Modeling Malaria Prevalence in Africa:

A linear mixed effects analysis accounting for missing data

Walter Kiprop Cheruiyot Sirma

I591/83839/2012

College of Biological & Physical Sciences, School of Mathematics,

University of Nairobi

This Project is submitted in Partial fulfillment of the requirements for the Degree of Master of Science in Social Statistics, University of Nairobi.

November, 2015

Declaration

This project report is my original work and has not been presented to any other institution for any academic award.

Signature …… Date ……...

Walter Kiprop Cheruiyot Sirma

This research project report has been submitted for examination with my approval as the University Supervisor.

Signature …… Date ……

Ms. Anne Wangómbe

Dedication

I dedicate this thesis to my family: Ivy, Joy, Becky and parents for their endless encouragement throughout my studies.

Acknowledgement

Certainly no list of acknowledgement will be complete. First, to God for unlimited blessings he has given unto us. I would like to thank my supervisor Ms. Anne Wang'ombe for her guidance throughout the thesis writing process. More particularly, the feedback he gave refined this work to the masterpiece it has turned out to be.

The knowledge and skills necessary to accomplish this task would not have been imparted in me were it not for the commitment of lecturers at the School of Mathematics, University of Nairobi. Their constant patience with my classmates and I as we tried to understand the statistical concepts deserves a special mention.

Much gratitude goes to; my entire family for their love, encouragement, and support throughout this phase of my academic journey and who have continuously inspired me to do what I believed in.

Finally I sincerely thank all my classmates and Dr. Taabbuk who have shared their thoughts and criticism about my work.

Abstract

Malaria is one of the leading causes of illness and mortality in Africa. A lot of resources have been, and continue to be channeled towards prevention, controlling and treatment of malaria. The World Health Organization receives annual data of reported malaria cases from all member countries including a number of African countries albeit some missing reports. The objective of this research is to use a statistical model to manipulate reported longitudinal data from selected African countries for the period 2000 to 2012 to estimate malaria disease burden. This model will also address the issue of missing data, which is a shortcoming in most Africa Countries.

The longitudinal nature of the data will be modeled using linear mixed-effects model using country specific intercepts and slopes. These models accounts for heterogeneity between countries with regards to magnitude of malaria. The model is also robust in cases where data is incomplete. As a final step, two ways of acknowledging and accounting for missing information (complete case analysis and multiple imputations) were explored and results compared with the direct likelihood approach whereby all the observed data is used for analysis.

Results indicated that the model is flexible enough to capture the variability in profiles of different countries, thus allowing for inference regarding the reported number of confirmed malaria cases for any given period per country.

In conclusion, the number of reported and confirmed malaria cases in Africa between the year 2000 and 2012 highly depends on the country under investigation. Some countries reported high cases while others reported fewer cases. The evolution of the number of cases over time has remained relatively constant. Therefore, choices on which countries have a higher malaria burden can be made with certainty, thus focusing more intervention resources to those countries as it's expected that for the lower risk countries, the rate of malaria infection will also remain relatively constantly low.

Keywords: Linear mixed-effects model, malaria, multiple imputation, Africa

Table of Contents

List of tables

List of figures

List of Abbreviations

WHO	World Health Organization
HIV	Human Immunodeficiency Virus
AIDS	Acquired Immunodeficiency Syndrome
NGOs	Non Governmental Organizations
KEMRI	Kenya Medical Research Institute
GDP	Gross Domestic Product
USAID	United States Agency for International Development
UK	United Kingdom
MI	Multiple Imputations
LMM	Linear Mixed Models
MLE	Maximum Likelihood Estimate
REML	Restricted Maximum Likelihood Estimation
SAS	Statistical Analysis System
MCAR	Missing Completely At Random
MAR	Missing At Random
AIC	Akaike Information Criteria
MCMC	MonteCarlo Multiple Chains

1 INTRODUCTION

1.1 Background

Malaria remains of a global public health concern. It is a vector borne disease that is most prevalent in tropical and sub tropical regions. It is a deadly infectious disease that mostly affects human species. The global impact of malaria on health, economic and social wellbeing is high. This is attributed to the reported morbidity and mortality rate, the cost of healthcare associated with malaria, the social capital required in combating the menace as well as the loss in man work hours and disability years due to the disease.

In fulfilling its mandate, the World Health Organization(WHO) routinely receives data on several malaria indicators on the number of reported and confirmed malaria cases, number of reported and confirmed malaria deaths, those receiving nets, pregnant women population reporting affecting by malaria, funding for malaria, country progress on malaria activities, availability of appropriate and recommended testing and treatment among other statistics (WHO, 2008). The data is usually provided by the relevant countries' departments of health at the end of every year. There are however instances where for some reason, a country fails to provide its malaria statistics to the WHO. This data is usually availed and stored in WHO database to inform malaria program implementation status in individual countries.

It is globally estimated that malaria kills one child in every 30 seconds, translating to about 3000 children deaths daily, (United Nations Children Fund, 2014). In the adult population, malaria related mortality rate almost nears that of Human Immunodeficiency Virus (HIV)/ Acquired Immunodeficiency Syndrome (AIDS) and tuberculosis. Even more worrying is an article by the Daily Telegraph which indicated that in the year 2014, despite the panic due to the Ebola outbreak, malaria killed 70 times more people than Ebola hence malaria is a bigger threat in Africa,(Nelson, 2014).

According to the WHO most of the deaths in Africa are largely preventable and treatable. The magnitude at which malaria kills the vulnerable has seen a lot of countries efforts geared towards reducing the spread of malaria mainly by putting up control and prevention measures.

The presence of malaria parasite, a female anopheles mosquito plays a major role in the transmission of malaria to humans. These mosquitoes breed mostly in paddy waters especially during rainy season, or on marshy grounds. The life cycle of a mosquito comprises of four stages: a new life begins when an adult mosquito lays eggs in stagnant water such as marshy areas or water pods. At the second stage, the eggs hatch into mosquito larvae which transform into pupae in their third phase of life. The pupa stage is interestingly a resting phase at which they do not feed but are still mobile. Finally an new adult emerges and floats on water for several days after which its ready to fly off, mate, feed amongst other tasks, (Kalman, 2004).

Mosquitoes feed on blood to survive. In the process of sucking blood form their hosts, they transmit protozoa that cause malaria from an infected host to other humans. Usually, the malaria causing agent is transmitted by female anopheles mosquito. The biological adaptations that specifically makes this species and gender of mosquito suited to transmission has been studied by several researchers, (Carter, 2001; Dimopoulos, Seeley, Wolf, & Kafatos, 1998).

Kenya government with support from private, Non Governmental Organizations (NGOs) and other stakeholders have been at the forefront in the fight against malaria using various interventions. Focus has also been directed to populations most at risk for example the provision of mosquito bed nets to pregnant mothers and infants(Mutuku et al., 2013). Moreover, a huge investment in medical research with regards to effective malaria vaccines by WHO researchers and other institutions in Kenya is an indicator of the public health importance of malaria in coming up with long term measures in combating the disease. For instance, The Kenya Medical Research Institute(KEMRI) and other research partners is at an advanced stage conducting clinical trials at various sites in Kenya; Kilifi in coastal and Siaya in western parts of Kenya(Agnandji et al., 2011; Moorthy, Good, & Hill, 2004).All this scientific work is being recognized by WHO.

1.2 Statistical model theoretical background

Linear mixed-effects models are flexible statistical models which are appropriate in modeling data that is collected severally from the same measurement unit. Data tables will be extracted from the WHO databases the reported malaria cases that were actually medically confirmed to be malaria. The resulting dataset is a longitudinal data at country level with missing information for specific countries. In this analysis, a unit is each country for whom annual malaria cases were

reported to the WHO. The challenge is however that some of the countries fail to remit their annual statistics due to several reasons that are beyond the scope of this analysis. Missing data however has an impact on the analysis and inference performed herewith. A statistical treatment for the missing data is therefore a point of focus for this analysis.

The objective of this study is to focus in formulating an appropriate linear mixed model for the extracted longitudinal data. The study also propose to fill in the missing data with plausible values by using complete case analysis and multiple imputations and compare how these two corrections influence the resulting estimates.

1.3 Case study data overview and analysis justification

The data used in this analysis was obtained from the website of the WHO by following the link http://apps.who.int/gho/data/node.main.A1364?lang=en. After some data cleaning and manipulations, the final dataset comprised of 43 out of 54 African countries, for which at least one reported and confirmed number of malaria cases available. The data was available from the year 2000-2012 by the time we extracted it although a recent check reveals that 2013 data has now been updated.

Although our initial interest was to analyze malaria cases in Kenya, the provided data was only an aggregate of all malaria cases reported per country for each year. Without a breakdown within the country for instance at county level, there was little statistical modeling that could be done just with Kenya's data. A careful examination of the data however revealed that there was high heterogeneity in the number of reported malaria cases in African countries across the years. This realization motivated us to explore statistical models that would capture this heterogeneity and more so, account for the fact that the data was expected to be highly correlated within a country.

Mixed effects models were the candidate models of choice. In particular, we chose linear mixed effects models rather than generalized linear models so that we could easily illustrate the concepts by transiting from the well known linear regression models to the linear mixed effects models. Generalized linear mixed effects models appropriate for this analysis would have entailed extending the Poisson regression model to account for the heterogeneity.

1.4 Problem Statement

According to WHO approximately 3.2 billion people in 2015 are at risk of malaria, statistics which is about half of the world's population. Sub-Saharan Africa accounts much of the malaria cases and deaths experienced. Sub-Saharan Africa continues to carry a disproportionately high share of the global malaria burden. In 2015, the region was home to 89% of malaria cases and 91% of malaria deaths.

Malaria burden cost Africa about twelve (12) billion in lost Gross Domestic Product (GDP) every year. This accounts for about 40% of all public health spending in Africa.

Surveillance of the disease is of create importance, since this will empower Health Programs and Policy makers to optimize on the interventions, based on the statistics to aid in resources allocation.

There is then an urgent need to monitor malaria statistics to enable a timely and effective malaria response in endemic regions, to prevent outbreaks and resurgences, to track progress, and to hold governments and the global malaria community accountable.

In fulfilling its mandate, WHO routinely receive data on several malaria indicators such as the number of reported and confirmed malaria cases deaths among other statistics. This then inform health program implementation status in individual countries through policy development and interventions recommendations as a move towards malaria control and elimination.

WHO in the recent years reported four countries as certified for having eliminated malaria: United Arab Emirates (2007), Morocco (2010), Turkmenistan (2010), and Armenia (2011). Thirteen countries also reported zero cases of malaria in the year 2014 while six countries reported fewer than ten cases of malaria.

There are however instances where for some reason, a country fails to provide its malaria statistics to the WHO. The extracted data is thus longitudinal at country level with missing information for specific countries.

The interest is therefore to formulate an appropriate linear mixed model for the longitudinal data since this is robust in cases of missing data, as seen on the extracted data that some profiles exhibit missing data scenarios.

With insight that failure to account for missing data may lead to misinformed inferences hence poor policy formulation and wrong interventions considered for health program implementation.

A comparison to establish if estimates with missing data may have influence on the inferences made using Linear mixed model, with complete case analysis & multiple imputations is proposed.

1.5 Research Questions

i. What are the implications of having missing data in the implementation of public health programs?

ii. What models are appropriate to enable programs and policy makers estimate the expected burden of a public health event and subsequently make inferences?

1.6 Objectives

1.6.1 General Objective:

The overall objective of this study is to apply appropriate statistical methodology for longitudinal malaria dataset, while examining the missing data patterns and resulting impact on statistical inference.

1.6.2 Specific objectives:

1. To fit an appropriate linear mixed-effects model to the malaria data in Africa using the data for 2000-2012.
2. To estimate the parameters using complete case analysis.
3. Provide a comparison of parameter estimates resulting from classical linear mixed-effects models with observed data, complete case analysis as well as the multiple imputations.

2 LITERATURE REVIEW

Malaria remains a global problem, thus the continued attention from the scientific community, government agencies, NGOs, and the population at large. A widespread campaign to control and possibly eradicate malaria globally by health and non health stakeholders is ongoing. Various strategies are being employed; for example the provision of insecticide treated nets, pesticides, environmental sanitization, equipping health facilities with test kits and adequate malaria medicine and research. Funding from both government and other agencies such as United States Agency for International Development (USAID), United Kingdom (UK) aid, Bill and Melinda Gates Foundation, Against Malaria Foundation amongst others have not only enabled the provision of protective gear but also funded research into vaccines and more effective treatment for malaria, (Agnandji et al., 2011; McCoy, Kembhavi, Patel, & Luintel, 2009).

Non-medical research into the impact of the interventions has also been a point of focus for academic researchers. Other findings documented that although mosquito nets were easily accessible among other interventions, their long term physical integrity was hampering their effectiveness(Mutuku *et al.*2013). Another community based cross-sectional survey in Kwale County to determine the physical condition of the nets and more importantly, identify the predictors of poor physical conditions of bed nets. To this end, a semi-parametric logistic regression was used and concluded that physical deterioration of nets was associated with higher use and washing frequency. Young and older children were using ineffective nets more than infants, highlighting the focus that had been placed on lactating mothers and infants by the interventions.(Hosmer, Jr., Lemeshow, & Sturdivant, 2013)

A decline reported in pediatric admissions on the coast of Kenya which could actually be attributed to malaria specific interventions that were being conducted in the hospital catchment areas of Kwale, Kilifi and Malindi between January 1997 and March 2007. In their analysis, they adjusted for seasonal variations in the rainfall and admission rates using different time series models with a 13-point moving average. To assess the effect of time on the reported malaria cases, they fitted a linear regression model using non-malaria cases as an additional covariate for predicting the malaria cases. By using this linear regression model, their assumption was that the model residuals were independent, an assumption which is usually invalid for longitudinal data of this nature, as they rightly acknowledged. Following additional tests they performed, they

indeed concluded that there was serial correlation in the dataset and hence they adjusted their model to an autocorrelation time-series with a lag of two months (Okiro *et al.* 2007).

Several longitudinal studies within the context of malaria have been done by many researchers. A longitudinal study conducted to describe the epidemiology of malaria infection amongst children in Western Kenya. Their dataset comprised of prospectively monitored children between June 1992 and July 1994 whereby blood smears were tested for malaria infection. They reported malaria prevalence of between 60% and 83%(Bloland *et al.*, 1999). In their statistical analysis, they applied generalized estimating equations to a binary outcome in order to account for correlation between outcomes of the same child. With this model, they were able to illustrate the effect of time (months) and age-group of children in the prevalence of malaria. Clinical follow up studies are often characterized by incomplete data due to patients failing to show up for some follow up visits, lost to follow up incidences as well as due to natural attrition. In their study however, there is no mention of the missing data and the treatment thereof.(Hardin & Hilbe, 2012; Ziegler, 2011)

There is abundance of literature for longitudinal analysis of malaria incidence data in Africa. Most of these studies however comprise of a binomial outcome which is presence or absence of malaria in an individual. They therefore employ the generalized estimating equations approach to account for dependency in the data(Bloland et al., 1999; Degefa et al., 2015). These models belong to the class of models for correlated data that are referred to as marginal models(Agresti, 2012; Geert Molenberghs & Verbeke, 2006). While there is nothing wrong statistically with these models, their approach aims at giving a global or the so called population averaged picture about the condition being studied. Thus, rather than making inference about a particular child who was followed up in the survey conducted by(Bloland *et al.*, 1999), the model provides inference for the entire children population in Western Kenya (or a region with similar demographic profile). Moreover, the outcome was dichotomous in nature as characterized by the zero prevalence of malaria in the patients. The resulting longitudinal model therefore extends the classical logistic regression model to accommodate repeated measurements per study subject.

The literature explored so far has mainly analyzed data collected at only one study area in the case of (Bloland *et al.*, 1999), or at least homogeneous study areas such as is the case for the analysis by(Okiro *et al.*(2007). More often than not, studies are conducted in heterogeneous

regions, sub populations or even countries. For instance, (Kleinschmidt 2001) was confronted with the challenge of analysing data from populations of northern most districts of KwaZulu Natal in South Africa, where there are strong heterogeneities in the incidence of malaria. Their approach entailed accounting for spatial correlations in the context of generalized linear mixed-effects models where the spatial effects were captured using appropriately defined random effects. This heterogeneity model provides a different dimension of information than we have seen in the previously presented literature since it allows for inference specific to different regions within the district. Spatial analysis of incidence data is a hot topic in disease mapping,(Kleinschmidt, 2001; Lai, So, & Chan, 2008).

Linear mixed-effects models allows for inclusion of random effects in the model makes it more special in that it allows for the hierarchical representation of findings. This is to say that the resulting random effects model are subject specific, thus allowing for subject specific inference to be performed.(Gałecki & Burzykowski, 2013; Verbeke & Molenberghs, 2009; West, Welch, & Galecki, 2014). Situations where such a model may be of interest are abundant in literature especially in medical applications. In clinical trials, patients are often followed up resulting in panel data. Linear mixed models have been applied to account for the heterogeneity amongst patients with regards to their CD4 count (Binquet, 2001; Hoffmann et al., 2013); to analyze survival and longitudinal outcomes simultaneously in the so called joint models (Rizopoulos, 2012) amongst other contexts. One of the challenges however is that longitudinal data often results in incomplete data,(Donald & Robert D., 2006; G Molenberghs, Fitzmaurice, Kenward, Tsiatis, & Verbeke, 2014). On a positive note, linear and generalized linear mixed models are robust to unbalanced data and can therefore often provide valid inferences even in the presence of unbalanced data,(Cnaan, Laird, & Slasor, 1997).

The topic of missing data has generated wide interest in recent times. Although initially researched upon by (RUBIN, 1976), recent literature on the subject can be found in(Little & Rubin, 2002; G Molenberghs et al., 2014; Shah, Laird, & Schoenfeld, 1997) amongst others. Several authors have presented illustrations of the problem and possible solutions in the context of longitudinal studies. One solution that stands out is the method of multiple imputation, (Carpenter & Kenward, 2012). Verbeke & Molenberghs, (2009) and Twisk & de Vente, (2002)

are a few authors who illustrated practical examples of multiple imputation in different settings of correlated data.

The literature gap that this study seeks to fill is quite clear. While there has been longitudinal analysis of binary malaria data, they often did not tackle the missing data analysis. Moreover, in this study we model incidence data which is not dichotomous. Such data calls for two possible approaches; either model counts of reported infections as having a Poisson distribution using the generalized linear mixed effects model or, perform a logarithmic transformation of the reported cases and analyze resulting data as a linear mixed effects model. We adopt the second approach for this analysis. To give a richer treatment of the problem at hand, multiple imputations to account for missing data is applied. Ad hoc treatment of the missing data by only analysing complete cases will also be considered depending on the number of available complete cases.

3 METHODOLOGY

Longitudinal malaria dataset for all WHO Africa member states will be extracted from the WHO database. Longitudinal data is observed when measurements are scheduled to be obtained from an individual at several time periods.

An important aspect of longitudinal data is that for a given unit of measurement (hereby referred to as the *subject*), at least one measurement is observed. In this thesis, longitudinal data results from obtaining the records of reported and confirmed malaria cases for countries in Africa as recorded by the World Health Organization. Thus, the subject in this case is each country in Africa for whom we have at least one year's statistics on the reported and confirmed malaria cases.

Formally, let Y_{ij} be the number of malaria cases in country $i = 1, 2...N$ in year $t_{ij} : j = 1, 2....J$. From the cleaned dataset, data was available for $N = 43$ countries out of the 54 African countries and $J = 13$ {Year 2000-2012}. Thus for a given country i, the vector of responses can be denoted as(Donald & Robert D., 2006);

$$\mathbf{Y}_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ . \\ . \\ . \\ . \\ Y_{iJ} \end{pmatrix} \qquad (1)$$

Subsequently, the complete vector of measurements for all the countries is denoted as follows;

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ . \\ . \\ . \\ \mathbf{Y}_N \end{pmatrix} \qquad (2)$$

The difference between the current dataset and that often used for linear regression analysis is that the components of the data vector are scalars since each subject contributes only one measurement to the data vector unlike in Equation (2) where the components are a vector of measurements from each subject. This distinction is important since it results to the breakdown of the theory on linear regression analysis in several ways: For one, the resulting model residuals from linear regression on such a dataset are no longer independent. This is clearly because measurements from a single country are correlated in that they follow a particular pattern and not just random. An illustration of longitudinal data resulting for the current case study is shown in Figure 1. While for linear regression analysis we could be having only line, here, a collection of profiles corresponding to each of the subjects is observed.

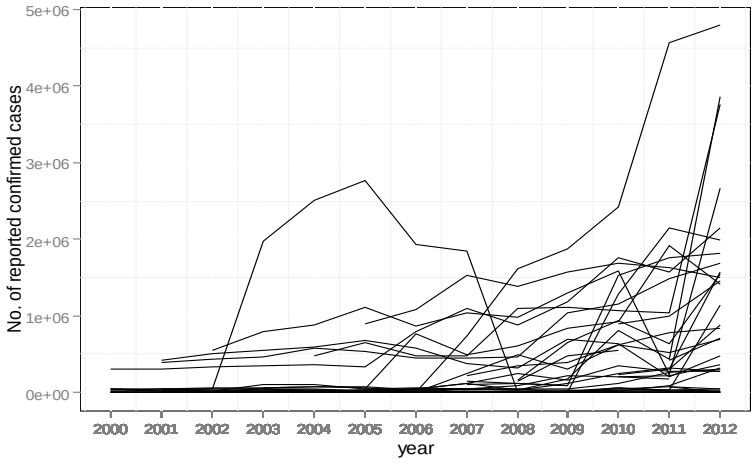


Figure 1: Country specific profiles for reported confirmed malaria cases between 2000 and 2012

The model ought to address the additional complexities introduced by the data structure. To this end, linear mixed-effects models an extension of the classical ordinary linear regression models are useful tools for handling Gaussian data from longitudinal studies as discussed in the next section.

3.1 Linear mixed-effects model (LMM)

For each country, a maximum of 13 data points is expected (for a country with all data available). Linear mixed models have been discussed by several authors including Liard and

Ware (1982) as well as Verbeke and Molenberghs (2000). The starting point of LMM is noting that from

Figure 1, a plot of the data per subject may reveal several patterns. For instance, the profile plot reveals the following.

i. The number of confirmed cases in the year 2000 varies a lot amongst countries.
ii. Although the evolution generally follows a similar pattern, there is variability in the evolution and more particularly, one of the country profiles (Tanzania) actually exhibits a quadratic profile.
iii. There is variability (fluctuations) in the measurements within a given country over time.

These aspects and more forms core of the linear mixed model analysis. An average profile can be obtained by averaging the observed number of cases for all the countries in those years, thus obtaining an average profile. Deviation of each country's measurement from the overall average profile is quantified by measurement error just like in linear regression analysis. To quantify variability between the subjects, random effects are introduced thus resulting in linear mixed-effects models. Linear mixed effects models can then be develop in a general setting.

3.1.1 Model formulation.

A linear mixed-effects model results by combining two computation steps of the two-stage modeling approach. Each of these two stages is described.

3.1.1.1 Stage 1

The development of a linear mixed-effects model can be performed in two stages(Verbeke & Molenberghs, 2009). In the first step, an ordinary linear regression model is fitted on each country's vector shown in Equation(1). This implies that at this stage, N linear equations are fitted each corresponding to a model for each of the countries. The linear regression model is denoted as;

$$\mathbf{Y}_i = \mathbf{Z}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i \qquad (3)$$

Where,

- $\mathbf{Z}_i$ is a $(n_i \times q)$ matrix of known covariates.
- $\boldsymbol{\beta}_i$ is a q dimensional vector of subject-specific regression coefficients
- $\boldsymbol{\varepsilon}_i \sim N(0, \sum_i)$ is the vector of measurement errors (residuals). In many situations, a simple structure for the covariance matrix for measurement error is assumed. Thus, rather than imposing a complex error structure, $\sum_i = \delta^2 I_{n_i}$ is used.

Estimation of the parameters of this model can be performed in any Statistical software that supports linear regression analysis. The parameter estimates can be obtained using ordinary least squares approach by minimizing the loss function

$$Q_i = \sum_{i=1}^{n_i} (y_{ij} - \bar{y}_i)^2 \qquad (4)$$

Note that the model in Equation (3) captures the variability *within* a subject (the so-called within-subject variability in multivariate statistics(Johnson & Wichern, 2002).

The resulting output for the model is thus a $(n_i \times q)$ matrix of regression coefficients which forms the input for the second stage of the model.

3.1.1.2 Stage 2

In the second stage, between subjects variability is modeled. This is achieved by relating the estimated coefficients β_i with known covariates. Thus,

$$\boldsymbol{\beta}_i = K_i \boldsymbol{\beta} + \mathbf{b}_i \qquad (5)$$

Where;

- K_i is a $(q \times p)$ matrix of known covariates. Additional covariates other than time can be adjusted for in this step.
- $\boldsymbol{\beta}$ is a p dimensional vector of regression coefficients.
- $\mathbf{b}_i \sim N(0, \mathbf{D})$ is a matrix of random effects indicating the deviation in individual subjects' measurements from the population average. **D** is the covariance matrix capturing the between-subject variability.

3.1.2 General linear mixed-effects model (LMM)

The two-stages can be combined and performed more efficiently in a single step. Thus, we substitute Equation (5) into Equation (3)resulting in;

$$\begin{aligned} &\mathbf{Y}_i = \mathbf{Z}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i; \boldsymbol{\beta}_i = K_i\boldsymbol{\beta} + \mathbf{b}_i \\ &\rightarrow \mathbf{Y}_i = \mathbf{Z}_i\left(K_i\boldsymbol{\beta} + \mathbf{b}_i\right) + \boldsymbol{\varepsilon}_i \\ &\rightarrow \mathbf{Y}_i = \mathbf{Z}_i K_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i \end{aligned} \qquad (6)$$

It is important to note that $\mathbf{X}_i = \mathbf{Z}_i K_i$ is a design matrix containing all the covariates of interest in the model. The final general linear mixed-effects model is hereby given as;

$$\begin{aligned} &\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i \\ &\mathbf{b}_i \sim N(0, \mathbf{D}) \\ &\boldsymbol{\varepsilon}_i \sim N(0, \Sigma_i) \end{aligned} \qquad (7)$$

All the parameters carry the same meaning as earlier defined. Additionally, the model in Equation (7) imposes an assumption that the random effects b_i and measurement errors ε_i are independent. The regression coefficients β comprises the fixed effects for the model. The model is linear in parameters (hence a linear model), contains both fixed and random effects (hence a mixed-effects) resulting in the terminology *linear mixed-effects model*(Donald & Robert D., 2006; Gałecki & Burzykowski, 2013; West et al., 2014).

3.1.2.1 *Additional properties of the linear mixed effects model*

The model in Equation (7) is referred to a hierarchical model due to the hierarchy exhibited in the variability structure. When ε_i is assumed to have a simple covariance structure as defined in Equation (3), then, conditional on the random effects,

$$\begin{aligned} \mathbf{Y}_i \,|\, \mathbf{b}_i &\sim N\left(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i, \sum_i\right) \\ \mathbf{b}_i &\sim N\left(0, \mathbf{D}\right) \end{aligned} \qquad (8)$$

Thus, the variance of the observed data can be decomposed into two components; *within-subject variability* and *between-subject variability*.

By using random effects, the between subject variability can be captured accordingly. The implied marginal model is denoted as;

$$\begin{aligned} \mathbf{Y}_i &\sim N\left(\mathbf{X}_i\boldsymbol{\beta}, V_i\right) \\ V_i &= ZDZ' + \sum_i \\ V_i &= ZDZ' + \delta^2{}_i \end{aligned} \qquad (9)$$

Also worth mentioning is that the hierarchical model shown in Equation (8) implies the marginal model (9) but the reverse is not true.

3.2 Parameter estimation and inference

3.2.1 Maximum Likelihood Estimation (MLE)

Unlike in the two-stage approach (or classical linear regression), estimation of the linear mixed model fixed effects parameters is performed using a maximum likelihood approach. For the marginal model in Equation(9), we (re)introduce the following notation;

$$\begin{aligned} &\beta: \text{ fixed effects vector} \\ &\alpha: \text{ vector of all variance components} \\ &\theta = \left(\beta', \alpha'\right): \text{ vector of all parameters in the marginal model} \end{aligned} \qquad (10)$$

The marginal likelihood function is given by:

$$L_{ML}\left(\theta\right) = \prod_{i=1}^{N}\left\{\left(2\pi\right)^{-\frac{n_i}{2}} \left|V_i\left(\alpha\right)\right|^{-\frac{1}{2}}\right\} \exp\left(-\frac{1}{2}\left(Y_i - X_i\beta\right)' V^{-1}\left(\alpha\right)\left(Y_i - X_i\beta\right)_i\right) \qquad (11)$$

When α is known, the Maximum Likelihood Estimate (MLE) of β is given by;

$$\hat{\beta}(\alpha)=\left(\sum_{i=1}^{n}(X_i'W_iX_i)^{-1}\sum_{i=1}^{n}(X_i'W_iy_i)\right);\ \text{where } W_i \text{ equals } V_i^{-1} \qquad (12)$$

Often, α is unknown and needs to be estimated. MLE $\hat{\alpha}_{ML}$ can then be obtained by maximizing Equation (11) with respect to α. Since both α and β are unknown, the algorithm maximizes the parameter vector $\theta=(\beta',\alpha')$ simultaneously.

3.2.2 Restricted Maximum Likelihood Estimation (REML)

In ordinary MLE theory, given μ and δ^2 are the population mean and variance respectively, if μ is known, the sample variance $\hat{\delta}^2=\sum_{i=1}^{n}(Y_i-\bar{Y})^2/N$ is an unbiased estimator for the population variance. For unknown μ however, $\hat{\delta}^2=\frac{N-1}{N}\sum_{i=1}^{n}(Y_i-\bar{Y})^2$ which is a biased estimate of δ^2 since $\hat{\delta}^2=\frac{N-1}{N}\delta^2$.

An unbiased estimate of δ^2 is the sample variance defined as $\hat{\delta}^2=\sum_{i=1}^{n}(Y_i-\bar{Y})^2/(N-1)$. This implies therefore in the context of linear mixed-effects models, in order to obtain an unbiased estimate of the population variance, a transformation of Y to eliminate μ from the likelihood is needed. This transformation is given by;

$$U=\begin{pmatrix} Y_1 & Y_2 \\ Y_2 & Y_3 \\ \vdots & \vdots \\ Y_{N-2} & Y_{N-1} \\ Y_{N-1} & Y_N \end{pmatrix}=A'Y\sim N\left(0,\delta^2A'A\right) \qquad (13)$$

With this transformation in place, the MLE of δ^2 is now the unbiased one $\hat{\delta}_{REML}{}^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 / (N-1)$ and is referred to as the REML estimate. We shall therefore use REML estimation for all the models fitted in this analysis since it is independent of the choice of A. A more comprehensive discussion of the REML in linear mixed-effects models is given by (Verbeke & Molenberghs, 2009). The models will be implemented in Statistical Analysis System (SAS) statistical Software, although almost every modern statistical package supports linear mixed-effects models.

3.3 Missing data analysis

With the advent of big data, there has been an increasing challenge of missing data. Moreover attrition is an inherent problem of longitudinally collected data especially if the time span between obtaining measurements for an observation is large(Twisk & de Vente, 2002). While some statistical models require balanced data, linear mixed models are robust to data imbalance. This means that, even with incomplete data, the models can still be applied without the underlying theory collapsing.

In some instances however, interest may be in addressing the issue of missing data one way or another. Ad hoc ways of dealing with missing data includes linear interpolation, substituting the missing values with an average of the two adjacent values, last observation carried forward, as well as complete case analysis. Little and Rubin (2002) presents a good overview of the problem of missing data and possible solutions. On the other hand, Molenberghs *et al.* (2014) developed a handbook of missing data methodology which we greatly borrow from in tackling missing data analysis. Three approaches to handling missing data will be considered in this thesis:

3.3.1 Complete case analysis

The most restrictive of the approaches is complete case analysis. As the name suggests, only subjects with measurements in each of the time points are used for the analysis. This is a strictly because the underlying assumption for complete case analysis is that the unobserved data is Missing Completely At Random (MCAR). In other words, *the mechanism generating the missingness is independent of both the observed and unobserved parameters of interest*. If indeed the missingness mechanism is MCAR, the resulting inference is unbiased. However, there are no formal tests for MCAR. Moreover, the fact that in nature missingness is rarely MCAR.

3.3.2 Direct likelihood: analysis of the data as-is

This is the output of the statistical models applied to the data without further considerations about the missing data. Therefore there are no additional computations or data manipulations necessary. The underlying assumption is that the data is Missing at Random (MAR). This is an unverifiable assumption: *given the observed data, the missingness mechanism does not depend on the unobserved data.* Likelihood inference based on direct likelihood approach is valid (under the assumption if MAR).

3.3.3 Multiple Imputation (MI)

The focus of this section is in performing Multiple Imputation to account for missingness in longitudinal data. The underlying assumption is still that the data generating mechanism is MAR. We decompose the density function for the data as follows;

Define,

$$R_{ij} = \begin{cases} 1, \text{ if } Y_{ij} \text{ is observed} \rightarrow \mathrm{Y}_i^o \\ 0, \text{ otherwise} \rightarrow \mathrm{Y}_i^m \end{cases}$$

$$f\left(y_i D_i \mid \theta, \psi\right) = f\left(y_i^o \mid \theta\right) f\left(M_i \mid Y_i^o, \psi\right) \qquad (14)$$

Where M_i denotes the time of missingness.

MI is a three-step process;

i. Create M complete dataset by filling in missing values with some estimates. The imputation values are obtained by first sampling them from a given distribution (and acknowledging the variability in these random variables accordingly).
ii. Perform standard analysis on each of the M datasets: in our case, this entails filling the linear mixed-effects model (7) in each of the datasets.
iii. Combine the results of the M analyses to perform inference on.

Theoretical justification of these steps can be obtained from Rubin, (1976), Little & Rubin, (2002) amongst other authors presented in the references.

3.3.3.1 Algorithm of multiple imputation for missing data

1) Draw a new parameter vector θ^* from its posterior distribution.
2) Draw Y_i^{m*} from $f\left(y_i^m \mid y_i^o, \theta^*\right)$
3) Using the new complete set of data $\left(Y^0, Y^{m*}\right)$, obtain an estimate of

   $\hat{\beta} = \hat{\beta}(Y) = \hat{\beta}\left(Y^o, Y^{m*}\right)$
4) Repeat the steps 1-3 M times thus obtaining $\hat{\beta}^m$ and within imputation variance

   $U^m = Var\left(\hat{\beta}\right), \quad m = 1, \cdots, M$
5) Pool the information obtained in step 4 to obtain the final estimate of the regression coefficients;

$$\hat{\beta}^* = \frac{\sum_{m=1}^{M} \hat{\beta}^m}{M}, \text{ where}$$
$$\left(\beta - \hat{\beta}^*\right) \sim N\left(0, V\right) \qquad (15)$$

The variance $V = W + \left(\frac{M+1}{M}\right) B$ can be decomposed into within-subject variance $W = \frac{\sum_{m=1}^{M} U^m}{M}$ and between-subject variance $B = \frac{\sum_{m=1}^{M} \left(\hat{\beta}^m - \hat{\beta}^{m*}\right)\left(\hat{\beta}^m - \hat{\beta}^{m*}\right)'}{M-1}$.

We shall presents results of these three models and compare their efficiency based on the estimated variance.

4 RESULTS

4.1 Exploratory data analysis

A crucial step in statistical modeling is exploring the data in order to gain insights into the information contained. This not only helps in making informed choices during the modeling steps, but also enables us to assess the model fit. The individual country profiles were presented in

Figure 1 whereby it was clear that the trends are not linear. Moreover, considering the fact that the domain of general linear mixed-effects model is the real line $(-\infty,\infty)$, we perform a logarithmic transformation as follows;

$$Y_{ij}' = \log\left(Y_{ij}+1\right) \tag{16}$$

To have an idea of the shape of the mean structure and the random effects to be included, a subject specific profile plot of the cases is shown. This is performed on the log-transformed data presented in Equation (16) resulting in the plot shown in Figure 2 .

From the figure, it is evident that there was much variability between countries in the year 2000. This therefore implies that there is need for random intercepts in the model. The evolution over the years was more or less constant as can be seen from the relatively flat profiles. The trends are not perfectly linear and therefore additional non-linear effects of time such as the quadratic and cubic slope coefficients will be explored.

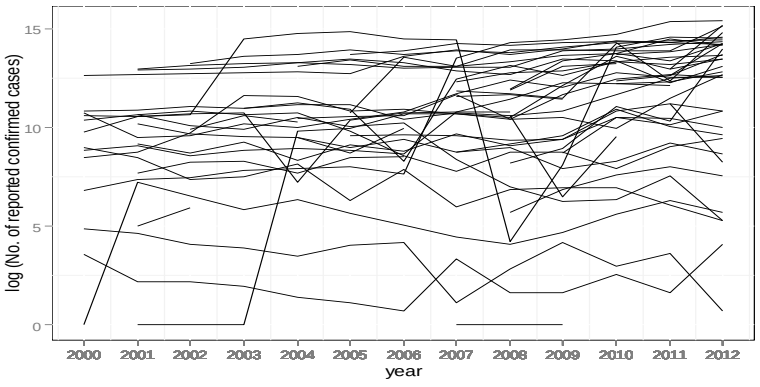


Figure 3: Subject-specific profiles of the log transformed malaria cases in Africa.

In longitudinal data analysis, correlation between observed outcomes is always expected and that's what differentiates the analysis of longitudinal data from ordinary linear regression. Linear mixed models account for correlation by including independent parameters. In order to decide on the covariance structure to model, exploration on the covariance structure in the current dataset as well as knowledge of the subject matter come in handy. A scatter plot matrix of the association between the log transformed number of cases across different years and all observations can be obtained.

Even better, a heat map of the correlation matrix gives a more appealing visual effect as seen in Figure 4. There were mainly strong correlations between blocks of measurements such as 2001, 2002 and 2003. Since there is no clear structure in the correlation over all the years, an unstructured covariance can be a good candidate. However, convergence issues may limit its usage hence the need to explore other covariance structures in the modeling and probably compare the models based on their Akaike Information Criteria (AIC)

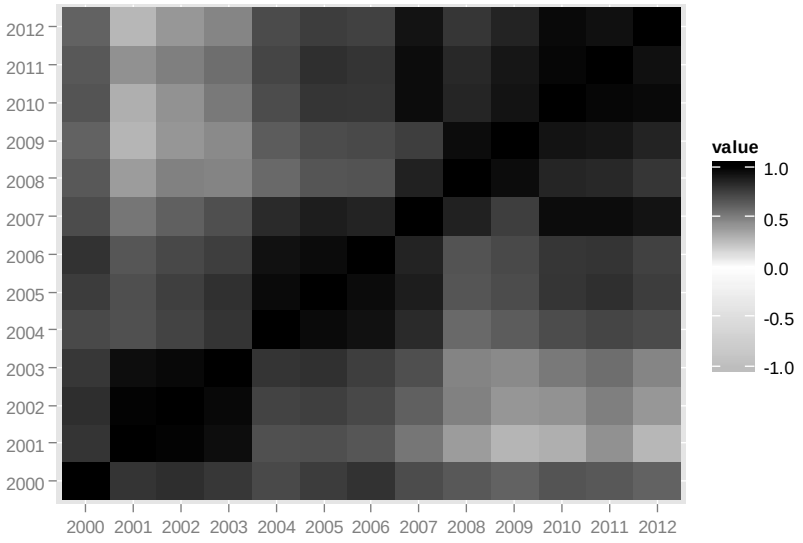


Figure 4: Heatmap of the correlation between the number of cases across the years.

The plot in Figure 3 also highlights the issue of missing data. There are many country profiles which are incomplete either missing data at the beginning, middle the end or a mix of missing

patterns. A country with no missing data should have 13 observations while for any country to be included in the analysis, at least one observation is necessary. A clearer picture of the extent of missing data in this analysis is presented in Figure 5. Only a few countries had all the data available in the 13 year period. These included Tanzania, Democratic Republic of Congo, Madagascar amongst others. As expected, the Northern horn of African countries including Somalia had sparse data mainly due to long term conflicts hence poor data collection by government agencies.

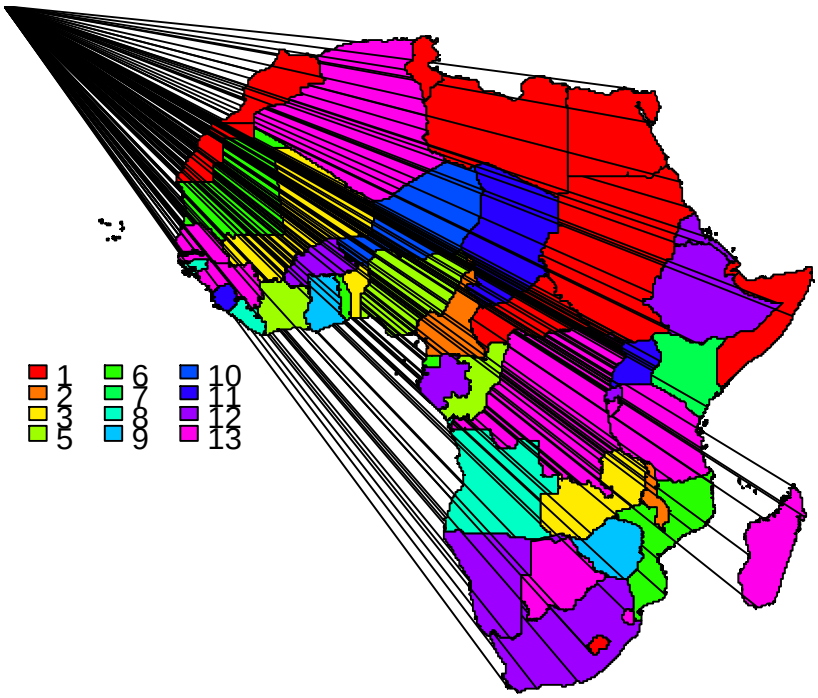


Figure 5: Reported data pattern for African countries.

Note: Each colour denotes the number of available observations

4.2 Linear mixed-effects model formulation and selection

The linear mixed effects model provides a mechanism for modellig the observed subject specific profiles. To begin with, the model comprises of two components; the fixed linear component, and the random effects. In this analysis, we allowed for linear, quadratic and cubic effects of time in the model. Random effects included the random intercept, linear, quadratic and cubic slopes.

The resultig model can be denoted as follows;

$$Y_{ij} = \left(\beta_1 + b_{1j}\right) + \left(\beta_2 + b_{2j}\right)t_{ij} + \left(\beta_3 + b_{3j}\right)t_{ij}^2 + \left(\beta_4 + b_{4j}\right)t_{ij}^3 + \varepsilon_{ij} \qquad (17)$$

Where β_k are the model intercept, fixed effects of linear, quadratic and cubic slope parameters respecively and b_{kj} are the corresponding random effects b_{1j}, b_{2j}, b_{3j} for $k = 1 \ldots 4$. An unstructured covariance for the random effects was fitted to begin with. This model however did not converge and therefore, the first step was to reduce the number of random effects to only have with an unstructured covariance between them which converged without issue.

The estimated variance of the quadratic random effect was still too low, an indication that we could possibly remove it and left with just a random intercept and random linear slope component. The AIC of the resulting model however, did not improve much as shown in Table 1, although this implies that we reduce the number of random effect parameters by three. More parsimonius models (models with few parameters) are always preferred.

Table 1:Akaike Information Criterion (AIC) fit statistic for different models

	Cubic mean	Quadratic mean	Linear mean
Cubic random effects	Did not converge	NA	NA
Quadratic random effects	1540.8	1532.3	NA
Linear random effects	1538.2	1529	**1521.2**
Random intercept only	1703.8	1694	1687.2

To clearly see the reduction in covariance parameters, consider the estimated covariance matrices for the models with intercept linear and quadratic random effects as well as for that with only intercept and linear slope parameter as shown in Table 2. A further reduction in the covariance structure could be obained by only having the random intercept in the model. This simple random effects model was not the most suitable for the data at hand as reveled in Table 2 whereby, the AIC was larger than models with more than one random effects parameters.

Table 2: Random effects covariance matrices for 3 and 2 random effects models

3 random effects model				
	Effect	$\mathbf{b_{1j}}$	$\mathbf{b_{2j}}$	$\mathbf{b_{3j}}$
$\mathbf{b_{1j}}$	Intercept	14.8664	-1.5394	0.03666
$\mathbf{b_{2j}}$	Linear slope	-1.5394	0.3883	-0.01119
$\mathbf{b_{3j}}$	Quadratic slope	0.03666	-0.01119	0.000472

2 random effects model			
	Effect	$\mathbf{b_{1j}}$	$\mathbf{b_{2j}}$
$\mathbf{b_{1i}}$	Intercept	13.3808	-0.9267
$\mathbf{b_{2i}}$	Linear slope	-0.9267	0.1789

Further steps in model selection entailed the reduction of the mean structure for each set of random effects. Starting with cubic time effects model, the time trend was simplified as shown in Table 1. For each model, the corresponding AIC is displayed. Model with only linear fixed effects (intercept and linear slope parameter) and linear randon effects (random intercept and random slope) had the lowest AIC value of **1521.2** hence the preferred model. The final model used for further analysis and the results are presented in the next section.

4.3 The fitted linear mixed-effects model

The best model based on the values of AIC is the linear model with fixed intercept and slope as well as random intercept and slope parameters. The final model is therefore presented below as-:

$$Y_{ij} = \left(\beta_1 + b_{1j}\right) + \left(\beta_2 + b_{2j}\right) t_{ij} + \varepsilon_{ij} \qquad (18)$$

This model implies that on the log scale, the number of cases follow a linear trend with intercept β_1 and slope β_2. Output resulting from this model is discussed next.

4.3.1 Random effects estimates

Fitting the model, there is two random effects parameters to be estimated that is the intercept and slope parameters. Different covariance structures can be imposed such as independence, unstructured, compound symetry, autoregressive covariance amongst others. The compound symmetry, autoregressive and independence in cases estimate two parameters while unstructured estimates three parameters.

The final random effects matrix is presented in Table 3. The variance of the random intercept is quite high $\left(\delta^2_{b_{1j}} = 13.3402\right)$ while for random slope, the variance is relatively low $\left(\delta^2_{b_{2j}} = 0.1802\right)$. This is consistent with the observation made on the individual profiles in Figure 3, where it was observed that there is greater variablility in the intercepts (number of cases reported in the year 2000). Equally the trend of the lines was found to be relatively constant (almost flat lines for all countries). This implies that the countries varied in the number of cases they reported in the year 2000, with some countries reporting high malaria cases while others reporting low malaria cases. Countries that reported high number of malaria cases in the year 2000, consistently reported high cases over the 13 year period, while those that reported low cases, consistently reported low cases over the same time period.

Table 3: Random effects covariance matrix for the final model

	Effect	$\mathbf{b_{1j}}$	$\mathbf{b_{2j}}$
$\mathbf{b_{1j}}$	Intercept	13.3402	-0.9267
$\mathbf{b_{2j}}$	Linear slope	-0.9267	0.1802

As illustrated in Equation (9), the variance of a linear mixed effects model is the sum of covariance matrices from the random effects and the model residual variance $\delta^2 I$. The estimated residual variance for the final model is $\delta^2 = 1.5420$. I is a 13×13 identity covariance matrix,

while $z = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, D = \begin{pmatrix} 13.3402 & -0.9267 \\ -0.9267 & 0.1802 \end{pmatrix}$.

By substituting these parameters in the formulae given in Equation (9), the final variance-covariance matrix for the number of cases observed across the 13 year period for any given country is obtained as shown in Table 4. Clearly, the covariance matrix is unstructured, implying that there is no assumption on the structure of the association between outcomes of a given country across the years.

Table 4: Covariance matrix for the final model

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
2000	14.882	12.414	11.487	10.560	9.633	8.707	7.780	6.853	5.926	5.000	4.073	3.146	2.219
2001		13.209	10.921	10.174	9.427	8.681	7.934	7.188	6.441	5.695	4.948	4.202	3.455
2002			11.896	9.788	9.222	8.655	8.089	7.523	6.956	6.390	5.824	5.258	4.691
2003				10.944	9.016	8.630	8.244	7.858	7.472	7.085	6.699	6.313	5.927
2004					10.352	8.604	8.398	8.192	7.987	7.781	7.575	7.369	7.163
2005						10.120	8.553	8.527	8.502	8.476	8.450	8.425	8.399
2006							10.249	8.862	9.017	9.171	9.326	9.480	9.635
2007								10.739	9.532	9.866	10.201	10.536	10.871
2008									11.589	10.562	11.077	11.592	12.107
2009										12.799	11.952	12.648	13.343
2010											14.370	13.703	14.579
2011												16.301	15.815
2012													18.593

From the estimated covariance matrix, the resulting correlation matrix is shown in Table 5. Interestingly, the resulting correlation structure shows that as time passes, the correlation between measurements reduces. This is more in line with the autoregressive covariance pattern whereby, measurements closer together in time are strongly correlated, while measurements further apart are less correlated.

Table 5: Correlation matrix for the fitted model

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
2000	1	0.8854	0.8633	0.8275	0.7761	0.7094	0.6299	0.5421	0.4513	0.3623	0.2785	0.202	0.1334
2001		1	0.8712	0.8462	0.8062	0.7508	0.6819	0.6035	0.5206	0.438	0.3592	0.2864	0.2205
2002			1	0.8578	0.831	0.7888	0.7326	0.6656	0.5925	0.5179	0.4454	0.3775	0.3154
2003				1	0.847	0.82	0.7784	0.7248	0.6634	0.5987	0.5342	0.4727	0.4155
2004					1	0.8406	0.8153	0.777	0.7292	0.676	0.6211	0.5673	0.5163
2005						1	0.8398	0.818	0.785	0.7447	0.7007	0.6559	0.6123
2006							1	0.8447	0.8273	0.8007	0.7684	0.7335	0.698
2007								1	0.8544	0.8416	0.8212	0.7963	0.7693
2008									1	0.8672	0.8584	0.8434	0.8248
2009										1	0.8813	0.8756	0.8649
2010											1	0.8954	0.8919
2011												1	0.9084
2012													1

4.3.2 Fixed effects estimates

To complete the linear mixed–effects specification, fixed effects are estimated. The model for fixed effects comprised of linear intercept and slope parameters as shown in Table 1.

Table 6: Fixed effects estimates for the final model

Effect	Estimate	Standard Error	P-value
Intercept	8.7687	0.6321	<.0001
Slope (year)	0.1909	0.07212	0.0085

These estimates indicate that both the intercept and slope parameters were statistically significant (p-value<0.05) implying that these parameters are significantly different from zero. The interpretation for these fixed parameters are;

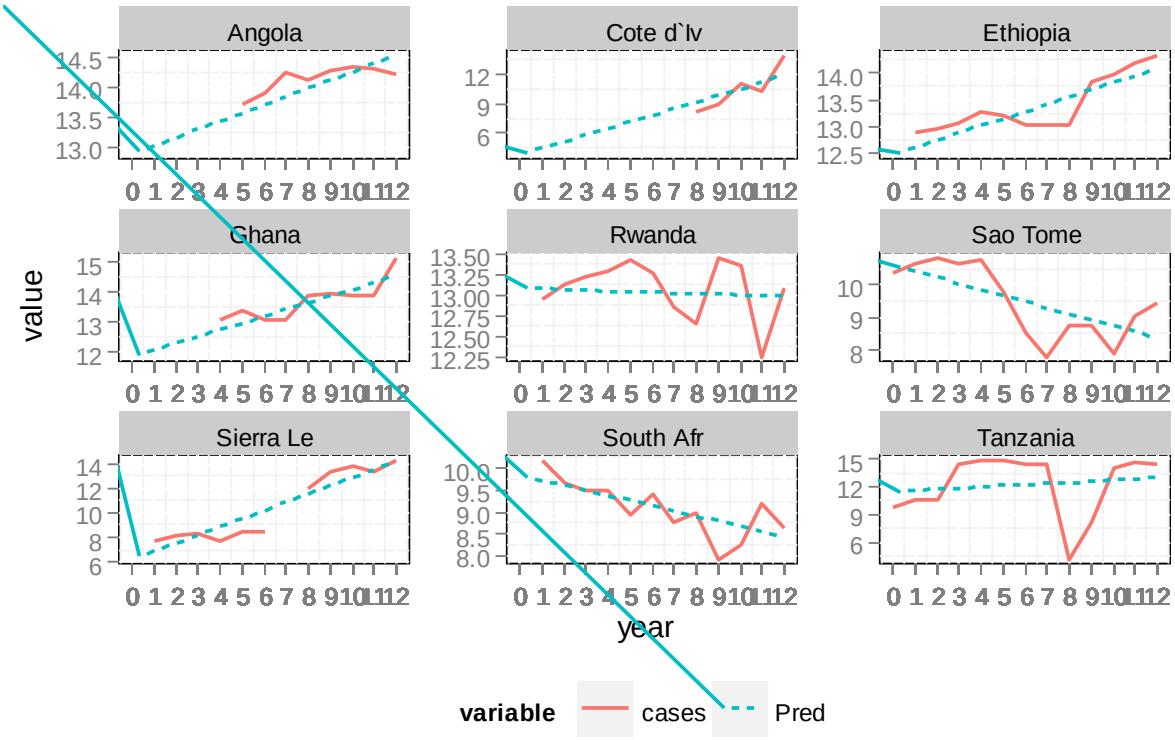
Fixed intercept: the average (log transformed) number of cases in the year 2000 for a country whose random intercept parameter is equal to zero. In other words, a country whose reported number of cases for the year 2000 is similar to the average of all countries had $\exp(8.7867) = 6,547$ reported malaria cases.

Fixed slope: the slope coefficient implies that, for each additional year, the expected number of cases for an average country (a country whose random effects are zero) is $\exp(0.1909) = 1.21$. This therefore implies that over the 13 year period, the expected number of malaria cases only increased by about $\exp(0.1909 * 13) = 12$ cases.

This, combined with the low varinace of random effects therefore explains the relatively flat slopes as seen in the individual profiles.

Figure 6 presents the fitted profiles for nine randomly selected countries. The model clearly captured the trend in the number of cases regardless of the number and pattern of missing data of a given country.

Figure 6: Predicted profiles for 9 randomly chosen countries. The model fits a regression line that best fits the data within a country.



4.4 Missing data analysis: complete cases analysis and multiple imputation

The final analysis for this thesis comprises of complete case analysis, whereby only countries with complete data are considered, and multiple imputation whereby we account for missing data by imputing alternative datasets and computing the parameter estimates of interest.

4.4.1 Complete case analysis

This is achieved by extracting the dataset from countries whose reported number of cases is available across all the 13 years under review. Out of 43 countries, only 11 countries had complete data for the 13 years period. These include; Algeria, Botswana, Burundi, Cape Verde, Democratic Republic of Congo, Guinea, Madagascar, Sao Tome, Senegal, Swaziland and Tanzania (as shown in the colour-coded map of Africa as well).

The subject specific profiles for these 11 countries are shown in Figure 7. The trend is similar to that observed previously in that, there was variations in the outcome in the year 2000 although the overall evolution remained constant.

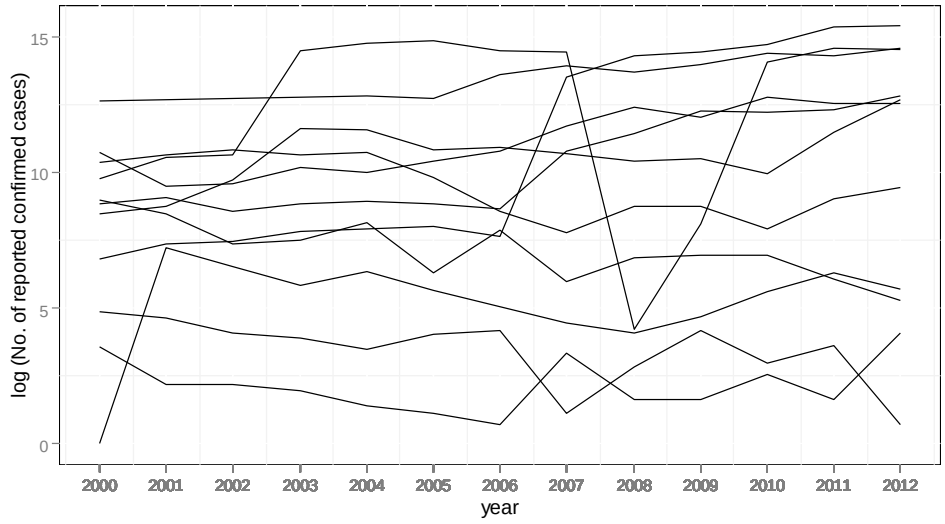


Figure 7: Complete case analysis subject specific profiles.subject specific profiles.

In order to make comparisons with other results discussed in this report, we refit the model specified in Equation (18) now with complete cases only. However, for ease of comparison, parameter estimates from the previous model, the complete case analysis and those resulting

from multiple imputation procedure are presented in one table for cross referencing. Results for complete case analysis will be discussed together with those of multiple imputations.

4.4.2 Multiple imputation

Multiple imputation entails first generating several imputed datasets (5 in this analysis). Imputation is performed by sampling using Bayesian MonteCarlo Multiple Chains (MCMC) sampling, the posterior estimate of the missing data is obtianed.For each new complete imputed dataset, the model specified in Equation (18) is fitted and parameter estimates for the fixed effects and random effects covariance matrix obtained in the second step. The final step in the analysis entails pooling together the parameter estimates for fixed and random effects from the five imputed datasets in order to obtain a single estimates for inference. The resulting estimates from multiple imputation is presented in Table 7 and Table 8.

Table 7: Fixed effects estimates for the three set of analyses conducted

	Direct likelihood	**Complete case**	**Multiple imputation**
Fixed effects	**Estimate (SE)**	**Estimate (SE)**	**Estimate (SE)**
Intercept	8.7687 (0.6321)	7.8962 (0.9865)	9.3204 (0.7103)
Slope (year)	0.1909 (0.0721)	0.1368 (0.0960)	0.1378 (0.0790)

Table 8: Random effects matrices for the three set of analyses conducted

	Direct likelihood	**Complete case**	**Multiple imputation**
Random effects	**Estimate**	**Estimate**	**Estimate**
b1j	13.340	10.135	20.217
b2j	0.180	0.090	0.247
Cov(b1j, b2j)	-0.927	0.021	-1.671
residual	1.542	2.071	3.175
rho	-0.598	0.022	-0.748

Looking at the fixed effects estimates from the three models, there was variation in the parameter estimate for the intercept across the three models. The estimate of slope coefficient is similar for the complete cases and multiple imputation although the direct likelihood approach (using all observed data) resulted in a slightly higher parameter estimate.The standard error of the fixed estimates were smallest for direct likelihood and largest for complete case analysis. Thus, under

the assumption of MAR mechanism, direct likelihood model using data as observed, is the most efficient model. In estmating the values of missing data complete case analysis performed poorer than multiple imputaion.

On the other hand, random effects estimates and the residual variance estimates were two variable across the models. There is need to conduct further research on the appropriatenes of the procedures used to conduct inference for random effects parameters resulting from multiple imputation process.

5 Discussion

This study focused on analysing observational data on the reported number of confirmed malaria cases in countries in Africa. The data was obtained from the WHO database for the period 2000-2012 for various countries. Thus, a longitudinal profile for the number of cases for each country with at least one year data on reported confirmed malaria cases was available. The focus of this study was to illustrate the use of appropriate methodology to aanalyse data resulting from longitudinal studies, especially in the presence of missing data.

To this end, a linear mixed model for continous data was proposed. The model comprises of fixed effects that capture the trend of an average subject and random effects that capture the heterogeneity between subjects in their responses relative to the average expecation. With linear mixed-effeects models, the correlation between outcomes of the same subject can be captured. In this analysis, a linear mixed model with only two fixed effects (intercept and linear effect of time) was the preffered model. This model suggests that the evolution of a country's number of reported cases is relatively constant over time. Moreover, the two random effects (random intercept and time) were highly correlated $(\rho = -0.596)$. The association between random effects implies that, for countries whose reported number of cases were very high, their evolution of the number of cases over the years was slower.

In developing the statistical model, choices had to be made for both the linear part of the model as well as the covariance structure. In linear mixed modelling, the variance matrix of the outcomes comprises of components from the residual variablity and the random effects variability. By chosing to impose an association on the random effects covariance, the resulting covariance matrix of observations can imply a particular covariance structure.

In this study, the final correlation matrix was autoregressive. More so, the trend of the association was that, measurements closer together in time were highly correlated while measurements further apart were less correlated. This implies that, there is little influence of time in the long run over the reported number of malaria cases. We can further conclude that the only component that determined the number of reported confirmed malaria cases was the country in question and not the year under review. Some countries were susceptible to higher malaria infestation while others were less susceptible.

One challenge of working with longitudinal data is the 'curse of missingness'. Often, followup studies results in information gaps due to several factors some known and preventable, such as inefficiencies in data management and reporting, while others are random and unknown. In order to deal with missing data, some assumptions ought to be put in place in order to perform inference using the resulting data. Where very strict assumptions of independence of the missing data from both the observed and unobserved cases (MCAR), the complete dataset can be analysed. The assumptions are however hard to verify and thus, less strict asssumptions of the missing data being independent of the observed data only (MAR) is more commonly applied.

By imputing missing observations, complete datasets are obtained from which inference can be performed. The imputation and inference mechanism ought however to account for the uncertainty introduced in the imputation process. This is achieved by using Bayesian principles of posterior means, which are a mix of the prior information and data likelihood. There is however need to perform further research especially sensitivity analysis as a way of evaluating the impact of multiple imputation in an analysis.

Finally, although the main goal of this analysis was to illustrate statistical methodology that come in handy in analysis of day to day data, we have shown that there is need to take into consideration the subject matter under investigation while performing statistical analysis. For instance, the choice of an appropriate distribution for the analysis led us to transforming the outcome with a logarithmic transform so as to obtain inference within the domain of the model. Moreovoer, by combining simple statistical skills such as graph generation and other exploratory techniques with the technical understanding of statistical modelling, we were able to formulate a starting model which formed the core of the analysis.

As a conclusion, for this particular study, there was high variability between countries in the year 2000 in their reported number of confirmed malaria cases. Over the years however, within a given country, the reported number of malaria cases were relatively constant, with only an increase of about 12 cases on average over the 13 year period. Multiple imputation can be used to generate estimates of the expected numebr of cases in years where a country fails to file its reports in order to have a more complete database.

For future research, I would propose to model the data using generalized linear mixed effects model (using the counts as a Poisson outcome rather than taking the logarithm) and comparing the resulting inference with the one presented here.

REFERENCES

Agnandji, C. et.al (2011). First results of phase 3 trial of RTS,S/AS01 malaria vaccine in African children. *The New England Journal of Medicine*. Massachusetts Medical Society. doi:10.1056/NEJMoa1102287 <http://dx.doi.org/10.1056/NEJMoa1102287>

Agresti, A. (2012). *Analysis of Ordinal Categorical Data*. John Wiley & Sons.

Binquet, C. (2001). Modeling Changes in CD4-positive T-Lymphocyte Counts after the Start of Highly Active Antiretroviral Therapy and the Relation with Risk of Opportunistic Infections The Aquitaine Cohort, 1996-1997. *American Journal of Epidemiology*, *153*(4), 386–393. doi:10.1093/aje/153.4.386

Bloland, P. et.al. (1999). Longitudinal cohort study of the epidemiology of malaria infections in an area of intense malaria transmission II. Descriptive epidemiology of malaria infection and disease among children. *Am J Trop Med Hyg*, *60*(4), 641–648.

Carpenter, J. et.al (2012). *Multiple Imputation and its Application*. John Wiley & Sons.

Carter, R. (2001). Transmission blocking malaria vaccines. *Vaccine*, *19*(17-19), 2309–2314. doi:10.1016/S0264-410X(00)00521-1

Cnaan, A. et.al (1997). Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Statistics in Medicine*, *16*(20), 2349–80.

Degefa, T. et.al (2015). Malaria incidence and assessment of entomological indices among resettled communities in Ethiopia: a longitudinal study. *Malaria Journal*, *14*(1), 24. doi:10.1186/s12936-014-0532-z

Dimopoulos, G. et.al (1998). Malaria infection of the mosquito Anopheles gambiae activates immune-responsive genes during critical transition stages of the parasite life cycle. *The EMBO Journal*, *17*(21), 6115–23. doi:10.1093/emboj/17.21.6115

Donald, H. et.al (2006). *Longitudinal Data Analysis*. Hoboken, NJ, USA: John Wiley & Sons, Inc. doi:10.1002/0470036486

Gałecki, A. et.al (2013). *Linear Mixed-Effects Models Using R: A Step-by-Step Approach*. Springer Science & Business Media.

Hardin, J. et.al (2012). *Generalized Estimating Equations, Second Edition*. CRC Press.

Hastie, T. J. et.al (1990). *Generalized Additive Models*. CRC Press.

Hoffmann, C. J. et.al (2013). CD4 count slope and mortality in HIV-infected patients on antiretroviral therapy: multicohort analysis from South Africa. *Journal of Acquired Immune Deficiency Syndromes (1999), 63*(1), 34–41. doi:10.1097/QAI.0b013e318287c1fe

Hosmer, D. W. et.al (2013). *Applied Logistic Regression*. John Wiley & Sons.

Johnson, R. A., et.al (2002). *Applied Multivariate Statistical Analysis, Volume 1*.

Kalman, B. (2004). *The Life Cycle of a Mosquito*. Crabtree Publishing Company.

Kleinschmidt, I. (2001). Use of Generalized Linear Mixed Models in the Spatial Analysis of Small-Area Malaria Incidence Rates in KwaZulu Natal, South Africa. *American Journal of Epidemiology*, *153*(12), 1213–1221. doi:10.1093/aje/153.12.1213

Lai. et.al (2008). *Spatial Epidemiological Approaches in Disease Mapping and Analysis*. CRC Press.

Little, R. et.al (2002). *Statistical Analysis with Missing Data: Little/Statistical Analysis with Missing Data*. *Statistical Analysis with Missing Data*. Hoboken, NJ, USA: John Wiley & Sons, Inc. doi:doi: 10.1002/9781119013563

McCoy, D. et.al (2009). The Bill & Melinda Gates Foundation's grant-making programme for global health. *Lancet*, *373*(9675), 1645–53. doi:10.1016/S0140-6736(09)60571-7

Molenberghs, G. et.al (2014). *Handbook of Missing Data Methodology*. Taylor & Francis.

Molenberghs, G. et.al (2006). *Models for Discrete Longitudinal Data*. Springer Science & Business Media.

Moorthy, V. S. et.al (2004). Malaria vaccine developments. *Lancet*, *363*(9403), 150–6. doi:10.1016/S0140-6736(03)15267-1

Mutuku, F. M. et.al (2013). Physical condition and maintenance of mosquito bed nets in Kwale County, coastal Kenya. *Malaria Journal*, *12*(1), 46. doi:10.1186/1475-2875-12-46

Nelson, F. (2014). Ebola may be gruesome but it's not the biggest threat to Africa.

Okiro, E. A. et.al (2007). The decline in paediatric malaria admissions on the coast of Kenya. *Malaria Journal*, *6*(1), 151. doi:10.1186/1475-2875-6-151

Organization, W. H. (2008). *World Malaria Report 2008*. World Health Organization.

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. doi:10.1093/biomet/63.3.581

Shah, A. et.al (1997). A Random-Effects Model for Multiple Characteristics with Possibly Missing Data. *Journal of the American Statistical Association*, *92*(438), 775–779. doi:10.1080/01621459.1997.10474030

Twisk, J. et.al (2002). Attrition in longitudinal studies. *Journal of Clinical Epidemiology*, *55*(4), 329–337. doi:10.1016/S0895-4356(01)00476-0

United Nations Children Fund. (2014). Fact Sheet: Malaria, A Global Crisis.

Verbeke, G.et.al (2009). *Linear Mixed Models for Longitudinal Data*. Springer Science & Business Media.

West, B. T. et.al (2014). *Linear Mixed Models: A Practical Guide Using Statistical Software, Second Edition*. CRC Press.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. CRC Press.

Ziegler, A. (2011). *Generalized Estimating Equations*. Springer Science & Business Media.