Estimating age-time-dependent malaria force of infection accounting for unobserved heterogeneity

Peer-reviewed author version

MUGENYI, Levicatus; ABRAMS, Steven & HENS, Niel (2017) Estimating age-time-dependent malaria force of infection accounting for unobserved heterogeneity. In: EPIDEMIOLOGY AND INFECTION, 145(12), p. 2545-2562.

DOI: 10.1017/S0950268817001297
Handle: http://hdl.handle.net/1942/24174

Estimating age-time dependent malaria force of infection accounting for unobserved heterogeneity

L. Mugenyi[1,2*], S. Abrams[2], N. Hens[2,3]

[1] *Infectious Diseases Research Collaboration, Plot 2C Nakasero Hill road, Kampala, Uganda*

[2] *Center for Statistics, Interuniversity Institute for Biostatistics and statistical Bioinformatics, UHasselt (Hasselt University), Diepenbeek, Belgium*

[3] *Centre for Health Economics Research and Modelling Infectious Diseases, Vaccine and Infectious Disease Institute, University of Antwerp, Antwerp, Belgium*

*Correspondence to be addressed to:

Levicatus Mugenyi (LM)

Hasselt University

Agoralaan 1

3590 Diepenbeek, Belgium

Email: *levicatus.mugenyi@uhasselt.be*

Running title: Estimating heterogeneity in malaria force of infection

Summary

Despite well-recognized heterogeneity in malaria transmission, key parameters such as the force of infection (FOI) are generally estimated ignoring the intrinsic variability in individual infection risks. Given the potential impact of heterogeneity on the estimation of FOI, we estimate this quantity accounting for both observed and unobserved heterogeneity. We used cohort data of children aged 0.5-10 years evaluated for the presence of malaria parasites at three sites in Uganda. Assuming a Susceptible-Infected-Susceptible model, we show how the FOI relates to the point prevalence, enabling the estimation of FOI by modeling the prevalence using a generalized linear mixed model. We derive bounds for varying parasite clearance distributions. The resulting FOI varies significantly with age and is estimated to be highest among children aged 5-10 years in areas of high and medium malaria transmission and highest in children aged below 1 year in a low transmission setting. Heterogeneity is greater between than within households and it increases with decreasing risk of malaria infection. This suggests that next to the individual's age, heterogeneity in malaria FOI may be attributed to household conditions. When estimating the FOI, accounting for both observed and unobserved heterogeneity in malaria acquisition is important for refining malaria spread models.

Keywords: Point prevalence, SIS compartmental model, Generalized linear mixed model, Clearance rate distribution

Introduction

Estimating the burden of malaria and evaluating the impact of control strategies, requires reliable estimates of transmission intensities [1]. Measures of malaria transmission intensity include the entomological inoculation rate (EIR), parasite prevalence and force of infection (FOI) [1-6]. The EIR is defined as the number of infectious bites per person per unit time [2, 7] whereas the FOI is defined as the number of infections per person per unit time [4] or the per capita rate at which a susceptible individual acquires infection [8, 9]. The malaria FOI counts all incident (that is, new) human malaria infections in a specified time interval regardless of clinical symptoms, and recurrent infections [4]. The EIR and FOI are related but differ; the EIR considers the number of infective bites delivered by the mosquito vector, whereas the FOI focuses on the infections acquired by the human host. In theory, there should be close relationship between the EIR and the FOI, especially in children with less developed immunity. In practice, however, there is a discrepancy between the two because not every infectious bite results in an infection due to various factors [10]. The efficiency of transmission can be estimated by taking the ratio of the two measures, i.e., the ratio of the EIR to the FOI, the number of infectious bites required to cause an infection [10]. A smaller ratio of the EIR to the FOI implies higher transmission efficiency. Most studies have shown that malaria transmission is highly inefficient [4]. Whereas more recently malaria FOI has been estimated from serological data [1, 11] by detecting past exposure to malaria infection, here we focus on estimating malaria FOI from parasitemia data [12-14].

Despite well-recognized heterogeneity in malaria transmission [15, 16], the FOI is often estimated ignoring intrinsic variability in the individual risk of malaria infection. Heterogeneity in malaria infection arises due to variability in risk factors, including environmental, vector, and host-related factors [17]. Taking these sources of heterogeneity into account [15, 17] in population-based epidemiological studies has been shown to be important [8].

Ronald Ross first published a mathematical model for malaria transmission in 1908 [16, 18]. This model was only firmly established in 1950 by the work of George Macdonald who used Ross's idea [16]. The "Ross-Macdonald" model describes a simplified set of concepts that serves as a basis for studying mosquito-borne pathogen transmission [16]. Using this concept, mathematical methods to estimate the FOI in relation to the EIR have been proposed by, e.g., Smith *et al.* [3, 4], Keeling and Rohani [19] and Aguas *et al.* [20]. Some of the parameters involved in these models are often unknown and should be estimated from data [21]. A solution proposed by Ross in 1916 is to iterate between two modelling frameworks, that is, mathematical and statistical models [21, 22]. The major difference in these two is that the mathematical models (*priori*) are based on differential equations describing the biological mechanism and causal pathway of transmission, whereas the statistical models (*posteriori*) start by the statistical analysis of observations and work backwards to the underlying cause [21]. These two frameworks complement each other and, here, we provide an explicit link between them.

In this paper, we use the well-known generalized linear mixed model (GLMM) framework [see, e.g., 19] to estimate the point prevalence accounting for both observed and unobserved heterogeneity and show how the FOI can be obtained from the point prevalence based on a mathematical Susceptible-Infected-Susceptible model. We derive an expression and easy-to-calculate bounds of the FOI for varying parasite clearance distributions. Our results can be used to refine mathematical malaria transmission models.

Methods

Source of data

The results in this paper are based on cohort data from children aged 0.5 to 10 years in three regions in Uganda; Nagongera sub-county, Tororo district; Kihihi sub-county, Kanungu district; and Walukuba sub-county, Jinja district. The data were collected as part of the Program for Resistance,

Immunology, Surveillance and Modelling of malaria (PRISM) study. The study regions are characterized by distinct transmission intensities. The EIR was previously estimated to be 310, 32 and 2.8 infectious bites per unit year, respectively, for Nagongera, Kihihi, and Walukuba [6]. The study participants were recruited from 300 randomly selected households (100 per region) located within the catchment areas. Data were routinely collected every 3 months (routine visits) and for non-routine clinical (symptomatic) visits. Individuals were tested for the presence of *Plasmodium* parasites using microscopy from August 2011 to August 2014 (3 years). All symptomatic malaria infections were treated with artemether-lumefantrine (AL) anti-malarial medications. More detailed information regarding the study design can be found in Kamya *et al.* [6]. Given that for clinical visits the sampling process is outcome-dependent (see discussion), the analysis here is restricted to the planned routine visits yielding unbiased estimates (simulation study, not shown).

The SIS model, point prevalence and FOI

A simplified version of malaria transmission can be described using the so-called Susceptible (S) - Infected (I) - Susceptible (S), or SIS, compartmental transmission model. This mathematical model classifies the population into two compartments, i.e., the susceptible (S) and the infected (I) class, which can be graphically depicted as shown in Fig 1.

Here, the rate λ(t) at which individuals leave the susceptible state S at time t and flow to the infected state I, as they are infected with malaria parasites, is referred to as the force of infection. Furthermore, γ represents the time-invariant clearance rate at which individuals regain susceptibility after clearing malaria parasites from their blood. Let $s(t)$ denote the proportion of susceptible individuals in the population and $i(t)$ the proportion of infected individuals at calendar time t, i.e., the (point) prevalence, then the following set of ordinary differential equations (ODEs) describes transitions in the compartmental SIS model:

$$\begin{cases} s'(t) = -\lambda(t)s(t) + \gamma i(t), \\ i'(t) = \lambda(t)s(t) - \gamma i(t). \end{cases} \quad (1)$$

As individuals are either susceptible to infection or malaria infected (at least in the aforementioned simplified SIS model), we have $s(t) = 1 - i(t)$. Substituting this expression for $s(t)$ in (1) yields:

$$\lambda(t) = \frac{i(t)\gamma + i'(t)}{1 - i(t)}, \quad (2)$$

where $i'(t)$ is the derivative of the point prevalence with respect to t. The force of infection $\lambda(t)$ can thus be estimated using an estimate for the prevalence $i(t)$ and the clearance rate γ.

Relaxing the assumption of an exponentially distributed parasite clearance distribution in the SIS model can be done by dividing the I compartment into J sub-compartments, such that infected individuals move from the first sub-compartment I_1 to the second I_2, and later to the J^{th} sub-compartment I_J during the different phases of clearing malaria parasites. Using identical rates γ for the transitions between these sub-compartments and for moving from I_J back to the S compartment results in an Erlang distribution with shape parameter J and rate γ for the time spent in all of the sub-compartments [23]. It is easily shown that equation (2) yields an upper bound for the FOI when compared to the aforementioned Erlang clearance distribution (see Appendix). A lower bound is readily obtained by taking $\gamma = 0$ in equation (2) (SI model - see Appendix). The FOI is thus bounded by $[\lambda_L(t), \lambda_U(t)] = \left[\frac{i'(t)}{1-i(t)}, \frac{i'(t)+\gamma i(t)}{1-i(t)}\right]$. Estimates for both the exponential assumption (upper bound) as well as the lower bound are presented in this paper. In order to estimate the prevalence $\pi(t) \equiv i(t)$, we use a generalized linear mixed model to account for individual- and household-specific clustering. This will enable us to explicitly model the observed and unobserved heterogeneity in the acquisition of malaria infection.

Generalized linear mixed model

Generalized linear mixed models (GLMMs) extend the well-known generalized linear models by explicitly taking into account (multiple levels of) clustering of observations [24].

Let Y_{ijk} denote the binary response variable indicating *parasitemia* in the blood (1 if parasites are present – malaria infected; and 0 if not – malaria uninfected) for the i^{th} individual nested in the j^{th} household at the k^{th} visit. Similarly, let X_{ijk} be a $(p+1) \times 1$ vector containing covariate information on p independent variables, and Z_{ijk} be a $q \times 1$ vector of information associated with q random effects. Given the subject-specific random effects $\boldsymbol{b}_{ij}$ and the covariate information X_{ijk}, the random variables $Y_{ijk}|X_{ijk}$ are assumed to be conditionally independent with conditional mean $\pi(X_{ijk}|\boldsymbol{b}_{ij}) = E(Y_{ijk}|X_{ijk}, \boldsymbol{b}_{ij}) = P(Y_{ijk} = 1|X_{ijk}, \boldsymbol{b}_{ij})$. The GLMM relates the conditional mean to the covariates X_{ijk} and Z_{ijk} as follows:

$$g[\pi(X_{ijk}|\boldsymbol{b}_{ij})] = g[P(Y_{ijk} = 1|X_{ijk}, \boldsymbol{b}_{ij})] = X_{ijk}{}^{T}\boldsymbol{\beta} + Z_{ijk}{}^{T}\boldsymbol{b}_{ij}. \quad (3)$$

Here, g is a monotonic link function (e.g., logit, cloglog and log); $\eta(X_{ijk} \mid \boldsymbol{b}_{ij}) = X_{ijk}{}^{T}\boldsymbol{\beta} + Z_{ijk}{}^{T}\boldsymbol{b}_{ij}$ is the linear predictor with $\boldsymbol{\beta}$ a vector of unknown regression parameters for the fixed effects; $\boldsymbol{b}_{ij} \sim N(0, \mathbf{D})$ a vector of subject-specific random effects for subject i in household j for which elements are assumed to be mutually independent; and $\mathbf{D}$ a $q \times q$ variance-covariance matrix [25]. Using equations (2) and (3), the FOI can be obtained using different link functions. Table 1 presents the prevalence and FOI when selecting either the logit, cloglog or log-link function in the GLMM.

Flexible parametric modeling

In a parametric framework such as the GLMM, fractional polynomials provide a very flexible modelling tool for the linear predictor $\eta(X_{ijk} \mid \boldsymbol{b}_{ij})$ [21, 26, 27]. In this paper, a GLMM using a fractional polynomial of degree one with regard to age, with power p selected from a grid (-3, -2, -1, -0.5, 0, 0.5, 1, 2, 3) using Akaike's information criterion (AIC), is used [28]. More precisely, we use

$$\eta(X_{ijk} \mid \boldsymbol{b}_{ij}) = \eta(a_{ijk}, l_{ij}|\boldsymbol{b}_{ij}) = \beta_0 + \beta_1 \text{age}_{ijk}^{p} + \beta_2 l_{ij} + b_{0i(j)} + b_{1i(j)}\text{age}_{ijk}^{p}, \quad (4)$$

where $b_{0i(j)}$is the nested random intercept and $b_{1i(j)}$ is the nested random slope for age. Nesting is done to explicitly acknowledge that individuals make up households. Furthermore, shifted year of birth: l_{ij}, defined as the child's birth year minus the birth year of the oldest child in the cohort (i.e., baseline year 2001), is used in the model to account for the (calendar) time effect since [calendar time] = [birth year] + [age]. The linear predictor (4) can be further extended to include additional covariates.

Age-time dependent force of infection

In equation (3), the conditional mean $\pi\left(X_{ijk}\middle|\boldsymbol{b}_{ij}\right)$ is the point prevalence conditional on the random and fixed effects. In this paper, we use the logit-link function, which enables easy calculation of the intra-cluster correlation coefficient (ICC) through an approximation indicating how much the elements within a cluster are correlated [24, 29, 30].

The age-time dependent FOI, conditional on random effects, is estimated by plugging in the parameter estimates obtained from the final fit in equation (2). More specifically, using a logit-link, the conditional age-time dependent FOI is estimated as follows:

$$\hat{\lambda}_{l_{ij}}\left(\mathrm{a}_{ijk}|\boldsymbol{b}_{ij}\right) = \hat{\gamma}e^{\hat{\eta}\left(a_{ijk}\,,l_{ij}|\boldsymbol{b}_{ij}\right)} + \hat{\eta}'\left(a_{ijk}\,,l_{ij}|\boldsymbol{b}_{ij}\right)\hat{\pi}_{l_{ij}}\left(a_{ijk}\,,l_{ij}|\boldsymbol{b}_{ij}\right), \qquad (5)$$

where $\hat{\gamma}$ is an estimate for the clearance rate and $\hat{\pi}_{l_{ij}}\left(a_{ijk}\,,l_{ij}|\boldsymbol{b}_{ij}\right)$ is the estimated age- and time-dependent conditional prevalence. For the lower boundary of FOI, $\hat{\gamma}e^{\hat{\eta}\left(a_{ijk}\,,l_{ij}|\boldsymbol{b}_{ij}\right)}$ is omitted in equation (5). In the above expression, an estimate for the clearance rate γ is required. Previously, Bekessy *et al*. [12] estimated annual clearance rates of 1.643, 0.584 and 0.986 years^{-1} for children aged less than 1 year, 1-4 years and 5-8 years, respectively. Later, Singer *et al*. [14] estimated these rates as 1.917, 1.425 and 2.364 years^{-1} for ages less than 1 year, 1-4 years and 5-8 years, respectively. Sama *et al*. [13] estimated a constant annual clearance rate of 1.825 years^{-1} by assuming an exponential distribution for infection duration or parasite clearance. Most recently, Bretscher *et al*.

[31] studied the parametric distributions of the infection durations using Ghanaian data, and concluded based on AIC that a Weibull distribution gave a better fit to the data followed by a gamma distribution, while an exponential one was performing worst. In this paper, we use both exponential and Erlang clearance distributions to derive estimates for the malaria FOI obtained based on the aforementioned clearance rates as distributional parameters.

Often, an investigator may wish to observe population averaged estimates. Under the random effects framework, this can be achieved by taking the expectation of the conditional estimates (e.g., the FOI in (5)) resulting into unconditional or marginal estimates. Using the logit-link function, the unconditional (population) force of infection is given by

$$\lambda_{l_{ij}}(\mathrm{a}_{ijk}) = E\left(\lambda_{l_{ij}}(\mathrm{a}_{ijk}|\boldsymbol{b}_{ij})\right) = E\left(\gamma e^{\eta(\mathrm{a}_{ijk}|\boldsymbol{b}_{ij})} + \eta'(\mathrm{a}_{ijk}, l_{ij}|\boldsymbol{b}_{ij}) * \pi_{l_{ij}}(\mathrm{a}_{ijk}, l_{ij}|\boldsymbol{b}_{ij})\right). \quad (6)$$

Calculation of the marginalized FOI in (6), requires integrating out the random effects, $\boldsymbol{b_{ij}}$ over their fitted distribution. This can be done using numerical integration techniques or based on numerical averaging [24].

Model selection

Model building was done using both AIC [32] and a likelihood ratio test for the random effects based on the appropriate mixture of chi-square distributions [33]. Backward model building was performed starting with the random effects and then the fixed effects. The covariates considered in the model building process included study site, age, time since enrollment, shifted birth year (i.e., shifted birth year = birth year – birth year of the oldest child), previous use of AL treatment, and the infectious status at the previous visit. The covariates 'time since enrollment' and 'shifted birth year' were generated to represent the calendar time, albeit we preferred the latter one since participants were not enrolled at the same time point.

Results

Of 989 children, recruited between August 2011 to August 2014, 334 (33.8%), 355 (35.9%) and 300 (30.3%) were from Nagongera, Kihihi and Walukuba, respectively. The baseline parasite prevalence among children aged below 5 years was 38.2%, 12.8% and 9.5% for Nagongera, Kihihi and Walukuba, respectively. The monthly parasite prevalence was higher in Nagongera (range: 26.7% to 68.4%) followed by Kihihi (range: 7.0% to 68.0%) and lastly by Walukuba (range: 0% to 42.9%). Other summary statistics are presented in Table 2. In general, the prevalence was higher among older children (5-10 years).

The parasite prevalence increases with age particularly for children less than 3 years of age and after 7 years of age a decrease is observed (Fig 2, panel A). The prevalence increases with calendar time in Kihihi with increasing variability, while it decreases in Walukuba, and slightly increases in Nagongera (Fig 2, panel B). These observations suggest a difference in malaria infection risk between the three study sites. Also, the infection risk seems to vary with age and calendar time and it tends to take different trends between sites indicating a possibility for a site-time interaction effect. The relationship with age seems to be non-linear. These observed effects were taken into consideration when building the GLMM.

The mean structure in our model consists of a fractional polynomial of age with power -1 (selected based on AIC) and the following covariates (based on significance testing at 5% significance level): shifted year of birth; infection status at previous visit and AL use; and study site. Goodness-of-fit of the final model was assessed using the ratio of the generalized Chi-square statistic to its degrees of freedom. A value of 0.74 was obtained, which is fairly close to 1, indicating that the variability in these data seems to be adequately modelled and little residual over-dispersion remains present [34].

The parameter estimates, standard errors and corresponding test results of the final GLMM fit are shown in Table 3. More details about the candidate models can be found in the Appendix (Tables A1

and A2) together with the fitted conditional and marginal prevalences for the different AL use categories (Fig A2). The results in Table 3 show an overall significant effect of age and shifted year of birth; the effect of age and shifted year of birth is non-significant and borderline significant, respectively, for Walukuba, whereas the effect of age is significant for Kihihi and Nagongera. Shifted year of birth is significant for Kihihi and non-significant for Nangongera. There is significant heterogeneity in the rate of acquiring malaria infection between households (Walukuba: variance = 2.80; Kihihi: variance=1.16; Nagongera: variance=0.21) and between household members (variance=0.24). The intra-household correlation coefficients are 0.44, 0.25 and 0.06 for Walukuba, Kihihi and Nagongera, indicating moderate, low and very low correlation within households, respectively. The intra-individual correlation coefficients are 0.04, 0.05 and 0.06 for Walukuba, Kihihi and Nagongera, respectively, indicating very low correlation in all sites.

Based on the final model fit and using equations (5) and (6) both the conditional (given the random effects) and marginal (population averaged) FOIs can be calculated provided that γ can be estimated. However, estimating γ from the same data is not possible due to an identifiability problem: two or more distinct values of γ give rise to the same (log)likelihood (see Fig A1 in the Appendix). Therefore, we use γ equal to the annual clearance rates given by Bekessy *et al*. [12] as 1.643, 0.584 and 0.986 $years^{-1}$ for children aged less than 1 year, 1-4 years and 5-10 years, respectively, to calculate the conditional and marginal FOIs. We further conduct a sensitivity analysis by considering different clearance rates ranging from 0 to 3 motivated by the ranges estimated by Bekessy *et al*. [12], Singer *et al*. [14], Sama *et al*. [13] and Bretscher *et al*. [31] (see Fig 5, top row). As discussed before, we also provide lower bounds for the FOI.

Fig 3 shows estimates for the marginal FOI together with the corresponding lower bound estimates. We focused on children who were born in the baseline year for graphical reasons. Similar plots were obtained (not shown) for other birth years. Estimates for the lower boundary of the FOI were higher

256 in Nagongera followed by Kihihi and Walukuba. For Nagongera and Walukuba, the lower bound for
257 the FOI was highest for children aged below 1 year and least in those aged 5-10 years, yet. In Kihihi,
258 it is highest among those aged 1-4 years.

259 Fig 3 further shows that in Nagongera and Kihihi, the estimates for the marginal FOI were highest
260 among children aged 5-10 years; yet in Walukuba it was highest among those aged below 1 year. The
261 values for the marginal FOI obtained using the upper boundary estimator, stratified by site, age group
262 and the previous infection status and use of AL are given in Table A3 in the Appendix. At the extreme,
263 the previously symptomatic children acquire up to 4 infections per year in Nagongera, and 8
264 infections per year both in Kihihi and Walukuba. Overall, the FOI is highest among the asymptomatic
265 children and smallest among previously symptomatic children across all age groups and sites (Fig 3
266 and Table 3A). Although Fig 3 clearly shows the impact of different distributional assumptions with
267 regard to the clearance time, the lower and upper bound estimates do not fully capture uncertainty
268 around the point estimates. In Table A4 of the Appendix, we show the 95% confidence bounds for the
269 age- and time-dependent force of infection.

270 Fig 4 (top row) shows the predicted conditional FOIs for 50 randomly selected individual profiles at
271 each of the three sites based on the lower boundary estimator for the FOI. For graphical purposes, we
272 focused on subjects who were symptomatic at the previous visit and who were born in the baseline
273 year. However, similar plots are obtained for other levels of the infection status at the previous visit
274 and for different birth years. Fig 4 (bottom row) shows the predicted marginal FOIs again based on
275 the lower boundary estimator, by age (continuous scale) and infection status at the previous visit and
276 past AL use. In general, the lower boundary estimator indicates that younger children have the
277 greatest FOI. In all sites, individuals that were asymptomatic at the previous visit have the highest
278 FOI, regardless of age. The depicted conditional FOI curves show that individuals have different

279 profiles, indicating substantial unobserved heterogeneity. The increasing trend in the FOI from 6
280 months of age is likely attributed to loss of maternal immunity in infants [35].
281
282 Fig 5 (top row) shows the marginal FOIs for different clearance rates from 0 up to 3 years^{-1} (y-axis).
283 For graphical purposes, and without loss of generality, we again focused on subjects who were
284 symptomatic at the previous visit and who were born in the baseline year. The colour gradient from
285 green (dark) to brown (light) in Fig 5 (top row) corresponds to an increasing FOI. The figure indicates
286 that in Nagongera and Kihihi, children who are below 1 year of age have a lower FOI (green colour)
287 regardless of the presumed clearance rate. Also, in Nagongera and Kihihi, the risk for malaria
288 infection increases with increasing clearance rate, except for the younger children less than 1 to 2
289 years. In Walukuba, the FOI increases with increasing clearance rate regardless of age.
290 Fig 5 (bottom row) shows how the FOI varied with age group (A, B and C) and calendar time among
291 subjects assumed to be symptomatic at the previous visit. In Kihihi, the risk of acquiring a new malaria
292 infection is slightly higher for children born in 2010 compared to those born in earlier years across
293 age groups but not for Nagongera and Walukuba. This would be expected since children born at a
294 later year are younger than those born at an earlier year, and hence are at a higher risk of infection.
295
296

Discussion

In this paper, we use data from a cohort study to estimate the malaria FOI among Ugandan children while accounting for observed and unobserved heterogeneity. The results clearly demonstrate the existence of heterogeneity in the acquisition of malaria infections, which is greater between households than between household members. These observations emphasize the claim by White *et al*. (2010)[17] that heterogeneity in malaria infection can arise due to several unobserved factors including environmental, vector, and host-related factors. This implies that estimating the malaria transmission parameters assuming homogeneity in the acquisition of infection may yield misleading results.

The findings were based on the use of a readily available statistical method, the GLMM, which takes into account heterogeneity between individuals and households in the acquisition of malaria infection. In particular, a fractional polynomial of age of degree 1 and power of -1, adjusted for the calendar time, by means of the so-called 'shifted birth year' (i.e., shifted birth year = birth year – birth year of the oldest child), and other covariates, was considered. The fractional polynomial was chosen because it provides a very flexible modelling tool while retaining the strength of a parametric function. The random slope effects for the fractional polynomial function of age resulted in negative estimates for the FOI, which are biologically implausible and therefore the random slopes were dropped. This could be perceived as a drawback of using the GLMM in combination with fractional polynomials and a more mechanistic approach in which heterogeneity is taken into account at different levels could prove valuable here (further research). When allowing for serial correlation in the model through the specification of an AR(1) correlation structure, the model failed to converge, indicating that too little information was available in the PRISM data to accommodate serial correlation, at least when assuming that the AR(1) assumption is appropriate. An in-depth investigation thereof is an interesting topic for further research.

Based on the SIS model, we derived an expression relating the FOI to the prevalence for infectious diseases such as malaria where we cannot assume lifelong immunity. This expression is an extension of the one proposed by Hens *et al.* (2012) for a so-called SIR model assuming lifelong immunity after recovery, an assumption, which is untenable for malaria. A compartmental model, which can account for temporally recovery due to prior use of treatment (induced immunity) or due to previous exposure to infection (acquired immunity), that is, Susceptible-Infected-Recovered(Treatment)-Susceptible (SIR(T)S), would potentially offer a better alternative compared to the more restrictive SIS model. However, an SIR(T)S model does not yield a closed-form expression for the point prevalence, and hence, for the force of infection. Nevertheless, the derivations are approximately valid for an SIR(T)S model with short recovery duration (derivations not included here). Consequently, we focused on the SIS model, albeit that we adjusted for the previous infection status and treatment in our model. The standard SIS compartmental model assumes that the clearance rate is exponentially distributed. We derived two estimators for the FOI, which provide a lower and upper boundary for the FOI based on different Erlang distributions for the clearance rate. The lower boundary approximately holds for a scenario in which the clearance rate is small compared to the FOI. Although mathematical models encompassing more complicated and more realistic transmission dynamics for malaria could be considered, we defer their treatment to future research in which we will combine Nonlinear Mixed Model (NLMM) methodology and numerical approaches for the estimation of the model parameters in the presence of unobserved heterogeneity.

The temporal inhomogeneity observed in the data is not in contradiction with the SIS model we used. Heterogeneity, age and temporal aspects are addressed in the GLMM, through the specification of random effects as well as age- and calendar time variables, whereas derivations from the SIS model under endemic equilibrium enable the estimation of the age- and time-dependent force of infection from the estimated age- and time-dependent parasite prevalence. Furthermore, estimation of the reproduction number can be done when focusing on the underlying mechanistic modelling of the FOI.

However, we deem this to be beyond the scope of this specific manuscript. Seasonality is not explicitly modelled here, however, inclusion of a covariate describing the amount of rainfall, due to the absence of a clear distinction between the different seasons, and based on additional information (not part of the PRISM data) would be an interesting topic for further research.

When the clearance rate is considered negligible, the rate at which children get infected is highest among those between 1 and 2 years. When the clearance rate is non-negligible, the infection rate is higher among children older than 5 years in areas with high and medium transmission (e.g., Nagongera and Kihihi) and higher in children below 1 year in areas with low transmission (e.g., Walukuba). In Kihihi, the FOI was least for children aged less than 1 year and it is observed to increase as children grow up from 6 months to 1 year. This could be explained by the fact that children lose maternal immunity in their first year of life [35], which puts them at an increased risk of malaria infection. The higher FOI among children aged 5 years and older could be explained by the fact that these children are often asymptomatic malaria cases and are rarely treated, which makes them reservoirs for infections. This finding conquers with the work by Walldorf *et al*. [36] who reported that children aged 6-15 years were at higher risk of (asymptomatic) infection compared to the younger ones. They concluded that older children represent an underappreciated reservoir of malaria infection and have less exposure to antimalarial interventions.

A higher risk was seen among children in Nagongera compared to those in Kihihi and Walukuba with no significant difference between the latter two sites. This could be explained by the fact that Nagongera is a predominantly rural area with many semi-structured houses and many mosquitoes compared to Walukuba or Kihihi as was noted by Kilama *et al.* (2014) [5]. Our results also demonstrated the importance of prior treatment in lowering infection risk due to the post treatment prophylactic effect of longer acting anti-malarials, such as artemether-lumefantrine (AL). For example, children who were previously treated with AL (the symptomatic malaria cases) had a lower

risk of getting re-infected compared to those who were asymptomatic or negative at the previous visit.

This study has two major limitations. First, the analysis was based on results of parasite prevalence determined by microscopy, which is less sensitive than molecular methods such as polymerase chain reaction (PCR) or loop-mediated isothermal amplification method (LAMP) [37, 38]. Thus, sub-microscopic infections would not have been detected. This could have resulted into lower estimates of the FOI. In addition, genotyping was not performed to distinguish new and recurrent infections. As a result, the FOI among individuals who were asymptomatic at the previous visit could have been overestimated. Secondly, the unscheduled clinical visits by the symptomatic individuals were triggered by the study outcome (i.e., parasitemia). This creates a dependency between the observation-time and outcome processes. This dependence, if not accounted for, has a potential to introduce bias in the model estimates and hence in the estimation of the FOI. This bias was avoided by dropping clinical visits and by using only routine data, though the infection status and use of treatment during clinical visits was accounted for in the model. This implies that the analysis used less data than was actually available. The latter limitation will be dealt with in future research by modelling both the outcome and the observation-time processes concurrently using a joint model [39, 40].

To conclude, we have used longitudinal data from a cohort of Ugandan children to estimate the malaria FOI accounting for both observed and unobserved heterogeneity. First, we show how the FOI relates to parasite prevalence assuming an SIS compartmental model and giving both lower and upper boundaries thereof by relaxing the exponential assumption with regard to the parasite clearance distribution. We estimated the parasite prevalence using a GLMM, whose estimates were used to obtain an estimate for the FOI. The malaria FOI was highest among children aged 1 to 2 years based on the lower boundary estimator, and it was higher among children older than 5 years in areas

of high and medium transmission based on the upper boundary estimator. In a low transmission setting, the FOI was highest in children aged below 1 year regardless of the boundary estimator for the FOI. The FOI varied between study sites highest in Nagongera and least in Walukuba. Heterogeneity increases with decreasing FOI and greater between households than household members. We recommend that estimating the malaria FOI should be done accounting for both observed and unobserved heterogeneity to enable refining existing mathematical models in which the FOI may be unknown.

Acknowledgements

Research reported in this publication was supported by the National Institute Of Allergy And Infectious Diseases of the National Institutes of Health under Award Number U19AI089674. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The PRISM project gratefully acknowledges the Uganda Ministry of Health, our research team, collaborators and advisory boards, and especially all participants involved. The support from the PRISM team including Moses Kamya, Grant Dorsey, Sarah Staedke and Dave Smith is highly appreciated. Support from the IAP Research Network P7/06 of the Belgian State (Belgian Science Policy) is gratefully acknowledged. This research was supported by the Antwerp Study Centre for Infectious Diseases.

Financial support

LM is supported by a grant of the Vlaamse Interuniversitaire Raad (VLIR). NH gratefully acknowledges support from the University of Antwerp scientific chair in Evidence-based Vaccinology, financed in 2009-2016 by a gift from Pfizer and in 2016 by GSK.

Conflict of interest

None

Ethical standards

The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008. The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional guides on the care and use of laboratory animals.

References

1. Corran P, et al. Serology: a robust indicator of malaria transmission intensity. Trends Parasitol 2007; 23: 575-82.
2. Onori E, Grab B. Quantitative Estimates of the evolution of a malaria epidemic in Turkey If remedial measures had not been applied. Bulletin of the World Health Organization 1980; 58: 321-326.
3. Smith DL, et al. The entomological inoculation rate and Plasmodium falciparum infection in African children. Nature 2005; 438: 492-5.
4. Smith DL, et al. A quantitative analysis of transmission efficiency versus intensity for malaria. Nature Communications 2010; 1.
5. Kilama M, et al. Estimating the annual entomological inoculation rate for Plasmodium falciparum transmitted by Anopheles gambiae s.l. using three sampling methods in three sites in Uganda. Malar J 2014; 13: 111.
6. Kamya MR, et al. Malaria transmission, infection, and disease at three sites with varied transmission intensity in Uganda: implications for malaria control. Am J Trop Med Hyg 2015; 92: 903-12.
7. Onori E, Grab B. Indicators for the forecasting of malaria epidemics. Bulletin of the World Health Organization 1980; 58: 91-98.
8. Coutinho FAB, et al. Modelling heterogeneities in individual frailties in epidemic models. Mathematical and Computer Modelling 1999; 30: 97-115.
9. Hens N, et al. Seventy-five years of estimating the force of infection from current status data. Epidemiol Infect 2010; 138: 802-12.
10. Smith DL, McKenzie FE. Statics and dynamics of malaria infection in Anopheles mosquitoes. Malar J 2004; 3: 13.
11. Von Fricken ME, et al. Age-specific malaria seroprevalence rates: a cross-sectional analysis of malaria transmission in the Ouest and Sud-Est departments of Haiti. Malar J 2014; 13: 361.
12. Bekessy A, Molineaux L, Storey J. Estimation of incidence and recovery rates of Plasmodium falciparum parasitaemia from longitudinal data. Bull World Health Organ 1976; 54: 685-93.
13. Sama W, Dietz K, Smith T. Distribution of survival times of deliberate Plasmodium falciparum infections in tertiary syphilis patients. Trans R Soc Trop Med Hyg 2006; 100: 811-6.

14. Singer B, Cohen JE. Estimating malaria incidence and recovery rates from panel surveys. Mathematical Biosciences 1980; 49: 273-305.
15. Smith TA. Estimation of heterogeneity in malaria transmission by stochastic modelling of apparent deviations from mass action kinetics. Malar J 2008; 7: 12.
16. Smith DL, et al. Ross, Macdonald, and a theory for the dynamics and control of mosquito-transmitted pathogens. Plos Pathogens 2012; 8.
17. White MT, et al. Heterogeneity in malaria exposure and vaccine response: implications for the interpretation of vaccine efficacy trials. Malaria Journal 2010; 9.
18. Ross R. Report on the prevention of malaria in Mauritius. New York: E. P. Dutton & Company, 1908.
19. Keeling MJ, Rohan P. Modeling infectious diseases in humans and animals. Princeton University Press, 41 William Street, Princeton, New Jersey 08540, 2008.
20. Aguas R, et al. Prospects for malaria eradication in Sub-Saharan Africa. Plos One 2008; 3.
21. Hens N., et al. Modeling infectious disease parameters based on serological and social contact data: A modern statistical perspective. Springer, 2012.
22. Ross R, Application of the theory of probabilities to the study of priori pathometry in *In: proceedings of the Royal Society of Landon. Series A, containing papers of a mathematical and physical character*. 1916.
23. de Smith MJ. Statistical Analysis Handbook - a web-based statistics resource. Winchelsea , UK.: The Winchelsea Press, 2015.
24. Molenberghs G, Verbeke G. Models for discrete longitudinal data. New York: Springer, Series in Statistics, 2005.
25. Zhang DW, Lin XH. Variance component testing in generalized linear mixed models for longitudinal/clustered data and other related topics. Random effect and latent variable model selection 2008; 192: 19-36.
26. Faes C, et al. Estimating herd-specific force of infection by using random-effects models for clustered binary data and monotone fractional polynomials. Journal of the Royal Statistical Society Series C-Applied Statistics 2006; 55: 595-613.
27. Shkedy Z, et al. Modelling age-dependent force of infection from prevalence data using fractional polynomials. Stat Med 2006; 25: 1577-91.
28. Akaike H. New look at statistical-model identification. Ieee transactions on automatic control 1974; Ac19: 716-723.
29. Musca SC, et al. Data with hierarchical structure: impact of intraclass correlation and sample size on type-I error. Front Psychol 2011; 2: 74.
30. Wu S, Crespi CM, Wong WK. Comparison of methods for estimating the intraclass correlation coefficient for binary responses in cancer prevention cluster randomized trials. Contemp Clin Trials 2012; 33: 869-80.
31. Bretscher MT, et al. The distribution of Plasmodium falciparum infection durations. Epidemics 2011; 3: 109-118.
32. Schwarz GE. Estimating the dimension of a model. Annals of Statistics 1978; 6 (2): 461–464.
33. Verbeke G, Molenberghs G. Linear mixed models for longitudinal data. Springer, 2000.
34. Wang J, Xie H, Fisher JH. Multilevel Models. Applications Using SAS. Library of Congress Cataloging-in-Publication Data: Mathematics Subject Classification, 2010.
35. Riley EM, et al. Do maternally acquired antibodies protect infants from malaria infection? Parasite Immunol 2001; 23: 51-9.
36. Walldorf JA, et al. School-Age Children Are a Reservoir of Malaria Infection in Malawi. PLoS One 2015; 10: e0134061.

501 37. Poschl B, et al. Comparative diagnosis of malaria infections by microscopy, nested PCR, and
502 LAMP in northern Thailand. Am J Trop Med Hyg 2010; 83: 56-60.
503 38. Coleman RE, et al. Comparison of PCR and microscopy for the detection of asymptomatic malaria
504 in a Plasmodium falciparum/vivax endemic area in Thailand. Malaria Journal 2006; 5.
505 39. Tan KS, French B, Troxel AB. Regression modeling of longitudinal data with outcome-dependent
506 observation times: extensions and comparative evaluation. Stat Med 2014; 33: 4770-89.
507 40. Rizopoulos D, Verbeke G, Molenberghs G. Shared parameter models under random effects
508 misspecification. Biometrika 2008; 95: 63-74.

509

510

511 **Table 1:** General structures for the FOI according to different link functions in a GLMM framework. η
512 refers to the linear predictor $\eta(\mathrm{X}_{ijk} \mid \boldsymbol{b}_{ij})$ and η' represents the derivative of the linear predictor with
513 respect to the predictor of interest.

Link function (g)	**Prevalence (π)**	**FOI (λ)**
logit	$\frac{e^{\eta}}{1+e^{\eta}}$	$\gamma e^{\eta} + \eta' \frac{e^{\eta}}{1+e^{\eta}}$
cloglog	$1 - e^{-e^{\eta}}$	$\gamma\left(e^{e^{\eta}} - 1\right) + \eta' e^{\eta}$
log	$1 - e^{-\eta}$	$\gamma(e^{\eta} - 1) + \eta'$

514

515

516 **Table 2**: Recruited number of children, baseline and monthly parasite prevalence, by study site and
517 age group

		Nagongera	Kihihi	Walukuba
< 5 years	Number	186	188	190
	Baseline prevalence† (%)	38.2	12.8	9.5
	Monthly prevalence† (%), range	27.4 – 54.7	7.0 – 64.7	0 – 32.0
5 – 10 years	Number	148	167	110
	Baseline prevalence† (%)	58.8	18.0	10.9
	Monthly prevalence† (%), range	26.7 – 68.4	8.3 – 68.0	0 – 42.9
Total	Number	334	355	300
	Baseline prevalence† (%)	47.3	15.2	10.0
	Monthly prevalence† (%), range	26.7 – 68.4	7.0 – 68.0	0 – 42.9

518 †*Parasite prevalence*

519

520

521 **Table 3.** Estimates of the fitted GLMM using a fractional polynomial of degree 1 for age and a logit-
522 link function.

Effect		Parameter	log OR (SE)	t-value	P	OR
Intercept		β_0	-3.04 (0.38)	-8.09	<0.001	
Study site (Reference = Walukuba)	Kihihi	β_1	0.86 (0.43)	2.01	0.045	2.36 (1.02–5.49)
	Nagongera	β_2	2.19 (0.40)	5.45	<0.001	8.94 (4.08–19.57)
Infection status at the previous visit (Ref=Negative and No AL treatment in past)	Negative + AL	β_3	-0.01 (0.10)	-0.05	0.956	0.99 (0.82–1.21)
	Symptomatic	β_4	-0.24 (0.10)	-2.30	0.022	0.78 (0.64–0.97)
	Asymptomatic	β_5	1.23 (0.12)	9.94	<0.001	3.43 (2.69–4.37)
Age^{-1}	Walukuba	β_6	-0.05 (0.83)	-0.06	0.948	0.95 (0.19–4.82)*
	Kihihi	β_7	-4.01 (0.87)	-4.62	<0.001	0.02 (0.003–0.10)*
	Nagongera	β_8	-1.75 (0.45)	-3.89	0.001	0.17 (0.07–0.42)*
Shifted year of birth†	Walukuba	β_9	-0.13 (0.06)	-2.00	0.045	0.88 (0.78 –1.00)
	Kihihi	β_{10}	0.11 (0.04)	2.58	0.010	1.12 (1.13–1.22)
	Nagongera	β_{11}	0.04 (0.03)	1.33	0.184	1.04 (0.98–1.10)

Variance components			**Variance**	**z-value**	
Variance for random intercepts for subjects		d_{11}	0.24 (0.07)	3.32	<0.001
Variance for random intercepts for households	Walukuba	d_{22}	2.80 (0.88)	3.20	0.001
	Kihihi	d_{33}	1.16 (0.28)	4.21	<0.001
	Nagongera	d_{44}	0.21 (0.08)	2.48	0.007

523 *† birth year – min(birth year); * note that the OR here should be interpreted at the* Age^{-1} *level*

524

525

526 **Fig 1:** A schematic diagram of the SIS compartmental model illustrating the simplified dynamics in
527 malaria transmission
528
529

530

531 **Fig 2**: Proportion of children infected with malaria parasites (parasitemia) in a cohort followed for 3
532 years, by study site (Nagongera, Kihihi and Walukuba) in Uganda based on data from August 2011 to
533 August 2014 with the size of the dots proportional to the number of observations. (A) observed
534 parasitemia varying with age; (B) observed parasitemia varying with calendar time.

535

536

537 **Fig 3:** The lower bound (green) for the marginal annual FOI and the difference between upper and
538 lower bound (yellow) with full bar showing the upper bound for the FOI, by study site, age group (A:
539 < 1 year, B: 1-4 years, and C: 5-10 years) and the infection status at the previous visit and past use of
540 AL (negative and no AL in the past (left column), negative and AL in the past (second left column),
541 symptomatic (second right column) and asymptomatic (right column)) for children assumed to be
542 born in the baseline year (2001). Top row: Nagongera, middle row: Kihihi, bottom row: Walukuba.

543

Fig 4: Top row: Individual-specific evolutions for the conditional annual FOI obtained using the lower boundary estimator, by study site for children assumed to be symptomatic at the previous visit and who were born in the baseline year (2001). Bottom row: The marginal annual FOI, obtained using the lower boundary estimator, by study site and the infection status at the previous visit and past use of AL (negative and no AL in the past (solid lines), negative and AL in the past (dotted lines), symptomatic (dash-dotted lines) and asymptomatic (long-dashed lines)). Left column: Nagongera, middle column: Kihihi, right column: Walukuba.

553

554 **Fig 5:** Top row: The marginal annual FOI (contour lines) considering different values for the clearance
555 rate ranging from 0 to 3 $years^{-1}$ by study site for individuals assumed to be symptomatic at the
556 previous visit and were born in the baseline year. Bottom row: The marginal annual FOI, obtained
557 using the upper boundary estimator, for individuals assumed to be symptomatic at the previous visit,
558 by study site, birth year (2001, 2004, 2007 and 2010) and by age group (A: < 1 year, B: 1-4 years, and
559 C: 5-10 years). Left panel: Nagongera, middle panel: Kihihi, right panel: Walukuba.

560

561

562

563

564

Appendix

Though, fractional polynomials are very flexible, they can result into negative estimates for the FOI whenever the estimated probability to be infected before age a is a non-monotone function [21, 27]. A solution to this is to define a non-negative FOI, $\lambda_l(\mathrm{a}_{ijk}|b_i) \geq 0$ for all a and to estimate $\pi_l(\mathrm{a}_{ijk}|b_i)$ under these constraints [27]. From Table 1, for a logit link function, the condition $\eta'(\mathrm{a}_{ijk}|b_i) \geq -\gamma/(1-\pi_l(\mathrm{a}_{ijk}|b_i))$ should be satisfied as to estimate a positive FOI. One option is to fit a constrained FP to ensure the above condition holds by applying a constraint on parameter estimates depending on the functional relationship with age. However, this approach becomes challenging especially if it involves constraining random effects. An alternative option is to find a probability of estimating a negative FOI using the model results. If this probability is considerably small, say less than 0.01, then one can consider the first option unnecessary. In this paper, the second option was applied. Indeed, all site-specific coefficients for age effect were negative (see Table 3), meaning that the site-specific derivatives for the linear predictors, $\eta'(\mathrm{a}_{ijk}|b_i) = \left(-(\hat{\beta}_6)a_{ijk}^{-2}, -(\hat{\beta}_7)a_{ijk}^{-2}, -(\hat{\beta}_8)a_{ijk}^{-2}\right) > 0$. This implies that the above condition always holds in our case since a_{ijk}^{-2}, γ and $(1-\pi_l(\mathrm{a}_{ijk}|b_i))$ are always positive. Therefore, the probability to estimate a negative FOI was zero.

For example, based on model results in Table 3, the conditional age-time dependent FOI for a subject from Walukuba, born in the baseline year (2001, that is, shifted year of birth = 0) and was symptomatic at the previous visit can be estimated as follows,

$$\hat{\lambda}_0(\mathrm{a}_{ijk}|b_i) = \hat{\gamma}\, Exp\left(\hat{\beta}_0 + \hat{\beta}_6 a_{ijk}^{-1} + \hat{\beta}_4 + \mathrm{b}_{1ij} + \mathrm{b}_{21j}\right) - (\hat{\beta}_6)a_{ijk}^{-2} * \hat{\pi}_0(\mathrm{a}_{ijk}|b_i) \qquad (7)$$

where $\hat{\beta}_0 = -3.04$, $\hat{\beta}_6 = -0.05$, $\hat{\beta}_4 = -0.24$, and $\hat{\pi}_0(\mathrm{a}_{ijk}|b_i)$ is the corresponding age-time conditional prevalence given as,

$$\hat{\pi}_0(\mathrm{a}_{ijk}|b_i) = \frac{Exp\left(\hat{\beta}_0+\hat{\beta}_6 a_{ijk}^{-1}+\hat{\beta}_4+\mathrm{b}_{1ij}+\mathrm{b}_{21j}\right)}{1+Exp\left(\hat{\beta}_0+\hat{\beta}_6 a_{ijk}^{-1}+\hat{\beta}_4+\mathrm{b}_{1ij}+\mathrm{b}_{21j}\right)} \qquad (8)$$

588 and $\hat{\gamma}$ is an estimate for the clearance rate. The conditional FOI for other sites given the infection
589 status at the previous visit and past use of AL can be estimated in a similar way.
590
591 **Marginalisation**
592 A sample of $M = 1000$ of the random affects vector $\mathrm{b}_i = (\mathrm{b}_{1i}, \mathrm{b}_{2sj})^T$, s = 1, 2, 3 (sites), was generated
593 from a multi-variate normal distribution, $\mathrm{N}(0, \hat{\mathrm{L}}\,\hat{\mathrm{L}}^T)$, where for example, for Walukuba, $\hat{\mathrm{L}} =$
594 $(0.49,\ 1.67)^{\boldsymbol{T}}$ whose elements are the square roots of $\hat{d}_{11}$ and $\hat{d}_{22}$, respectively as given in Table 3. A
595 fine grid of age, $a = 0.5$ to 11 with interval 0.1 years (the age range in the data, though extrapolation
596 is possible) was considered. For example, the marginalized FOI at each age value in the grid, again
597 considering a subject from Walukuba, born in the baseline year and was symptomatic at the previous
598 visit is calculated as in (9).

599 $$\hat{\lambda}_0(\mathrm{a}) = \frac{1}{1000}\sum_{i=1}^{1000}\left(\hat{\gamma}\, Exp(\hat{\beta}_0 + \hat{\beta}_6\mathrm{a}^{-1} + \hat{\beta}_4 + \mathrm{b}_{1i} + \mathrm{b}_{21i})\right) - (\hat{\beta}_6)\mathrm{a}^{-2} * \hat{\pi}_0(\mathrm{a}), \qquad (9)$$

600 where $\hat{\pi}_0(\mathrm{a})$ is the corresponding marginalized prevalence given by

601 $$\hat{\pi}_0(\mathrm{a}) = \frac{1}{1000}\sum_{i=1}^{1000}\left(\frac{Exp(\hat{\beta}_0+\hat{\beta}_6\mathrm{a}^{-1}+\hat{\beta}_4+\mathrm{b}_{1i}+\mathrm{b}_{21i})}{1+Exp(\hat{\beta}_0+\hat{\beta}_6\mathrm{a}^{-1}+\hat{\beta}_4+\mathrm{b}_{1i}+\mathrm{b}_{21i})}\right) \qquad (10)$$

602 Extensions to estimate the marginal averages at different birth years, for different study sites and for
603 different infection statuses at the previous visit, are straightforward. The SAS macro performing the
604 numerical averaging for a case of $\hat{\gamma} = 1.643$ is attached in the Appendix.
605

A general S(I)$_J$(R)S system

Let s, i and r represent the proportion susceptible, infected and recovered, respectively. Also, let μ represent the natural birth rate assumed to be equal to the natural death rate, β the transmission rate, γ the clearance rate and σ the recovery rate.

System:

$$\begin{aligned}
\frac{ds}{dt} &= \mu - \beta si + \sigma r - \mu s \\
\frac{di_1}{dt} &= \beta si - \gamma i_1 - \mu i_1, \\
\frac{di_2}{dt} &= \gamma i_1 - \gamma i_2 - \mu i_2, \\
&\vdots \\
\frac{di_J}{dt} &= \gamma i_{J-1} - \gamma i_J - \mu i_J, \\
\frac{dr}{dt} &= \gamma i_J - \sigma r - \mu r
\end{aligned} \tag{11}$$

where $i = \sum_{j=1}^{J} i_j$

Rewriting the system collapsing the infectious classes into i:

$$\begin{aligned}
\frac{ds}{dt} &= \mu - \beta si + \sigma r - \mu s, \\
\frac{di}{dt} &= \beta si - \gamma i_J - \mu i, \\
\frac{dr}{dt} &= \gamma i_J - \sigma r - \mu r,
\end{aligned} \tag{12}$$

Simplifying the model to an **S(I)$_J$S** system:

$$\begin{aligned}
\frac{ds}{dt} &= \mu - \beta si + \gamma i_J - \mu s, \\
\frac{di}{dt} &= \beta si - \gamma i_J - \mu i,
\end{aligned} \tag{13}$$

yields (replacing $\frac{di}{dt}$ by i', $\lambda = \beta i$ and $s = 1 - i$)

$$i' = \lambda(1 - i) - \gamma i_J - \mu i, \tag{14}$$

and thus

$$\lambda = \frac{i' + \gamma i_J + \mu i}{1 - i} \approx \frac{i' + \gamma i_J}{1 - i}, \tag{15}$$

expressing time dependency,

$$\lambda(t) = \frac{i'(t)+\gamma i_J(t)+\mu i(t)}{1-i(t)} \approx \frac{i'(t)+\gamma i_J(t)}{1-i(t)}, \quad (16)$$

since $\mu i(t) \ll \gamma i_J(t)$. Let's look at the factor $\gamma i_J(t)$. In case $J = 1$, $\gamma i_J(t) = \gamma i(t)$. In case $J > 1$, $\gamma i_J(t) < \gamma i(t)$. This gives us a lower and upper boundary for our force of infection.

$$[\lambda_L(t), \lambda_U(t)] = \left[\frac{i'(t)}{1-i(t)}, \frac{i'(t)+\gamma i(t)}{1-i(t)}\right]. \quad (17)$$

These formulas readily extend to the age-heterogeneous case since we do not explicitly model the underlying transmission mechanism.

Table A1: Overview of the fractional polynomial model selection.

Power	-3	-2	**-1**	-0.5	0	0.5	1	2	3
AIC	7202.3	7178.6	**7150.0**	7152.9	7154.4	7160.9	7171.2	7190.6	7204.9

Table A2: Overview of model building (number of observations in each case equal to 8645).

Model	**Log-likelihood**	**AIC**	**BIC**
$a^{-1} * S + l * S + S + PT + PT * S + \mathrm{b}_{1ij} + \mathrm{b}_{2j} * \mathrm{S}$	-3199.09	6442.17	6525.75
$a^{-1} * S + l * S + S + PT + PT * S + \mathrm{b}_{2j} * \mathrm{S}$	-3208.24	6458.48	6538.26
$a^{-1} * S + l * S + S + PT + PT * S + \mathrm{b}_{1ij} + \mathrm{b}_{2j}$	-3213.90	6467.80	6543.78
$\boldsymbol{a^{-1} * S + l * S + S + PT + \mathrm{b}_{1ij} + \mathrm{b}_{2j} * \mathrm{S}}$	**-3204.93**	**6441.86**	**6502.64**
$a^{-1} + l * S + S + PT + \mathrm{b}_{1ij} + \mathrm{b}_{2j} * \mathrm{S}$	-3210.56	6449.12	6502.31
$a^{-1} * S + l + S + PT + \mathrm{b}_{1ij} + \mathrm{b}_{2j} * \mathrm{S}$	-3209.73	6447.45	6500.64
$a^{-1} + l + S + PT + \mathrm{b}_{1ij} + \mathrm{b}_{2j} * \mathrm{S}$	-3211.87	6447.74	6493.32

S= study site, P=Infection status at previous visit, T=treatment with AL at previous infection, PT=combination of P and T. Note that P and T were collinear (sign of T changes whenever P is included with T)

641 **Table A3:** Maximum values for the marginal annual FOI by study site, previous infection status and
642 use of AL, and by age group.
643

Site	Previous infection status and use of AL	Maximum annual FOI		
		< 1 year	1 – 4 years	5 – 10 years
Nagongera	Negative, No AL	3.99	4.21	8.49
	Negative, AL	4.45	4.80	9.69
	Symptomatic	2.21	2.07	4.14
	Asymptomatic	7.73	9.21	18.70
Kihihi	Negative, No AL	5.35	24.95	64.82
	Negative, AL	1.46	4.64	11.78
	Symptomatic	1.06	3.23	8.11
	Asymptomatic	4.62	20.25	52.56
Walukuba	Negative, No AL	18.01	6.65	11.28
	Negative, AL	20.07	7.41	12.58
	Symptomatic	8.02	2.95	5.01
	Asymptomatic	98.24	36.34	61.66

644

Fig A1: Plots for log-likelihood verses the clearance rate (left panel) and force of infection verses the clearance rate (right panel) obtained after fitting 1000 models to the data according to $\pi = \frac{\lambda}{\lambda+\gamma}(1 - e^{-(\lambda+\gamma)a})$ as given by Pull and Grab (1974) by choosing values for the annual clearance rate on a grid of 0.1 to 2.0 with a step size of 0.0019.

651

652 **Fig A2:** Top row: Individual-specific evolutions for the conditional prevalence, by study site for
653 children assumed to be symptomatic at the previous visit and were born in the baseline year (2001).
654 Bottom row: Average evolutions for marginalized prevalence, by study site and the infection status
655 at the previous visit and past use of AL (negative and no AL in the past (solid lines), negative and AL
656 in the past (dotted lines), symptomatic (dash-dotted lines) and asymptomatic (long-dashed lines)).
657 Left panel: Nagongera, middle panel: Kihihi, right panel: Walukuba.

658
659

660 **Table A4:** Marginal FOI and the 95% confidence bounds for the age- and time-dependent marginal
661 annual FOI by study site, previous infection status and use of AL, and by age group for children born
662 in the baseline year (2001).
663

Infection status at the previous visit and past use of AL	**Age in years**	**Nagongera** Marginal annual FOI (95% CI) x1000	**Kihihi** Marginal annual FOI (95% CI) x1000	**Walukuba** Marginal annual FOI (95% CI) x1000
Lower bound				
Negative and no AL in the past	<1	143.78 (141.16 - 146.39)	9.27 (8.52 - 10.01)	10.20 (9.75 - 10.65)
	1-4	53.69 (53.20 - 54.19)	22.69 (22.34 - 23.04)	0.95 (0.92 - 0.97)
	5-10	8.57 (8.53 - 8.62)	7.24 (7.17 - 7.31)	0.09 (0.09 - 0.09)
Negative and AL in the past	<1	137.35 (134.84 - 139.87)	7.64 (7.28 - 8.00)	10.72 (10.27 - 11.18)
	1-4	51.67 (51.19 - 52.14)	20.09 (19.82 - 20.35)	0.99 (0.97 - 1.02)
	5-10	8.29 (8.24 - 8.33)	6.59 (6.52 - 6.65)	0.10 (0.09 - 0.10)
Symptomatic	<1	105.62 (103.73 - 107.51)	6.26 (5.98 - 6.54)	9.58 (9.18 - 9.98)
	1-4	41.4 (41.02 - 41.79)	16.91 (16.70 - 17.12)	0.89 (0.87 - 0.91)
	5-10	6.83 (6.79 - 6.87)	5.70 (5.65 - 5.75)	0.09 (0.08 - 0.09)
Asymptomatic	<1	426.73 (420.32 - 433.14)	24.87 (23.68 - 26.06)	22.88 (22.14 - 23.63)
	1-4	123.3 (122.22 - 124.39)	55.20 (54.57 - 55.81)	2.11 (2.07 - 2.14)
	5-10	16.86 (16.78 - 16.93)	15.69 (15.57 - 55.83)	0.20 (0.20 - 0.20)
Upper bound				
Negative and no AL in the past	<1	234.51 (229.74 – 239.28)	12.22 (11.16 – 13.28)	309.33 (285.17 – 333.49)
	1-4	224.99 (223.32 – 226.65)	61.66 (60.13 – 63.20)	112.84 (109.37 – 116.31)
	5-10	445.73 (442.83 – 448.62)	161.20 (157.32 – 165.08)	191.40 (186.57 – 196.22)
	<1	223.74 (219.15 – 216.36)	10.03 (9.54 – 10.53)	322.88 (298.09 – 347.66)

Negative and AL in	1-4	214.75 (213.15 – 216.36)	51.65 (50.91 – 52.39)	117.76 (114.2 – 121.31)
the past	5-10	424.49 (421.70 – 427.29)	131.29 (129.73 – 132.85)	199.73 (194.78 – 204.67)
Symptomatic	<1	170.65 (167.29 – 174.0)	8.22 (7.84 – 8.60)	246.55 (231.99 – 261.10)
	1-4	164.17 (163.02 – 165.32)	42.74 (42.17 – 43.30)	89.53 (87.46 – 91.61)
	5-10	320.14 (318.21 – 322.07)	107.71 (106.53 – 108.89)	151.64 (148.75 – 154.52)
Asymptomatic	<1	741.36 (728.10 – 754.61)	32.81 (31.15 – 34.46)	1134.5 (1034.6 – 1234.4)
	1-4	717.36 (712.29 – 722.31)	159.84 (157.55 – 162.14)	417.93 (403.49 – 432.36)
	5-10	1532.75 (1523.4 – 1542.1)	429.11 (423.66 – 434.55)	711.13 (691.0 – 731.26)

664

```
***** SAS MACRO *****

*GLIMMIX code
proc glimmix data=Cohortfulldata2 method=laplace  NOCLPRINT;
   class hhid id siteid(ref="1") pinfectstatusandAL(ref="0");
   model parasitemia =  fpcohortage*siteid yearshift*siteid siteid pinfectstatusandAL/ dist=bin oddsratio
link=logit solution;
   random intercept/ subject = hhid group=siteid solution;
   random intercept / subject = id(hhid) solution;
   COVTEST/ WALD;
run;
**Numerical averaging
**Considering children born between 2001 to 2014 as they appear in the data;
al.(1976) are
data numaveragingprevfoinc;
do site =1 to 3 by 1; *study sites 1(walukuba),2(kihihi),3(nagongera);
 do pinfect =1 to 4 by 1; *infection status 1(negative+no AL), 2(negative+AL), 3(symptomatic), 4(asymptomatic);
  do subject=1 to 1000 by 1; *generate 1000 samples;
  bi1=rannor(123); bi2=rannor(123); bi3=rannor(123); bi4=rannor(123); *used seed=123 to generate from standard
normal;
  d11=0.24;d22=2.80;d33=1.16;d44=0.21;*variances from the final fit, elements in D;
  rd11=d11**0.5;rd22=d22**0.5;rd33=d33**0.5;rd44=d44**0.5; *sqrt(S2) to be used in Cholesky decomposition;
  r1=rd11*bi1; r2=rd22*bi2; r3=rd33*bi3; r4=rd44*bi4; *using U+sqrt(S2)*rannor(seed): Note elements in here are
sqrt of elements in D;
  do a=0.5 to 11 by 0.1; *generate 1000 samples at each age point in the grid;
     do L=0 to 13 by 1; *Repeat the above process for each value of birth year shift (L=year of birth - 2001);
          *Parameter estimates;
B0=-3.04;B1=0.86;B2=2.19;B3=-0.01;B4=-0.24;B5=1.23;B6=-0.05;B7=-4.01;B8=-1.75;B9=-0.13;B10=0.11;B11=0.04;
ap=1/a; *Power of age, age-1;
          *Linear Predictors;
          lp11=B0+B6*ap+B9*L+r1+r2; lp12=B0+B6*ap+B9*L+B3+r1+r2;
          lp13=B0+B6*ap+B9*L+B4+r1+r2;lp14=B0+B6*ap+B9*L+B5+r1+r2;
          lp21=B0+B7*ap+B10*L+B1+r1+r3; lp22=B0+B7*ap+B10*L+B1+B3+r1+r3;
          lp23=B0+B7*ap+B10*L+B1+B4+r1+r3;lp24=B0+B7*ap+B10*L+B1+B5+r1+r3;
          lp31=B0+B8*ap+B11*L+B2+r1+r4; lp32=B0+B8*ap+B11*L+B2+B3+r1+r4;
          lp33=B0+B8*ap+B11*L+B2+B4+r1+r4;lp34=B0+B8*ap+B11*L+B2+B5+r1+r4;
          *Derivative of linear predictor;
          lpder1=-(B6)*(ap*ap); lpder2=-(B7)*(ap*ap); lpder3=-(B8)*(ap*ap);
          *Prevalence;
     if site=1 and pinfect=1 then pi=exp(lp11)/(1+exp(lp11));
```

```
704         if site=1 and pinfect=2 then pi=exp(lp12)/(1+exp(lp12));
705         if site=1 and pinfect=3 then pi=exp(lp13)/(1+exp(lp13));
706         if site=1 and pinfect=4 then pi=exp(lp14)/(1+exp(lp14));
707         if site=2 and pinfect=1 then pi=exp(lp21)/(1+exp(lp21));
708         if site=2 and pinfect=2 then pi=exp(lp22)/(1+exp(lp22));
709         if site=2 and pinfect=3 then pi=exp(lp23)/(1+exp(lp23));
710         if site=2 and pinfect=4 then pi=exp(lp24)/(1+exp(lp24));
711         if site=3 and pinfect=1 then pi=exp(lp31)/(1+exp(lp31));
712         if site=3 and pinfect=2 then pi=exp(lp32)/(1+exp(lp32));
713         if site=3 and pinfect=3 then pi=exp(lp33)/(1+exp(lp33));
714         if site=3 and pinfect=4 then pi=exp(lp34)/(1+exp(lp34));
715         **FOI;
716         *Clearance rate of 1.643 for children <1 year as given by Bekessy et al.(1976) is demonstrated, a
717    similar code can easily be adopted for ages 1-4 years and 5-10 years.;
718         if site=1 and pinfect=1 and a<1 then foi=1.643*exp(lp11)+ lpder1*exp(lp11)/(1+exp(lp11));
719         if site=1 and pinfect=2 and a<1 then foi=1.643*exp(lp12)+ lpder1*exp(lp12)/(1+exp(lp12));
720         if site=1 and pinfect=3 and a<1 then foi=1.643*exp(lp13)+ lpder1*exp(lp13)/(1+exp(lp13));
721         if site=1 and pinfect=4 and a<1 then foi=1.643*exp(lp14)+ lpder1*exp(lp14)/(1+exp(lp14));
722         if site=2 and pinfect=1 and a<1 then foi=1.643*exp(lp21)+ lpder2*exp(lp21)/(1+exp(lp21));
723         if site=2 and pinfect=2 and a<1 then foi=1.643*exp(lp22)+ lpder2*exp(lp22)/(1+exp(lp22));
724         if site=2 and pinfect=3 and a<1 then foi=1.643*exp(lp23)+ lpder2*exp(lp23)/(1+exp(lp23));
725         if site=2 and pinfect=4 and a<1 then foi=1.643*exp(lp24)+ lpder2*exp(lp24)/(1+exp(lp24));
726         if site=3 and pinfect=1 and a<1 then foi=1.643*exp(lp31)+ lpder3*exp(lp31)/(1+exp(lp31));
727         if site=3 and pinfect=2 and a<1 then foi=1.643*exp(lp32)+ lpder3*exp(lp32)/(1+exp(lp32));
728         if site=3 and pinfect=3 and a<1 then foi=1.643*exp(lp33)+ lpder3*exp(lp33)/(1+exp(lp33));
729         if site=3 and pinfect=4 and a<1 then foi=1.643*exp(lp34)+ lpder3*exp(lp34)/(1+exp(lp34));
730    output;
731        end;
732      end;
733    end;
734    end;
735    end;
736    run;
737    *sort data;
738    proc sort data= numaveragingprevfoinc; by a site pinfect L;run;
739    *Get means;
740    proc means data= numaveragingprevfoinc; var pi foi; by a site pinfect L; output out=outpifoinc; run;
741    *Keep data for marginalized means;
742    data marginalizedprevandfoinc; set outpifoinc; where _stat_='MEAN'; run;
```