of cumulative probabilities given by:

$$Logit\left[P\left(Y_{ij} \leq k \middle| x_{ij}\right)\right] = \alpha_k + \boldsymbol{X_{ij}'\beta}\ , \qquad for\ \ k = 1, \dots \dots, K-1 \qquad (\mathbf{6})$$

where Y_{ij} is the multinomial response with K categories, α_k are the intercepts for each logit, $\boldsymbol{X_{ij}'}$ and $\boldsymbol{\beta}$ are vectors of explanatory variables and slope parameters respectively. An important assumption for the PO model is assuming a common slope for each logit. Therefore, before making inferences based on this model, this assumption should be tested first. A restriction in implementation is that softwares such as SAS can only allow independence correlation structure with the multinomial distribution. Thus, in this part of the analysis, an independence correlation structure was assumed.

3.3 Subject-Specific Models

When interest is in the marginal population-averaged models to describe the relationships of the covariates to the dependent variable for an entire population, marginal models as discussed in section 3.2 are preferred. However, in most biomedical and biological data problems, interest often lies in understanding the response of individual patient characteristics

and how this response is influenced by a given set of possible covariates (Myers *et al.,2010)*. This proves even to be essential when individual interventions may be necessary. Subject specific models are useful in such cases. Subject specific models differ from the marginal models by inclusion of parameters that are specific to clusters or subjects within a population. Consequently, random effects are directly used in modeling the random variation in the dependent variable at different levels of the data.

3.3.1 Generalized Linear Mixed Model (GLMM)

Generalized Linear Mixed models (GLMM) are part of the broad family of subject-specific models. In general using the same notations as for GEE, generalized linear mixed models are defined as (Molenberghs and Verbeke, 2006);

$$E[Y_{ij}|\boldsymbol{b}_i] = \mu_i, \quad \eta(\mu_i) = \boldsymbol{X'}_{ij}\boldsymbol{\beta} + \boldsymbol{Z'}_{ij}\boldsymbol{b}_i, \qquad i = 1,..,n \qquad (7)$$

where μ_i is the mean response vector conditional on the random effects $\boldsymbol{b}_i$ for subjects in cluster $\boldsymbol{i}$ and $\boldsymbol{Z}_i$ is the design matrix for the random effects. The random effects $\boldsymbol{b_i}$ are assumed to follow a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{D}$.

Equation (7) can be extended to account for the three-level random-effects leading to the following expression

$$Logit\left[P\left(Y_{ijk} = 1 \middle| v_i, u_{ij}\right)\right] = \boldsymbol{X}_{\boldsymbol{ij}}^{'}\boldsymbol{\beta} + v_i + u_{ij} \qquad (8)$$

Where k and j are indices for child and household respectively.

The random effects $v_i \; and \; u_{ij}$ are assumed to be independently distributed as $N(\mathbf{0}, \sigma_{\boldsymbol{c}}^{\boldsymbol{2}})$ and $N(\mathbf{0}, \sigma_{\boldsymbol{w}}^{\boldsymbol{2}})$, respectively. Variance components, $\sigma_{\boldsymbol{c}}^{\boldsymbol{2}}$ and $\sigma_{\boldsymbol{w}}^{\boldsymbol{2}}$, address between-cluster and within-cluster variation, respectively.

4. Results

4.1 Exploratory Data Analysis

A total of 4372 children from the 375 selected clusters in Mainland Tanzania were eligible for anaemia testing. Of the eligible children approached for anaemia testing, 236(5%) refused to take part. Approximately equal number of eligible males and females took part in the anaemia testing, 2078(50.24%) and 2058(49.79%) respectively. Table 1 presents basic descriptive information that summarizes the associations between the determinant factors and anaemia. Overall, 68.28% of children had anaemia (Hb <11 g/dl), with males and females having approximately 3 percent of severe anaemia. On the other hand, 85% of eligible children lived in rural areas, 68% of which had either of the severity level of illness. Approximately 4 percent of children in urban areas had severe anaemia. Additionally, information on the socio-economic status and size of the household are important because they provide information about the welfare of the household. About 17% and 13% of eligible children lived in households belonging to the fourth (richer) and fifth (richest) wealth indexes respectively, whereas the rest of the children who lived in households in the first three levels (poorest thru middle) of socio-economic status were almost equally distributed. In the former, children experienced 2 percent and 4 percent respectively of severe anaemia, meanwhile in the latter children experienced 3 percent, 2 percent and 3 percent respectively of severe anaemia. According to (TDHS, 2011), the mean household size is 5. Sixty three percent of eligible children lived in households with size greater than 5, approximately 3 percent of which experienced severe anaemia. Furthermore, over 6 in 10 eligible children lived in households located within 5km from the nearest health facility and about 12% their households were located more than 10km from the nearest health facility. Children in the former group experienced about 3 percent of severe anaemia, while those in the latter experienced about 4 percent. Overall, 62% of households did not own a bednet. Among children in household with mosquito net (treated or untreated), 12% slept under an instecticide-treated mosquito net (ITN) the night before the survey. Approximately 4 percent of children in the treated or untreated mosquito net group experienced severe anaemia, whereas 2 percent of children in the no mosquito net group experienced severe anaemia. Moreover, it is worth noting that, one in four Tanzanian women is un educated, three percent of children aged 6-59 months whose mothers have little or no education, had severe anaemia. Exclusive breastfeeding is recommended during the first 6 months of a child's life because it limits exposure to disease agents as well as providing all of the nutrients that a child needs.

Ninety eight percent of children were exclusively breastfed, approximately 3 percent of which had severe anaemia.

Table 1: Descriptive statistics and determinant factors of anaemia in children under five in Mainland-Tanzania

Variable		Number of Children	Severity of illness (%)		
			Mild	Moderate	Severe
Child gender:					
	Male	2078(50.24)	533(25.65)	851(40.95)	66(3.18)
	Female	2058(49.76)	557(27.06)	765(37.17)	66(3.21)
Type of residence:					
	Urban	621(15.1)	163(26.25)	259(41.71)	24(3.86)
	Rural	3515(84.9)	927(26.37)	1357(38.61)	94(2.67)
Socio-economic status (SES):					
	Poorest	1028(25.11)	274(26.65)	407(9.88)	34(3.31)
	Poorer	903(22.06)	263(29.13)	358(39.65)	20(2.21)
	Middle	907(22.15)	224(24.70)	356(39.25)	28(3.09)
	Richer	725(17.71)	186(25.66)	292(40.28)	15(2.07)
	Richest	531(12.97)	143(26.93)	203(38.23)	21(3.95)
Household size(Hh-size):					
	≤5	1512(36.56)	373(24.67)	601(39.75)	41(2.71)
	>5	2624(63.44)	717(27.32)	1015(38.68)	77(2.93)
Distance to health facility:					
	<5Km	2643(64.04)	744(28.25)	1114(42.15)	83(3.14)
	5-10Km	971(23.53)	321(33.06)	406(41.81)	39(4.02)
	>10Km	513(12.43)	135(26.32)	197(38.40)	22(4.29)
Mosquito net:					
	Treated	481(11.86)	125(25.56)	216(44.91)	19(3.95)
	Untreated	1044(25.75)	257(24.62)	451(43.20)	40(3.83)
	None	2530(62.39)	708(27.98)	949(37.51)	59(2.33)
Mother's education:					
	No education	1018(24.71)	273(26.82)	415(40.77)	33(3.24)
	Primary incomplete	2950(71.62)	775(26.27)	1135(38.47)	81(2.75)
	Primary complete	133(3.23)	38(28.57)	58(43.61)	4(3.01)
	Secondary+	18(0.44)	4(22.22)	8(44.44)	0(0)
Duration of breast feeding:					
	≤ 6 month	67(1.62)	17(25.37)	23(34.33)	3(4.48)
	>6 month	4069(98.38)	1073(26.37)	1593(39.15)	115(2.83)

The graphs of proportions of children with anaemia by the covariates were made to have a hint of how anaemia is associated to these covariates. About 90% and 80% of male and female children in the age group (6-11) months respectively, experienced anaemia (Hb level<11 g/dl) (Figure 1, panel a). Generally, the proportions for males with anaemia are observed to be higher at all age group points compared to their female counterparts, these

proportions seem to be relatively decreasing over higher age groups with children in the age group 48-59 months having the lowest cases. Figure 1, panel b shows the proportions of children with anaemia (Hb level<11 g/dl) across the 21 regions of Mainland Tanzania. A variation in anaemia proportions among children could be hinted across the 21 regions in Mainland Tanzania. While 80 percent of children in Ruvuma and Lindi regions are anaemic, less than 4 in 10 children are anaemic in Manyara region. Furthermore, the proportions of children with anaemia were observed to be high at all age group points for the children who had malaria at the time of the anaemia testing as compared to those who did not; an intimation of high association between anaemia and malaria. Figure 1, panel c depicts the association of anaemia and malaria over age groups. Moreover, one can assess the association of anaemia and malaria over age groups by further extending the binary response variable to multinomial responses (Figure 1, panel d). It can be observed that proportions of mild anaemic children were slightly increasing over age groups, whereas the reverse is observed for moderate anaemic group. Between (36-47) and (48-59) age groups there is a possibility of interaction between the moderate and mild levels of anaemia. On the other hand the proportions of severely anaemic children group remain relatively constant below the other two groups of children with anaemia over all age groups.

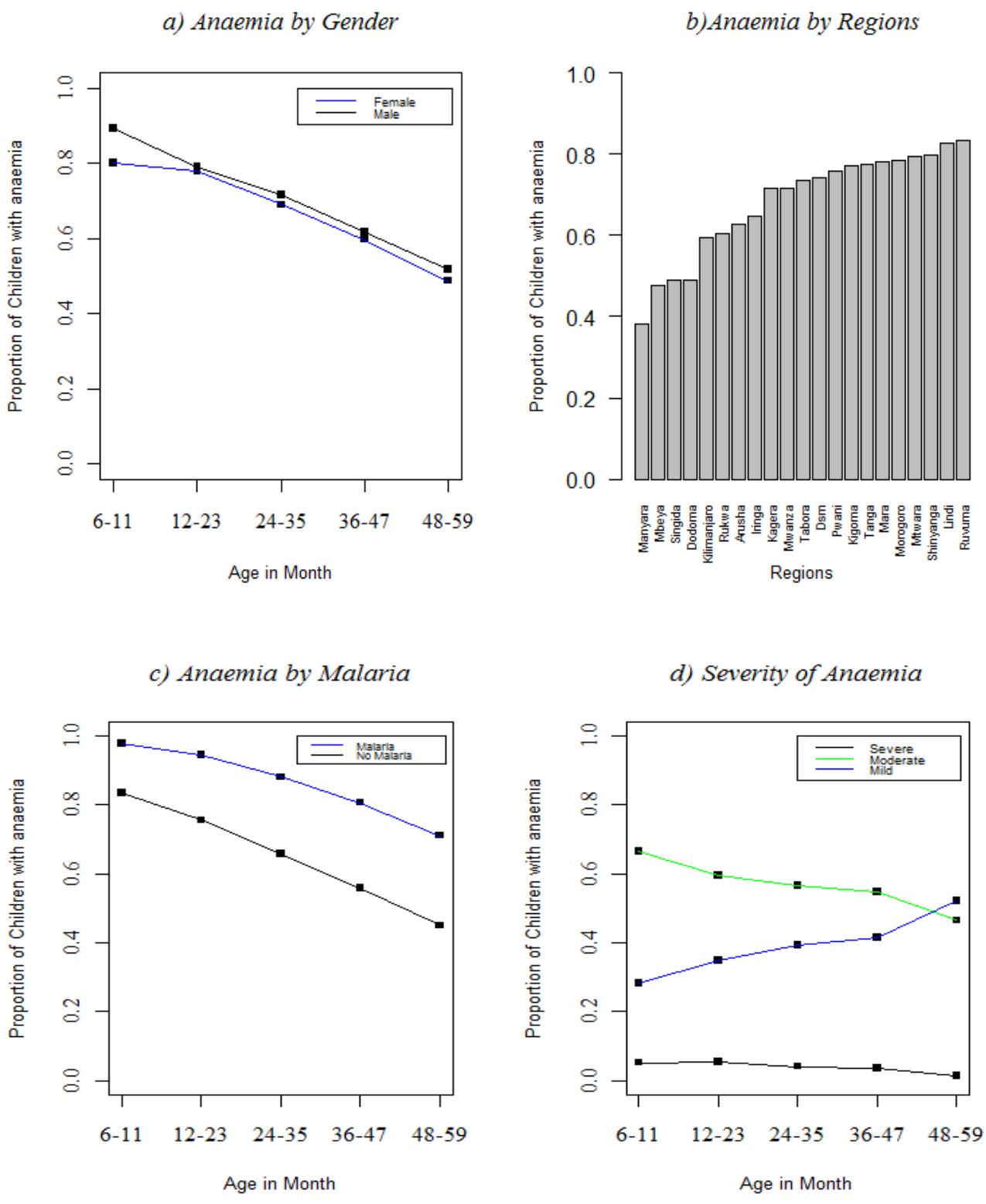

Figure 1: *Proportion of children with anaemia by gender, region, malaria and levels of anaemia over age groups, panels (a-d) respectively.*

Moreover, one can use correlation to quantifying the level of association between the response variable and the covariates as well as the associations between the covariates themselves. Based on the spearman coefficient, a strong correlation (0.854) was found between Socio-economic status and household size. On the other hand, the weakest correlation (-0.0077) was between age and household size.

4.2 Statistical Analysis

Model Building

Model selection is a vital part of data analysis strategy which leads to a search of "best" model. With this, we mean selecting the best subset of the covariates from the available covariates in the data. Subsections 4.2.1 and 4.2.2 explain how model selection was done in this report.

4.2.1 Marginal Model

In a situation, when the likelihood function cannot be fully specified, e.g., as in the GEE case, the Akaike's Information Criterion (AIC) cannot be directly applied for model selection procedures. Instead, one can use the modified Akaike's Information Criterion (QIC) which is based on the quasi-likelihood function (McCullagh and Nelder, 1989). Additionally, QIC is also applicable in selecting a working correlation structure under GEE settings.

Firstly, under the GEE, model building strategy started by fitting a model containing all possible covariates in the data and two-way interaction terms. This was done by considering three different working correlation assumptions (exchangeable, independence and unstructured). In order to select the important factors related to anaemia, the backward selection procedure was used. The strategy is called backward because we are working backward from our largest starting model to a smaller final model. In this case the procedure was used to remove interactions as well as main effects with non-significant p-values($p - value > 0.05$) . This means that the variables that did not contribute to the model based on the highest p-value were eliminated sequentially and each time a new model with the remaining covariates was refitted, until we remained with covariates necessary for answering our research question. None of the interaction terms were found to be significant. It turned out that the model with age, malaria, residence type, gender and household size as covariates was found to be the most parsimonious model. This model is considered as the best model

because the corresponding QIC value is the smallest and this is true for all three correlation structures. Finally, as a customary, comparison of empirical and model based standard errors for the parameter estimates obtained based on the three working correlation assumptions (in this study exchangeable, independence and unstructured) was performed. Exchangeable working correlation assumption was found plausible since the two standard errors were closest. Our proposed final model is given as;

$$\text{logit}(\pi) = \beta_0 + \beta_1 \text{age} + \beta_2 \text{Urban} + \beta_3 \text{Male} + \beta_4 \text{Malaria} + \beta_5 \text{Hh}_{\text{size}} \qquad (9)$$

Additionally, using the five selected covariates, a two-level ALR model which provides information about pairwise association of observations between two different individuals within the same household was fitted. Later this model was extended to a three-level ALR model to accommodate the association of pairs of responses from two different households within the same cluster. Based on the QIC values of 4172.13 and 4119.332 for the two and three-levels ALR models respectively, it was concluded that the three-level ALR model was a better model in explaining the population-averaged association between anaemia and the selected predictor variables. Thus, our interpretation will rely on the three level ALR model.

Table 2 presents parameter estimates and their corresponding empirically corrected standard errors along side the p-values from GEE and three-level ALR. Each parameter β_j reflects the effect of factor $X_j$on the log odds of having anaemia, statistically controlling for all other factors. Overall, parameter estimates under ALR are slightly less than or equal to those of GEE. The slight differences in parameter estimates could be attributed to the fact that ALR takes the association into account, whereas GEE treats the association as a nuisance parameter.

The ALR analysis with three-level (the right half of Table 2) suggests that malaria is significantly related to anaemia, it was observed that children who had malaria at the time of the anaemia testing had $exp\,(1.103) = 3.01$ times higher odds of anaemia than children who did not have malaria. Moreover, children who lived in urban areas had $exp\,(0.298) = 1.35$ times higher odds of anaemia than their rural children counterparts. Males aged 6 to 59 months had $exp\,(0.232) = 1.26$ times higher odds of anaemia than their female counterparts. Age (in month) was found to be negatively associated with Anaemia. This implies that adjusting for other predictor variables in the model, an additional unit increase in age (month) reduces the odds of being anaemic by 4 percent. The situation is different for size of the household, whose estimated odds ratio is $exp\,(0.026) = 1.03$. This implies that

an additional unit increase in household size (number of members) increases the odds of being anaemic for children aged 6-54 months by 3 percent.

Table 2: Parameter estimates (empirically corrected standard errors) from GEE and three-level ALR

			GEE		ALR	
Effect	Level	Parameter	Estimate (s.e)	p-value	Estimate (s.e)	p-value
Intercept		β_0	1.488(0.126)	<.0001	1.576(0.134)	<.0001
Age		β_1	-0.039(0.003)	<.0001	-0.039(0.003)	<.0001
Residence	Urban	β_2	0.330(0.111)	0.0168	0.298(0.111)	0.0072
Gender	Male	β_3	0.222(0.078)	0.004	0.232(0.074)	0.0017
Malaria		β_4	1.240(0.129)	<.0001	1.103(0.134)	<.0001
Hh Size		β_5	0.036(0.011)	0.0007	0.026(0.0104)	0.0137
Alpha 1		α_0			0.938(0.171)	<.0001
Alpha 2		α_1			0.347(0.074)	<.0001

Note: Hh-Size = Household size

Table 2 also provides the estimated constant log odds ratios (alpha 1 and 2). These two measures provide information about association between individuals within a household and the association between individuals from different households within same cluster respectively. This means that, the estimated pairwise odds ratio relating two responses from the same household was 2.55 (95% CI: 1.83, 3.57). The estimated pairwise odds ratio relating pairs of responses from two different households within the same cluster was 1.41 (95% CI: 1.22, 1.64). These associations were found to be highly significant (p-value <0.0001).

The Proportional Odds Model

Marginal ordinal logistic regression model follows from an extension of the GEE model to accommodate ordering of the levels in anaemia. Referring to equation 2, anaemia status was categorized using a four-point ordinal scale: 1 = severe, 2 = moderate, 3 = mild, 4 = Non anaemic.

Proportional odds assumption: Before inference could be made, the assumption of common slope for proportional odds model had to be tested first. The score test for the null hypothesis of constant slope gave a chi- square=6.1980, df=10, p-value=0.7984. We therefore fail to reject the null hypothesis (i.e., that the PO assumption holds) and can proceed to examine the model output. Table 3 summarizes the results of the fitted marginal ordinal logistic regression model. Considering the estimated cumulative odds ratios and 95% confidence intervals for the odds ratios for the selected predictor variables, the results suggest that age, malaria, type of residence, gender and household size are significant predictors in the cumulative logit model for anaemia as their 95% confidence intervals for the odds ratios do not include 1.

Table 3: Estimated Cumulative Logit Regression Model

				Cum. Odds Ratio
Effect	Level	Parameter	Estimate(s.e)	OR [95% CI]
Intercept 1		α_1	-3.209(0.153)	0.04 [0.03 , 0.05]
Intercept 2		α_2	0.261(0.104)	1.30 [1.06 , 1.59]
Intercept 3		α_3	1.476(0.107)	4.38 [3.55 , 5.39]
Age		β_1	-0.039(0.002)	0.96 [0.96 , 0.97]
Residence	Urban	β_2	0.380(0.092)	1.46 [1.22 , 1.75]
Gender	Male	β_3	0.261(0.066)	1.30 [1.14 , 1.47]
Malaria		β_4	1.284(0.099)	3.61 [2.98 , 4.38]
Hh Size		β_5	0.031(0.008)	1.03 [1.01 , 1.05]

Note: Hh Size= Household Size

The model coefficient for the AGE predictor is negative, suggesting that increasing age (in month) is related to decreasing anaemia status (*higher valued* categories on this rating scale). This means that adjusting for other predictor variables in the model, the estimated cumulative odds ratio is 0.96, suggesting that the cumulative odds of children being in non-anaemic category relative to the more severe grades of anaemia decrease by approximately 4% for an additional unit increase in age (month). It is worth noting that, this interpretation holds across the entire range of anaemia status, from "severe" to "non-anaemic" as figure 2 indicates. This follows from the fact that the proportional odds assumption was not rejected. Furthermore, the model coefficient for the household size (Hh size) predictor is positive, suggesting that increasing household size is positively related to increasing anaemia risk (*lower valued* categories on this rating scale). This means that, the estimated cumulative odds ratio is 1.03, suggesting that the cumulative odds of children being in a more severe grade of anaemia category relative to a less severe grade of anaemia (the odds of being either severely, moderately or mildly anaemic relative to the odds of being non anaemic) increases by approximately 3% for each additional unit increase in household size (holding other variables fixed). Moreover, the results indicate that the logarithm of the odds of severely, moderately or mildly anaemic for males is estimated to be 0.261 times the logarithm of the odds for females. The estimated cumulative odds ratio comparing males with females is $exp(0.261) = 1.30$. This means that, in this sample, the cumulative odds that males will be in a more severe grade of anaemia is 30% more than that of females of the same age. Similarly, children with malaria were over thrice as likely to have a more severe grade of anaemia compared to children who did not have malaria. This is in line with observation from the exploratory results. The estimated cumulative odds ratio for comparing children who live in urban areas to their children counterpart in rural areas is 1.46.

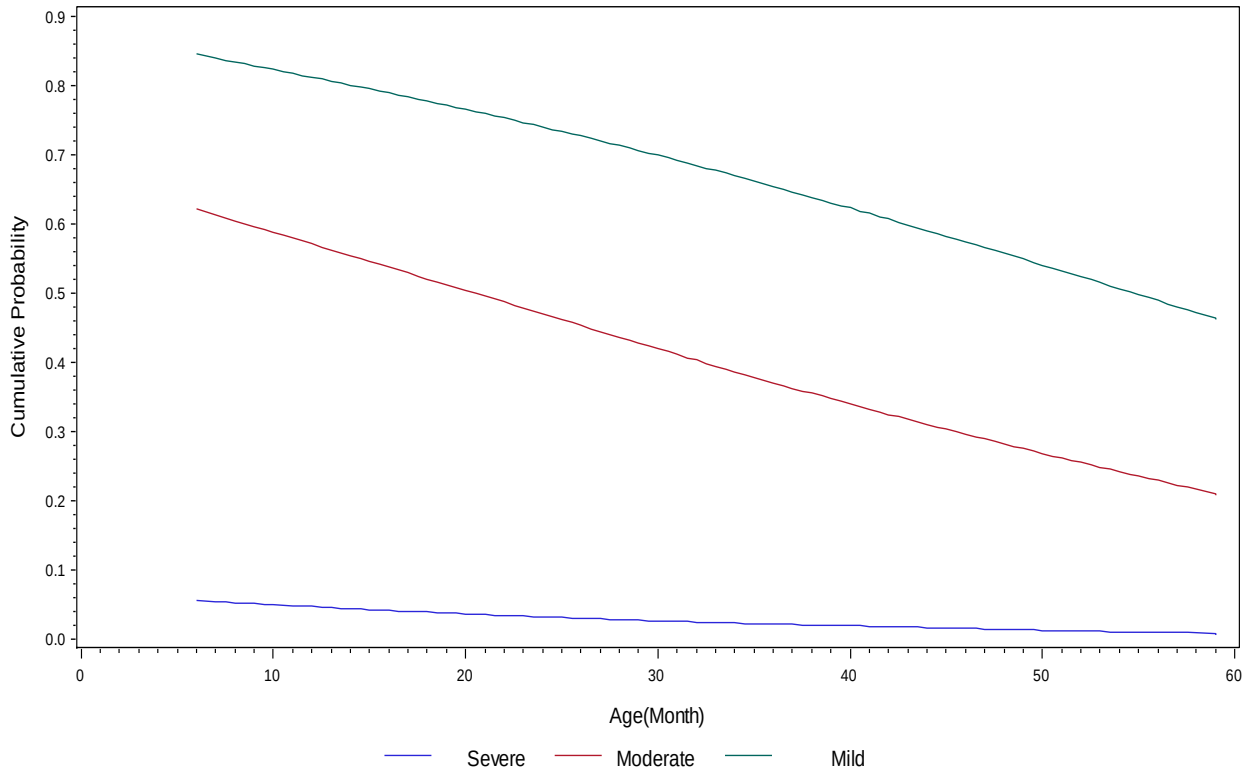

Figure 2: Cumulative Probabilities for Anaemia status with age

4.2.2 Subject-Specific

A different approach to account for clustering is by using random model components such as random intercepts. Under the GLMM, model fitting began by adoption of the marginal model covariates, which was then extended by allowing two-way interaction terms of the selected covariates. Additionally, the model also included the random effects-in this case, random intercepts to address the between-cluster and within-cluster variations. These were introduced in the generalized linear mixed model due to the fact that the probability of having anaemia possibly varies for individuals within the same household as well as individuals in different households. This was done by using restricted pseudo-likelihood implemented with the SAS GLIMMIX procedure. The model was fitted under the three possible working assumptions, namely, independence, compound symmetry and unstructured. However, the unstructured and compound symmetry assumptions turned out not to be feasible due to convergence issues even after removing the interaction terms. These two random intercepts were assumed to be independent. This restricted our selection set to only main effects models. Thus, results for GLMM were also qualitatively similar to those from GEE and ALR.

$$Logit\left[P\left(Y_{ijk}=1\middle|v_i,u_{ij}\right)\right]$$
$$=\beta_0+\beta_1 age_{ijk}+\beta_2\text{Malaria}_{ij}+\beta_3\text{Urban}_{ij}+\beta_4\text{Male}_{ijk}+\beta_5\text{Hh\_size}_{ijk}+v_i$$
$$+U_{ij} \qquad (10)$$

In order to decide on the best of the two random effects models, two models were fitted, one with the two random intercepts (between and within clusters variations) and one with one random intercept(within cluster variation). One can use the REML pseudo-likelihood ratio test (LRT) to compare these two models (Myers *et al., 2010)*. Let l_{full}^{R} = -2 Res Log Pseudo-Likelihood= 4701.3, denote the restricted log-pseudo-likelihood value of the full (complex) model. Let l_{red}^{R} = -2 Res Log Pseudo-Likelihood= 4742.31, denote the restricted log-pseudo-likelihood value of the reduced model. The appropriate likelihood ratio test statistics, Λ , is

$$\Lambda = l_{full}^{R} - l_{red}^{R} = 41.01$$

We compared this to a χ_1^2 and divided the p-value by 2. The resulting p-value is < 0.001 and so we conclude that, the model with two random intercepts should be used to address the between-cluster and within-cluster heterogeneity in the data. The parameter estimates and standard errors of the subject-specific model are presented in Table 4.

A corresponding interpretation can be performed based on parameter estimates from the right half of Table 4 of the generalized linear mixed model as was the case in the three-level ALR. It should be kept in mind however that the two models are different and their comparability may not be meaningful as they have different parameter interpretations. That is, in the GLMM framework, parameter interpretation is based on subject specific unlike in the GEE and ALR where parameters are treated as population averages. Parameter estimates obtained from GLMM are generally bigger in absolute values than those from GEE and ALR.

Table4: Parameter estimates (standard errors) for two and three-level GLMM

			Two-Level		**Three- level**	
Effect	Level	Parameter	Estimate (s.e)	p-value	Estimate (s.e)	p-value
Intercept		β_0	1.541(0.147)	<.0001	1.873(0.154)	<.0001
Age		β_1	-0.041(0.002)	<.0001	-0.04(0.003)	<.0001
Residence	Urban	β_2	0.412(0.151)	<.0001	0.492(0.18)	<.0001
Gender	Male	β_3	0.173(0.053)	0.0024	0.29(0.062)	0.0041
Malaria		β_4	1.215(0.109)	0.0032	1.135(0.141)	0.0019
Hh Size		β_5	0.082(0.023)	0.0011	0.054(0.019)	0.0009
Variance		σ_c^2	0.303(0.072)			
Variance		σ_c^2			0.201(0.085)	
		σ_w^2			0.300 (0.087)	

Note: HH-Size= Household Size

To illustrate the difference in interpretation, consider the effect of malaria on the probability of being anaemic using the generalized linear mixed model. The result shows that the estimated odds of having anaemia was $exp\,(1.135) = 3.11$ (95% CI: 2.36, 4.09) times higher if a particular individual(child) had malaria at the time of the anaemia testing than if this person did not have malaria. The interpretation of other predictor variables can be done in a similar way.

4.2.3 Data Missingness

Table A1 (in the appendix) provides the parameter estimates and their corresponding standard errors along with the odds ratios and 95% confidence intervals for the odds ratios for the fitted three-level ALR model (in which the probability of being anaemic is modeled as a function of the selected covariates) using the Multiply-imputed dataset. Results from Table A1 (appendix) can be compared with the results from the right half portion of Table 2. From these two tables, it can be observed that the estimated standard errors for the MI estimation of the ALR model are generally less than or equal to the standard errors of the coefficients from the analysis of the standard ALR. The slightly smaller standard errors from the MI analysis reflect the fact that the imputation has recovered additional statistical information from the cases that were excluded by procedure from the analysis of the non-imputed data. Furthermore, models fitted using standard ALR and ALR with Multiple Imputation approaches provide comparable parameter estimates. That is, in this particular model, the differences are not large enough to require remedial action to assess the statistical significance of the individual predictors in the final model selection.

5. Discussions and Conclusions

This project was aimed at identifying socio-demographic determinant factors of anaemia among children aged 6-59 months in Mainland Tanzania. Children with higher probability of occurrence of these determinant factors would be inferred to be most likely to experience anaemia as haemoglobin (Hb) concentration below the normal level is often associated with anaemia. As a preliminary analysis, various summary statistics as well as graphical techniques were employed to explore the association between the response variable of interest and available covariates. It should be noted that there is inconsistency in the conclusion from the analysis of various summary statistics which might be due to the fact that they make use of varying amount of information which determines the power of their inferences. Thus, the analysis was extended to other statistical methods to account for the clustered nature of correlated observations. The data were then analyzed using two model families: (1) marginal models (GEE, ALR and POM), and (2) random effects model (Generalized linear mixed model). All the four models led to the same conclusion that age (in month), gender, type of residence, household size and malaria were found to be significantly related to any type of anaemia (Hb level below normal). Age had a negative effect while household size had a positive effect indicating that children from households with a greater number of members had higher probabilities of being diagnosed with haemoglobin (Hb) concentration below the normal level. On the other hand, children who had malaria had the highest probabilities of developing anaemia.

Although we fitted both model families in the same analysis, it should be kept in mind that the two model families are rather different, and that the parameters have to be interpreted differently. Indeed, in practical situations the choice on which model family to use is guided by the research question.

Furthermore, the three-level alternating logistic regression under the marginal model family further indicated a strong significant association between any two pairs of responses from the same household as well as pairs of observations from two households within the same cluster. It worth noting that variable region did not appear in the final model, however, the significance of measures of associations and the presence of type of residence in the final model can provide information about within region variation of anaemia.

The problem of missing data is one that is almost unavoidable. Incorrect or failure of accounting for missing data can lead to invalid inference and misleading study conclusions. This may in turn have detrimental effects in real life decision making. In this study, multiple

imputation technique was used to take into account the missingness. However, the differences of results from models fitted using standard ALR and ALR with Multiple Imputation were not large enough to require remedial action to reverse decision concerning the statistical significance of the individual predictors in the final model.

6. Recommendations

At the clinical level, based on the highly significant effect of malaria on anaemia, it would then be important for this stratum of individuals to receive effective treatment and examination to guarantee timely intervention if needed. This is important for both conditions in reducing the burden of child mortality due to anaemia and malaria. This follows due to strong association between anaemia and malaria, this means that patients especially children under five years of age brought to health facilities with malaria should be checked carefully for anaemia. Additionally, in this analysis, we have studied how the risk of being anaemic depends on age of a child, type of residence, gender, malaria and household size. However, it is worth noting that the probability of being anaemic, that is, having haemoglobin (Hb) concentration below the normal level could be affected by other factors such as nutritional deficiencies, hookworm infections and inherited red blood cell disorders. Investigation of such factors could be recommended in future studies. However, challenges may lie on the side of resources made available and possibly means of collecting these factors.

7. References

Agresti, A. (2002). *Categorical Data Analysis*, 2nd Edition. New York: John Wiley & Sons.

Diggle, P.J., Liang, K.Y. and Zeger, S.L.(1994). *Analysis of Longitudinal Data*. Oxford: Claredon Press.

Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models* , 2nd edition. London: Chapman & Hall.

Ministry of Health and Social Welfare (MOHSW) Tanzania. (2006). *National guidelines for malaria diagnosis and treatment 2006.* Dar es Salaam, Tanzania: NMCP.

Molenberghs, G. and Verbeke, G. (2006). *Models for Discrete Longitudinal Data*. New York: Springer-Verlag.

Myers, R.H., Montgomery, D.C., Vining, G. G., and Robinson, T.J. (2010). *Generalized Linear Models with Application in Engineering and the Sciences*. 2nd edition. New Jersey: John Wiley & Sons.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations, *Biometrics* **57**, 120-125.

Schellenberg, D., Schellenberg, J., Mushi, A., de Savigny, D., Mgalula, L., Mbuy, C. and Victora, C.G. (2003). The silent burden of anaemia in Tanzanian children: a community based study. *Bulletin of the World Health Organisation* **81**,581 – 590.

Tanzania National Bureau of Statistics and ICF Macro. (2011). *2010 Tanzania Demographic and Health Survey: Key Findings*. Calverton, Maryland, USA: NBS and ICF Macro.

UN. (2009). The Millenium Development Goals report.

World Health Organisation. (2011). WHO. *Haemoglobin concentrations for the diagnosis of anaemia and assessment of severity.* Vitamin and Mineral Nutrition Information System. Geneva, Vol.11, issue.1. Available at: http://www.who.int/vmnis/indicators/haemoglobin.pdf, accessed on [15/07/2011].

Zeger, S.L., Liang, K.Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrika* **42**, 121-130.

8. Appendix

Table A1: Estimated Three-Level ALR (Multiple Imputation Analyses)

			MI(n=4,372)	
Effect	Level	Parameter	Estimate (s.e)	OR [95% CI]
Intercept		β_0	1.817 (0.116)	6.15 [4.90 , 7.72]
Age		β_1	-0.04 (0.003)	0.96 [0.95 , 0.97]
Residence	Urban	β_2	0.532 (0.075)	1.70 [1.47 , 1.97]
Gender	Male	β_3	0.31 (0.107)	1.36 [1.11 , 1.68]
Malaria		β_4	0.986 (0.128)	2.68 [2.08 , 3.44]
Hh Size		β_5	0.028 (0.0103)	1.03 [1.01 , 1.05]
Alpha 1		α_0	0.887(0.183)	2.43 [1.69 , 3.48]
Alpha 2		α_1	0.301(0.098)	1.35 [1.11 , 1.63]

Note: MI= Multiple Imputation

Selected SAS and R Codes:

```
/Figure 1: Panel a and d
male=read.table('D:\\Thesis\\graph\\gender.txt',header=T,sep=",")
# Subseting gender Data
fm <- male[which(male$gender=='f'),]
ml <- male[which(male$gender=='m'),]
anm<-read.table('D:\\Thesis\\graph\\anaemia.txt',header=T,sep='\t')
# Subseting anaemia Data
severe<- anm[which(anm$status=='s'),]
moderate<- anm[which(anm$status=='md'),]
mild<- anm[which(anm$status=='mld'),]
win.graph()
par(mfrow=c(1,2))
# Figure A:Graph of anaemia by gender
plot(male$month,male$prop,type="n",xlab="Age in Month",ylim=c(0,1),
ylab="Proportion of Children",main = '') Anaemia by Gender")
points(fm$month,fm$prop, pch=".",cex=6);lines(fm$month,fm$prop,col="blue",lty=1,lwd=1)
points(ml$month,ml$prop, pch=".", cex=6);lines(ml$month,ml$prop,col="black",lty=1,lwd=1)
legend (3.5,1, c("Female","Male"),cex=0.5, col=c("blue","black"),lty=1:2)
# Figure D: Severity of anaemia by age group
plot(anm$agegroup,anm$prop,type="n",xlab="Age in Month",ylim=c(0,1),
ylab="Proportion of Children",font.main=3,main = "d) Severity of Anaemia ")
points(severe$agegroup,severe$prop,pch=".",cex=6);
lines(severe$agegroup,severe$prop,col="black",lty=1,lwd=1)
points(moderate$agegroup,moderate$prop,pch=".",cex=6);
lines(moderate$agegroup,moderate$prop,col="green",lty=1,lwd=1)
points(mild$agegroup,mild$prop,pch=".",  cex=6);
lines(mild$agegroup,mild$prop,col="blue",lty=1,lwd=1)
legend (3.5,1, c("Severe","Moderate","Mild"),cex=0.4, col=c("black","green","blue"),lty=1:2)
```

```
Fitting PO model: Severity of illiness*/
proc genmod data=thesis.child5;
```

```
/*Fitting ALR model: */
proc genmod data=thesis.child5
```

```
class CASEID Anaemia(ref="4")
residence (ref="2")
malaria_status(ref="0")
gender(ref="2")/param=ref
ref=first;
weight V005;
model Anaemia = age residence
gender malaria_status  Hh_size/
dist=multinomial link=clogit
type3;
repeated subject= CASEID
/type=ind corrw modelse;
run;

/*Fitting GEE model: */
proc genmod data=thesis.child5
descending;
weight V005;
class CASEID residence (ref="2")
gender(ref="2")
malaria_status(ref="0")/
param=ref ref=first;
model resp = age residence
gender malaria_status  Hh_size /
dist=bin link=logit type3;
repeated subject= CASEID/type=
exch modelse;
run;

descending;
weight V005;
class CLID HhID CASEID gender
(ref="2") residence (ref="2")
malaria_status(ref="0")/
param=ref ref=first;
model resp = age residence
gender malaria_status  Hh_size /
dist=bin link=logit type3;
repeated subject= HhID(CLID) /
logor=nest1 subclust=CASEID
modelse;
run;
```

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Socio-Demographic determinants of anemia among children aged 6-59 months in mainland Tanzania

Richting: **Master of Statistics-Biostatistics**
Jaar: **2011**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Cosmas, Stephano

Datum: **12/09/2011**