Stats 212 - Final Exam

Federica Zoe Ricci - Student ID: 93086057

6/11/2020

Abstract

In this work, we try to understand whether the probability of manifesting malaria among children in school age is related to their age, their sex, the type of food that they are given at school and to seasonality effects. We analyze data from a 2 year longitudinal study on 12 schools in Kenya, including approximately 6000 observations on 500 children. We use a logistic regression model, obtaining similar results both with a GEE and with a GLME approach. Our model includes a mixture of cosine functions to model the seasonality of the data. The results indicate that age is an important factor in determining the probability to get malaria, even for children over 5 years. We also found that sex and feeding interventions did not substanctially impact the probability of getting malaria among the observed children.

1. Introduction

Malaria is a disease counting more than 200 million cases every year and disproportionally affecting developing countries [1]. In particular, African countries in 2018 accounted for over 90% of all malaria cases [1]. Worldwide, women during pregnancy and children aged under 5 years constitute the most vulnerable subjects to malaria [1]. In this study, we analyze data from a randomized field experiment involving 12 schools in Embu, Kenya [2]. This experiment was conducted between 1998 and 2001 with the main goal to ascertain a causal relationship between consumption of animal source food and children growth and cognitive outcomes. The measurements that were collected at regular intervals over the course of the experiment include the prevalence of malaria among the schools' children. In the present study, we analyze the evidence provided by that data with respect to the relationship between malaria and age, gender and food supplements. We also investigate the presence of a seasonality effect, estimate the fraction of children with malaria in the population under study and provide a prediction for the proportion of affected children in schools in the control group. Our work is organized as follows: the remainder of this section details some background information that was relevant to our analysis and model choices. Section 2 outlines the main characteristics of the dataset. Section 3 describes the methodology. Section 4 presents the results. Section 5 concludes.

1.1 Background Information

Previous research observed that the probability of manifesting malaria and the disease's severity depend upon age [3] and pointed to age as an important factor influencing the vulnerability to malaria infection among children [4][5]. Much of the literature has focused on children under 5, who are the most likely to develop severe complications such as cerebral malaria, hypoglycemia and severe anaemia [6]. However, *"increasing success in lowering the level of malaria transmission in many previously highly endemic areas will result in children acquiring immunity to malaria later in life than has been the case in the past"* [7] and there is therefore interest in deepening our understanding of the prevalence of malaria among children in school-age.

The relationship between nutrition and malaria has proven complex to unfold, with some studies coming to seemingly opposing conclusions [8]. As Nyakeriga puts it in [5], *"on the one hand, malaria may cause malnutrition, whereas on the other hand, malnutrition itself may modulate susceptibility to the disease."*. Indeed, protein malnutrition has been associated with greater malaria morbidity and nutrients like vitamin A and zinc have been shown to reduce clinical malaria attacks [9][10], but other reports have suggested that

malnutrition can also be protective of severe forms of malaria [11][12]. Moreover, there has been evidence of a possible interaction between nutrition and age, but only among younger children until 2 years of age [5].

Recent research based on large cross-sectional data found no significant differences in the prevalence of malaria among children aged 6 to 59 months according to the child's sex [13]. In contrast, according to other studies, difference may exists among adults due to the mediation of gender roles [14].

As for seasonality, the literature agrees that *"malaria transmission is both seasonal and heterogeneous"* [15]. It is therefore important to include controls for seasonality when malaria is the outcome variable of interest. However, we did not find a clear recommendation on the appropriate way to model seasonality in contexts like that of our data, so we followed a simple yet flexible approach detailed in [16] which consists in including sine and cosine functions in the regression model.

2. Data

Our dataset consists of 6773 observations over a period of 24 months, during which measurements were collected in the course of 15 periodical visits (the first year monthly and the second year bimonthly) on 502 different children from 12 Kenyan schools in the municipality of Embu. This data was collected within an experiment in which the 12 schools were randomized to 4 feeding interventions: Calorie, Milk, Meat and Control. Feeding consisted of mid-morning snacks and was carried out at school [2]. Measures collected include time-invariant variables (children's ID, school, treatment group, age at baseline visit, sex) and time-variant variables (visit, relative month and year, malaria, other morbidities and their severity - none/mild/severe). Table 1 displays summary information on the characteristics of the schools, grouped according to their treatment group, at the baseline visit and detailed between boys and girls: number of observations, children's age (mean and standard deviation), proportion of children with malaria and with mild or severe morbidities. The number of observations in each treatment group is similar (111-129). We notice that the age of children at baseline across treatment groups is between 7 and 8 years, with girls tending to be younger than boys. The proportion of children with malaria is between 0.2 and 0.38 and seems similar across sexes. Schools in the Meat group recorded a higher proportion of boys and girls with malaria than schools in the other groups at the baseline visit. Mild and severe morbidities were commonly observed in all treatment groups in about 70% of boys and girls.

In our study, 8 observations had to be removed from the original dataset since the information on the age of the child at the baseline visit was missing. There do not seem to be patterns of missing data worthy of attention with the exception that we found a missingness of observations for the last visit (visit 15) at the school level: 4 schools have no observation in the course of the last visit and 1 school has only 2 observations; the remaining schools have virtually the same number of observations for the 14th and the 15th visits. Of the schools for which the 15th visit is lacking, 2 are in the Meat group, 2 are in the Calorie group, 1 is in the Milk group. The observations for these 5 schools in the 3 visits prior to the 15th one were similar to the ones of the other schools (see Appendix A). Otherwise, for the most part, the data was collected on almost all subjects for all visits and there is only slight attrition over time, which we did not see as clearly associated to the prevalence of malaria or to other observed characteristics.

While analyzing the characteristics of the data, the aspect that we thought was the most relevant for our study was the extreme variation in the proportion of children with malaria between schools, especially at the baseline visit and possibly (but less clearly) over time. This is shown in Figure 1. From this figure, we can see the great differences at baseline visit and notice that there is one school (school 11) that mantains a proportion of 0 children with malaria almost throughout all visits. This school is relatively small (about 20 students) and it is in the control group. This aspect of the data suggested us that it would be important to account for the heterogeneity between schools in our statistical model. From the figure, we can also observe that, as mentioned before, 4 schools disappear after visit 14.

Table 1: Treatment groups at baseline visit

	N.	Age[1]	Malaria[2]	Morbidities[2a]
Calorie				
boys	65	7.6 (1.56)	0.20	0.69
girls	64	7.15 (1.16)	0.38	0.73
total	129	7.38 (1.39)	0.29	0.71
Meat				
boys	60	8.08 (1.47)	0.35	0.67
girls	59	7.73 (1.04)	0.32	0.78
total	119	7.91 (1.28)	0.34	0.72
Milk				
boys	66	7.53 (1.17)	0.29	0.74
girls	59	7.28 (1.29)	0.25	0.71
total	125	7.41 (1.23)	0.27	0.73
Control				
boys	58	7.63 (1.31)	0.29	0.74
girls	53	7.05 (0.89)	0.30	0.64
total	111	7.35 (1.16)	0.30	0.69

[1] mean (sd)
[2] cases/total
[a] severe or mild morbidities

3. Methodology

We analyze the data in order to: (1) understand the relationship between the probability of getting malaria and age, sex, feeding interventions and time, (2) assess the effect of seasonality in the probability of getting malaria, (3) estimate the fraction of children with malaria in the population under study and the proportion of "typical" affected children in schools in the control group.

For question (1), we started by looking at the descriptive statistics (proportion of children with malaria by age, gender and feeding interventions). From the literature (Section 1.1) we expect that age, time and possibly feeding interventions could be a relevant factor, while we do not have clear expectations regarding sex. We then selected a statistical model. Three aspects were most relevant to the choice of our model: (a) that the outcome variable is binary (at each visit, a child is classified as either having or not having malaria), (b) that our data is longitudinal (the same child is observed over time) and (c) that children are clustered into schools which display heterogeneous prevalence of malaria. In light of (a), we opted for a logistic regression. In response to (b), we chose to fit a GLM GEE model with an unstructured pairwise correlation[1] and a Generalized Linear Mixed Effect Model (GLME) featuring a random intercept for each subject. In view of (c), we included fixed effects for schools in the GLM GEE model and random intercepts for schools in the GLME. The models were fitted using the R functions *geeglm* (*geepack* library) and *glmer* (*lme4* library).

For question (2), we adopt the approach detailed in [16], which consists in including sine and cosine functions in the regression model, finding an appropriate time period or mixture of time periods. For the choice of the time period, since we are not interested in doing inference about the sine and cosine coefficients, we simply opted for the period that would yield the best fit to the data pattern. We started by using perhaps the most intuitive time period, that is 12 months. As the fit seemed to show room for improvement, we tried with a shorter time period (9 months) which showed to fit the data better. We then found that the inclusion of a mixture of cosine functions that combined a 9 month and a 3 month period allowed for a further improvement.

[1] We wanted to model the within-subject association using separate pairwise log odds ratios, but, in practice, we were not able to implement the GLM GEE model in that way because the function ordgee never worked on this data, always causing R to crash.

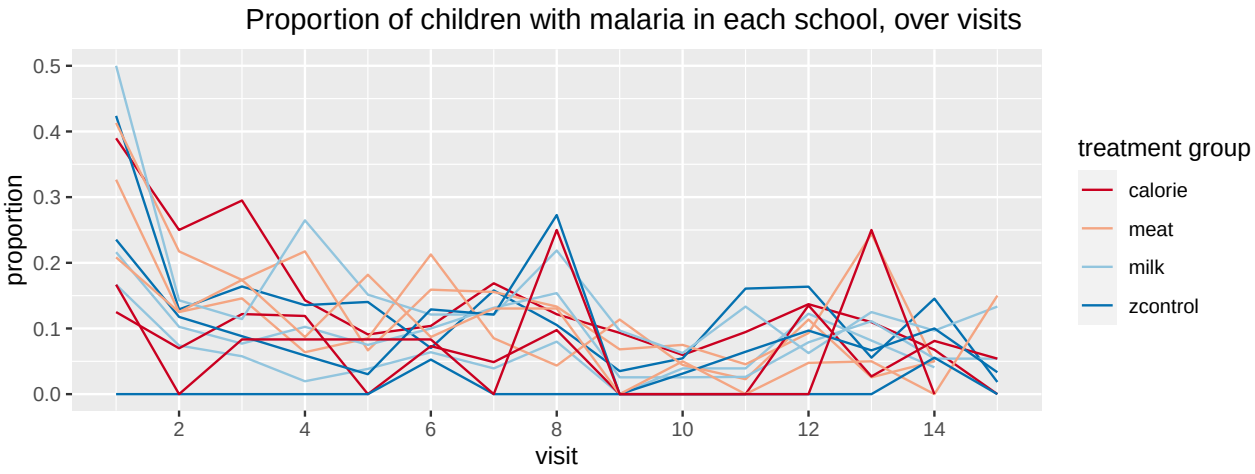


Figure 1: Heterogeneity of malaria proportion across schools

We return to this in the next section.

For a child i, in month j and in school k, the final model that was selected for the GLM GEE is:

$$\log\left(\frac{\mu_{ijk}}{1-\mu_{ijk}}\right) = \beta_0 + \beta_{school_k} + \beta_{age}\text{Age at baseline}_i + \beta_{month}\text{Month}_j + \beta_{sin}\sin_j + \beta_{cos}cos_j + \beta'_{sin}\sin'_j + \beta'_{cos}\cos'_j, \tag{1}$$

where

$$\sin_j = \sin\left(\frac{2\pi\text{Month}_j}{9}\right), \quad cos_j = \cos\left(\frac{2\pi\text{Month}_j}{9}\right), \quad \sin'_j = \sin\left(\frac{2\pi\text{Month}_j}{3}\right), \quad \cos'_j = \cos\left(\frac{2\pi\text{Month}_j}{3}\right).$$

The final GLME model selected is similar, but replaces the school fixed effects β_{school_k} with random intercept for schools b_{0k} and includes a random intercept for subjects.

Moving on to question (3), we are interested in estimating the fraction of children with malaria in our population. Since we only observe data from 12 schools that were not randomly selected, statistically speaking the population to which our findings can be extended is that of Kenyan children comparable to the ones who attend these schools, while our findings cannot easily be extended to all of Kenya's children. One reason is that there is no way for us to determine if the distribution of socio-economic characteristics (which prove very important in determining a child's probability of contracting malaria [13]) in our sample well represents the socio-economic characteristics of the families of the totality of Kenya's children. For the estimation of the fraction of children with malaria in this population, we use a GEE GLM model as in equation (1) but without age at baseline and schools fixed effects. The reason is that the results of a GEE GLM have precisely an interpretation in terms of population-averaged coefficients, and the model without child-specific variables somewhat averages out these characteristics and provides us with an estimate at the population level. Instead, for the proportion of "typical" affected children in schools in the control group we use the estimates from the GLME including random effects for both subjects and schools but excluding fixed effects for age at baseline. In both cases, we build confidence bands around the predicted probabilities by using a Sheffe's s-method correction for multiple testing (margin of error$_j$ $= \sqrt{pF_{p,df.res,0.95}} * SE(x_j^T \beta x_j)$), where p is the number of fixed effects in the model and $df.res$ the degrees of freedom of the residual, x_j is a vector with the values of included covariates in time j).

4. Results

In this section we describe the results of our analysis, beginning from the inference on age at baseline, sex, time and treatment coefficients.

4.1 Age at baseline, Sex, Time and Feeding Interventions

Figure 2 shows the proportion of children with malaria over time across all schools (A), according to children's age at baseline (B) and according to children's sex (C).

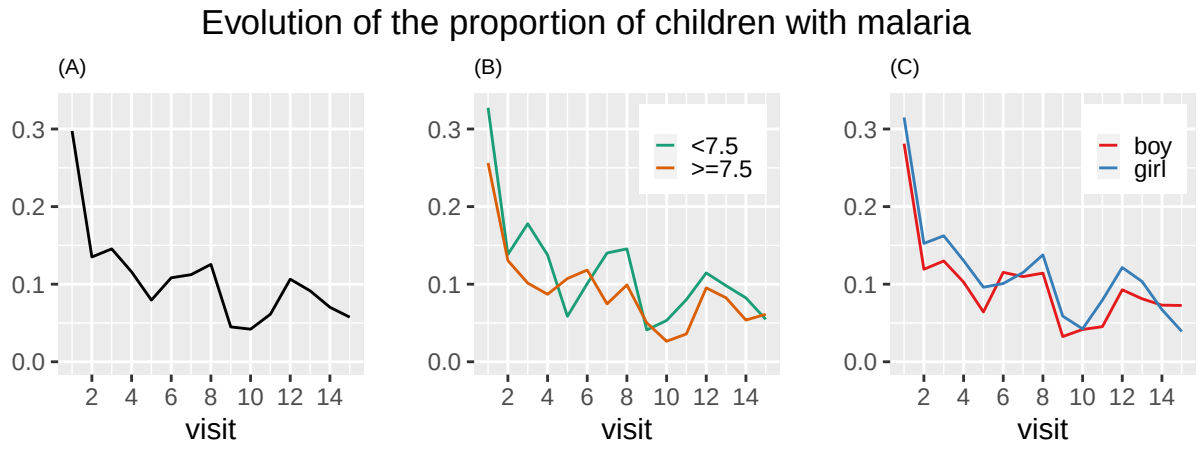


Figure 2: Proportion of children with malaria over time

Figure 2 (A) demonstrates a decrease in the proportion of children with malaria over time, with an oscillation that suggests a potential seasonality effect. Figure 2 (B) shows that younger children (those younger than the sample mean age of 7.5 years) tended to have higher probability of contracting malaria. Figure 2 (C) displays a weak tendency of girls to have higher probability of contracting malaria, but we also recall from Section 2 that girls tend to be younger in our dataset, so there may be counfounding. Figure 1 in Section 2 showed the proportion of children with malaria over time for each school, and did not display an evident "treatment effect" on the probability of contracting malaria over time.

To test whether the effects of age at baseline, sex, time, treatment and seasonality are clearly distinguishable from 0, we fit a full GLM GEE model including fixed effects for age at baseline, sex, time (in months) and seasonality (as described in Section 3) and an interaction between treatment and time. We compare this model with nested models in which each of the covariates of interest is excluded. A comparison of these models was done with a Wald Test (Table 2) and with a QIC approach (Table 3). According to the Wald Test, the coefficients for sex and for the interactions of treatment and time are not statistically distinguishable from zero with a 95% confidence, while those for age at baseline, month and seasonality are. The QIC approach converges to the same conclusions, with the final model described in Section 3 (without sex and treatment) featuring the best results.

Table 4 displays the estimated coefficients for the final GLM GEE model, as well as their standard errors, p-value and Sheffe's S-method corrected confidence intervals. The estimated coefficients with the final GLME model were similar, and are reported in Appendix (B).

Our analysis suggests that age may be an important factor for determining the probability of contracting malaria, even among children older than 5 years. The population-averaged log odds ratio to manifest malaria between a child in some school and month and a child in the same school and in the same month but one

Table 2: Wald tests for hypotheses of interest

	d.f.	χ^2	$\Pr(\geq \chi^2)$
sex = 0	1	0.5217211	0.4701
age at baseline = 0	1	16.8101500	0.0000
treatment group*month = 0	3	3.0864464	0.3785
month = 0	4	72.4095329	0.0000
seasonality params. = 0	7	57.9875835	0.0000

Table 3: QIC results for various nested model specifications

	QIC	QICu	Quasi Lik	CIC	params	QICC
full model	4374.480	4377.058	-2166.529	20.71064	22	4379.372
Nested models						
w/o sex	4373.540	4375.912	-2166.956	19.81410	21	4378.356
w/o age at baseline	4401.938	4405.208	-2181.604	19.36472	21	4406.753
w/o treatment effects	4371.240	4374.681	-2168.340	17.27937	19	4375.903
w/o month	4462.676	4465.532	-2214.766	16.57245	18	4467.264
w/o seasonality	4424.098	4426.441	-2198.221	13.82810	15	4428.463
final model	4370.281	4373.532	-2168.766	16.37464	18	4374.869

Table 4: Estimated population-averaged regression coefficients (geeglm)

	Estimate	Std. Error	p-value	Conf. Interv.
Intercept	0.0683642	0.3340470	0.8378422	(-1.73, 1.86)
School 2	-0.1342662	0.1610007	0.4043103	(-1, 0.73)
School 3	-1.1026599	0.2252572	0.0000010	(-2.31, 0.11)
School 4	-0.0422864	0.2200472	0.8476092	(-1.23, 1.14)
School 5	-0.1756156	0.1902186	0.3558869	(-1.2, 0.85)
School 6	-0.5516579	0.2209480	0.0125328	(-1.74, 0.64)
School 7	-1.0792254	0.2470988	0.0000126	(-2.41, 0.25)
School 8	0.2025389	0.2025177	0.3172599	(-0.89, 1.29)
School 9	-0.4241796	0.2060732	0.0395525	(-1.53, 0.68)
School 10	-0.4227011	0.2410348	0.0794831	(-1.72, 0.87)
School 11	-3.2570688	0.7462005	0.0000127	(-7.27, 0.75)
School 12	-0.9833309	0.4040790	0.0149532	(-3.16, 1.19)
Age at baseline	-0.1794368	0.0402731	0.0000084	(-0.4, 0.04)
Month	-0.0593523	0.0070355	0.0000000	(-0.1, -0.02)
Sin 9 months	0.2760583	0.0546566	0.0000004	(-0.02, 0.57)
Cos 9 months	0.2306441	0.0605561	0.0001397	(-0.09, 0.56)
Sin 3 months	-0.0058504	0.0571169	0.9184165	(-0.31, 0.3)
Cos 3 months	-0.2200915	0.0590809	0.0001951	(-0.54, 0.1)

year older are estimated to be -0.179 (odds ratio 0.455). The rationale which is perhaps the most consistent with the literature is that older children have higher chances to have developed immunity, as compared to younger children. Time also seems a very important factor. This may again be for the effect of aging, and thus developing immunity, but it may also be related to the presence of the intervention or to other policies implemented in concurrence with the intervention. The population-averaged log odds ratio to manifest

malaria between a child in some school and month and a child in the same school and with the same age at baseline, but one month later are estimated to be -0.0594 (odds ratio 0.485).

4.2 Sesonality effect

Figure 3 shows the results of different specifications of the time periods for the seasonality effect.

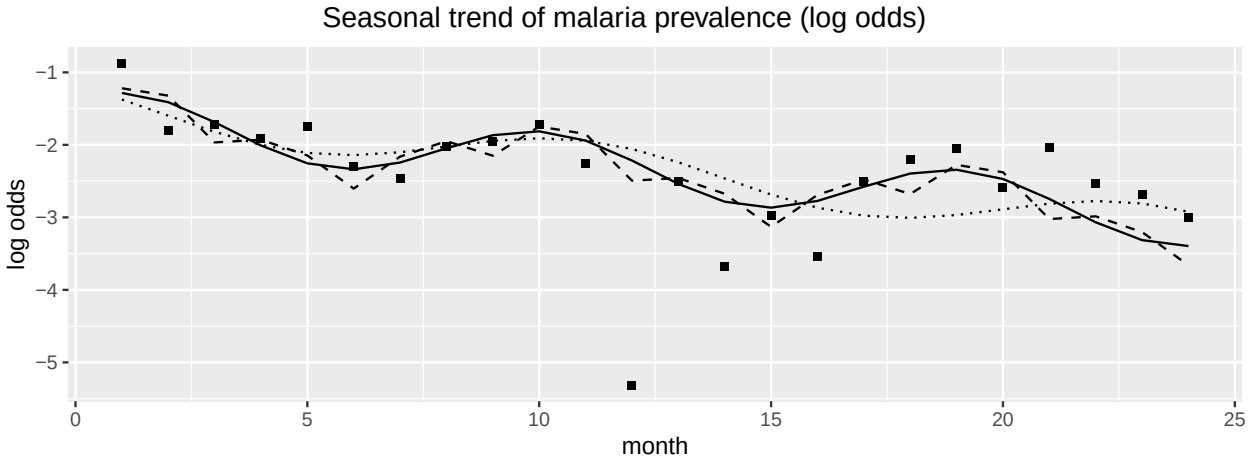


Figure 3: Log odds of the probability of contracting malaria over time. Square symbols: observed log odds, Dotted line: 12 months period. Solid line: 9 month period. Dashed line: Mixture of 9 and 3 month periods.

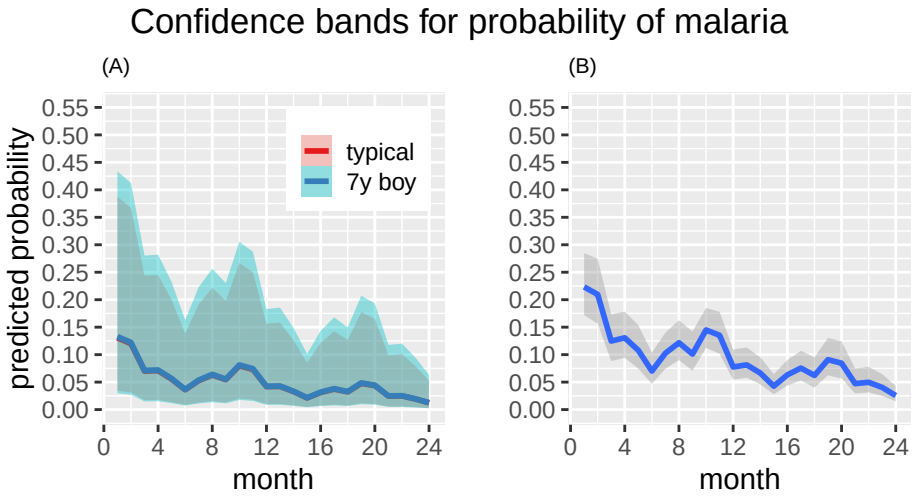
The observed vs. fitted data clearly shows that there is a seasonal pattern. As malaria is spread by mosquitoes, its spread is very sensitive to rains and we speculate that the appropriate mixture of periods may depend on the peculiar pattern of rainfall seasons in certain years. In this case, the 9 month period seems to fit the data much better than the 12 month period, and the mixture of 9 and 3 month periods seems to provide a slight improvement.

4.3 Predictions

Figure 4 shows estimated predictions and their confidence bands at the "typical child" level and for a 7 year old boy in the control group (A) and at the population level (B). In order to estimate the probabilities of malaria specifically for the control group and for a 7 year old boy, Plot (A) was obtained by first fitting a GLME model with a fixed effect for the control group and then also including a fixed effect for sex. We notice that the point estimate for the predicted probability of malaria for a typical child in the control group is generally smaller than the population estimated fraction of children with malaria (as the fixed effect for the control group was estimated to be negative), but has a much larger confidence interval. The greater uncertainty in the GLME model is due to much larger standard errors for all time-invariant fixed effects.

5. Discussion

In this work, we analyzed data on malaria prevalence on school-age children (over 6 years old) from a randomized field experiment in 12 schools in Kenya. We observed a great heterogeneity between schools, that we accounted for in our models through the use of fixed or random effects. We found that the different feeding interventions did likely not affect, on average, the probability of children to manifest malaria over



(A): Predicted probability for a typical child (red) and for a typical 7−year old boy (blue) in the control group
(B) Estimated fraction of children with malaria in the population

Figure 4: Estimated probabilities

time. This is in line with previous literature that found an interaction between nutrition and malaria only on children under 2 years of age [7]. Moreover, while a descriptive analysis initially suggested a possibly higher probability for girls to get malaria, we found no evidence of significant differences between girls and boys after accounting for age at baseline visit, school, time and seasonal effects. In contrast, age, time and seasonality were important factors in determining the probability of contracting malaria for children in this dataset.

The dataset used has several limitations. For example, we lacked information on the socio-economic characteristics of the children's families. We are also unsure regarding the precision with which malaria was measured, as we don't know whether it was measured through a blood analysis or only by parents' self-report.

In view of the importance of addressing malaria among school-age children, future research could explore other type of interventions and whether there are interactions between interventions' success and schools' and families' characteristics.

References

[1] World Health Organization. (2019). World malaria report 2019.

[2] Siekmann, J. H., Allen, L. H., Bwibo, N. O., Demment, M. W., Murphy, S. P., & Neumann, C. G. (2003). Animal source foods to improve micronutrient nutrition and human function in developing countries. *The Journal of Nutrition*, Volume 133, 3972S-3980S.

[3] Dondorp, A.M., Lee, S.J., Faiz, M.A., Mishra, S., Price, R., Tjitra, E., Than, M., Htut, Y., Mohanty, S., Yunus, E.B. and Rahman, R. (2008). The relationship between age and the manifestations of and mortality associated with severe malaria. *Clinical Infectious Diseases*, Volume 47 (Issue 2), 151-157.

[4] Schumacher, R. F., & Spinelli, E. (2012). Malaria in children. *Mediterranean journal of hematology and infectious diseases*, Volume 4 (Issue 1).

[5] Nyakeriga, A. M., Troye-Blomberg, M., Chemtai, A. K., Marsh, K., & Williams, T. N. (2004). Malaria and nutritional status in children living on the coast of Kenya. *The American journal of clinical nutrition*, Volume 80 (Issue 6), 1604–1610.

[6] World Health Organization. (2015). Guidelines for the treatment of malaria. World Health Organization.

[7] Nankabirwa, J., Brooker, S. J., Clarke, S. E., Fernando, D., Gitonga, C. W., Schellenberg, D., & Greenwood, B. (2014). *Malaria in school-age children in Africa*: an increasingly important challenge. Tropical Medicine & International Health, Volume 19 (Issue 11), 1294-1309.

[8] Das, D., R. F. Grais, E. A. Okiro, K. Stepniewska, R. Mansoor, S. Van Der Kam, D. J. Terlouw, J. Tarning, K. I. Barnes, and P. J. Guerin. (2018). Complex interactions between malaria and malnutrition: a systematic literature review. *BMC medicine*, Volume 16 (Issue 1), 186.

[9] Caulfield, L. E., Richard, S. A., & Black, R. E. (2004). Undernutrition as an underlying cause of malaria morbidity and mortality in children less than five years old. The American journal of tropical medicine and hygiene, 71(2_suppl), 55-63.

[10] Anuraj H. Shankar (2000), Nutritional Modulation of Malaria Morbidity and Mortality, *The Journal of Infectious Diseases*, Volume 182, (Issue 1 Supplement), 37–S53.

[11] Genton, B., Al-Yaman, F., Ginny, M., Taraika, J., & Alpers, M. P. (1998). Relation of anthropometry to malaria morbidity and immunity in Papua New Guinean children. The American journal of clinical nutrition, Volume 68 (Issue 3), 734-741.

[12] Snow, R. W., Byass, P., Shenton, F. C., & Greenwood, B. M. (1991). The relationship between anthropometric measurements and measurements of iron status and susceptibility to malaria in Gambian children. Transactions of the Royal Society of Tropical Medicine and Hygiene, Volume 85 (Issue 5), 584-589.

[13] Afoakwah, C., Deng, X., & Onur, I. (2018). Malaria infection among children under-five: the use of large-scale interventions in Ghana. *BMC public health*, Volume 18 (Issue 1), 536.

[14] Tolhurst, R., Nyonator, F. K. (2006), Looking within the household: gender roles and responses to malaria in Ghana, *Transactions of The Royal Society of Tropical Medicine and Hygiene*, Volume 100 (Issue 4), 321–326.

[15] Selvaraj, P., Wenger, E. A., & Gerardin, J. (2018). Seasonality and heterogeneity of malaria transmission determine success of interventions in high-endemic settings: a modeling study. *BMC infectious diseases:*, Volume 18 (Issue 1), 1-14.

[16] Stolwijk, A. M., Straatman, H., & Zielhuis, G. A. (1999). Studying seasonality by using sine and cosine functions in regression analysis. *Journal of epidemiology and community health*, Volume 53 (Issue 4), 235–238.

Appendix A

Table 5: Summaries of visits 12,13 and 14 for schools with no observation in visit 15 (Missing = TRUE) and schools that were normally observed in visit 15 (Missing = FALSE)

Visit	Missing	Malaria (proportion)	N. schools	N. observations	Age at baseline (mean)	Morbidities (proportion)
12	FALSE	0.08	7	232	7.53	0.61
12	TRUE	0.09	5	219	7.52	0.68
13	FALSE	0.06	7	227	7.53	0.55
13	TRUE	0.14	5	210	7.53	0.73
14	FALSE	0.08	7	228	7.54	0.60
14	TRUE	0.04	5	213	7.51	0.60
15	FALSE	0.06	7	224	7.54	0.57
15	TRUE	0.00	1	2	6.70	0.00

Appendix B

Table 6: Estimated population-averaged regression coefficients (glmer)

	Estimate	Std. Error	p-value	Conf. Interv.
Intercept	-0.4695233	0.4010138	0.2416628	(-1.97, 1.04)
Age at baseline	-0.1983319	0.0469396	0.0000239	(-0.37, -0.02)
Month	-0.0610238	0.0063537	0.0000000	(-0.08, -0.04)
Sin 9 months	0.2979604	0.0569590	0.0000002	(0.08, 0.51)
Cos 9 months	0.2382826	0.0584480	0.0000457	(0.02, 0.46)
Sin 3 months	-0.0033643	0.0583865	0.9540510	(-0.22, 0.22)
Cos 3 months	-0.2251324	0.0590711	0.0001383	(-0.45, 0)
Sd random intercept (ID)	0.7322620	NA	NA	NA
Sd random intercept (School)	0.6076841	NA	NA	NA