DESIGN-BASED APPROACH TO MODELLING OF MALARIA INCIDENCE IN NIGERIA.

Olawore, Sunday Adebayo[1,*] and Ekong, Aniefiok Henry[2]

[1]Oyo State College of Agriculture and Technology, Igboora, Oyo State.
[2]Federal Medical Centre Abeokuta, Ogun State.

[*]Correspondence email: sundayolawoore1978@gmail.com

Abstract

Most available datasets on malaria in Nigeria used in literatures, were limited to small localised population or small sample size of cohorts. Hence, inference were not scalable to the entire country. Interest in malaria modelling is on population estimates and in order to improve cost-effectiveness and representational accuracy in large-scale surveys, complex sampling designs such as multistage sampling with uneven selection probabilities are often employed. To avoid biased estimations, specific, weighted analytical techniques are applied to take into consideration complex data structures. Such techniques, like resampling or Taylor series linearization, give estimates about the population, instead of the sample as seen in model-based analysis. The literature on modelling malaria in Nigeria used survey datasets that were collected from complex survey designs, yet scarce literature accounted for the complex design in their analyses. The present study examined the effects of complex design in modelling malaria incidence in Nigeria, using dataset from the 2021 Nigeria Malaria Indicator Survey (NMIS). The effects of sampling weights and survey information in the analysis of malaria datasets were examined and compared with model-based approach. The analysis capturing the survey design produced a different result from the model-based approach, effectively showing the difference in the sample and population estimates, for which the latter is the object of surveys such as the 2021 NMIS. Hence, this paper adds to the literature on country-wise malaria incidence analysis and how survey design is incorporated in the analysis to meet the objective of such survey.

Keywords: complex design; design-based; malaria incidence; model-based; weights

1. Introduction

According to the World Health Organisation (WHO) report on malaria in Nigeria 2022, Nigeria had an estimated 68 million cases and 194,000 deaths due to the disease in 2021. Nigeria has the highest burden of malaria globally, accounting for nearly 27% of the global malaria burden. The risk of transmission exists throughout the country, all year round. According to the WHO report, despite the progress that Nigeria has achieved, the country is faced with several important challenges that hamper its fight against malaria (WHO, 2023). Of the many challenges, the focus of this research is on the statistical modelling interventions that have been hitherto carried out to help alleviate this disease, through prediction and surveillance.

Most available datasets on malaria used in most literatures, were limited to small localised population (patients within a local hospital or state), or small sample size of cohorts. Hence, the inferences were not scalable to the entire country. Large-scale interventions funded by international organizations, such the Nigeria Malaria Indicator Survey (2021 NMIS), which was conducted in 2021 with the support from the Demographic and Health Survey (DHS), are used to gather big datasets on malaria (NMEP, 2022). The aim of such a survey is to estimate characteristics of real, finite population, for example, the proportion of malaria incidence in the population of Nigeria. The sampling design is thus taken into account in inferences. Inference about these real-population parameters is termed design-based analysis (DBA). In contrast, if the object of inference is the process that gave rise to a real population, such inference is termed model-based analysis (MBA). The coefficients of a model fit from MBA, to the whole population are subject to statistical uncertainty (Fox and Weisberg, 2018).

For the 2021 NMIS, based on sampling probabilities, sampling weights were calculated independently for each stage and cluster in the two-stage stratified cluster survey. Another name for the sampling strategy is complex designs. In order to improve cost-effectiveness and representational accuracy in large-scale surveys, stratified, clustered, and multistage techniques with uneven selection probabilities are used as complex sampling designs. To avoid biased estimations, they need specific, weighted analytical techniques to take into consideration complex data structures, like resampling techniques or Taylor series linearization.

However, the modelling of Nigeria malaria incidences in the literature have not accounted for the sampling design in the analyses of malaria datasets, especially when the dataset was from the 2021 NMIS. Collins and Duffy (2022) used a system of ordinary differential equations to study the effective reproduction rate of disease in Nigeria. The study used a deterministic model on a time series values of incidences of malaria, without recourse to the longitudinal factors for the period the data was collected. The limitation of Collins and Duffy (2022) is that a deterministic model is best for well-defined, stable systems, whereas dynamics of malaria contains uncertainty, risk, and complex, real-world variability. Dambazau et al. (2025) also used a system of ordinary differential equations to study the effective reproduction rate of disease in Kano state, Nigeria. The paper also added to its limitation as reliance on localised secondary data, which introduced inaccuracies in parameter estimates.

Segun et al. (2020) fiited Possion and ARIMA models on the incidence of malaria and weather factors in Abuja Nigeria using time series dataset. Ogbuagada et al. (2022) used vector autoregressive (VAR) time series model to account for caases of malaria incidence in Jimeta metropolis of Adamawa State. Ozodiegwu et al. (2023) used EMOD v2.20, an agent-based model of P.falciparum transmission that comprises a model of temperature-dependent vector lifecycle and vector population dynamics, coupled to a model of human disease and immunity, and intervention effects to simulate malaria transmission and intervention impact. They used datasets from 2010 Malaria Indicator Survey (MIS), but without recourse to the survey sampling design.

Jimoh et al. (2024) used autoregressive integrated moving average (ARIMA) model to predict malaria incidents in the Northeastern zone of Nigeria, as time series dataset.

Okunlola et al. (2021) fitted a Bayesian binomial distribution on malaria prevalence using the Demographic and Health Survey (DHS) of Nigeria conducted in 2018. Isiko et al. (2024) fitted a logistic regression on the incidence of malaria using the 2021 NMIS dataset. Their aim was to identify factors associated with the risk of malaria. However, the complex sampling design was not accounted for in their analysis, resulting in the limitations of the interpretation of their results and not representative of the study population, nor valid for the survey purpose. Okunlola et al. (2024) fitted hierarchical beta regression to forecast the malaria incidence and assess the effects of livestock using the Nigerian Demographic and Health Survey's Geospatial Covariate Datasets. Akeboi et al. (2025) fitted a Bayesian hierarchical spatio-temporal Poisson model to DHIS2 under five malaria case data from 2014 to 2023 in Kano and Lagos States. The challenges with routine DHIS2 data are reporting completeness and data quality, as noted by Akeboi et al. (2025), and these can not be alleviated by high-frequency, geographically disaggregated information in the DHIS2 data as they claimed.

The literature on modelling malaria in Nigeria used survey datasets that were collected from complex survey designs, yet scarce literature accounted for the complex design in their analyses. The impact of complex sampling designs on data analysis results have been discussed in the literature on design-based modelling. Analysing stratified and unequal probability datasets as a simple random sample (SRS) dataset overestimates the standard errors, while taking a cluster sample analysis as an SRS dataset underestimates the standard errors (Lumley, 2004). According to Lumley (2004), it should be noted that while the design and misspecification effects do not agree (qualitatively) for uneven probability sampling, they do agree for stratification and clustering. The issue of design effect, where complex designs typically have higher variance than SRS for the same sample size also abound. The consequences of these is that inference from the malaria modelling studies in Nigeria cannot be generalised to the country and the estimates are biased.

The present study examined the effects of sampling design in modelling malaria incidence in Nigeria using large scale dataset from the 2021 NMIS. The effects of sampling weights and survey information in the analysis of malaria datasets from complex design surveys (DBA) were examined and compared with MBA where the survey information was not accounted for.

2. Material and method

Study sampling design of the motivating dataset

The 2021 Nigeria Malaria Survey Indicator was constructed by the National Malaria Elimination Programme assisted by the National Bureau of Statistics (NBS) and National Population Commission (NPC) was implemented by the Global Fund to Fight AIDS, TB and Malaria (GFATM). The 2021 NMIS used the sample frame for the proposed 2023 Population and Housing

Census (PHC) of the Federal Republic of Nigeria. Administratively, Nigeria is divided into states. Each state is subdivided into local government areas (LGAs), each LGA is divided into wards, and each ward is divided into localities. Localities are further subdivided into convenient areas called census enumeration areas (EAs). The primary sampling unit (PSU), referred to as a cluster unit for the 2021 NMIS, was defined on the basis of EAs for the proposed 2023 PHC (NMEP, 2022).

A two-stage sampling strategy was adopted for the 2021 NMIS. In the first stage, 568 EAs were selected with probability proportional to the EA size. The EA size is the number of households residing in the EA. The result was a total of 568 clusters throughout the country, 195 in urban areas and 373 in rural areas. A complete listing of households in these clusters was conducted between 26 August and 18 September 2021, with the resulting lists of households serving as the sampling frame for the selection of households in the second stage. GPS dongles were used to capture coordinates during household listing in the 2021 NMIS sample clusters.

In the second stage's selection process, 25 households were selected in each cluster via equal probability systematic sampling. The 2021 NMIS was also designed to provide baseline information on malaria parasite prevalence and other malaria indicators for Bonny Island, which lies in Rivers State. There is an effort to declare the island malaria free in the near future, and it was critical to have baseline information. Oversampling was done for Bonny Island to be able to measure a decline in rapid diagnostic test (RDT) malaria prevalence from 22.3% (the prevalence in Rivers State according to the 2018 Nigeria Demographic and Health Survey [NDHS]) to zero. A total of 25 clusters were selected from Rivers State, with eight clusters selected from Bonny Island and the remaining 17 from the other LGAs in the state.

Three questionnaires were used in the 2021 NMIS: the Household Questionnaire, the Woman's Questionnaire, and the Biomarker Questionnaire. Theis study focuses on information from the Woman's and the Biomarker Questionnaires. The Woman's Questionnaire was used to collect information from all eligible women age 15–49. These women were asked questions on the following topics: Background characteristics (including age, education, and media exposure), Reproduction (birth history and child mortality), Pregnancy and intermittent preventive treatment, Fever in children, and Malaria knowledge and beliefs. The Biomarker Questionnaire was used to record results of anaemia and malaria rapid diagnostic testing of children age 6–59 months.

Blood samples for biomarker testing were collected via finger or heel pricks from children age 6–59 months. Each field team included one medical laboratory scientist (biomarker specialist) who carried out the anaemia and malaria testing and prepared the blood smears. Malaria testing using a rapid diagnostic test (RDT) using the same drop of blood, taken from the same finger or heel prick that was used for anaemia testing, was tested immediately using the Nigeria-approved SD BIOLINE Malaria Ag P.f. (HRP-II)™ RDT. This qualitative test detects the histidine-rich protein II antigen of *Plasmodium falciparum* in human whole blood. The *P. falciparum* parasite, transmitted by the Anopheles mosquito, is the major cause of malaria in Nigeria. RDT results were available in 15 minutes and recorded as either positive or negative, with faint test lines considered positive and the results were provided to the child's parent or guardian verbally and in written form and were recorded on the Biomarker Questionnaire.

The 2021 Nigeria Malaria Indicator Survey (NMIS) is a representative probability sample designed to produce estimates for the country as a whole, for urban and rural areas separately, for each of the six geographic zones, and for each of the country's 36 states and the Federal Capital Territory (FCT). In addition, Bonny Island in Rivers State was oversampled, so the Bonny Island local government area (LGA) sample can be considered as representing a baseline survey to measure the impact of a planned project to transform the island to a malaria-free zone. Table A.1, in 2021 NMIS report, shows the distribution of the population by state and by type of residence according to 2021 NPC population projections (NMEP, 2022), which is used here to compute the finite population correction (fpc) for coefficient estimates. It was the NPC population projections for 2021

Since the 2021 NMIS sample was a two-stage stratified cluster sample selected from the sampling frame, sampling weights were calculated based on sampling probabilities separately for each sampling stage and each cluster. The 1st stage sampling probability of the *i*th cluster in stratum *h* was P_{1hi} and the 2nd stage sampling probability within the *i*th cluster (households) was P_{2hi}. a_h was *h*-th stratum clusters sum, M_{hi} *i*-th cluster households sum, and $\sum M_{hi}$ households sum in each stratum. The probability of selecting the *i*th cluster was given as $\frac{a_h M_{hi}}{\sum M_{hi}}$. It was taken that b_{hi} was the proportion of households in relative to total number of households in enumeration area (EA) *i* in stratum *h* if the EA was selected, otherwise, $b_{hi} = 1$. So the probability of sampling cluster *i* was $P_{1hi} = \frac{a_h M_{hi}}{\sum M_{hi}} b_{hi}$. With L_{hi} taken as the number of households listed in cluster *i* in stratum *h*, and g_{hi} as the number of households selected in the cluster. The 2nd stage probability of selection for each household was $P_{2hi} = \frac{g_{hi}}{L_{hi}}$. The overall probability of sampling household in *i*-th cluster of *h*-th stratum was $P_{hi} = P_{1hi} \times P_{2hi}$. The sampling weight for a household in cluster *i* of stratum *h* was the inverse of its overall probability of selection, given as $W_{hi} = 1/P_{hi}$. The total observations then becomes

$$N = \sum_{h=1}^{H} \sum_{i=1}^{a_h} \sum_{j=1}^{M_{hi}} \sum_{m=1}^{L_{hi}} n_{mjih} \tag{1}$$

The design weights were then adjusted for household and individual nonresponse to determine the sampling weights for households and women. Individual nonresponse caused differences between household and individual sampling weights. The final sampling weights were normalized so that the total number of unweighted cases was equal to the total number of weighted cases at the national level for both household and individual weights. Normalized weights are relative weights that can be used to estimate averages, proportions, and ratios, but not population totals or pooled data (NMEP, 2022).

Complex Model- and Design-based Approaches

Skinner and Wakefield (2017) contrasted model- and design-based inferences for survey data, noting that randomness from randomization concerned with probability sampling, characterized design-based method. While the model-based approach is defined with randomness from a model that is assumed to have generated the population values y_i. Under the design-based approach, the values in population $y_1, \dots, y_N$, are fixed constants, from which a random sample *S*, of units selected. With *N* known and w_i, is as defined in section above, a standard weighted *HT* estimator of the population mean is

$$\bar{y}_{HT} = \frac{\sum_{i \in S} w_i y_i}{N} \tag{2}$$

The key motivation for such weighting is to remove bias, as explained in detail by Haziza and Beaumont (2017). An alternate mean estimator under the design approach is the Hajek estimator

$$\bar{y}_{HJ} = \frac{\sum_{i \in S} w_i y_i}{\hat{N}} \tag{3}$$

where $\hat{N} = \sum_{i \in S} w_i$, vindicating w_i's interpretation.

Under model-based approach, y_i are taken realization of population values Y_i , $i = 1, \dots, N$, which are assumed to be defined by some specified model. The population is taken as drawn from a hypothetical superpopulation. The predictor for the model-based approach mean estimator is

$$\bar{Y}_U = \frac{\sum_{i \in U} Y_i}{N} \tag{4}$$

The predictor $\hat{\bar{Y}}$ is sample mean $\hat{\bar{Y}}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$. The change of notation to $i = 1, \dots, n$ means that the set *S* from *N* is no longer important (Skinner and Wakefield, 2017). The model-based method might not be affected by the intricate sampling plan. Actually, there are two main ways that it does so. First, a sophisticated sampling plan will rely on the population's structure, such as clustering or stratification.

Model-based predictor for population totals

$$\bar{Y}_U = \frac{1}{N}\sum_{k \in U} Y_k$$

$$\mu = E_M(Y_k), \quad \mathrm{var}_M(Y_k) = \sigma^2$$

$$E_M\left[\hat{\bar{Y}}_n - \bar{Y}_U\right] = \frac{1}{n}\sum_{k=1}^{n} E_M(Y_k) - \frac{1}{N}\sum_{k=1}^{N} E_M(Y_k) = 0$$

$$E_M\left[\left(\hat{\bar{Y}}_n - \bar{Y}_U\right)^2\right] = E_M\left[\left(\frac{1}{n}\sum_{k=1}^{n} Y_k - \frac{1}{N}\sum_{k=1}^{N} Y_k\right)^2\right] = \left(1 - \frac{n}{N}\right)\frac{\sigma^2}{n} \tag{5}$$

Design-based by Horvitz-Thompson estimator for population mean with d_k design weight

$$\bar{y}_{HT} = \frac{\sum_{k\in S} d_k y_k}{N}$$

$$E_S[\bar{y}_{HT}] = \frac{1}{N} E_S\left[\sum_{k\in S} d_k y_k\right] = \frac{1}{N} E_S\left[\sum_{k\in U} d_k I_k y_k\right] = \frac{1}{N}\sum_{k\in U} \pi_k^{-1} E_S[I_k] y_k = \bar{y}_U$$

$$\widehat{\mathrm{var}}_S(\bar{y}_{HT}) = \frac{1}{N^2}\sum_{k\in S}\sum_{l\in S} \frac{\Delta_{kl}}{\pi_{kl}} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l} \tag{6}$$

In contrast to the model-based analysis, where values are typically considered to be independent from a hypothetical infinite population, designs typically sample without replacement. The design-based approach uses estimation techniques including jackknife, bootstrap, and linearization, whereas the model-based approach frequently uses sandwich estimation.

Design based-generalised linear model for malaria incidence

The focus of this paper is on design-based inference for malaria incidence (DBA). When units are sampled with unequal probability it is necessary to give them correspondingly unequal weights in the analysis. The malaria incidence from the 2021 NMIS for children under 5 years is modelled using the logistic regression, which is a member of the generalized linear models (GLM) class. Assuming then that incidence of malaria (y_{mjih}) in a child follows a Bernoulli distribution and there are covariate information, $\boldsymbol{x}^T$ collected from the survey, then the model can be written in terms of the logistic link function as

$$\mathrm{logit}\left\{P\left(y_{mjih} = 1 \middle| \boldsymbol{x}^T\right)\right\} = \exp\left\{\frac{P\left(y_{mjih} = 1 \middle| \boldsymbol{x}^T\right)}{\left(1 - P\left(y_{mjih} = 1 \middle| \boldsymbol{x}^T\right)\right)}\right\} = \boldsymbol{x}^T\boldsymbol{\beta}. \tag{7}$$

The parameter vector is estimated by approximation of the likelihood function of the sampled finite population with likelihood function by observed sample and sampling weights, w_{mjih}, (Archer et al., 2007), and the resulting pseudo-log-likelihood function is given as

$$\begin{aligned}\ell(\boldsymbol{\beta}) = \sum_{h=1}^{H}\sum_{i=1}^{a_h}\sum_{j=1}^{M_{hi}}\sum_{m=1}^{L_{hi}} & w_{mjih}\{y_{mjih} \times \ln\left[P\left(y_{mjih} = 1\middle|\boldsymbol{x}^T\right)\right] \\ & + \left(1 - y_{mjih}\right) \\ & \times \ln\left[P\left(y_{mjih} = 1\middle|\boldsymbol{x}^T\right)\right]\}.\end{aligned} \tag{8}$$

The variance estimators of the parameters are obtained through Taylor linearization method implemented in the *survey* R package (Lumley, 2024), see Lumley (2004) for details of the method. The method follows that the inverse-probability weighting gives a contribution to the statistic from each PSU and as PSUs are sampled independently within cluster, the variance of the cluster's statistic can be estimated by the empirical variance of the PSU statistic. Then, adding up estimates for each of cluster and the second level of strata gives a total and an estimate of its variance (Lumley, 2004).

3. Application Data exploration and Data analysis

3.1 Exploratory analysis of 2021 NMIS data

Comparison of design-based analysis (DBA) and model-based analysis (MBA) on the 2021 NMIS dataset. The incidence of malaria was collected on children age 6–59 months, by questions from parents if they were told child had malaria, after blood samples of child was tested for malaria using the Nigeria-approved SD BIOLINE Malaria Ag P.f. (HRP-II)™ RDT. The options to the question were "Yes", "No", "don't know". For the purpose of this study, responses of "don't know" were filtered out. So only the binary respone of "Yes/No" were used for the analysis as dependent variable. The predictors used in the analysis included "If respondent had heard/seen malaria message", "Age of child (5 -year Group)", "Highest Education Level of respondent", "Water Source", "If respondents have mosquito net for sleeping", "Main floor material", "Main wall material", "Main roof material", "Frequency of using internet", "Number of living children", "Sex of child", "Current age of child", "Type of mosquito net slept under", "Geopolitical Region", "Wealth index combined" and "Type of place of residence". The proportion of children to have been reported to have had malaria are presented in Table 1 from DBA and MBA. All analyses were done in R software (R Core Team, 2024); DBA was carried using the *survey* package (Lumley, 2024), while MBA was done using *stats* package (R Core Team, 2024).

Table 1: Proportion of Malaria incidence in children from the 2021 NMIS

	DBA	**MBA**
child_malariaNo	0.5393	0.5468
child_malariaYes	0.4607	0.4532

It can be seen from Table 1 that the there were 46.1% of children with malaria with DBA as against 45.3% reported with MBA. This means that 46.1% was the proportion of children with malaria representing the country-population dynamics, based on DBA. However, the interpretation from

MBA is that 45.3% child malaria is based on a random model that is assumed to have generated the population values, from a hypothetical superpopulation. Safe to say that the 46.1% is a representation of the population with the particular characteristics of the country of Nigeria.

The similarity between the DBA and MBA estimates suggests consistency in the prevalence distribution regardless of the analytical method used. However, the difference in both estimates, when scaled to the finite population is significant and can affect nationwide interventions to the heterogeneous locations of the country. Additionally, the relatively high proportion of malaria-positive children highlights the continued burden of malaria among children in Nigeria. These findings are consistent with previous national malaria surveys that show malaria remains one of the leading causes of morbidity among children under five years of age.

Comparing the incidence of malaria in children residing in rural area versus urban area using DBA and MBA, showed that DBA had higher proportion in rural area, while the proportion was approximately the same for children in urban area. Table 2 showed these proportions in details, indicating a clear disparity between rural and urban areas, with 35.4% and 34.3% malaria incidences in rural areas with DBA and MBA respectively. While 10.7% and 11.0% incidences were reported in urban areas with DBA and MBA respectively.

Table 2: Proportion of Children with malaria by residence type

residence	**child has malaria**	**DBA**	**MBA**
Rural	No	0.4048	0.4024
Rural	**Yes**	**0.3535**	**0.3426**
Urban	No	0.1345	0.1444
Urban	**Yes**	**0.1072**	**0.1107**

Once, again the difference in both inferential methods is highlighted in the estimates of Table 2. These results indicate that malaria prevalence is substantially higher in rural areas than in urban areas. Several factors may explain this pattern and may include the fact that rural communities often have limited access to healthcare services, lower availability of preventive tools such as insecticide-treated nets, and environmental conditions that promote mosquito breeding such as stagnant water and poor drainage systems. In contrast, urban areas generally benefit from improved housing, better sanitation infrastructure, and greater access to healthcare facilities. Addressing these disparities will require targeted malaria control strategies that prioritize rural populations.

The proportions of children with malaria across the six geo-political zones in Nigeria were presented in Table 3, as calculated using DBA and MBA. It can be seen that the region with the highest proportion of malaria incidence was North-West with 19.2% and 21.9% respectively using DBA and MBA. The region with the least proportion of malaria was South-West Nigeria with 2.2% and 2.6% respectively by DBA and MBA. The differences in the proportions by DBA and MBA is seen in all the regions. The proportion of children with malaria by region with DBA and

MBA was graphically shown in Figure 1, highlighting the differences in the values for each region as produce by DBA and MBA.

Table 3: Proportion of children with malaria by region

Region	**child has malaria**	**DBA**	**MBA**
North Central	No	0.0661	0.0613
North East	No	0.1020	0.0921
North West	No	0.2047	0.2398
South East	No	0.0514	0.0330
South South	No	0.0821	0.0703
South West	No	0.0405	0.0428
North Central	Yes	**0.0628**	**0.0599**
North East	Yes	**0.0729**	**0.0802**
North West	Yes	**0.1922**	**0.2187**
South East	Yes	**0.0538**	**0.0393**
South South	Yes	**0.0498**	**0.0361**
South West	Yes	**0.0218**	**0.0264**

The higher malaria prevalence observed in the northern regions may be attributed to a combination of climatic conditions, socioeconomic challenges, and health system factors. Northern Nigeria typically experiences environmental conditions that favour mosquito breeding, including seasonal rainfall patterns and higher temperatures. Additionally, socioeconomic inequalities and limited access to healthcare services may contribute to increased malaria vulnerability in these regions.

In contrast, southern regions—particularly the South West—tend to benefit from better infrastructure, improved housing conditions, and stronger access to healthcare services, which may help reduce malaria transmission.

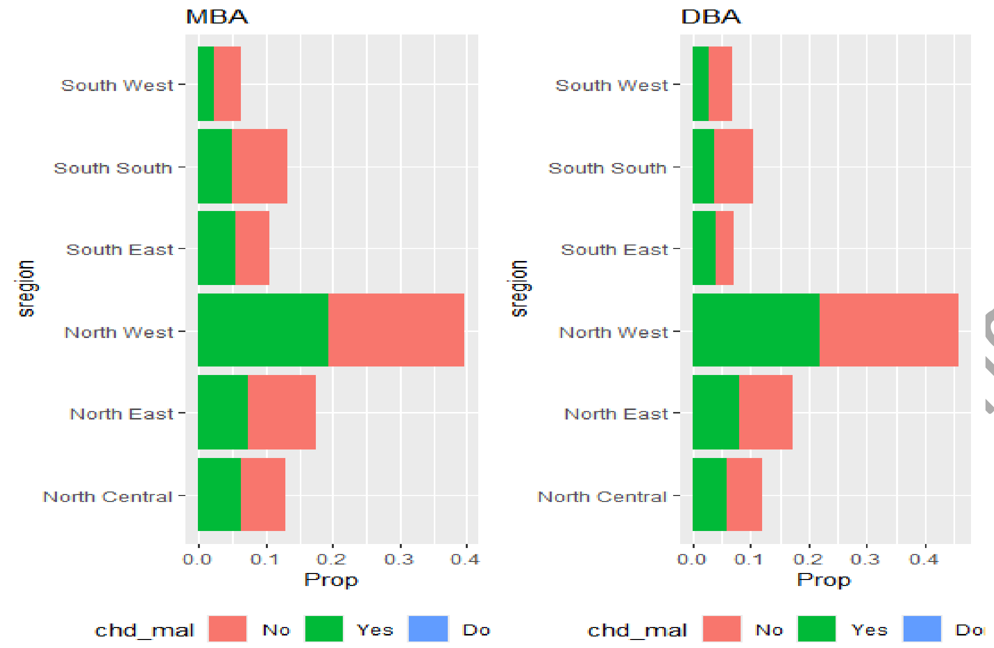

Figure 1: Chart of Proportion of children with malaria by region

Similarly, the comparison of the estimates of malaria incidences in children using DBA and MBA was made across the 36 states of the Federation and the Federal Capital Territory (FCT). The differences between DBA and MBA were shown in Figure 2. Top three states with the highest proportions in order according to DBA were Katsina, Kebbi and Bauchi with 5.2%, 4.8% and 4.4% respectively. With MBA the top three states with highest proportions were Kebbi, Sokoto and Bauchi with 4.2%, 3.6% and 3.0%. The differences in proportions of malaria incidences across states using DBA and MBA showed that inferences depending MBA proved that inferences from such studies of malaria using MBA that are not representative of the finite populations are biased inferences, since the sampling design used in collecting the dataset were not accounted for in the analysis.

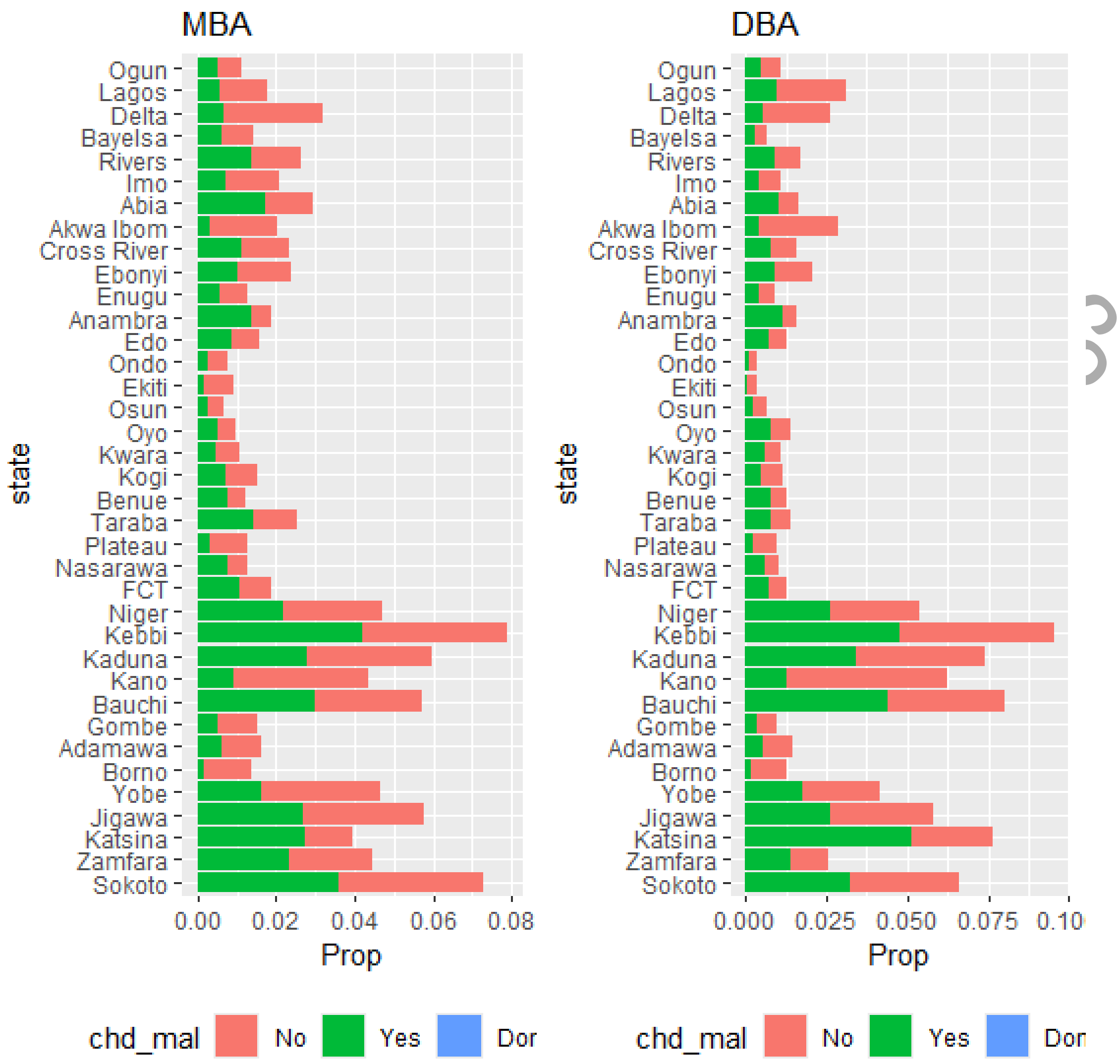

Figure 2: Chart of Proportion of children with malaria by states

3.2 Design-based Logistic Regression analysis of 2021 NMIS data

In this section logistic regression under DBA was applied to the 2021 NMIS data with the predictors listed in section 3.1. Logistic regression under MBA was also applied to the same dataset and compared with the DBA results. The Analysis of Deviance Table (Type II tests) was used to compare the effects of the predictors under DBA and MBA and the results were presented in Table 4. Exposure to malaria messages was highly significant in both models ($p < 0.001$), indicating that individuals who had heard or seen malaria-related information were significantly associated with differences in malaria outcomes. This finding highlights the importance of public health communication campaigns in improving malaria awareness and preventive practices. Also, the age of the child was also a highly significant predictor in both models ($p < 0.001$). This suggests that malaria risk varies across different age groups of children, which may reflect differences in immunity development, exposure patterns, and caregiving practices.

In the DBA model, several additional factors were statistically significant. These include the highest education level of caregivers (p = 0.035), the number of living children in the household (p = 0.046), the material used for house walls (p < 0.001), and geopolitical region (p = 0.014). Education may influence health-seeking behaviour and knowledge of malaria prevention methods, while housing materials may affect exposure to mosquito entry.

In the MBA model, water source (p = 0.026), roof material (p = 0.024), wall material (p < 0.001), and geopolitical region (p < 0.001) were significant predictors. These results emphasize the role of environmental and housing conditions in malaria transmission. Poor housing structures often provide entry points for mosquitoes and may increase the likelihood of mosquito bites during night-time.

Interestingly, variables such as sex of the child, type of residence, wealth index, and mosquito net ownership were not statistically significant in the models. This suggests that while these factors may influence malaria risk indirectly, their effects were not strong enough to reach statistical significance within the model framework.

Comparing both DBA and MBA using the Akaike information Criterion (AIC), showed the AIC values for DBA and MBA reported were respectively 4960.9 and 4998.46. The Type II Wald Chi-square tests from analysis of deviance table, comparing the DBA to MBA showed that DBA was preferred (Working 2logLR = 0.27584, p-value = 0.59525, df = 1).

Table 4: Logistics Model Effects using DBA and MBA

	DBA			**MBA**		
Predictors	stats	df	p-value	LR Chisq	DF	p-value
Heard/seen malaria message	63.362	1	**0.00000**	47.398	1	**0.00000**
Age (5 -year Group)	4.424	6	0.80141	6.006	6	0.42247
Highest Education Level	15.064	3	**0.03542**	2.729	3	0.43527
Water Source	37.607	16	0.15805	27.391	15	**0.02571**
Have mosquito net for sleeping	5.622	1	0.08920	0.033	1	0.85608
Main floor material	17.458	9	0.11073	9.724	9	0.37329
Main wall material	87.691	14	**0.00006**	70.254	14	**0.00000**
Main roof material	22.744	14	0.39723	26.216	14	**0.02430**
Frequency of using internet	8.270	3	0.14546	7.658	3	0.05364
Number of living children	6.627	1	**0.04573**	1.557	1	0.21209
Sex of child	0.770	1	0.46514	0.021	1	0.88381
Current age of child	19.700	1	**0.00003**	18.678	1	**0.00002**
Type of mosquito net slept under	6.029	2	0.13013	3.152	2	0.20683
Geopolitical Region	32.622	5	**0.01368**	37.257	5	**0.00000**
Wealth index combined	8.791	4	0.25534	7.479	4	0.11265
Type of place of residence	0.547	1	0.59553	0.060	1	0.80622

On the modelling of malaria incidence in children, the findings of both approaches highlight the persistent burden of malaria among children in Nigeria and emphasize the importance of socioeconomic, environmental, and geographic factors in shaping malaria risk. First, the high overall proportion of malaria cases demonstrates that malaria continues to pose a major public health challenge despite ongoing control interventions. This underscores the need for sustained malaria prevention strategies including insecticide-treated nets, indoor residual spraying, prompt diagnosis, and effective treatment. Second, the clear rural–urban disparity suggests that malaria control programs should prioritize rural communities where vulnerability remains high. Improving rural healthcare access, strengthening community health education, and expanding preventive intervention coverage could significantly reduce malaria transmission.

Third, the regional variations observed in this study highlight the need for region-specific malaria control strategies. Northern regions with higher malaria prevalence may require intensified interventions and greater resource allocation to effectively address the burden. Finally, the significant influence of housing characteristics and environmental factors suggests that malaria control efforts should also incorporate improvements in housing quality and environmental management. Such structural interventions can complement existing biomedical and behavioural strategies in reducing malaria transmission.

4. Conclusion

The aim of the 2021 Nigeria Malaria Indictor Survey (2021 NMIS), among others, was to estimate malaria characteristics of real, finite Nigerian population. The sampling design of 2021 NMIS which was a two stage stratified cluster sampling design, thus requires that inferences from the dataset should take into the complex design. This inferential approach about real-population parameters, termed design-based analysis (DBA) is applied to the 2021 NMIS to model malaria incidences in children in Nigeria, and the results were contrasted with model-based analysis (MBA), where inference is focused on the process that gave rise to a real infinite population. As have been noted earlier, the coefficients of a model fit from MBA are subject to statistical uncertainty (Fox and Weisberg, 2018). This uncertainty in the model estimates and the fact that the inference is solely about the model that produced a hypothetical infinite population, rather than a finite population from which a sample was drawn and estimation made from that sample. Inferences on survey data, usually entails individuals in the sample being weighted in inverse proportion to their probability of selection, producing unbiased estimates of population characteristics. Weights may also be used to compensate for differential rates of global nonresponse to the survey. DBA comprises the uses population data from surveys to adjust survey-sampling weights. Many works in the literature of malaria stochastic modelling of malaria incidence have been based on MBA. This paper highlights the contrasts between both approaches and the case for concern about inferences on malaria from MBA for finite populations like Nigeria.

In summary, the results from the DBA and MBA on 2021 NMIS dataset demonstrate that malaria among children remains a significant public health concern in Nigeria. Rural residence, geopolitical region, housing conditions, exposure to malaria messages, and child age were key factors influencing malaria incidence. The findings highlight the need for targeted and context-specific malaria control policies that address environmental, socioeconomic, and behavioural determinants of malaria risk.

The analysis of the 2021 NMIS data indicated that malaria remains a major public health issue among children. Rural residence and northern geopolitical regions show higher proportions of malaria cases compared with urban and southern regions. Logistic regression analysis further reveals that factors such as exposure to malaria information, housing conditions, water source, education level, number of children, and the age of the child significantly influence malaria occurrence. These findings highlight the need for targeted malaria control strategies that focus on rural communities, high-burden northern regions, and improvements in housing and environmental conditions. Strengthening malaria awareness campaigns and expanding preventive interventions such as insecticide-treated nets and improved sanitation may help reduce malaria prevalence among children.

Future malaria interventions should focus on strengthening rural health systems, improving housing conditions, expanding health education campaigns, and prioritizing high-burden regions. Such comprehensive approaches will be essential for reducing malaria prevalence and improving child health outcomes in Nigeria.

References

Akeboi, O., Nwafia, E., Godwin, I. et al. (2025). Spatio-temporal modelling of malaria risk in Kano and Lagos States using local health system data. Discov Public Health 22, 536 (2025). https://doi.org/10.1186/s12982-025-00945-4

Archer, K.J., Lemeshow, S. and Hosmer, D.W. (2007) Goodness-of-Fit Tests for Logistic Regression Models When Data Are Collected Using a Complex Sampling Design. Computational Statistics and Data Analysis , 51, 4450-4464. http://dx.doi.org/10.1016/j.csda.2006.07.006

Collin, O.C. and Duffy, K. J. (2022). A mathematical model for the dynamics and control of malaria in Nigeria. Infectious Disease Modelling, 7 (4), pp. 728-741. https://doi.org/10.1016/j.idm.2022.10.005

Dambazau et al. (2025). Malaria transmission dynamics and the impact of treatment interventions in Kano State, Nigeria: a compartmental model approach. PAMJ-One Health. 2025;18:11. [doi: 10.11604/pamj-oh.2025.18.11.48342]

Fox, J. and Weisberg, S. (2018). *An R Compaanion to Applied Regression*. Sage Publication.

Haziza, D., and Beaumont, J.-F. 2017. Construction of Weights in Surveys: A Review. *Institute of Mathematical Statistics*, *Vol. 32*, *N*, 206-226 (21 pages).

Isiko, I., Nyegenye, S., Bett, D.K. et al. (2024). Factors associated with the risk of malaria among children: analysis of 2021 Nigeria Malaria Indicator Survey. Malar J 23, 109 (2024). https://doi.org/10.1186/s12936-

024-04939-6

Jimoh, A. A, Salawu, S. A., Ojuope, K. I. and Ahmed, M. O. (2024) A Model for Estimation of Malaria Prediction in North-East Zone, Nigeria, European Journal of Computer Science and Information Technology, 12 (7), 36-53

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, *009* (issue i08).

Lumley, T. (2024). Survey: analysis of complex survey samples. *R package version 4.4*.

National Malaria Elimination Programme (NMEP) [Nigeria], National Population Commission (NPC) [Nigeria], and ICF 2022. Nigeria Malaria Indicator Survey final report 2021. Abuja, Nigeria, and Rockville: NMEP, NPC, and ICF.

Ogbuagada, S. O., Okolo, A., Torsen, E. and John, O. T. (2022). Multivariate time series analysis in modelling malaria cases in Jimeta metropolis of Adamawa state, Nigeria. FUDMA Journal of Sciences (FJS) Vol. 6 No. 3, June, 2022, pp 62 - 69

Okunlola, O. A., Oyeyemi, O. T. and Lukman, A. F. (2021). Modeling the relationship between malaria prevalence and insecticide-treated bed net coverage in Nigeria using a Bayesian spatial generalized linear mixed model with a Leroux prior. Epidemiol Health 2021;43:e2021041. https://doi.org/10.4178/epih.e2021041

Okunlola, O., Oloja, S., Ebiwonjumi, A. et al. (2024). Vegetation index and livestock practices as predictors of malaria transmission in Nigeria. Sci Rep 14, 9565 (2024). https://doi.org/10.1038/s41598-024-60385-z

Ozodiegwu, I.D., Ambrose, M., Galatas, B. et al. (2023). Application of mathematical modelling to inform national malaria intervention planning in Nigeria. Malar J 22, 137 (2023). https://doi.org/10.1186/s12936-023-04563-w

R Core Team (2024). _R: A Language and Environment for Statistical Computing_. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Segun, O. E., Shohaimi, S., Nallapan, M., Lamidi-Sarumoh, A. A., and Salari, N. (2020). Statistical Modelling of the Effects of Weather Factors on Malaria Occurrence in Abuja, Nigeria. International Journal of Environmental Research and Public Health, 17(10), 3474. https://doi.org/10.3390/ijerph17103474

Skinner, C., and Wakefield, J. 2017. Introduction to the design and analysis of complex survey data. *Statistical Science*, *32*(2), 165–175. https://doi.org/10.1214/17-STS614

WHO (2023). Report on malaria in Nigeria 2022. Brazzaville: WHO Regional Office for Africa; 2023.