Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences
School for Information Technology

Master of Statistics

Master's thesis

Modeling the relationship between age and malaria prevalence: A comparison of mechanistic and empirical methods

AGNES KALUNDE NGUKU

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics, specialization Biostatistics

SUPERVISOR :
dr. Thomas NEYENS

SUPERVISOR :
Dr. Emanuele GIORGI
Prof. dr. Peter DIGGLE

2018
2019

Maastricht University

Faculty of Sciences
School for Information Technology
Master of Statistics

Master's thesis

Modeling the relationship between age and malaria prevalence: A comparison of mechanistic and empirical methods

AGNES KALUNDE NGUKU

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics, specialization Biostatistics

SUPERVISOR :

dr. Thomas NEYENS

SUPERVISOR :

Dr. Emanuele GIORGI

Prof. dr. Peter DIGGLE

Contents

List of Figures

List of Tables

Acknowledgements

First, I offer my gratitude to VLIR for offering a scholarship that enabled me to pursue my Master's degree in a world-class university. Without financial help, I could not have enrolled in the program. Therefore, I dedicate this Thesis to VLIR, for offering me a beacon of hope that has illuminated my path to achieve my academic dreams. I would also like to thank Hasselt University, particularly the faculty of Biostatistics, whose members have greatly assisted me while researching my thesis project. The institution has provided me with the resources that I required such as a well-equipped library that enabled me to find pieces of literature that guided my study. Additionally, I thank my thesis supervisors, Dr. Emanuele Giorgi and Dr. Thomas Neyens, for without their help, I could not have completed the project proficient. Their doors were always open when I had problems with the project. Although they ensured that the paper was my own work, they always steered me in the right direction. Moreover, their careful reading and insightful comments guided me in writing this thesis. Therefore, I wish to share my success with them, bearing in mind that any mistake is mine.

I am grateful to my friends, who motivated me in the course of the project. Thank you for bearing with me for all the parties that I have missed! Finally, I thank my family, particularly my mum, who instilled in me the values of hard work, integrity and perseverance, which I relied on heavily in this study. May God always bless you.

Abstract

Malaria is a leading epidemic that has high prevalence and burden. This study acknowledges that there are many parasites which include but not limited to *Plasmodium vivax*, *Plasmodium ovale*, *Plasmodium malariae* and *Plasmodium falciparum*. However, in this study the focus is on the latter. Although past studies have provided data on this parasite from different regions, such data has not been standardized with respect to age. Therefore the current study sought to compare mechanistic and empirical methods in modelling the relationship between age and malaria prevalence. Data were collected in cross-sectional studies of two high-altitude communities in Kenya that encompassed 3352 participants representing different age groups. A rapid diagnostic test was used to scan if a participant suffered from malaria or not. A generalized linear model with splines as basis functions and Pull and Grab model were used to fit the data and to test the models to determine the so-called stability and generality to new data. Generalized linear models estimates the probability of a given event occurring when the values of explanatory variables have been provided. Pull and Grab model utilizes differential equations based on the relationship of variable with the response variable to estimate the parameters. The results indicated that both models captured the association between the prevalence of malaria and age effectively.

Cross validation with 10 folds were used to validate the models. AIC and MSE criteria selected splines with 4 degrees of freedom while the standardized MSE selected Pull and Grab model. The comparison of the two models was done through simulating 1000 data sets utilizing the Pull and Grab model. The average MSE, AIC and standardized MSE values with their corresponding 95% confidence intervals were computed. The three methods suggested Pull and Grab model as the best performing model. However, there was an overlap in the confidence intervals indicating that the difference between the two models was not statistically significant. In conclusion, a researcher could use splines with appropriate degrees of freedom since it performed well even when the data were simulated utilizing Pull and Grab model. Splines could be chosen over Pull and Grab model due to the fact that for the latter a lot of computation is needed in coding differential equations for each variable included in the analysis.

The current study is significant since it could enable the public health department to guesstimate the occurrence of malaria in a given community once the age groups of the participants are known. However, it is crucial for other similar studies to be conducted utilizing the Pull and Grab or spline function to explain the residual variation that was unaccounted for in the current study due to factors such as net use, anti-malarial drug use and socioeconomic status.

Keywords: *Prevalence, Standardize, Pull and grab, Splines, MSE*

1 Introduction

Malaria epidemic is a primary cause of mortality globally. Since the year 2012, $627,000$ people succumbed to the disease out of the 201 million malaria cases (Murray et al., 2012). Possibly Africa experiences the highest morbidity due to harsh climatic conditions and low access to quality health care. Approximately 80% of the reported infections occurred in the continent (Organization et al., 2005). Moreover, in 2010, 24% of the infant mortality cases were attributed to malaria. There has been significant decrease in child mortality due to the improvement in access to health care facilities; this has reduced the incidence of the disease by 54% from 2000 to 2012 as pointed out by Murray et al. (2012). However, the spread of the disease has increased as the incidence rate has risen in Kenyan highlands, with an altitude of more than 1500m, where the disease was previously rare. According to the Organization (2003); Organization et al. (2005), cases of malaria have been reported in approximately 95 countries with 80% of the cases confined to only 15 states, which are mostly located in Africa. The vulnerable groups are children who are younger than 5 years and expectant mothers who have high morbidity and mortality. Approximately $306,000$ children succumbed to malaria with 67% of the demises happening in African nations. Nearly, 90% of children mortality could be attributed to the illness.

Malaria is among the leading diseases that have high mortality and morbidity in Kenya (Machini et al., 2016). Carlos et al. (2014) valued that the incidence of the disease is approximately 6.7 million and the mortality is 4000 per year, which substantially increases the burden of malaria in the country. Greenhouse et al. (2011) observed that children were most likely to have a high incidence and prevalence of malaria. Most of the studies have focused on children who are below five years, which indicates that it is critical to obtain similar data for the school-aged population since they have the highest incidence and prevalence (Beadle et al., 1995). Additionally, children who are above the age of 5 years are likely to have a higher threat of malaria because they could fail to sleep under a treated net (Brooker et al., 2009; Noor et al., 2009).

The prevalence of malaria could be defined as the quantity of the population that has been infected by a malaria parasite in a given sample. Typically, the surveys to determine prevalence involve the collection of thick and thin blood samples from volunteer participants. Next, the blood samples are stained with Giemsa and microscopically examined to determine the count parasites. Use of Giemsa staining procedure is recommended for screening of malaria parasites primarily due to its high reliability. Most of the past researchers have conducted cross-sectional studies; however, longitudinal; research could also be performed to specifically relate individual events with subsequent antibody responses. In Africa, data regarding the prevalence of malaria could be obtained from the health ministries of the respective countries.

Recent studies in Kenya indicate that over 600 surveys regarding the prevalence and incidence of the disease have been conducted across the country. Such data could be useful in identifying the patterns of endemicity which could improve the efficacy of the preventive and curative policies adopted by the government to reduce the commonness of malaria.

The difficult-to-combat malaria has been attributed to its wide variation in epidemiology and clinical manifestation (Bloland et al., 2001). Malaria is transferred indirectly from an infected person to the next through mosquito bites. The five species of mosquitoes that are vectors for the illness belong to the genus *Anopheles*. The parasites that cause malaria when they are introduced into the bloodstream belong to the genus *Plasmodium* (Fana et al., n.d.; Organization et al., 2005). The spread of *Plasmodium falciparum* has been interrelated to the morbidity and mortality of malaria in Africa (Snow et al., 1998).

Bio-statisticians, in collaboration with health care facilities, have plotted accurate maps indicating the variation in the concentration of the parasite depending on climate and other environmental factors, which influence the epidemiology of malaria (Craig et al., 1999; Rogers et al., 2002; Snow et al., 1999). Such maps have been utilized as resourceful tools for the prevention, management, and estimation of the burden of malaria as they provide a potent visual instrument to recognize zones where intended interventions and measures are most likely to have the maximum effect. The primary limitation of the maps is that they do not detail the prevalence of malaria, which is crucial in mitigating the incidence of the disease. Consequently, it is crucial to develop such maps based on empirical field observation, which could help identify areas with a high prevalence of the illness that should be prioritized in the allocation of preventive and curative resources. Thus the maps could substantially reduce the strain of the disease in Africa and evaluation of programs that have been implemented to curb the illness.

Currently, *Plasmodium falciparum* parasite rate (PfPR), the portion of population with parasites detectable in their blood, is not reported in standardized format. Consequently, most of the studies report crude PfPR estimates without grouping them by the age of participants. Subsequently the different age ranges over which studies reports PfPR makes it hard to compare prevalence estimates at different ages and regions. However, PfPR has well-known pattern as a function of transmission intensity and age. It is known that PfPR increases during infancy and early childhood (Gupta et al., 1999), before settling to a plateau in older children, and further decreases in adolescents and adults as their malaria immunity improves (Baird et al., 1991). Since the likelihood of a person testing positive for malaria varies across the age of the participants, age is considered as a confounder variable. A statistical way to minimize the confounding effect attributed to age is to standardize the age. It is also crucial to note that the crude PfPR estimates are also a function of the age distribution of the sample population. Moreover, the crude PfPR does not contain information about the

age distribution of the participants, which compels the researcher to infer it.

The *Plasmodium falciparum* parasite (PfPR), is the most frequently used approach for quantifying the intensity of the incidence of malaria. Several techniques for obtaining standardized PfPR data from crude statistics have been developed. Smith et al. (2007) achieved this aim by training and validating different candidate algorithms. Modified Pull and Grab equations was utilized to come up with the algorithm, which was ranked the highest in standardizing the crude PfPR. The algorithm facilitates age-standardization of PfPR. Consequently, the technique used allows the comparison of different populations even if the age of the participants could vary substantially. Smith et al. (2007) analysis focused on a target age group of 2 to 10 year old children, while discarding data for other age groups. However, the approach did not use a substantial portion of the data that weakened its ability to make accurate predictions especially in regions with scarce data.

The predominant confounding factor for mortality from *Plasmodium falciparum* is the age of individuals. The claim can be supported by several factors that could be either classified as biological or behavioral. For instance, past studies (see Snow et al. (1997)) have revealed that early exposure to high densities of the parasite could lead to the development of immunity when the children get older. In contrast, the low transmission of the disease among children could lead them to develop low immunity as they age, which could render them susceptible to the illness even after they mature. Consequently, in the current study, it is predicted that the lowest incidence of malaria will be observed after a reduction in transmission. The primary reason for this is that when the transmission rate is low, the older children who had been exposed to the disease have developed immunity. Furthermore, younger children who have not developed immunity will be less susceptible to disease because transmission rate is low in the general population. Consequently, as pointed out by Okiro et al. (2009), after the reduction in the initial exposure there emerges a cohort of older children with low susceptibility to malaria.

Age has remained a critical factor in the treatment and management of the disease. Past researchers have not prioritized the statistical methods of confounder control for investigating the correlation between age and incidence of malaria due to two primary reasons. First, severe malaria is quite rare in states where the disease is not considered as an endemic. Secondly, most of the senior citizens are immune, which reduces age-based data that are available for research. Since our study is based in Kenya which is an endemic country it is critical to investigate the relationship between age and commonness of malaria. It is crucial to note that low incidence among children could result in an increase in the infection rate of the older generations due to the delay of the development of blood-stage immunity, which is termed as age-shift. This indicates that epidemiologies such as malaria lead to endemic stability rather than epidemics.

The paper is divided into several sections. Section 1 describes background information and study objectives. A data description is discussed in section 2. On the other hand, section 3 provides detailed description of the methods adopted in analyzing data and inference. Finally, section 4 and section 5 presents the interpretation of results and conclusion, respectively.

1.1 Malaria parasites

Plasmodium parasites cause malaria. They mainly target the red blood cells of humans leading to severe illness, which could be fatal if the appropriate intervention is not provided fast. Typically, the parasites are transferred from an infected person to a healthy person via *Anopheles* mosquito bites. The vectors are mainly active in the dawn, evening and night. Broadly, malaria could be classified into four types that correspond to the four types of mosquito species that transmit the pathogens. The four types of malaria parasites are *Plasmodium falciparum*, *Plasmodium vivax*, *Plasmodium malariae* and *Plasmodium ovale*. The predominant vector species in Kenya is the *P. falciparum*, which is responsible for approximately $80 - 90\%$ of the infections. However, note that the other three species are also found in the country. According to Kouznetsov et al. (1996), *P. falciparum* is the most dangerous vector that should receive special attention to mitigate cases of malaria.

1.2 Life cycle of the malaria parasite

Human malaria (*Plasmodium* parasite) is transmitted from an infected person to another person by *Anopheles* mosquitoes, as shown in Figure 1.

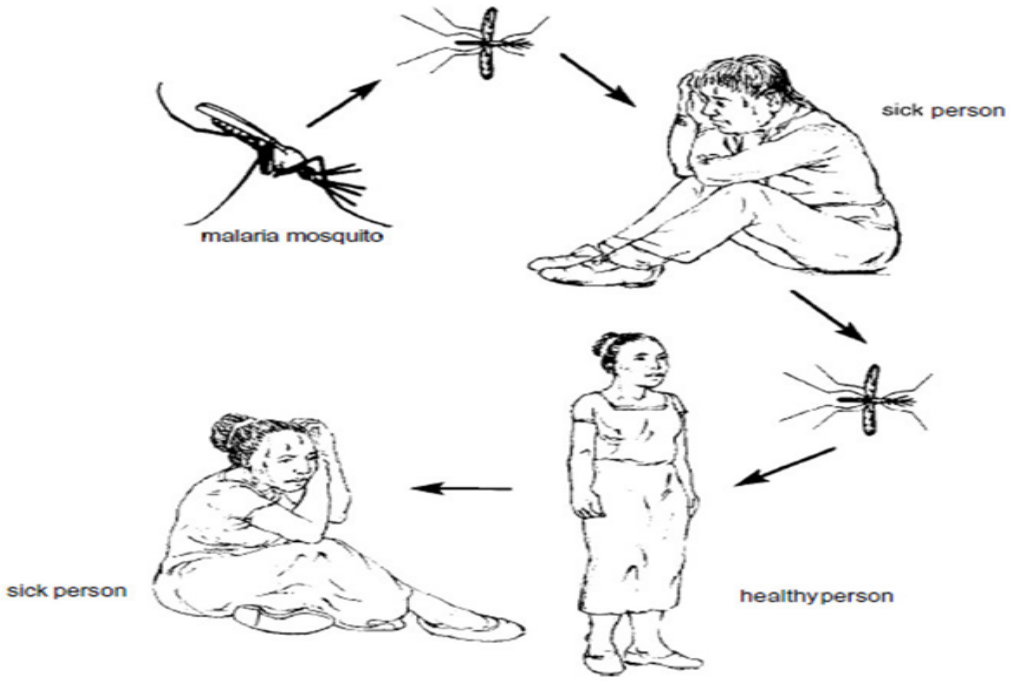

Figure 1: How malaria is transmitted from one person to another. (WHO, 1996, *Malaria: A Manual for Community Health Workers)*

The malaria parasite spreads by infecting two types of hosts: the female *Anopheles* mosquitoes and humans. The parasite is not transmitted from humans to animals and vice versa. Instead, it is transferred from infected people to others through the mosquito vector. The spread of the illnesses commences when a mosquito bites a person and draws some of his or her blood, which could contain the parasite. Once the parasites enter the body of the mosquito, they multiply and mature within $10 - 14$ days. Consequently, at the expiry of the incubation period, the organisms could leave the vector and enter into the bloodstream of other individuals infecting them with malaria. A single mosquito could bite several people leading to a high incidence of the illness in villages that do not use precautionary measures such as sleeping under treated mosquito nets.

The development of malaria in humans occurs through two stages, which take place in the liver and red blood cells. The process commences when a mosquito that has bitten an infected person bites a healthy individual, and the malaria parasites get transferred from the saliva of the vector to the new host. The parasites migrate to the liver and infect its cells within the first 30 minutes. The *Plasmodium* parasites then begin to multiply and develop, leading to the rapture of the liver cells, which releases the parasite to the bloodstream. The liver stage lasts for $6 - 15$ days, after which they infect red blood cells where further replication and development continues. Again, after the number of the parasites increases significantly, they cause the red blood cells to rapture, discharging them to the blood where they can infect other cells over several replication cycles. Most of the pathological and clinical manifestation

of the illness occurs at the red blood cell stage when most of the symptoms of malaria could be observed. The two stages are visualized in Figure 2.

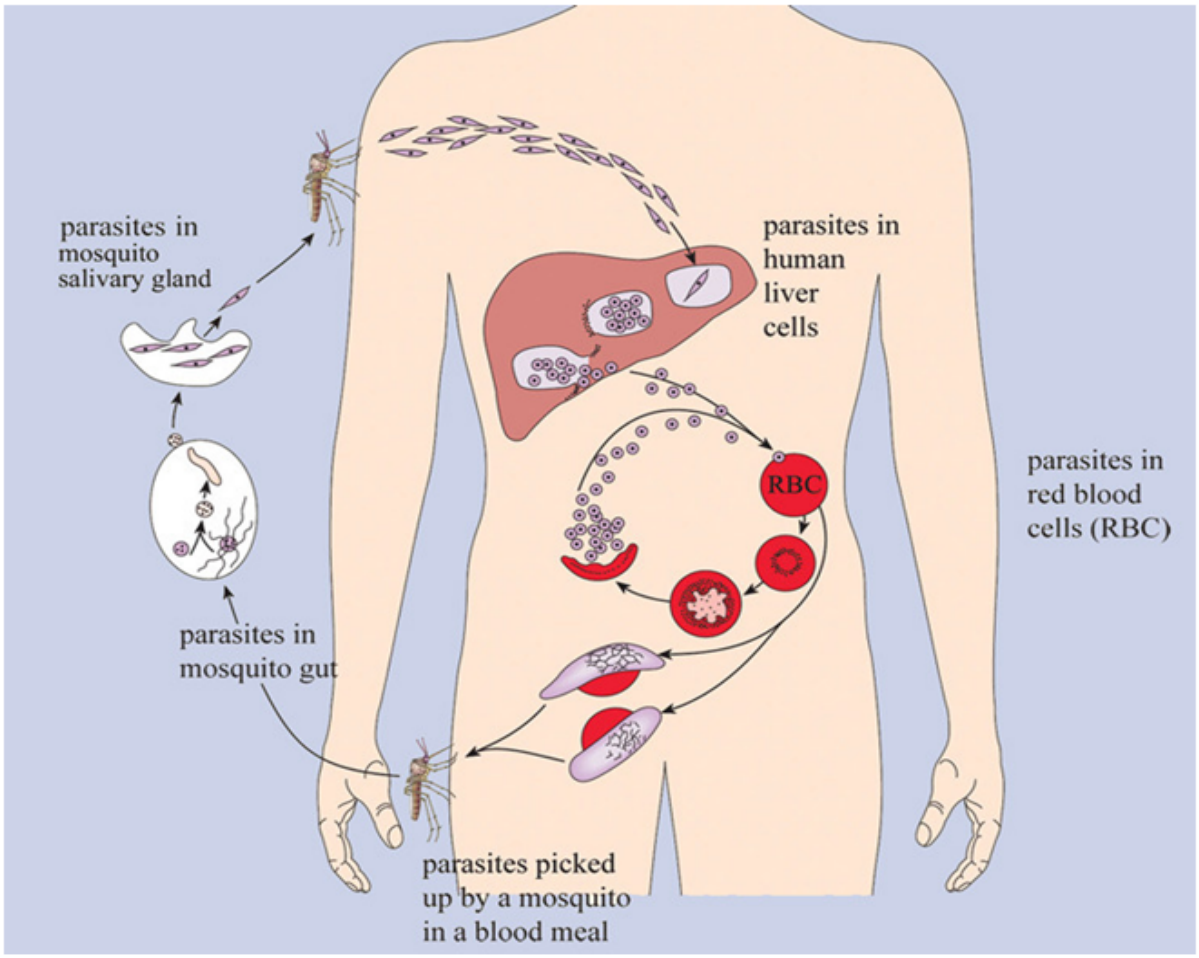

Figure 2: Life cycle of the malaria parasite. (WHO, 1996, *Malaria: A Manual for Community Health Workers)*

The female *Anopheles* picks up *gametocytes*, which are a type of blood stage parasites during blood meal on an infected person. The parasites then multiply in the mosquitoes and continue to develop, and within $10 - 18$ days they are located in the salivary gland of the vector. Consequently, the mosquito could spread the parasite through subsequent bites of healthy people. Although replication and development of the *Plasmodium* parasites occur in the body of the mosquito, they do not harm the insect. Besides the transmission of malaria through mosquitoes, the illness could also spread through other mechanisms. For instance, several studies have indicated that mother-to-child transmission could occur during pregnancies. Moreover, the sharing of syringes and blood transfusion could also spread malaria. However, the infections resulting from mechanisms other than through the mosquito vector are sporadic and could be considered unimportant in the prevention and control of malaria.

Different malaria parasites have varying incubation periods. The incubation period could be

defined as the period from the time that a person is infected to the time that manifestation occurs. The incubation period in days for *Plasmodium falciparum* is $7 - 14$.

2 Data description

The data that were analyzed for the study were obtained from Rachuonyo South and Kisii Central districts located in Nyanza, Kenya. The altitude of the regions ranged $1400 - 1600$m (Stuckey et al., 2012). The predominant tribe in Rachuonyo was Luo, while Kisii was Kisiis. The three dominant mosquito species that spread malaria in the region are *Anopheles funestus*, *An. arabiensis* and *Plasmodium falciparum*. The data set was collected from 8204 individuals who were enrolled in concurrent schools. Additionally, community cross-sectional studies were carried out in 46 schools in the western Kenyan highlands. The community study considered all the individuals (compounds) who were distributed within a 600m radius from each school. The radius of 600m was selected to reduce the likelihood of the overlap of the recruitment area. The specific geographic locations were determined by utilizing a personal digital assistant (PDAs) that was equipped with the global positioning system (GPS). The study samples included all compounds who were above the age of 6 months.

In both surveys, participants were requested to offer blood samples, which were obtained by a finger prick. A rapid diagnostic test (RDT) was employed to test the presence of malaria parasites in the blood sample of participants. The method is normally used in place of microscopy services, which could be more time consuming and expensive (Ayele et al., 2012). Additional data were obtained by issuing questionnaires, which were designed according to the guidelines proposed by Gitonga et al. (2010) who sought to determine the participants' economic status, travelling tendencies, and the use of malaria preventive measures. The people who were found to be infected with malaria were prescribed altemethere-lumefantrine, which is a commonly used malaria drug in the country.

Additionally, individuals who were anaemic were also offered haematinics to comply with the national guidelines for the treatment and management of malaria. In order to comply with the autonomy of the patients, children who were positively identified to be suffering from malaria were not given drugs directly; instead, their parents were asked to visit the schools. The chief aim of our study was to determine the causality between age and prevalence of malaria. The study focused on community cross-sectional studies since the school surveys may be biased due to two primary reasons. Some of the students could be absent which could reduce the likelihood of the sample reflecting the population truly. Secondly, children with similar socioeconomic status are likely to be enrolled in a particular school. The outcome of interest was RDT results. Thus, the response variable is binary as it can only take two states:

positive or negative for malaria.

3 Methods

Past research studies fitted the RDT result data using parametric methods. The models were opted for because they are a reliable tool for modelling the causality between covariates and response variables. Typically the function depicting the relationship is unknown or partly unspecified, which leads to the preference of semi-parametric additive models (Ayele et al., 2014). Hastie and Tibshirani (1987) recommended generalized additive models (GAM) which incorporates a nonlinear component into the additive predictor link scale. This resulting model structure is widely applied in scientific researches that seek to determine parametric nonlinear regression relationship. Using parametric methods for covariates whose relationship to the response is of undetermined functional form might lead to wrong conclusion since the assumption of linearity does not always hold (Hastie and Tibshirani, 1987). Here, one often turns to nonparametric approaches as they enhance the flexibility of the researcher in determining the structure and connection of the data.

The data utilized in the current study were analyzed by fitting it into a generalized linear model (GLM). The GLM was selected because it could generalize linear regression by associating predictor and response variables through a link function and permitting the size of each parameter to be a function of its predicted value. Additionally, the GLM model could also be extended to accommodate random and mixed effects, correlated data, relaxing distributional assumptions, and allowing the use of semiparametric linear predictors (Ayele et al., 2014). An example of the GLM is the logistic regression model, which is mainly used to model binary data.

The *Plasmodium falciparum* parasite rate (PfPR) that is commonly used to report the index of malaria transmission increases after birth and plateaus after individuals develop immunity, before declining in senior citizens. Past studies involving populations with different age groups typically report PfPR as an average, which indicates that age is a crucial variable when harmonizing the reported raw PfPR data. Consequently, the heterogenization confounds simple comparisons of PfPR studies, which have been performed in different places and periods. According to Smith et al. (2007), a so-called Pull and Grab model is useful in standardizing the estimates of the PfPR, which is utilized in the current study in the analysis of data.

3.1 Generalized linear models

In the current study, the binary logistic regression model, which belongs to the family of GLM was used to fit the data. The model estimates the probability of a given event occurring when the values of the explanatory variables have been provided. Let Y be a binary response variable defined by the following signum function.

$$Y_i = \begin{cases} 1 & \text{if RDT is positive in observation } i \\ 0 & \text{if RDT is negative in observation } i \end{cases}$$

Let A_i represent the age of the observation i. In this instance observation i represents a person to be tested for malaria. Additionally, for i, Y_i and π_i denote the indicator for RDT and likelihood of RDT respectively. Therefore, the odds of a participant being positive are $\frac{\pi_i}{1-\pi_i}$. Consequently, the model could be expressed as:

$$\log\left[\frac{\pi_i}{1-\pi_i}\right] = \beta_0 + f(A_i), \tag{1}$$

where $\pi_i = \Pr(Y_i = 1|A_i)$ represents the conditional probability (given the predictors) that the response equals 1 and $f(A_i)$ represents a nonlinear function. In the current study $f(A_i)$ was fitted using splines, and Pull and Grab, which are discussed in the subsequent sections.

3.2 Splines

A general family of transformations to predictors is utilized to ensure that non-linearity is considered in the data analysis. The flexibility of the selected family ought to be flexible enough to adapt to various shapes but should not overfit the data. Basis function denotes a family of transformations that can fit together to a general shape expressed in equation 2.

$$f(A) = \sum_{j=1}^{J} \beta_j b_j(A), \tag{2}$$

where $b_j(A)$ are the basis functions which are fixed and known. There are essentially an infinite number of basis functions to choose from. For instance, wavelets or Fourier series can be used to construct basis functions. As an illustration, we consider polynomial basis functions of the form:

$$b_j(A) = A^j \qquad j \geq 0, \tag{3}$$

this results in a polynomial of variable A given as:

$$f(A) = \beta_0 + \beta_1 A + \beta_2 A^2 + \beta_3 A^3 + \cdots = \sum_{j=0}^{\infty} \beta_j A^j. \tag{4}$$

Splines were utilized in the development of basis functions primarily because of their high flexibility. The popularity of the splines has increased in the last decade because their flexibility simplifies data analysis. Moreover, they have high generality and stability, which could improve the performance of the selected model. According to Wahba (1990), and Wood (2017), splines are restrictive which could simplify data analysis and facilitate parametric estimations.

The primary weakness of the methodology mainly occurs in the selection of the number of knots that are utilized in the development of the spline function. Goepp et al. (2018) claim that inappropriate choice of the number of knots could reduce the performance of the model due to overfitting or underfitting. Overfitting occurs when many data points are selected where the variation of the given function is relatively low. In contrast, underfitting could arise when a few data points are chosen when many points are present. However, the weakness could be mitigated by choosing data points that minimize the residual sums of squares to properly fit the spline function, which could enhance the performance of the predictive model.

Hastie and Tibshirani (1987) discuss several ways of selecting the number of knots, which could enhance the performance of the model. Generally, large numbers of knots are selected where the function is believed to vary rapidly while fewer knots are chosen where the function seems to be more stable. This reduces the complexity of the data analysis. Moreover, one can specify the degree of freedom and then utilize suitable computer application to place the corresponding number of knots at uniform quantiles automatically. Alternatively, one could also arbitrarily select a different number of knots and settle on those that produce the closest fit curve. However, it could be necessary to utilize a more precise technique referred as cross validation which divides data into several sets, where a portion is used to train the model, and the reserved data is used to test the accuracy of its prediction. The choice of the approach selected depends on the desired accuracy level of the data analysis.

The application of the least square and simple regression in the splines method is based on several assumptions. According to Ma et al. (2015), the primary assumption of using splines is that the average change in the predicted variable is proportional to the variation in the predictor variable. However, the assumption does not always apply since the change in the predicted variable could result from other factors other than the variation of the predictor variable. Moreover, an additional error could occur when the mean change is equal for all the variables. Nonetheless, the errors could be mitigated by utilizing the least square method, which determines if the spline function is properly fitted using several primary criteria. For instance, trends of dependent and independent variables that have extremely high or low values, which are treated as outliers, signpost the lack of proper fit of the data into the model.

3.3 Pull and Grab model

The current study utilizes the Pull and Grab model developed by Smith et al. (2007) because their algorithm has been deemed to be reliable to standardize the age, which could transform an estimate of PfPR over any age range into a PfPR over a standard age range. In order to apply the algorithm, we assume that the function $G(A)$ denotes the true prevalence and A corresponds to age. Additionally, let $H(A)$ represent the sensitivity of RDT, which is a standard approach for estimating PfPR as a function of age. The $H(A)$ function was selected since the sensitivity decreases with the increase in age because the blood-stage immunity improves reducing the parasite densities to a point where they could not be detected by the RDT test (Mayor et al., 2007). Therefore, the apparent PfPR, which could be found using the results of RDT, is $p(A) = G(A)H(A)$.

The $G(A)$ and $H(A)$ functions were developed by Pull and Grab (1974), based on the advancement of the previous work of the Ross model Ross (1911) and Muench (1959). The variation of PfPR with age could be expressed as the function:

$$d\pi/dA = h(1-\pi) - r\pi, \tag{5}$$

where h denotes the force of infection in terms of effective contacts per unit of time and susceptible individual, and A is the age. In this study A will be measured in years; consequently h, also called the parasitological inoculation rate, will be a yearly rate. r is the rate at which infections clear and π represents the fraction of the population that tested positive in the RDT. Consequently, when $\pi(0) = 1$, this equation could be expressed as

$$G(A) = \pi'(1 - e^{-dA}). \tag{6}$$

$\pi' = h/(h+r)$ is the PfPR at equilibrium. Additionally, $d = h + r$ represents the rate at which PfPR approaches the equilibrium. In the current study a three parameter approach was utilized for $H(A) : 1 - s[1 - \min(1, e^{-q(A-\alpha)}]$. Therefore, commencing at age A, the function declines from 1 to $1 - s$. Moreover, the parameter q describes the decline from 1 to $1 - s$ as a function of age. Consequently, this function could be regarded as the decrease in detecting an active infection. However, the decrease in the PfPR could be attributed to the immunity, which contributes to the real decline in h or a real increase in r. Note that in the current study the biological reasons for the decline in the function are not relevant for achieving standardization. The modified Pull and Grab model was fitted using the technique of maximum likelihood estimation (MLE), though other estimation methods are probable. The underlying reason for using MLE is due to its applicability in complicated estimation problems and its capacity to produce more robust parameter estimates (Meng and Rubin, 1993).

The estimation of h and r can be done by parametrizing π' to $1/1 + \frac{r}{h}$. However the approach yields an unidentifiable model. An unidentifiable model is one where different inputs could produce an identical probability distribution. Consequently, there will be different parameter values that will be associated with maximum likelihood of any set of the observed data. Consequently, it is crucial to ensure that the model we use is identifiable. In order to make our model identifiable, it is essential to estimate the ratio of r and h.

Table 1: Algorithm Parameters

d	The slope, the rate that PfPR approaches the plateau
π'	PfPR in older children
α	The age when PfPR begins to decline
$1 - s$	The asymptotic sensitivity of RDT
q	The rate that PfPR declines with age, after age α

3.4 Model Validation and Selection

Model validation is conducted to ascertain model stability by considering how well it could be generalized to new data and assessing if the model has low bias and variance that could arise when it picks up too much noise (Wang and Zheng, 2013). According to Alpaydin (2014), model validation could be considered as the process of evaluating a trained model using a set of testing data. Training data set (TDS) is obtained from the available data to enhance the generalization capacity of the training model (Alpaydin, 2009). Finally, the technique is utilized to compare models and select the one with optimal performance. Cross validation is the most commonly applied technique that involves dividing the available data into two sets that are used to train and validate the model (Hastie and Tibshirani, 1987). The splitting of the data set should be conducted randomly to ascertain that the two sets have a similar distribution.

The quality of regression models could be quantified using multiple statistical metrics (Chai and Draxler, 2014). Firstly, R-squared (R2) represents the correlation between observed and predicted values. A high value of the parameter indicates effectiveness of the model. Secondly, the root mean squared error (RMSE) could be applied in the evaluation of the mean prediction error that could be generated by the model when predicting the outcome for a given observation. Consequently, it computes the average deviation between the observed values and predicted values as shown by the formula below:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{7}$$

Effective models usually have low RMSE. Finally, the mean absolute error (MAE), could be used instead of RMSE. The primary strength of MAE is that it is more sensitive to outliers than RMSE. MAE indicates the absolute difference between the observed and predicted results as given in the formula below:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{8}$$

A low MAE could indicate that the model has high performance. It is important to realize that various validation techniques are chosen based on the number of splits of the data set as discussed in the subsequent subsections.

3.4.1 Validation Set Approach

The validation set approach entails dividing the available set of observations into two parts, a training set and a validation set or hold-out set (Hastie and Tibshirani, 1987). One of the sets is used to train the model while the other is utilized for testing. The method could only be used when a large data set is available. The primary weakness of the approach is that it is built on a portion of the available data set, which could leave some of the important data leading to higher bias (James et al., 2013). Based on the data that are used in the training set or the validation set the test error rate could have high variability. When working with small data under the validation set it is more important to assign more data points to train the model than to test it. The primary reason for this is that most models perform better when they have been trained with a large data set to avoid underestimating the error rate in the validation set. However, the approach could also increase the risk of overfitting the data. The approach is preferred because it is simple and reduces the data analysis time and resources. However, the simplicity of the method is often overlooked due to the instability of the data validation method, which inclines either adopting Leave One Out Cross Validation or the K-fold method.

3.4.2 Leave One Out Cross Validation (LOOCV)

Under the LOOCV approach the number of folds or subsets is equal to the number of observations that are contained in the data set. Consequently, only one data point is reserved while the others are used to train the model. Additionally, the reserved data are used to assess the test error associated with the prediction values that are observed. The overall prediction error is determined by computing the mean of all test errors of the prediction that have been calculated. The approach is appropriate when working with small data sets primarily

because all the data points are used to train or test the model, which reduces the incidence of bias. However, LOOCV, suffers from the inherent weakness of high execution time when n is extremely large since the process is repeated for all the data points as purported by James et al. (2013). Additionally, model performance is tested against one data point at each iteration which could increase the variation of the predicting error due to the points that could be considered as outliers. Consequently, it is imperative to find an alternative method that does not suffer from the problem, such as the K-fold cross validation.

3.4.3 K-fold Cross Validation

The K-fold cross validation entails randomly dividing the available data set into k folds or groups that are nearly equal in size. The first fold $k1$ is used to train the model while the other data are employed to validate the model. Additionally, the mean squared error (MSE) is calculated for each of the observations that are predicted in the validation set. Consequently, the number of MSE is equal to the folds. Therefore, the reliability of the model could be determined by finding the average of MSE. According to Hastie and Tibshirani (1987), the formula can be derived as shown below:

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i \tag{9}$$

The K-fold validation is considered as an efficient approach for estimating accuracy of a given model. The primary benefit of the approach over LOOCV is the reduction in the computation time since the process is repeated k times instead of n times. Moreover, the method is superior to LOOCV because it yields more precise approximations of the test error rate. The main problem associated with the cross validation method is the determination of the value of k. Selecting a low value of k could increase the likelihood of the model being biased while selecting a high value of k could lead to a high variability of the test error rate even though it could be less biased. However, there is a golden rule of selecting the optimal value of k. In particular, 5 or 10 folds should be selected, as these values have been shown to minimize bias of the test error rates (James et al., 2013).

3.4.4 Information Criteria

Several factors could make cross-validation inappropriate in training and testing the model. For instance, lack of adequate data to fit in different models depending on the technique used, or the computation could be cumbersome. Therefore, it could be better to utilize the technique of Information Criteria (IC). The primary objective of IC is to apply analytical

techniques that could minimize overfitting. Several IC exist that are based on different statistical assumptions. The current study utilizes the Akaike Information Criteria (AIC), which was developed by Akaike (1998), as a technique for comparing different models on a particular outcome. AIC was selected because it creates balance between overfitting and underfitting, which would substantially improve the performance of a given model. Although AIC is based on information theory, it could be perceived as a criterion that selects the model that has a good fit to the actual data with few parameters, under the heuristic approach. Mathematically it can be expressed as

$$\text{AIC} = -2(ln(ll)) + 2k, \tag{10}$$

where ll is the likelihood of the data given a model and k is the number of free parameters in the model, which is often referred to as degrees of freedom or the number of estimable parameters. According to Leeb and Pötscher (2009), the constant two remains due to historical reasons. AIC scores, ΔAIC is the difference between the best model (with the smallest AIC) and each of the models. Consequently, the model that has a zero ΔAIC should be selected.

The second order information criterion (AIC_C) which takes sample size into consideration, as well as maximizing the model complexity when working with small data sets could also be used.

$$\text{AIC}_C = AIC + \frac{2k(k+1)}{n-k-1}, \tag{11}$$

where n is the sample size and k is the degrees of freedom. AIC is sufficient when the value of n is substantially larger than k. The primary advantage of the AIC_C is that it has high generality, which makes most of the data scientists to prefer using it over the AIC. However, the primary weakness of both IC scores is that they are ordinal; thus, do not have standalone meaning.

3.5 Simulation

In order to compare the performance of the Pull and Grab model and the splines, a simulation study was conducted. It involves empirical computer studies where real-life events are approximated by a given model (Freeman and Avery, 2017). The data to be utilized in the simulation is obtained from the available data using pseudo-random sampling. The validity and reliability of simulation is high because some of the parameters of interest are known and utilized in the data generation process. Finally, modelling could be used to determine the bias of a given model by determining the degree it conforms to the available life data. Therefore, modelling is used to circumnavigate the problem of generalization that is based on too few or too many trials.

The primary advantage of the technique is that it reduces the cost used in studies since random numbers are generated which eliminates the expenses of data collection. Secondly, it consumes less time than the traditional data collection technique that allows researchers to run more trials that could increase the accuracy of the model. Thirdly, simulation results could accurately approximate the real-life if the hypotheses are properly formulated (Freeman and Avery, 2017). Finally, simulations could identify high-risk areas that could necessitate further testing. The primary weakness of the model is that it cannot perfectly reflect all the dynamic factors that could influence the conclusion of the simulation mainly due to the application of few data points or bias of the researcher. The weakness could be overcome by working with relatively large data points and quantifying the uncertainty of the model and the high-risk areas that involve outlying values.

The precision of simulation could be undermined by several factors. For instance, the choice of the data generation model (DGM) has to be consistent with the statistical model and the issue being investigated. Additionally, the sample size used to generate data for simulation could influence the quality of the simulation. According to Pateras et al. (2018), the choice of DGM determines the results of the simulation studies involving few trials to a greater extent than those that have many data points. Consequently, the precision of the simulation model is dependent on the quality of planning, coding, analysis and interpretation of the results.

In the current study, the Pull and Grab model was utilized to simulate 1000 data sets. The probability distribution to be utilized in data generation process was computed employing Pull and Grab equations. At initial stage, the analysis of data utilizing Pull and Grab model resulted in parameter estimates which were then plugged in the equations to ensure high validity and reliability of the simulation study. The average MSE, AIC and standardized MSE values with their corresponding 95% confidence interval were computed.

4 Results and Interpretation

4.1 Exploratory Data Analysis

Data exploration is the initial step in analysis. It is conducted to provide an insight about the data structure and most plausible implications to be considered during model fitting (Perer and Shneiderman, 2008). The current study utilized community cross-sectional studies that involved 3352 individuals. The proportion of the population in terms of the age of the group is summarized in the graph contained in Figure 3. From the figure, it can be deduced that the largest age group was $16 - 100$ years, which consisted of 1566 participants. The second most populous age group was $5 - 15$ years, which comprised of 1104 individuals. The smallest age

group was $0-4$ years which consisted of 682 children. Therefore, the current study included participants who belonged to different age groups.

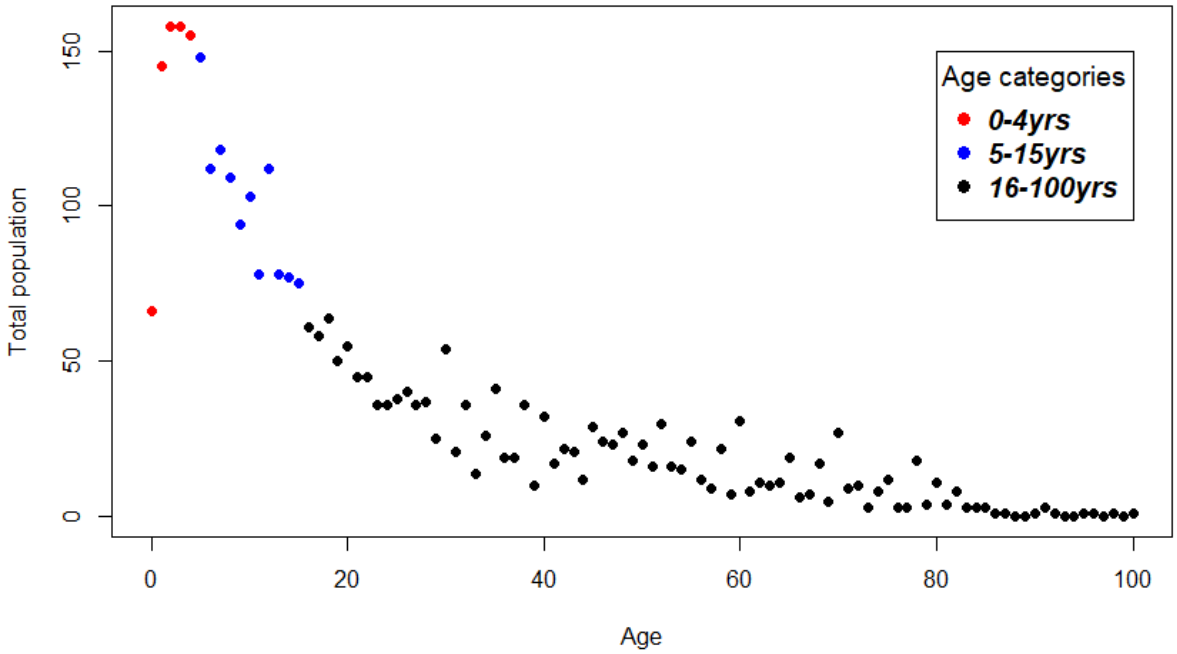

Figure 3: Population in each age category

Figure 4 provides a scatter diagram of the malaria prevalence for average age. From the figure, it can be deduced that from age $0-4$ years, the prevalence of malaria initially increases at a slow rate before increasing exponentially from age 5 years. Thereafter, the prevalence decreases substantially before plateauing in adults. The scatter plot also indicates that the relationship between the malaria prevalence and age is a non-linear one.

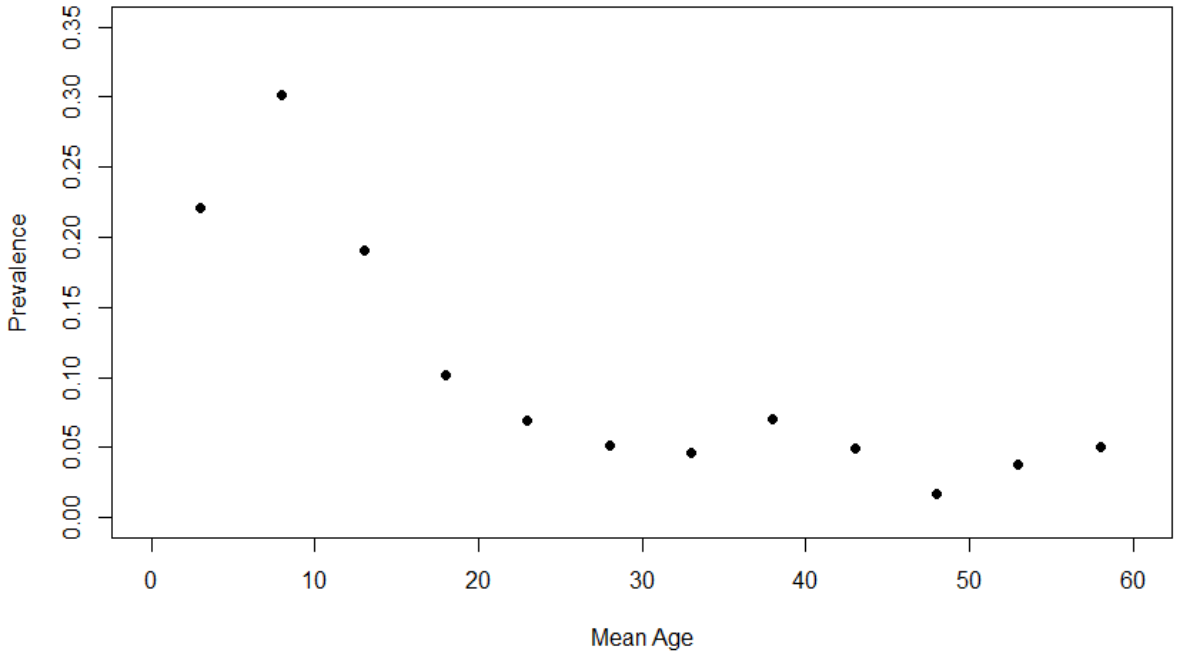

Figure 4: Relationship of prevalence and age

4.2 Splines and Pull and Grab Model

As pointed out earlier, the first step of performing spline regression is settling on the number of knots that will be used in training and testing the spline model. In the current study, 5 years and 15 years were used in developing the spline function. This was biologically motivated by the fact that malaria prevalence rises during infancy and early childhood settles to a plateau in older children, and declines in adolescents and adults as malaria immunity develops. The primary intention is to place the knots where the function seems to change rapidly. Therefore, 5 and 15 years were selected since the prevalence coefficient could rapidly vary in such regions.

The maximum likelihood estimation technique was utilized in fitting the Pull and Grab model. The parameter α, which is the age when PfPR begins to decline was estimated as (2.0936×10^{-7}), which is very insignificant. A likelihood ratio test was performed to determine if it was possible to set the limit value of α to zero. Therefore, the hypothesis test was conducted involving the following null and alternative hypotheses.

$$H_0 : \alpha = 0$$

$$H_1 : \alpha \neq 0$$

The p-value that was obtained at 5% level of significance was statistically insignificant. Therefore, it was not possible to reject the null hypothesis, hence leading to the conclusion

that the value of α could be set to zero. Nonetheless, it was possible to estimate the other parameters, and the results are summarized in Table 2.

Table 2: Parameter Estimation

Parameter	
ratio.r.h	1.4273×10^{-7}
d	0.1527
q	0.1317
$1 - s$	0.0416

Figure 5 visualizes the relationship between age and the malaria prevalence using splines (blue) and Pull and Grab model (red). As it can be viewed from the figure, both models captures the association between the prevalence of malaria and age effectively. It can be deduced from the plot that the prevalence rises during early childhood and infancy settles to plateau in older children, and declines in adolescents and adults.

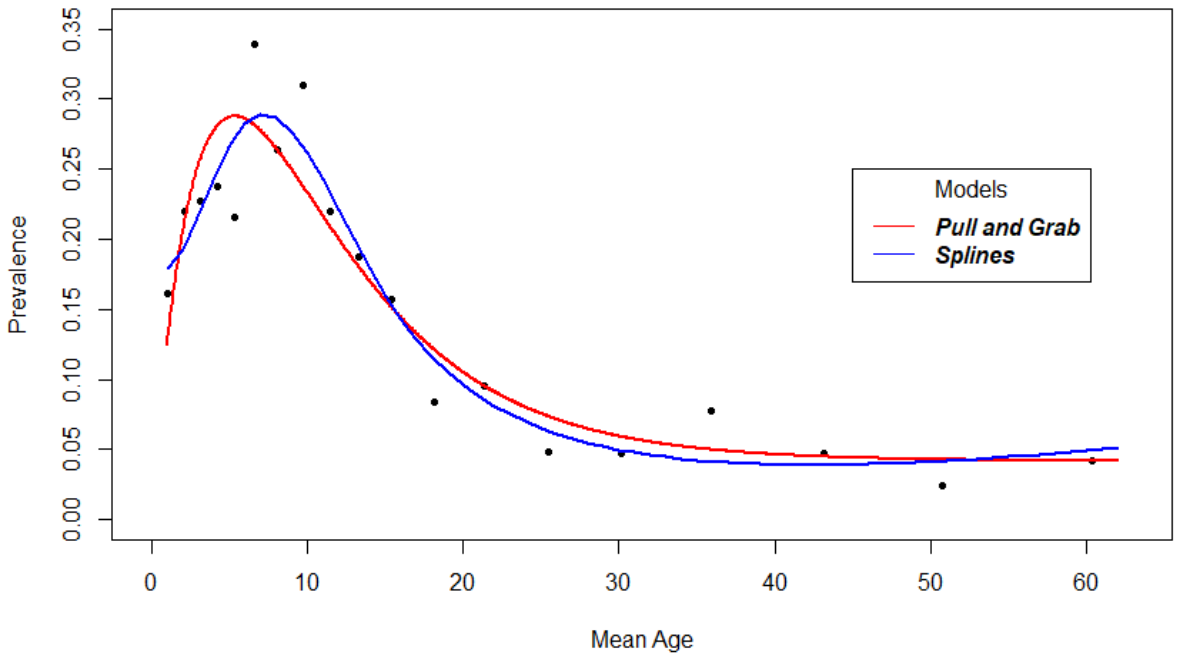

Figure 5: Relationship of prevalence and age using splines and Pull and Grab model

4.3 Model Validation and Selection

Out of the three methods of cross-validation discussed in section 3.4, the K-fold was utilized due to two reasons. It was more effective method for estimating accuracy of a given model and less tasking than the LOOCV. Additionally, 10 folds were selected since they had been proved to be effective in enhancing the performance of models in several empirical studies. Due to the need for the model to be sensitive to outlying values, we utilized the Mean squared error (MSE) metric to quantify the overall quality of the model. The model with low MSE indicated that it had high performance. Akaike information criteria (AIC) was also utilized to compare the performance of several models and select the one that had optimal performance. Ideally, the model with the lowest AIC should be selected. Also, MSE was standardized by averaging the MSE of the different age groups. The age was binned by 5 years through age 15, and the remaining data were treated as one age group. Table 3 summarizes the validation results of different models. k is the number of parameters in each model. Based on the two criteria, the splines with 4 degrees of freedom was selected, while standardized MSE selected the Pull and Grab model.

Table 3: AIC and MSE values

Model	k	AIC	MSE	MSE Standardized
Splines				
df=1	4	2691.304	0.1238	0.1509
df=2	4	2691.304	0.1238	0.1509
df=3	4	2691.304	0.1238	0.1509
df=4	5	**2651.715**	**0.1219**	0.1513
df=5	6	2657.412	0.1220	0.1518
Pull and Grab	4	2659.634	0.1221	**0.1504**

4.4 Simulation

The average AIC values and their corresponding 95% confidence intervals of the simulated data sets were calculated for splines with different degrees of freedom and Pull and Grab model. The results of the computation are summarized in table 4. The analysis of the results indicated that Pull and Grab model had the lowest AIC value, hence it had high performance. Consequently, at 95% confidence interval there was an overlap with the splines with 5 degrees of freedom which had the lowest AIC, indicating that there is no statistically significant difference between the two models.

Table 4: AIC values

Model	k	AIC	95% CI
Splines			
df=1	4	2683.493	(2550.391, 2816.626)
df=2	4	2683.493	(2550.391, 2816.626)
df=3	4	2683.493	(2550.391, 2816.626)
df=4	5	2646.585	(2512.649, 2777.214)
df=5	6	2644.440	(2512.194, 2779.037)
Pull and Grab	4	**2642.726**	**(2512.184, 2775.096)**

The MSE and standardized MSE values were computed to determine how well the different models fitted the given data. The average MSE and standardized MSE values with their corresponding 95% confidence interval for the 1000 simulated data sets are summarized in table 5. The analysis of the results indicates that the Pull and Grab model had the lowest MSE and MSE standardized values, hence it had high performance. However, at 95% confidence interval there was an overlap with the splines with lower MSE values that is splines with 1, 2 or 3 degrees of freedom. Therefore, it can be concluded that there is no statistically significant difference between the two models.

Table 5: MSE and MSE Standardized values

Model	MSE (95% CI)	MSE Standardized
Splines		
df=1	0.1363 (0.1100, 0.1628)	0.1506
df=2	0.1363 (0.1100, 0.1628)	0.1506
df=3	0.1363 (0.1100, 0.1628)	0.1506
df=4	0.1381 (0.1116, 0.1648)	0.1509
df=5	0.1383 (0.1118, 0.1651)	0.1511
Pull and Grab	**0.1214 (0.1141, 0.1288)**	**0.1483**

5 Discussion and Conclusion

The primary aim of conducting the research was to compare nonlinear models for predicting the relationship between malaria prevalence and age. Data were fitted into the Pull and Grab model and the generalized linear model that utilized splines as the basis function. The results indicated that both models captured the relationship between the prevalence of malaria and age effectively. The stability and the generality of the two models to new data were assessed

using the cross validation technique of k-fold. A 10 fold was utilized because past studies have verified that it has high validity and reliability (Hastie and Tibshirani, 1987). The primary reason for selecting MSE was to quantify the variation between the predicted and observed values, which indicates the overall quality of the performance of the selected model. Additionally, MSE is sensitive to outlying values. On the other hand, AIC was utilized to compare the performance of the models and select the one that had optimal performance. The standardized MSE was calculated by averaging the difference between the predicted and observed values for different age groups. Age was binned by 5 years through age 15, and the remaining data were treated as one age group. The primary motivation for standardizing MSE was to ease the comparison of prevalence of malaria between different ages and regions. The AIC and MSE criteria selected splines with 4 degrees of freedom while standardized MSE selected the Pull and Grab model.

In the comparison between the performance of the splines and the Pull and Grab, the latter model was used to simulate 1000 data sets. AIC, MSE and standardized MSE values were calculated with their corresponding 95% confidence interval. The three criteria indicated that the Pull and Grab model had the optimal performance. However, it is crucial to note that there was an overlap in confidence intervals hence concluding that there was no significant difference between the two models. In conclusion, both the Pull and Grab model and the splines could be used to model the relationship between the malaria prevalence and age, since even when the data were simulated utilizing Pull and Grab model, splines performed well.

The findings of the study contribute to the comprehension of how prevalence of malaria varies with age. The prevalence increases during infancy and early childhood because at this stage they are more vulnerable due to the fact that it is their first infection. It continues to increase up to a peak level where it remains fairly constant before the onset of adolescence. This trend however starts declining as age increases into adolescents and adults. The reason for such decline is that as the children grow older they develop immunity to the disease, that is they become less susceptible to the disease. Prior studies have failed to quantify the relationship between the two variables using high performance models such as splines. Therefore, the current study could assist public health planners and practitioners in estimating the number of people that are likely to be infected with malaria in a given community once their age is known. Consequently, adequate resources could be utilized in the prevention and management of malaria, which could reduce the burden of the disease. Finally, it is recommendable for a researcher to use splines with appropriate degrees of freedom since it performed well even when the data were simulated utilizing Pull and Grab model. Splines could be chosen over Pull and Grab model due to the fact that for the latter a lot of computation is needed in coding differential equations for each variable included in the analysis.

However, there is need to conduct studies that focus on other factors that influence the prevalence of malaria, such as net use, access to effective anti-malarial drugs and social economic status. Such studies could explain some of the residual variations that remained unaccounted for at the conclusion of the current study.

References

Akaike, H. (1998), Information theory and an extension of the maximum likelihood principle, *in* 'Selected papers of hirotugu akaike', Springer, pp. 199–213.

Alpaydin, E. (2009), *Introduction to machine learning*, MIT press.

Alpaydin, E. (2014), *Introduction to machine learning*, MIT press.

Ayele, D. G., Zewotir, T. T. and Mwambi, H. G. (2012), 'Prevalence and risk factors of malaria in ethiopia', *Malaria Journal* **11**(1), 195.

Ayele, D. G., Zewotir, T. T. and Mwambi, H. G. (2014), 'Semiparametric models for malaria rapid diagnosis test result', *BMC public health* **14**(1), 31.

Baird, J. K., Jones, T. R., Danudirgo, E. W., Annis, B. A., Bangs, M. J., Basri, P. H. and Masbar, S. (1991), 'Age-dependent acquired protection against plasmodium falciparum in people having two years exposure to hyperendemic malaria', *The American journal of tropical medicine and hygiene* **45**(1), 65–76.

Beadle, C., McElroy, P. D., Oster, C. N., Beier, J. C., Oloo, A. J., Onyango, F. K., Chumo, D. K., Bales, J. D., Sherwood, J. A. and Hoffman, S. L. (1995), 'Impact of transmission intensity and age on plasmodium falciparum density and associated fever: implications for malaria vaccine trial design', *Journal of Infectious Diseases* **172**(4), 1047–1054.

Bloland, P. B., Organization, W. H. et al. (2001), Drug resistance in malaria, Technical report, Geneva: World Health Organization.

Brooker, S., Kolaczinski, J. H., Gitonga, C. W., Noor, A. M. and Snow, R. W. (2009), 'The use of schools for malaria surveillance and programme evaluation in africa', *Malaria journal* **8**(1), 231.

Carlos, M., Elizabeth, A., Elizabeth, L., George, O., Juddy, K., Jukes, M. C., Katherine, E., Kiambo, N., Margaret, M., Simon, J. et al. (2014), 'Impact of intermittent screening and treatment for malaria among school children in kenya: a cluster randomized trial', *Policy Research Working Paper Series* .

Chai, T. and Draxler, R. R. (2014), 'Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature', *Geoscientific model development* **7**(3), 1247–1250.

Craig, M. H., Snow, R. and le Sueur, D. (1999), 'A climate-based distribution model of malaria transmission in sub-saharan africa', *Parasitology today* **15**(3), 105–111.

Fana, S. A., Bunza, M. D. A., Anka, S. A., Imam, A. U. and Nataala, S. U. (n.d.), 'Prevalence and risk factors associated with malaria infection among pregnant women in a semi-urban community of north-western nigeria', *Infectious diseases of poverty* **4**.

Freeman, L. J. and Avery, K. M. (2017), 'Modeling and simulation validation'.

Gitonga, C. W., Karanja, P. N., Kihara, J., Mwanje, M., Juma, E., Snow, R. W., Noor, A. M. and Brooker, S. (2010), 'Implementing school malaria surveys in kenya: towards a national surveillance system', *Malaria journal* **9**(1), 306.

Goepp, V., Bouaziz, O. and Nuel, G. (2018), 'Spline regression with automatic knot selection', *arXiv preprint arXiv:1808.01770* .

Greenhouse, B., Ho, B., Hubbard, A., Njama-Meya, D., Narum, D. L., Lanar, D. E., Dutta, S., Rosenthal, P. J., Dorsey, G. and John, C. C. (2011), 'Antibodies to plasmodium falciparum antigens predict a higher risk of malaria but protection from symptoms once parasitemic', *Journal of Infectious Diseases* **204**(1), 19–26.

Gupta, S., Snow, R. W., Donnelly, C. and Newbold, C. (1999), 'Acquired immunity and postnatal clinical protection in childhood cerebral malaria', *Proceedings of the Royal Society of London. Series B: Biological Sciences* **266**(1414), 33–38.

Hastie, T. and Tibshirani, R. (1987), 'Generalized additive models: some applications', *Journal of the American Statistical Association* **82**(398), 371–386.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013), *An introduction to statistical learning*, Vol. 112, Springer.

Kouznetsov, R. et al. (1996), *Malaria: a manual for community health workers*, World Health Organization.

Leeb, H. and Pötscher, B. M. (2009), Model selection, *in* 'Handbook of Financial Time Series', Springer, pp. 889–925.

Ma, S., Racine, J. S. and Yang, L. (2015), 'Spline regression in the presence of categorical predictors', *Journal of Applied Econometrics* **30**(5), 705–717.

Machini, B., Waqo, E., Kizito, W., Edwards, J., Owiti, P. and Takarinda, K. (2016), 'Trends in outpatient malaria cases, following mass long lasting insecticidal nets (llin) distribution in epidemic prone and endemic areas of kenya', *East African Medical Journal* **93**(10), 10–15.

Mayor, A., Aponte, J. J., Fogg, C., Saúte, F., Greenwood, B., Dgedge, M., Menendez, C. and Alonso, P. L. (2007), 'The epidemiology of malaria in adults in a rural area of southern mozambique', *Malaria journal* **6**(1), 3.

Meng, X.-L. and Rubin, D. B. (1993), 'Maximum likelihood estimation via the ecm algorithm: A general framework', *Biometrika* **80**(2), 267–278.

Muench, H. (1959), *Catalytic models in epidemiology*, Vol. 2, Harvard University Press Cambridge, MA.

Murray, C. J., Rosenfeld, L. C., Lim, S. S., Andrews, K. G., Foreman, K. J., Haring, D., Fullman, N., Naghavi, M., Lozano, R. and Lopez, A. D. (2012), 'Global malaria mortality between 1980 and 2010: a systematic analysis', *The Lancet* **379**(9814), 413–431.

Noor, A. M., Kirui, V. C., Brooker, S. J. and Snow, R. W. (2009), 'The use of insecticide treated nets by age: implications for universal coverage in africa', *BMC Public Health* **9**(1), 369.

Okiro, E. A., Al-Taiar, A., Reyburn, H., Idro, R., Berkley, J. A. and Snow, R. W. (2009), 'Age patterns of severe paediatric malaria and their relationship to plasmodium falciparum transmission intensity', *Malaria journal* **8**(1), 4.

Organization, W. H. (2003), *Diet, nutrition, and the prevention of chronic diseases: report of a joint WHO/FAO expert consultation*, Vol. 916, World Health Organization.

Organization, W. H. et al. (2005), 'The world health report: 2005: make every mother and child count'.

Pateras, K., Nikolakopoulos, S. and Roes, K. (2018), 'Data-generating models of dichotomous outcomes: Heterogeneity in simulation studies for a random-effects meta-analysis', *Statistics in medicine* **37**(7), 1115–1124.

Perer, A. and Shneiderman, B. (2008), Integrating statistics and visualization: case studies of gaining clarity during exploratory data analysis, *in* 'Proceedings of the SIGCHI conference on Human Factors in computing systems', ACM, pp. 265–274.

Pull, J. and Grab, B. (1974), 'A simple epidemiological model for evaluating the malaria inoculation rate and the risk of infection in infants', *Bulletin of the World Health Organization* **51**(5), 507.

Rogers, D. J., Randolph, S. E., Snow, R. W. and Hay, S. I. (2002), 'Satellite imagery in the study and forecast of malaria', *Nature* **415**(6872), 710.

Ross, R. (1911), *The prevention of malaria*, John Murray; London.

Smith, D. L., Guerra, C. A., Snow, R. W. and Hay, S. I. (2007), 'Standardizing estimates of the plasmodium falciparum parasite rate', *Malaria journal* **6**(1), 131.

Snow, R., Craig, M., Deichmann, U. and Le Sueur, D. (1999), 'A preliminary continental risk map for malaria mortality among african children', *Parasitology today* **15**(3), 99–104.

Snow, R. W., Gouws, E., Omumbo, J., Rapuoda, B., Craig, M., Tanser, F., Le Sueur, D. and Ouma, J. (1998), 'Models to predict the intensity of plasmodium falciparum transmission: applications to the burden of disease in kenya', *Transactions of the Royal Society of Tropical Medicine and Hygiene* **92**(6), 601–606.

Snow, R. W., Omumbo, J. A., Lowe, B., Molyneux, C. S., Obiero, J.-O., Palmer, A., Weber, M. W., Pinder, M., Nahlen, B., Obonyo, C. et al. (1997), 'Relation between severe malaria morbidity in children and level of plasmodium falciparum transmission in africa', *The Lancet* **349**(9066), 1650–1654.

Stuckey, E. M., Stevenson, J. C., Cooke, M. K., Owaga, C., Marube, E., Oando, G., Hardy, D., Drakeley, C., Smith, T. A., Cox, J. et al. (2012), 'Simulation of malaria epidemiology and control in the highlands of western kenya', *Malaria journal* **11**(1), 357.

Wahba, G. (1990), *Spline models for observational data*, Vol. 59, Siam.

Wang, H. and Zheng, H. (2013), *Model Validation, Machine Learning*, Springer New York, New York, NY, pp. 1406–1407.
URL: *https://doi.org/10.1007/978-1-4419-9863-7_233*

Wood, S. N. (2017), *Generalized additive models: an introduction with R*, Chapman and Hall/CRC.

6 SOFTWARE CODES

```
###Fitting binary logistic regression model###;
model1=glm(I(RDT)~bs(Age,knots = c(5,15)),data=thesis,family = "binomial")
probability<-predict(model1,newdata=data.frame(Age=1:62),type="response")
A <- thesis$Age
y <- thesis$RDT
age.breaks <- quantile(A,seq(0,1,0.05))
age.class <- cut(A,breaks=age.breaks)
prop.class <- tapply(y,age.class,mean)
age.mean.class <- tapply(A,age.class,mean)
plot(age.mean.class,prop.class,pch=20,
     xlim=c(0,63),ylim=c(0,0.35),
     xlab="Mean Age",ylab="Prevalence")
```

```
lines(1:62,probability,type="l",col="red",lwd=2)

###Fitting Pull and Grab Model###
A <- thesis$Age
y <- thesis$RDT

llik <- function(theta) {
  ratio.r.h <- exp(theta[1])
  b <- exp(theta[2])
  c <- exp(theta[3])
  s <- exp(theta[4])/(1+exp(theta[4]))
  P.A <- (1/(1+ratio.r.h))*(1-exp(-b*A))
  F.A <- 1-s*(1-sapply(A,function(x) min(1,exp(-c*x))))
  prob <- F.A*P.A
  out <- sum(y*log(prob/(1-prob))+log(1-prob))
  return(out)
}
theta.start <- rep(0,4)
estim<- nlminb(theta.start, function(x) -llik(x),
               control=list(trace=1))
p.A <- function(A,theta) {
  ratio.r.h <- exp(theta[1])
  b <- exp(theta[2])
  c <- exp(theta[3])
  s <- exp(theta[4])/(1+exp(theta[4]))
  P.A <- 1/(1+ratio.r.h)*(1-exp(-b*A))
  F.A <- 1-s*(1-min(1,exp(-c*(A))))
  prob <- F.A*P.A
}
p.A <- Vectorize(p.A,"A")
age.breaks <- quantile(A,seq(0,1,0.05))
age.class <- cut(A,breaks=age.breaks)
prop.class <- tapply(y,age.class,mean)
age.mean.class <- tapply(A,age.class,mean)
plot(age.mean.class,prop.class,pch=20,
     xlim=c(0,63),ylim=c(0,0.35),xlab="Mean Age", ylab="Prevalence")
age.set <- seq(1,62,length=1000)
p.A.set <- p.A(age.set,estim$par)
```

```
lines(age.set,p.A.set,col=2,lwd=2)

###K-FOLD CROSS VALIDATION for splines###
##Splines with df=i vary the df from 1 to 5###
set.seed(1993)
n <- nrow(thesis)
K <- 10
block.ID <- rep(1:K,each=n/K)
block.ID <- block.ID[sample(1:n)]
MSE1=matrix(NA,K)
for(i in 1:K){
  ind.block <- which(block.ID==i)
  training <- thesis[-ind.block,]
  test <- thesis[ind.block,]
  glmfit1=glm(I(RDT)~bs(Age,df=i),data=training,family = "binomial")
  predict1 <- predict(glmfit1,newdata=test,type="response")
  MSE1[i]=mean((test$RDT-predict1)^2)
}
mean(MSE1)
###K-FOLD CROSS VALIDATION FOR PULL AND GRAB MODEL###
set.seed(1993)
n <- nrow(thesis)
K <- 10
block.ID <- rep(1:K,each=n/K)
block.ID <- block.ID[sample(1:n)]
theta.hat.i <- matrix(NA,ncol = 4,nrow = K)
MSE=matrix(NA,K)
for(i in 1:K){
  ind.block <- which(block.ID==i)
  training <- thesis[-ind.block,]
  test <- thesis[ind.block,]
  # Fit Pull and Grab with training
  A <- training$Age
  y <- training$RDT
  llik <- function(theta) {
    ratio.r.h <- exp(theta[1])
    b <- exp(theta[2])
    c <- exp(theta[3])
```

```
    s <- exp(theta[4])/(1+exp(theta[4]))
    P.A <- (1/(1+ratio.r.h))*(1-exp(-b*A))
    F.A <- 1-s*(1-sapply(A,function(x) min(1,exp(-c*x))))
    prob <- F.A*P.A
    out <- sum(y*log(prob/(1-prob))+log(1-prob))
    return(out)
  }
  theta.start <- rep(0,4)
  estim.i <- nlminb(theta.start, function(x) -llik(x),
                    control=list(trace=1))
  theta.hat.i[i,] <- estim.i$par

}
# Predict the test
p.A <- function(A,theta) {
  ratio.r.h <- exp(theta[1])
  b <- exp(theta[2])
  c <- exp(theta[3])
  s <- exp(theta[4])/(1+exp(theta[4]))
  P.A <- 1/(1+ratio.r.h)*(1-exp(-b*A))
  F.A <- 1-s*(1-pmin(1,exp(-c*(A))))
  prob <- F.A*P.A
  MSE <- mean((test$RDT-prob)^2) # MSE
  MSE
}

MSE <- rep(NA,K)
for(i in 1:K){
  ind.block <- which(block.ID==i)
  test <- thesis[ind.block,]
  MSE[i] <- p.A(test$Age,theta.hat.i[i,])
}
mean(MSE)

# LRT for alpha equals zero
A <- thesis$Age
y <- thesis$RDT
llik <- function(theta) {
```

```
  ratio.r.h <- exp(theta[1])
  b <- exp(theta[2])
  c <- exp(theta[3])
  s <- exp(theta[4])/(1+exp(theta[4]))
  alpha <- exp(theta[5])
  P.A <- 1/(1+ratio.r.h)*(1-exp(-b*A))
  F.A <- 1-s*(1-sapply(A,function(x) min(1,exp(-c*(x-alpha)))))
  prob <- F.A*P.A
  out <- sum(y*log(prob/(1-prob))+log(1-prob))#positive
  return(out)
}
theta.start <- rep(0,5)
estim2 <- nlminb(theta.start, function(x) -llik(x),
                 control=list(trace=1))


llik0 <- function(theta) {
  ratio.r.h <- exp(theta[1])
  b <- exp(theta[2])
  c <- exp(theta[3])
  s <- exp(theta[4])/(1+exp(theta[4]))
  P.A <- (1/(1+ratio.r.h))*(1-exp(-b*A))
  F.A <- 1-s*(1-sapply(A,function(x) min(1,exp(-c*x))))
  prob <- F.A*P.A
  out <- sum(y*log(prob/(1-prob))+log(1-prob))
  return(out)
}
theta.start <- rep(0,4)
estim0 <- nlminb(theta.start, function(x) -llik0(x),
                  control=list(trace=1))

1-pchisq(-2*(estim2$objective-estim0$objective),1) #p-value from LR test

###CROSS VALIDATION FOR DIFFERENT AGE RANGES###
##SPLINES WITH df=i vary the df from 1 to 5 for all age ranges###
age1<-thesis[which(thesis$Age>=0& thesis$Age<=4.9999),]
age2<-thesis[which(thesis$Age>=5& thesis$Age<=9.9999),]
age3<-thesis[which(thesis$Age>=10& thesis$Age<=14.9999),]
```

```
age4<-thesis[which(thesis$Age>=15& thesis$Age<=101.9999),]
set.seed(1993)
n <- nrow(age1)
K <- 10
block.ID <- rep(1:K,each=n/K)
block.ID <- block.ID[sample(1:n)]
MSE_age1=matrix(NA,K)
for(i in 1:K){
  ind.block <- which(block.ID==i)
  training <- age1[-ind.block,]
  test <- age1[ind.block,]
  glmfit1=glm(I(RDT)~bs(Age,df=i),data=training,family = "binomial")
  predict1 <- predict(glmfit1,newdata=test,type="response")
  MSE_age1[i]=mean((test$RDT-predict1)^2)
}
mean(MSE_age1)

###Pull and grab validation using different age ranges###
set.seed(1993)
n <- nrow(age1)
K <- 10
block.ID <- rep(1:K,each=n/K)
block.ID <- block.ID[sample(1:n)]
theta.pull.i <- matrix(NA,ncol = 4,nrow = K)
MSEpull1=matrix(NA,K)
for(i in 1:K){
  ind.block <- which(block.ID==i)
  training <- age1[-ind.block,]
  test <- age1[ind.block,]
  # Fit Pull and Grab with training
  A <- training$Age
  y <- training$RDT
  llik <- function(theta) {
    ratio.r.h <- exp(theta[1])
    b <- exp(theta[2])
    c <- exp(theta[3])
    s <- exp(theta[4])/(1+exp(theta[4]))
    P.A <- (1/(1+ratio.r.h))*(1-exp(-b*A))
```

```
    F.A <- 1-s*(1-sapply(A,function(x) min(1,exp(-c*x))))
    prob <- F.A*P.A
    out <- sum(y*log(prob/(1-prob))+log(1-prob))
    return(out)
  }
  theta.start <- rep(0,4)
  estimpull.i <- nlminb(theta.start, function(x) -llik(x),
                    control=list(trace=1))
  theta.pull.i[i,] <- estimpull.i$par

}
# Predict the test
p.A <- function(A,theta) {
  ratio.r.h <- exp(theta[1])
  b <- exp(theta[2])
  c <- exp(theta[3])
  s <- exp(theta[4])/(1+exp(theta[4]))
  P.A <- 1/(1+ratio.r.h)*(1-exp(-b*A))
  F.A <- 1-s*(1-pmin(1,exp(-c*(A))))
  prob <- F.A*P.A
  MSEpull1 <- mean((test$RDT-prob)^2) # MSE
  MSEpull1
}

MSEpull1 <- rep(NA,K)
for(i in 1:K){
  ind.block <- which(block.ID==i)
  test <- age1[ind.block,]
  MSEpull1[i] <- p.A(test$Age,theta.pull.i[i,])
}
mean(MSEpull1)

###SIMULATION STUDY###
###Calculating probability using the pull and grab equations
P.A <- (1/(1+ratio.r.h))*(1-exp(-b*A))
F.A <- 1-s*(1-sapply(A,function(x) min(1,exp(-c*x))))
prob <- F.A*P.A
###P.A1
```

```
ratio.r.h <- 1.4273* 10^(-7)
b <- 0.1527
P.A1 <- (1/(1+ratio.r.h))*(1-exp(-b*A))
###FA
c<-0.1317
s<-0.9584
F.A1 <- 1-s*(1-sapply(A,function(x) min(1,exp(-c*x))))
probab<-P.A1*F.A1

y<-rbinom(nrow(thesis),1,probab)

set.seed(19902986)
x<-replicate(1000,{y<-rbinom(nrow(thesis),1,probab)})
z<-data.frame(x)
A<-thesis$Age
sim<-cbind(z,A)

###Fitting models using simulated data###
###Splines with different degrees of freedom vary df=i from 1 to 5###
N<-1000
AIC1=c(1:N)
for (i in 1:N) {
model2=glm(I(sim[,i])~bs(sim$A,df=i),family = "binomial")
AIC1[i]<-model2$aic
}
hist(AIC1)
mean(AIC1)
(conf1<-quantile(AIC1,probs=c(0.025,0.975)))

##Pull and Grab model
N<-1000
A <- thesis$Age
aic.hat.i<-c(1:N)
for(i in 1:N){
llik <- function(theta) {
  ratio.r.h <- exp(theta[1])
  b <- exp(theta[2])
  c <- exp(theta[3])
```

```
  s <- exp(theta[4])/(1+exp(theta[4]))
  P.A <- (1/(1+ratio.r.h))*(1-exp(-b*A))
  F.A <- 1-s*(1-sapply(A,function(p) min(1,exp(-c*p))))
  prob <- F.A*P.A
  out <- sum((x[,i])*log(prob/(1-prob))+log(1-prob))
  return(out)
}
theta.start <- rep(0,4)
simulated.i<- nlminb(theta.start, function(p) -llik(p),
               control=list(trace=1))
aic.hat.i[i]<-simulated.i$objective
}
hist(AIC)
mean(aic.hat.i)
AIC<-(aic.hat.i*2)+8
mean(AIC)
(conf<-quantile(AIC,probs=c(0.025,0.975)))

###K-FOLD CROSS VALIDATION for splines###
##Splines with df=i vary i from 1 to 5###
set.seed(1993)
n <- nrow(sim)
N<-1000
K <- 10
block.ID <- rep(1:K,each=n/K)
block.ID <- block.ID[sample(1:n)]
MSE2=matrix(NA,ncol=N,nrow=K)
for (j in 1:N) {
for(i in 1:K){
  ind.block <- which(block.ID==i)
  training <- sim[-ind.block,]
  test <- sim[ind.block,]
  glmfit1=glm(I(sim[,j])~bs(sim$A,df=i),data=training,family = "binomial")
  predict1 <- predict(glmfit1,newdata=test,type="response")
  MSE2[i,j]=mean(((test[,j])-predict1)^2)

}
}
```

```
mean(MSE2)
(confint1<-quantile(MSE2,probs=c(0.025,0.975)))
###MSE standardized for different splines
##SPLINES WITH df=i vary i from 1 to 5 for all age ranges###
set.seed(1993)
n <- nrow(age1)
N<-1000
K <- 10
block.ID <- rep(1:K,each=n/K)
block.ID <- block.ID[sample(1:n)]
MSE11=matrix(NA,ncol=N,nrow=K)
for (j in 1:N) {
  for(i in 1:K){
    ind.block <- which(block.ID==i)
    training <- age1[-ind.block,]
    test <- age1[ind.block,]
    glmfit1=glm(I(age1[,j])~bs(age1$A,df=i),data=training,family = "binomial")
    predict1 <- predict(glmfit1,newdata=test,type="response")
    MSE11[i,j]=mean(((test[,j])-predict1)^2)
  }
}
mean(MSE11)
(confint11<-quantile(MSE11,probs=c(0.025,0.975)))

###PULL AND GRAB MODEL VALIDATION
set.seed(1993)
n <- nrow(sim)
K <- 10
N<-1000
a<-list()
block.ID <- rep(1:K,each=n/K)
block.ID <- block.ID[sample(1:n)]
theta.hat2.i <- data.frame()
MSE=matrix(NA,N)
MSEK=matrix(NA,nrow=K,ncol=N)
for (j in 1:N) {
  for(i in 1:K){
    ind.block <- which(block.ID==i)
```

```
  training <- sim[-ind.block,]
  test <- sim[ind.block,]
  # Fit Pull and Grab with training
  A <- training$A
  y <- training[,j]
  llik <- function(theta) {
    ratio.r.h <- exp(theta[1])
    b <- exp(theta[2])
    c <- exp(theta[3])
    s <- exp(theta[4])/(1+exp(theta[4]))
    P.A <- (1/(1+ratio.r.h))*(1-exp(-b*A))
    F.A <- 1-s*(1-sapply(A,function(x) min(1,exp(-c*x))))
    prob <- F.A*P.A
    out <- sum(y*log(prob/(1-prob))+log(1-prob))
    return(out)
  }
  theta.start <- rep(0,4)
  estim.2 <- nlminb(theta.start, function(x) -llik(x),
                    control=list(trace=1))
  a[[1]]<- estim.2$par
  list<-t(sapply(a, unlist))
  theta2 <-data.frame(list)
  theta.hat2.i<-rbind(theta.hat2.i,theta2)
  p.A <-function(A,theta) {
    theta <- as.numeric(theta)
    ratio.r.h <- exp(theta[1])
    b <- exp(theta[2])
    c <- exp(theta[3])
    s <- exp(theta[4])/(1+exp(theta[4]))
    P.A <- 1/(1+ratio.r.h)*(1-exp(-b*A))
    F.A <- 1-s*(1-pmin(1,exp(-c*(A))))
    prob <- F.A*P.A
    MSE <- mean((test[,j]-prob)^2) # MSE
    return(MSE)
  }
  MSEK[i,j] <- p.A(test$A, theta2)
}
MSE[j]<-mean(MSEK[,j])
```

```
}
mean(MSE)
(ci<-quantile(MSE,probs=c(0.025,0.975)))

###Pull and Grab for all different age ranges
set.seed(1993)
n <- nrow(age1)
K <- 10
N<-1000
a<-list()
block.ID <- rep(1:K,each=n/K)
block.ID <- block.ID[sample(1:n)]
theta.hat2.i <- data.frame()
MSEA1=matrix(NA,N)
MSEA1K=matrix(NA,nrow=K,ncol=N)
for (j in 1:N) {
  for(i in 1:K){
    ind.block <- which(block.ID==i)
    training <- age1[-ind.block,]
    test <- age1[ind.block,]
    # Fit Pull and Grab with training
    A <- training$A
    y <- training[,j]
    llik <- function(theta) {
      ratio.r.h <- exp(theta[1])
      b <- exp(theta[2])
      c <- exp(theta[3])
      s <- exp(theta[4])/(1+exp(theta[4]))
      P.A <- (1/(1+ratio.r.h))*(1-exp(-b*A))
      F.A <- 1-s*(1-sapply(A,function(x) min(1,exp(-c*x))))
      prob <- F.A*P.A
      out <- sum(y*log(prob/(1-prob))+log(1-prob))
      return(out)
    }
    theta.start <- rep(0,4)
    estim.2 <- nlminb(theta.start, function(x) -llik(x),
                  lower = 0, upper = Inf,
                      control=list(trace=1))
```

```
    a[[1]]<- estim.2$par
    list<-t(sapply(a, unlist))
    theta2 <-data.frame(list)
    theta.hat2.i<-rbind(theta.hat2.i,theta2)
    p.A <-function(A,theta) {
      theta <- as.numeric(theta)
      ratio.r.h <- exp(theta[1])
      b <- exp(theta[2])
      c <- exp(theta[3])
      s <- exp(theta[4])/(1+exp(theta[4]))
      P.A <- 1/(1+ratio.r.h)*(1-exp(-b*A))
      F.A <- 1-s*(1-pmin(1,exp(-c*(A))))
      prob <- F.A*P.A
      MSEA1 <- mean((test[,j]-prob)^2) # MSE
      return(MSEA1)
    }
    MSEA1K[i,j] <- p.A(test$A, theta2)
  }
  MSEA1[j]<-mean(MSEA1K[,j])
}
mean(MSEA1)
(c2<-quantile(MSEA1,probs=c(0.025,0.975)))
```

7 Appendix

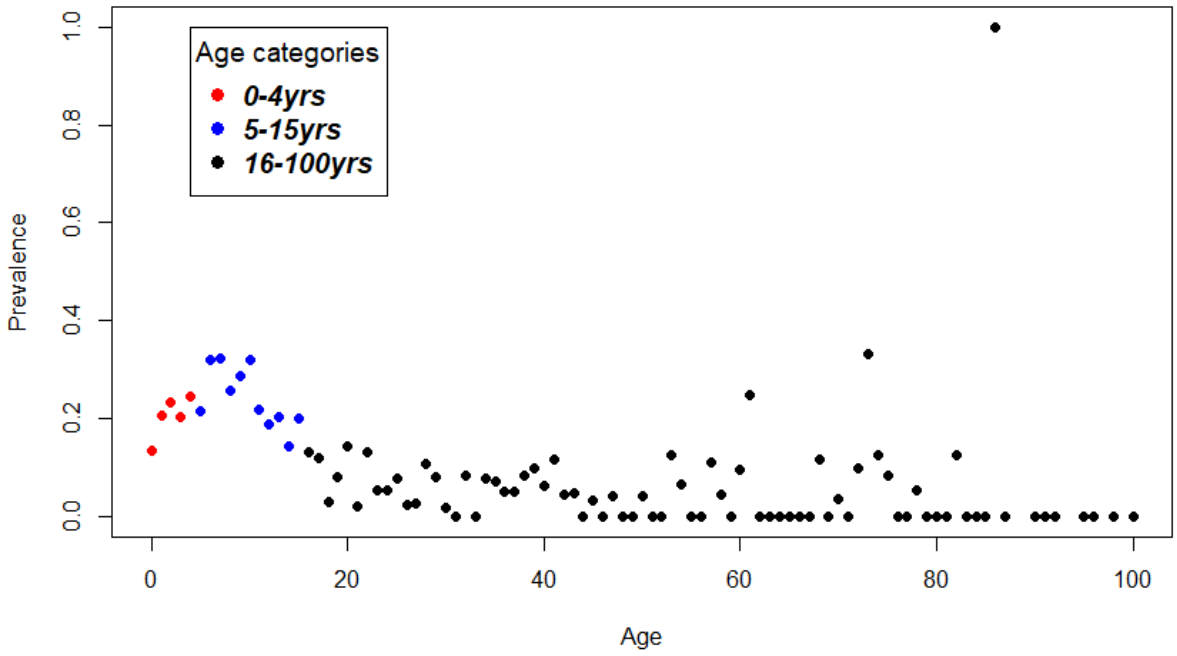

Figure 6: Relationship of prevalence and age