Epidemiology and Infection

cambridge.org/hyg

A unified and flexible modelling framework for the analysis of malaria serology data

Irene Kyomuhangi and Emanuele Giorgi

CHICAS, Lancaster Medical School, Lancaster University, Lancaster, UK

Original Paper

Cite this article: Kyomuhangi I, Giorgi E (2021). A unified and flexible modelling framework for the analysis of malaria serology data. *Epidemiology and Infection* **149**, e99, 1–10. https://doi.org/10.1017/S0950268821000753

Received: 21 December 2020
Revised: 30 March 2021
Accepted: 30 March 2021

Author for correspondence:
Irene Kyomuhangi,
E-mail: i.kyomuhangi@lancaster.ac.uk

Abstract

Serology data are an increasingly important tool in malaria surveillance, especially in low transmission settings where the estimation of parasite-based indicators is often problematic. Existing methods rely on the use of thresholds to identify seropositive individuals and estimate transmission intensity, while making assumptions about the temporal dynamics of malaria transmission that are rarely questioned. Here, we present a novel threshold-free approach for the analysis of malaria serology data which avoids dichotomization of continuous antibody measurements and allows us to model changes in the antibody distribution across age in a more flexible way. The proposed unified mechanistic model combines the properties of reversible catalytic and antibody acquisition models, and allows for temporally varying boosting and seroconversion rates. Additionally, as an alternative to the unified mechanistic model, we also propose an empirical approach to analysis where modelling of the age-dependency is informed by the data rather than biological assumptions. Using serology data from Western Kenya, we demonstrate both the usefulness and limitations of the novel modelling framework.

Introduction

Despite the significant progress made in the control of malaria worldwide, this still remains a significant public health threat in many countries, particularly in Sub-Saharan Africa [1]. Even with the decline of malaria prevalence in endemic countries [2], there are still challenges that require robust mechanisms for monitoring malaria transmission and evaluation of elimination efforts [1].

Classical methods of estimating malaria risk rely on the detection of the *Plasmodium* parasite in humans and mosquito populations. *Plasmodium falciparum (Pf)* is the most prevalent malaria parasite in Africa, while *Plasmodium vivax (Pv)* dominates in the Americas and South East Asia [1]. Parasite prevalence is determined by the proportion of infected individuals at the time of data collection [3, 4], while the entomological inoculation rate (EIR) is the rate at which individuals are bitten by infectious mosquitoes [5]. Both of these measures may vary over time due to the joint effect of several environmental factors, and the precision with which they can be estimated is often low, particularly in low transmission settings [3, 4]. Additionally, the collection of entomological data is labour-intensive, expensive and excludes the recruitment of children, due to ethical considerations [6–8].

Several studies have shown the utility of serological markers as a viable alternative for estimating transmission intensity. Because of the persistence of antibodies, serological markers (1) provide information on cumulative exposure to the pathogen over time, (2) smooth out the effect of seasonality in transmission, and (3) allow estimation of transmission intensity with more feasible sample sizes even in low transmission settings [3, 8–10].

Antibody responses to blood-stage malaria parasites provide protection against clinical disease, however this response does not confer sterile immunity, therefore individuals remain susceptible to repeated infections [11, 12]. In malaria endemic settings, antibody levels generally increase as individuals become older, are boosted by repeated infection and decay in the absence of re-infection [4, 13]. Using existing knowledge on the dynamics of transmission, malaria serology models aim to derive a measure of transmission which can be used to monitor trends in endemic areas over time.

The most commonly used approach to estimate malaria transmission intensity is based on the classification of individuals as seronegative and seropositive which is then used as the input of a reversible catalytic model (RCM), to estimate the seroconversion rate, which quantifies the rate at which individuals convert from seronegative to seropositive [4, 8, 9]. Assuming latent seronegative and seropositive distributions in the sample, mixture models fitted to the antibody distribution are used in order to identify optimal thresholds for the classification of individuals into seropositives and seronegatives [4, 14]. The major drawback of this approach is that it can generate biased estimates of transmission intensity as a result of the misclassification, especially among inconclusive cases whose probabilities of belonging to either group are

CAMBRIDGE UNIVERSITY PRESS

close to 50% [15, 16]. Bollaerts *et al.* [15] and Hens *et al.* [16] propose a 'direct' method of estimating seroprevalence from continuous antibody measurement using an underlying mixture model, which avoids the use of thresholds and thus the bias arising from the misclassification of individuals. In those publications, the direct method is applied to Salmonella and Varicella-Zoster virus antibody data. This approach has not been applied to analyse malaria serology data and, in this paper, we propose a modelling framework that is inspired by Hens *et al.* [16].

In addition to the seroconversion rate, boosting rates, i.e. the rate at which antibody levels are acquired, can also be used as a marker for transmission intensity [4, 17, 18]. Antibody acquisition models (AAMs) have been developed as an alternative approach to RCMs, and do not involve the use of thresholds but instead rely on the full antibody measurements in order to estimate boosting rates. However, in the context of malaria serology, current formulations of the AAM assume that the antibody measurements follow a log-Gaussian distribution, clearly an invalid assumption in the case of a bi-modal distribution arising from the mixing of the seropositive and seronegative populations [17].

RCMs and AAMs that have been applied to the analysis of malaria serology data make strong assumptions on the temporal dynamics of transmission, which are generally restricted to the following patterns: constant transmission, a sharp stepwise drop in transmission and a linear drop in transmission [4, 17–19]. The validity of these assumptions is often questionable, and more flexible functional forms for the variation of transmission over time have not been considered in the context of malaria serology.

In this paper, we develop a unified mechanistic model for the analysis of malaria serology data which combines the properties of mixture models, RCMs and AAMs in order to reliably estimate malaria transmission intensity. We also show that the additional flexibility brought by this novel model allows a better description of temporal dynamics of malaria transmission. In addition to this, we present an alternative empirical approach to account for the age-dependency of the antibody distributions and use this approach to validate the unified mechanistic model.

The structure of the paper is as follows. Section 'Existing models' provides an overview of current models for malaria serology analysis. Section 'A unified mechanistic model for the analysis of malaria serology data' introduces a unified mechanistic model and outlines an alternative empirical approach that can be used to analyse malaria serology data. In section 'Analysis of malaria serology data from Western Kenya', we apply this new framework to cross-sectional antibody data from Western Kenya, and section 'Discussion' is a discussion of the results. Finally, section 'Conclusion' provides a summary and conclusion.

Existing models

Mixture models

In the context of malaria and other infectious diseases, mixture models are developed under the assumption that the population of interest is indeed a mixture of latent seropositive and seronegative populations [4, 20]. More formally, let Y_i denote the log-transformed antibody measurement for the i-th individual. Let S^+ and S^- be a shorthand notation for 'seropositive' and 'seronegative' classifications, respectively. Assuming independent and identically distributed realisations for a sample of n individuals, we write the density function of Y_i as

$$f(y_i) = \prod_{i=1}^{n} [(1-p) f_{S^-}(y_i; \mu_{S^-}, \sigma^2_{S^-}) + p f_{S^+}(y_i; \mu_{S^+}, \sigma^2_{S^+})], \quad (1)$$

where f_{S^+} is a univariate log-Gaussian distribution with mean μ_{S^+} and variance $\sigma^2_{S^+}$ for the S^+ population, and analogously for S^-; finally, p is the probability of being S^+.

Let C_i and C_i^* denote the random variables representing classification based on the mixture model and true classification of the i-th individual, respectively. One approach is to define a seropositivity threshold, usually $\mu_{S^-} + 3\sigma_{S^-}$, above which C_i is S^+, and S^- if below [4, 15, 16, 19, 21]. An alternative, more elaborate, approach is to first calculate the probability of belonging to group C_i^*, conditional on the antibody measurement $Y_i = y_i$, i.e.

$$\begin{aligned} P(C_i^* = S^+ | y_i) &= \frac{p f_{S^+}(y_i; \theta_{S^+})}{(1-p) f_{S^-}(y_i; \theta_{S^-}) + p f_{S^+}(y_i; \theta_{S^+})} \\ P(C_i^* = S^- | y_i) &= 1 - P(C_i^* = S^+ | Y_i = y_i). \end{aligned} \quad (2)$$

Based on two probability thresholds, c^- and c^+, the classification C_i is

$$C_i = \begin{cases} S^- & \text{if } P(C_i^* = S^- | Y_i = y_i) \le c^- \\ I & \text{if } c^- < P(C_i^* = S^- | Y_i = y_i) < c^+ \\ S^+ & \text{if } P(C_i^* = S^+ | Y_i = y_i) \ge c^+ \end{cases}, \quad (3)$$

where I is an additional classification label introduced to denote inconclusive cases. In serology analysis, a common approach is to exclude these cases, depending on the type of disease, and report the proportion of inconclusive cases [15, 16, 22].

In malaria serology, most studies favour the first threshold-based approach that does not introduce the classification for inconclusive cases [17–19, 23–25]. This is likely due to the nature of antibody responses to malaria infections which result in a large proportion of 'inconclusive' cases, as reported by Sepúlveda *et al.* [4].

However, both of these threshold-based approaches are prone to misclassification, which can create bias in estimating epidemiological parameters [4, 15, 16]. Furthermore, current applications of mixture models in malaria serology analysis do not take into account the age-dependence of antibody levels, and assume that the mixing of S^+ and S^- is the same across all ages, which may further exacerbate the issue of misclassification.

The two component mixture Gaussian models also do not account for antibody boosting upon re-exposure to malaria parasites. Sepúlveda *et al.* [4] present an extension to the traditional mixture model where more components are added in order to account for this boosting effect. These components can be interpreted as varying degrees of malaria exposure; unexposed, once exposed, twice exposed, etc. Assuming a known number of components, say K, the sampling distribution is given by

$$f(y_i) = \prod_{i=1}^{n} \left[\sum_{k=1}^{K} p_k f_k(y_i; \theta_k) \right]. \quad (4)$$

The number of components K is then treated as an additional parameter to estimate using the profile likelihood. However, the interpretation of the components of the model is problematic due to ambiguity about classification rules, particularly when

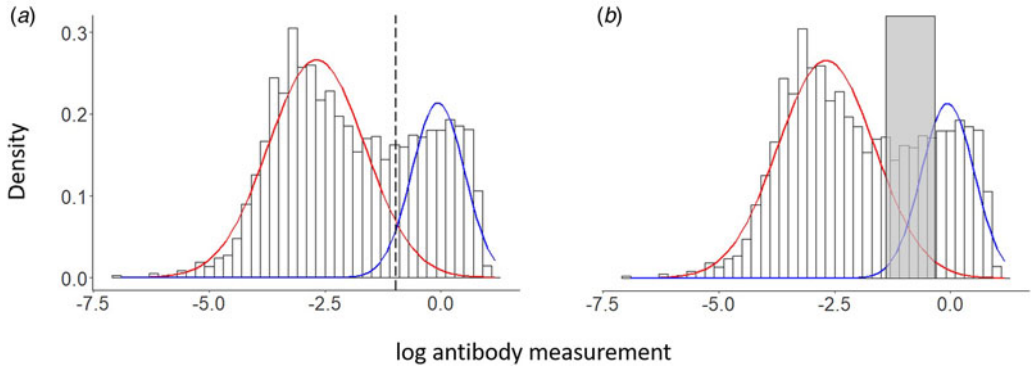

Fig. 1. An illustration of the mixture model showing the bi-modal distributions for the S^- (red) and S^+ (blue) populations. The dotted line in (a) shows the seropositivity threshold $\mu_{S^-} + 3\sigma_{S^-}$, above which individuals are classified as S^+. The grey rectangle in (b) shows the inconclusive cases as defined by equation (3). In this case, the probability thresholds c^- and c^+ have been set to 90%. Individuals below this grey region are classified as S^-, while individuals above this region are classified as S^+. These data are taken from the *Pf* AMA1 analysis in section 'Analysis of malaria serology data from Western Kenya'.

component means are close together. This approach also further compounds the problem of inconclusive cases as they occur across multiple components.

Reversible catalytic models

Following the dichotomisation of the continuous antibody measurements through the application of a mixture model, the resulting S^+ and S^- outcomes are modelled using an RCM. A common assumption of the RCM is that individuals are born S^- and, after becoming S^+ upon exposure to malaria, can revert to S^- in the absence of exposure. This mechanistic approach is illustrated in Figure 2a. Since antibody data are assumed to represent the cumulative exposure of individuals during their lifespan, the age of individual prior to the sample collection is used as a proxy for historical time.

Let $\lambda(a)$ denote the seroconversion rate for an individual at age a and ω the seroreversion rate. According to the RCM, the temporal dynamics that regulate the proportion of S^+ individuals of age a, hence $p(a)$, are expressed by the following differential equation

$$\frac{\mathrm{d}p}{\mathrm{d}a} = \lambda(a)(1 - p(a)) - \omega p(a). \tag{5}$$

In the above equation, $\lambda(a)$ is a measure of the underlying transmission intensity which is associated with the gold standard indicator of transmission, the EIR [8], while ω is typically fixed and assumed to be constant [4]. However, some authors Bosomprah [19] and Akpogheneta *et al.* [26] suggest that ω may be age-dependent. Sepúlveda *et al.* [4] argue that the malaria serology data often carry little information in the estimation of ω, a problem which will persist also in our novel modelling framework. Hence, throughout this paper, we shall make the working assumption of a constant ω. Note that the reciprocals of λ and ω estimates, i.e. $1/\lambda$ and $1/\omega$, indicate the estimated number of years within which seroconversion and seroreversion would occur, respectively.

Three transmission profiles have so far been proposed to model the seroconversion rate $\lambda(a)$. The simplest assumes a constant transmission, hence $\lambda(a) = \lambda$ for all a. In this case, the differential equation in (5) gives the following solution

$$p(a) = \frac{\lambda}{\lambda + \omega}(1 - \mathrm{e}^{-(\lambda+\omega)a}). \tag{6}$$

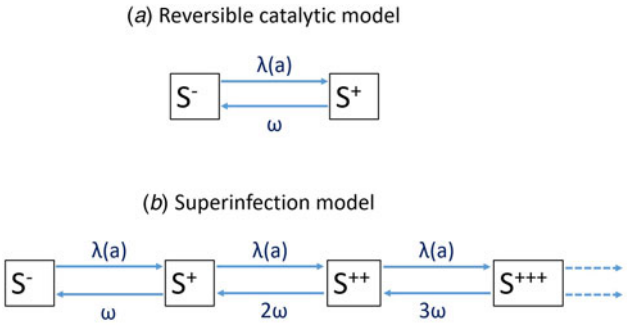

Fig. 2. (a) This figure is a representation of the reversible catalytic model (RCM) where individuals transition between seronegative (S^-) and seropositive (S^+) states through the SCR, $\lambda(a)$and the SRR, ω. (b) This figure is a representation of the superinfection model (SIM) where individuals can have their antibodies 'boosted' through increasing seropositive ($S^{+\cdots}$) states depending on the cumulative exposure to malaria parasites.

In the equation above, the proportion of S^+ at older ages reaches a maximum value of about $\lambda/(\lambda + \omega)$. In other words, in a cohort of an initially malaria-naive population, $p(a)$ will ultimately reach a plateau at which the number of individuals seroconverting is the same as the number of individuals seroreverting [4, 8]. However, these assumptions may be too stringent as they ignore changes in transmission that may be due, for example, to the introduction of control interventions [4, 20, 21].

To tackle this issue, one approach is to assume a transmission profile with a sharp drop in transmission at the time of intervention. In this model, two transmission rates are estimated: λ_1 and λ_2 which represent the transmission rates before and after the drop, respectively. An alternative approach to account for control interventions is to assume a linear reduction in the seroconversion rate $\lambda(a)$, rather than a step-change as we have just illustrated. However, in this case, the differential equation in (5) cannot be solved analytically and numerical procedures must instead be used.

In the study by Yman *et al.* [17], the two transmission profiles that do not assume a constant $\lambda(a)$ provide a better fit to the data. However, assumption of a step-change or linear drop in $\lambda(a)$ may be inappropriate in the presence of major or prolonged malaria outbreaks within the historical time-frame considered. In general, the validity of any of these profiles is dependent on a variety of

factors, including intervention history, climate and vector characteristics. More recently, Varela *et al.* [27] propose a model where the number of times that λ changed in the past, which is also estimated from the data.

Where seropositivity is defined using the traditional two-component Gaussian mixture model, there is still the issue of how to account for antibody boosting due to repeated exposure to malaria parasites. Bosomprah [19] suggests an extension to the RCM, which involves creating more seropositive classes in a superinfection model (SIM), similar to the multi-component mixture model described by Sepúlveda *et al.* [4]. In this framework, a seronegative individual can transition to the first seropositive class, S^+, upon first exposure, and subsequently to a higher seropositive class S^{++} upon re-exposure, and so on, as illustrated in Figure 2b. The SIM also faces challenges with interpretation of results where initial exposure and boosting between the multiple seropositive classes may be conflated [3, 4].

Antibody acquisition models

An alternative modelling approach to estimate transmission intensity is to use AAMs [17, 18]. Unlike RCMs, AMMs use the full antibody measurements without requiring any dichotomisation of the data. More specifically, AAMs are used to estimate the boosting rate, i.e. the rate at which antibodies are acquired, a marker for transmission intensity [4, 17, 18, 28]. Let $\mu(a)$ denote the average antibody level in the general population of individuals of age a. Assuming that following exposure to parasites, $\mu(a)$ is boosted at a rate $\gamma(a)$ and assuming a constant decay rate r, we can express this mechanism through the following differential equation

$$\frac{\mathrm{d}\mu}{\mathrm{d}a} = \gamma(a) - r\mu(a). \tag{7}$$

We can then use the above equation to infer changes in average antibody levels as a function of age a. Finally, in order to fit (7) using likelihood-based methods of inference, the antibody levels of individuals at age a are assumed to follow a log-Gaussian distribution with mean $\mu(a)$ and variance σ^2 [4, 17, 18].

Similar to the way the way seroconversion rates have been modelled in RCMs (section 'Reversible catalytic models'), previous studies have considered three transmission profiles for the specification of $\gamma(a)$. The simplest approach assumes that $\gamma(a) = \gamma$ is constant which leads to the following solution of (7)

$$\mu(a) = \frac{\gamma}{r}(1 - \mathrm{e}^{-ra}). \tag{8}$$

Similarly to RCMs, extensions of the AAM assumes either a step-change or linear reduction in the boosting rate γ; see Sepúlveda *et al.* [4], Yman *et al.* [17] and Weber *et al.* [18] for more details.

Direct comparison of γ and λ from the AAM and RCM, respectively, may not be possible as these estimate different serological indicators. However, Yman *et al.* [17] find that the AAMs provide a more consistent fit to age-dependent antibody data compared to RCM fit to age-dependent seroprevalence data. Additionally, AAMs provide better precision in parameter estimation and appear to be more robust to sample size reduction. It has been found that AMMs often provide a good fit to serological data in high to moderate transmission settings, where a large proportion of individuals may be seropositive [17], or where an antigen is highly immunogenic, leading to high seropositivity to its antibody in the population [18].

A unified mechanistic model for the analysis of malaria serology data

In this section, we develop a statistical modelling framework which extends the standard mixture model outlined in section 'Mixture models' to incorporate both the RCM and AAM dynamics and provides a more flexible approach to model time changes in the seroconversion rate and boosting rate. In this unified framework, the mixing probabilities – i.e. probability of belonging to the S^+ and S^- populations – are modelled based on the RCM, while the means of the two latent S^+ and S^- distributions are informed by AAM dynamics.

To avoid the need of solving complex differential equations, we re-express (5) with a discrete-time difference equation, i.e.

$$p(a) - p(a-1) = \lambda(a)(1 - p(a)) - \omega p(a)$$

or, equivalently,

$$p(a) = \frac{\lambda(a) + p(a-1)}{1 + \lambda(a) + \omega}.$$

Assuming that $\lambda(0) = 0$, and by iteratively applying the above expression, we then obtain

$$p(a) = \sum_{h=1}^{a} \frac{\lambda(h)}{\prod_{k=h}^{a}(1 + \lambda(h-k) + \omega)}. \tag{9}$$

This allows us to specify any function for $\lambda(a)$ without being constrained to three options described in section 'Reversible catalytic models'. The above describes the proportion of S^+ individuals who are aged a, $p(a)$, as a weighted sum of transmission intensities occurring in all the years since birth, $\lambda(h)$, with weights decreasing exponentially as we move further back in time from the time of data collection.

We apply this same idea to the AAM, allowing for temporally varying $\gamma(a)$. More specifically, by using a discrete-time dynamic we re-write (7) as

$$\mu(a) - \mu(a-1) = \gamma(a) - r\mu(a)$$

or, equivalently,

$$\mu(a) = \frac{1}{1+r}(\gamma(a) + \mu(a-1)).$$

By applying the above expression iteratively and assuming that $\gamma(0) = 0$, we obtain that

$$\mu(a) = \sum_{h=1}^{a} \gamma(h)\left(\frac{1}{1+r}\right)^{a-h+1}. \tag{10}$$

Similar to the interpretation of (9), in this expression, the mean antibody level at age a, $\mu(a)$, is given by weighted sum of all the boosting rates since birth, $\gamma(h)$, and the weights given are exponentially decaying. The assumptions of $\lambda(0) = 0$ and $\gamma(0) =$

0 may not be strictly valid, however, this is a pragmatic choice since the true boosting and seroconversion rates at birth are not known but are expected to be close to zero on account of underdeveloped immune responses to malaria in infants who rely on maternal antibodies up to 9 months after birth [28–30].

To model the temporal changes in $\lambda(h)$ and $\gamma(h)$, in the absence of a detailed information on intervention history, a pragmatic approach is to use a log-linear regression in the years before the time of data collection, which is expressed as

$$\begin{aligned}\lambda(h) &= \exp\{l_0 + l_1(a-h)\}\\ \gamma(h) &= \exp\{g_0 + g_1(a-h)\},\end{aligned} \tag{11}$$

where h corresponds to a given age of an individual before the time of collection and, thus, $a-h$is the years before the time of data collection. Finally, l_0, l_1, g_0 and g_1 are regression parameters to estimate (Fig. S1 of the Supplementary material further illustrates the mechanism of this approach).

Assuming $\mu(a_i)$ in (10) to be the mean level of antibodies in the S^- population, the density function of the resulting mixture model using the 'direct' approach is

$$\begin{aligned}f(y_i) = \prod_{i=1}^{n} [&(1-p(a_i))f_{S^-}(y_i;\, \mu(a_i),\, \sigma^2_{S^-})\\ &+p(a_i)f_{S^+}(y_i;\, \delta\mu(a_i),\, \sigma^2_{S^+})],\end{aligned} \tag{12}$$

where $\delta > 1$ is a multiplicative factor accounting for the higher mean levels of antibodies in the S^+ population. In the 'direct' approach, we utilise the underlying structure of the mixture distribution in order to estimate transmission parameters in the unified mechanistic model, thus avoiding dichotomisation of the antibody measurements while accounting for age dependency of the mean and probabilities of the mixture. The resulting structure of the unified mechanistic model is summarised in Figure 3(a).

When analysing cross-sectional data, estimation of the model in (12) can be problematic because of the large number of parameters to estimate. In the absence of a large amount of data, the approach we follow in this paper is to consider two models, one assuming a time-varying seroconversion rate and a constant boosting rate, and a second where the reverse is assumed. Comparison between the two models is then carried out based on a goodness-of-fit index, such as the Akaike Information Criterion (AIC).

Another simplification that we introduce in the maximisation of the likelihood function is to fix the seroreversion rate ω. In practice, we found that using numerical optimisation with a continuous ω was unstable as a result of a very flat likelihood surface.

Alternative empirical approaches to model age-dependency

When the interest is in describing the effect of age on the distribution of antibody data, an empirical, rather than mechanistic approach, may provide a better statistical solution. Additionally, the empirical approach outlined in this section can be used to validate the unified mechanistic model by assessing the discrepancy between the age distributions generated by the two modelling approaches.

To this end, we modify the framework introduced in the previous section by replacing the modelling of mixing probability based on RCMs, and the mean level of antibodies based on AAMs, with their empirical counterparts. More specifically, we model the age-dependency in $\lambda(a)$ and $p(a)$ using a log-linear and logit-linear regression as

$$\begin{aligned}\log\left\{\frac{p(a)}{1-p(a)}\right\} &= \alpha_1 + f_1(a)\\ \log\{\mu(a)\} &= \alpha_2 + f_2(a),\end{aligned} \tag{13}$$

where $f_1(a)$ and $f_2(a)$ are functions that can be specified with the aid of simple graphical tools, such as scatter plots. The resulting structure of the empirical model is summarised in Figure 3b, and we give examples of this in the application of section 'Analysis of malaria serology data from Western Kenya'.

Analysis of malaria serology data from Western Kenya

We analyse data collected from a cross-sectional survey conducted in Rachuonyo South District, in the western Kenyan highlands, in 2011. At the time, malaria transmission in Rachuonyo South was described as generally low but highly heterogeneous, with an average of 14.8% malaria prevalence [31]. Transmission was characterised as seasonal, following peaks in rainfall, typically between March–June and October–November [31, 32].

Most malaria was attributed to *Pf*, with predominant vector species being *Anopheles gambiae s.s.*, *A. arabiensis* and *A. funestus* [33, 34]. Malaria control interventions at the time included the distribution of long-lasting insecticide-treated nets which had been ongoing for many years, and indoor residual spraying which started in 2009 [34]. Further details of the study design and data collection can be found in Bousema *et al.* [31, 34]

In the study, finger prick blood was collected from all participants on filter paper and used to detect total immunoglobulin G antibodies against the blood-stage *Pf* antigen apical membrane antigen 1 (*Pf* AMA1) using the enzyme-linked immunosorbent assay. Optical density (OD) values were obtained for this antigen and are the outcome that we model in this analysis, which we restrict to individuals between 1 and 16 years of age. Children under 1 year old are excluded from the analysis due to the effect of maternal antibodies, which are present at birth, and are believed to wane between 6 and 9 months [9, 17, 35]. The upper age range of 16 years is selected to exclude older individuals whose antibody levels may exhibit a noisier distribution and thus hinder the ability of the model to detect changes in transmission in the recent past from the time of data collection [17].

The data-set consists of $n = 9549$ children. Figure 4 shows the age and OD distributions of the individuals included in the analyses.

We fit both unified mechanistic and empirical models to the *Pf* AMA1 antibody data using the maximum likelihood method of estimation. To obtain 95% confidence intervals (CIs) for the model parameters estimates, we use parametric bootstrap. In this procedure, parameter estimates from the respective models are used to generate 1000 replicate datasets. For each of the datasets, we refit the model and re-extract the parameter estimates in order to construct the bootstrap distribution, and therefore the CIs. We also account for the truncated nature of the antibody distributions, due to the exclusion of individuals under age 1 and over age 16, by using truncated log-Gaussian distributions. The upper limit of the truncation is estimated for each age group as the maximum observed value of OD.

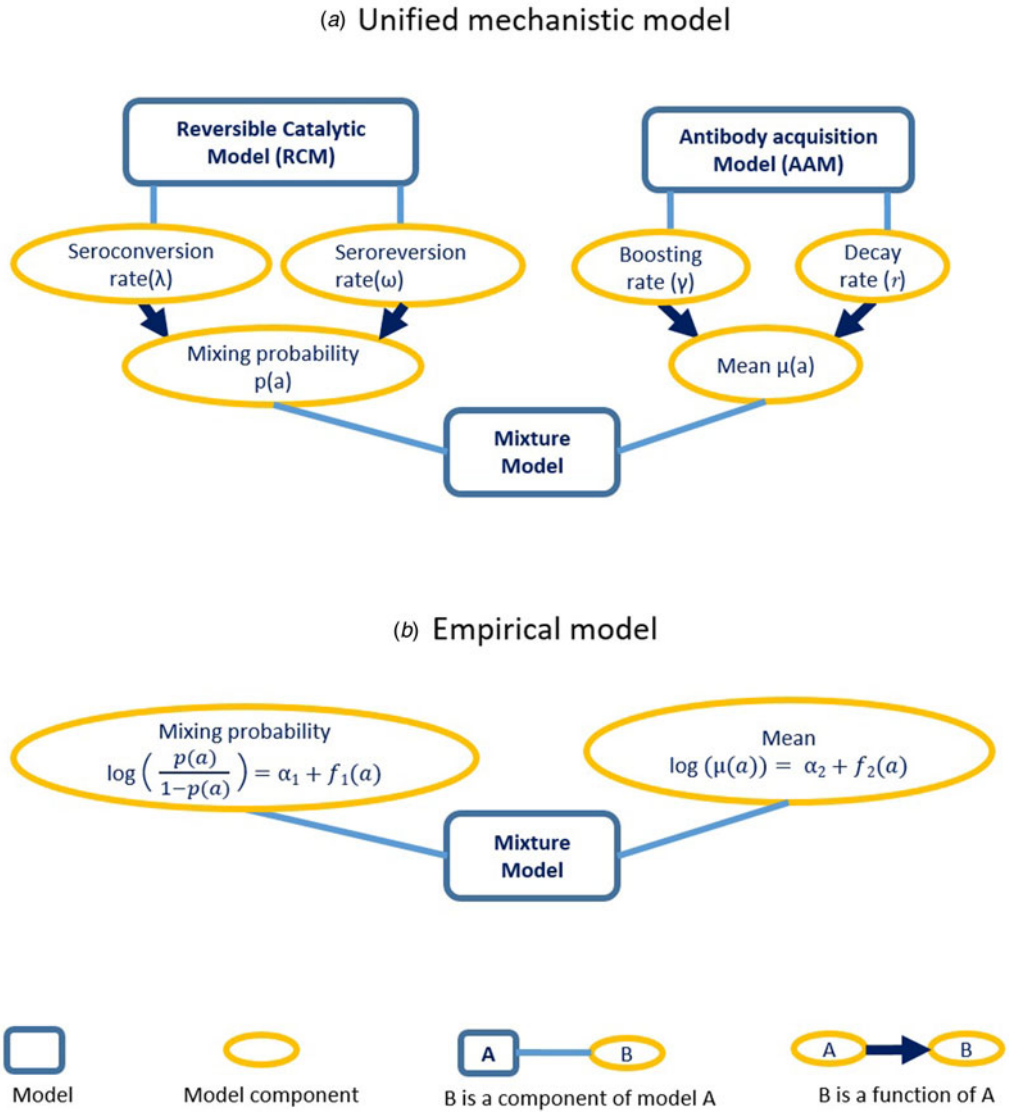

Fig. 3. (a) This figure is a representation of the unified mechanistic model, showing how the reversible catalytic model and antibody acquisition model are incorporated into the mixture model for antibody data. (b) This figure is a representation of the empirical model used to model age-dependence in the mixing probabilities and mean antibody level.

Based on the comparison between the AIC values (see Table 2 in the Supplementary material), preliminary analysis of the *Pf* AMA1 data shows that a unified mechanistic model that assumes a time-varying seroconversion rate $\lambda(a)$ and a constant boosting rate γ provides a better fit to the data than a model where the reverse assumptions is made (i.e. constant λ and time varying $\gamma(a)$). We let ω take three values, namely 0.01, 0.5 and 1, hence assuming that seroreversion events among individuals would occur between 1 and 100 years [8, 9, 26]. In what follows, we present results for the best performing value for ω, i.e. $\omega = 0.01$.

To summarise, the unified mechanistic model parameters to estimate via maximum likelihood are the following: $l_0$and l_1 which are related to the seroconversion rate λ as described by (9) and (11); boosting rate γ and decay rate r from (10); and the mixture distribution parameters δ, $\sigma^2_{S^-}$ and $\sigma^2_{S^+}$ from (12).

For the empirical model, $\mu(a)$ and the mixing probability are modelled according to (13), and are informed by Figure 5. We apply a linear spline with a knot at age 10, based on the empirical trend for $\mu(a)$ observed in Figure 5a, to give

$$\mu(a) = \exp\{\beta_1 + \beta_2 a + \beta_3(a-10)I(a>10)\}, \qquad (14)$$

where $I(a > 10)$ is an indicator function that takes value 1 if $a > 10$, and 0 otherwise. Based on Figure 5b, we introduce the log-transformed age as a logit-linear predictor for $p(a)$, such that

$$p(a) = \frac{\exp\{\tilde{\beta}_0 + \tilde{\beta}_1 \log a\}}{1 + \exp\{\tilde{\beta}_0 + \tilde{\beta}_1 \log a\}}. \qquad (15)$$

Thus, the model parameters to estimate for the empirical model are: the regression coefficients β_1, β_2 and β_3 in (14), and $\tilde{\beta}_0$ and $\tilde{\beta}_1$ in (15); and, as in the unified model, δ, $\sigma^2_{S^-}$ and $\sigma^2_{S^+}$.

Results of this analysis indicate strong evidence of age-dependency for the mixing probabilities of *Pf* AMA1. Figure 6 shows a bi-modal antibody distribution between ages 5 and 10,

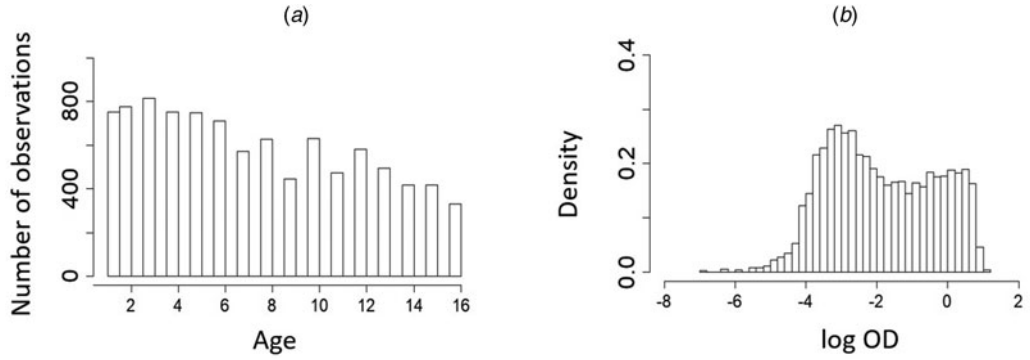

Fig. 4. Descriptive plots of the age distribution (a) and the log OD distribution (b) of individuals aged 1–16, who are included in the *Pf* AMA1 antibody analysis.

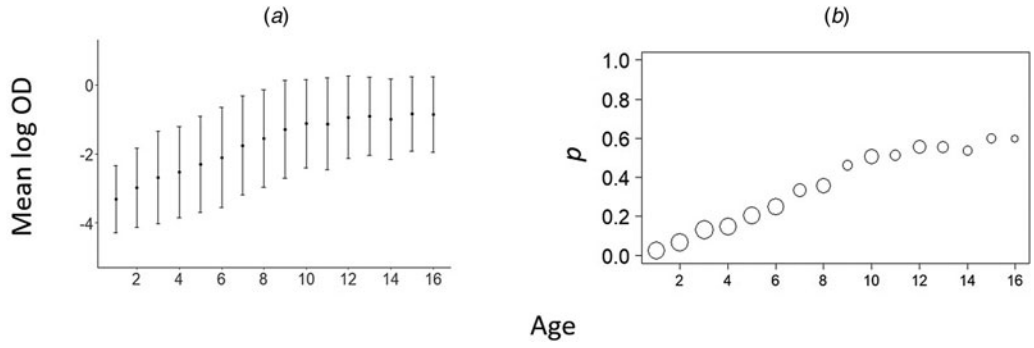

Fig. 5. Exploratory analysis of the Rachuonyo South District *Pf* AMA1 antibody data. (a) This figure shows the geometric mean OD across age while (b) shows the proportions of S^+ individuals, p, as defined by (1), using the seropositivity threshold (i.e. $\mu_{S^-} + 3\sigma_{S^-}$). The circle sizes in (b) are proportional to the sample size in each age group.

which is less evident in younger and older individuals. Both the empirical and mechanistic models are able to capture the increase in the means of antibodies for the S^+ and S^- distributions, with younger children having generally lower antibody levels than older individuals.

By comparing the fitted density functions of mixture distributions between the mechanistic and empirical models for *Pf* AMA1 (Fig. 6), we notice that, while there is a general agreement between the two models, there are visible discrepancies at certain ages. These are more evident in very young individuals at age 1, and in older children from around age 8 onward, where the empirical model indicates a more noticeable peak for the S^- distribution.

Finally, the estimates for δ and σ^2_{s+} from the unified mechanistic and empirical models are comparable, with largely overlapping 95% confidence intervals (Table 1).

With regards to $\lambda(h)$, Figure 7 shows the estimated changes in this parameter in the 16 years before data collection. The results indicate a decrease in transmission in recent years.

Finally, based on the AIC, we note that the unified mechanistic model is larger, suggesting that inferences from the mechanistic model should be drawn with caution. This is because the mechanistic model may not provide an equally good description of the antibody distribution across all ages as shown by the discrepancies between the red and blue lines of Figure 6. However, because the differences between the models are not substantial, we believe that the unified mechanistic model does provide useful insights into time variations of the seroconversion and boosting rates, for which the empirical model does not provide any information.

Discussion

We have introduced a unified mechanistic model which (1) avoids the dichotomisation of continuous antibody data and (2) provides a more flexible way for modelling antibody distributions while allowing for the joint estimation of seroconversion and boosting rates, namely $\lambda(a)$ and $\gamma(a)$, respectively.

The additional flexibility is obtained by modelling the age-dependency of antibody distributions and the temporal variations in $\lambda(a)$ and $\gamma(a)$ which are informed by RCM and AAM, respectively. The disadvantages of dichotomising continuous data into binary data, a common practice in the standard use of RCMs, are well established. Dichotomisation can lead to the loss of information which affects the ability to reliably recover regression relationships and the precision of parameter estimates [36–40]. The proposed unified modelling framework in this paper avoids this problem by making use of the full continuous antibody distribution.

As an alternative approach to the mechanistic framework, we have proposed the use of an empirical approach where the age dependency is informed by the data rather than by biological assumptions. The choice between the unified and empirical models may depend on the research context. The mechanistic approach allows for the estimation of $\lambda(a)$ and $\gamma(a)$ that may be of intrinsic scientific interests, whilst the empirical model does not provide any information on these. In our application, the empirical model provided a better fit to and, hence a better description of, the antibody distributions for different ages, although the discrepancies between the fitted antibody distributions of the empirical and unified models were minimal.

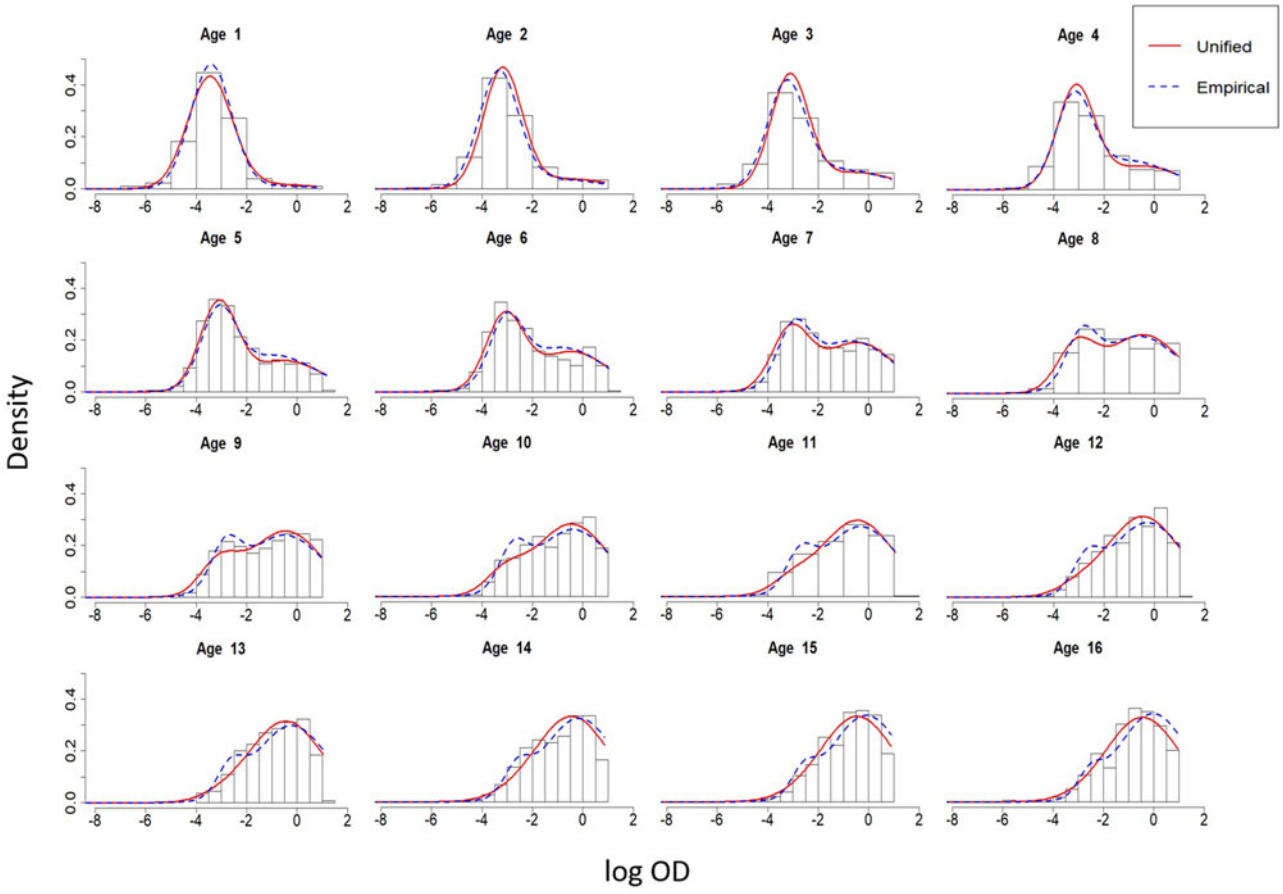

Fig. 6. Age-dependent mixture distributions of *Pf* AMA1 antibodies for individuals 1–16 years of age in Rachuonyo South District. The red line indicates distributions derived from the unified mechanistic model, while the blue dotted line indicates distributions derived from the alternative empirical model.

Table 1. Maximum likelihood estimates with associated 95% CIs (within brackets) for the unified mechanistic model (UFM) and empirical model (EM), fitted to the *Pf* AMA1 antibody data

Equation	Parameter	UFM	EM
Equations (9) and (11)	l_0	−2.696 (−2.627, −2.397)	
	l_1	0.246 (0.202, 0.264)	
Equation (10)	γ	−1.5 (−1.687, −1.291)	
	r	3.806 (3.122, 4.754)	
Equation (12)	δ	31.086 (27.637, 37.837)	28.348 (25.265, 34.197)
	σ^2_{s-}	$2.506 \cdot 10^{-3}$ ($2.169 \cdot 10^{-3}$, $2.914 \cdot 10^{-3}$)	$1.895 \cdot 10^{-3}$ ($1.613 \cdot 10^{-3}$, $2.288 \cdot 10^{-3}$)
	σ^2_{s+}	23.977 (15.783, 46.364)	36.063 (23.244, 70.104)
Equation (14)	β_1		−3.141 (−3.191, −3.087)
	β_2		0.052 (0.046, 0.058)
	β_3		−0.021 (−0.032, −0.005)
Equation (15)	$\tilde{\beta}_0$		−3.031 (−3.194, −2.69)
	$\tilde{\beta}_1$		2.005 (1.915, 2.188)
	AIC	29 791.910	29 711.460

The Akaike Information Criterion (AIC) is also reported.

One of the main issues of the proposed unified modelling framework is that it requires a large amount of data in order to reliably estimate the model parameters. In cases where the separation between the seronegative and seropositive populations is weak, this may result in very uncertain estimates. For example, additional analysis of the antigen *Pf* $MSP1_{19}$ showed limited

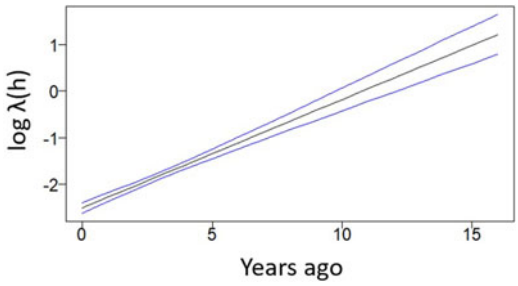

Fig. 7. Changes in λ over historical time as derived from the unified mechanistic model fitted to *Pf* AMA1 antibody data. The blue lines indicate 95% CIs. 'Years ago' corresponds to $(a - h)$ as described in (11).

evidence of a bi-modal distribution or age dependency in the mixture distribution, making the estimation of the proposed model unfeasible. More generally, mixture models may be difficult to estimate, especially in areas of high transmission where a great majority of the population is seropositive [3, 17]. Additionally, the seroreversion rate ω may also difficult to estimate in this scenario and, for this reason, is often fixed [4]. This is one of the main limitations in RCMs, which also applies to the unified mechanistic model. Generally, to alleviate the problem of over-parametrisation, further simplification of the model may be considered by, for example, assuming a constant $\lambda(a)$. In such scenarios, however, we believe selection between models should also be guided by scientific, a not purely statistical, judgement, while also taking into consideration the levels uncertainty inherent to each model.

More complex functional forms for modelling time-changes in $\lambda(a)$ and $\gamma(a)$ than a log-linear regression, as used in this paper, could also be considered. For example, polynomials and smoothing splines would be a natural choice to increase the flexibility of the model. Alternatively, contextual knowledge on events that may have significantly impacted transmission in the past, such as interventions and malaria outbreak, may also be used to inform the modelling of $\lambda(a)$ and $\gamma(a)$. However, the increased flexibility comes at the cost of an increased model complexity which may make the model very difficult, if not impossible, to estimate.

Conclusion

We have proposed a unified modelling framework for the analysis of malaria serology data which allows for the joint estimation of seroconversion and boosting rates. Our framework makes the best possible use of the data by avoiding the dichotomisation of the continuous antibody measurements, a common practice in the analysis of malaria serology data. More importantly, the unified framework allows to critically assess and evaluate assumptions on the heterogeneity of biological indicators of malaria transmission using a principled likelihood-based framework.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S0950268821000753.

Data. R scripts for the implementation of the unified mechanistic and empirical models are available on request from the authors.

Acknowledgements. We thank all those who contributed to the collection of data included in this paper, specifically the survey participants in Kenya, and the KEMRI/CDC research team. Thanks to Dr Lindsey Wu, Professor Chris Drakeley and Dr Gillian Stresman for useful discussions on this work.Irene Kyomuhangi is a Commonwealth Scholar, funded by the UK government

Conflict of interest. None.

References

1. **World Health Organization** *et al.* (2019) World malaria report 2019.
2. **The Malaria Atlas Project**. Available at https://malariaatlas.org/ (Accessed 11 December 2019).
3. **Pothin E *et al.*** (2016) Estimating malaria transmission intensity from *Plasmodium falciparum* serological data using antibody density models. *Malaria Journal* **15**, 79.
4. **Sepúlveda N *et al.*** (2015) Current mathematical models for analyzing anti-malarial antibody data with an eye to malaria elimination and eradication. *Journal of Immunology Research,* **2015**, Article ID 738030. https://doi.org/10.1155/2015/738030.
5. **Kilama M *et al.*** (2014) Estimating the annual entomological inoculation rate for *Plasmodium falciparum* transmitted by *Anopheles gambiae* sl using three sampling methods in three sites in Uganda. *Malaria Journal* **13**, 111.
6. **Smith T *et al.*** (2004) Relationships between the outcome of *Plasmodium falciparum* infection and the intensity of transmission in Africa. *The American Journal of Tropical Medicine and Hygiene* **71**, 80–86.
7. **Tusting LS *et al.*** (2014) Measuring changes in *Plasmodium falciparum* transmission: precision, accuracy and costs of metrics. *Advances in Parasitology* **84**, 151–208.
8. **Corran P *et al.*** (2007) Serology: a robust indicator of malaria transmission intensity? *Trends in Parasitology* **23**, 575–582.
9. **Drakeley CJ *et al.*** (2005) Estimating medium-and long-term trends in malaria transmission by using serological markers of malaria exposure. *Proceedings of the National Academy of Sciences* **102**, 5108–5113.
10. **Bousema T *et al.*** (2010) Serologic markers for detecting malaria in areas of low endemicity, Somalia, 2008. *Emerging Infectious Diseases* **16**, 392.
11. **Long CA *et al.*** (2017) Immune responses in malaria. *Cold Spring Harbor Perspectives in Medicine* **7**, a025577.
12. **Cockburn IA *et al.*** (2018) Malaria prevention: from immunological concepts to effective vaccines and protective antibodies. *Nature Immunology* **19**, 1199–1211.
13. **Akpogheneta OJ *et al.*** (2010) Boosting antibody responses to *Plasmodium falciparum* merozoite antigens in children with highly seasonal exposure to infection. *Parasite Immunology* **32**, 296–304.
14. **Kwiatkowski DP** (2005) How malaria has affected the human genome and what human genetics can teach us about malaria. *The American Journal of Human Genetics* **77**, 171–192.
15. **Bollaerts K *et al.*** (2012) Estimating the population prevalence and force of infection directly from antibody titres. *Statistical Modelling* **12**, 441–462.
16. **Hens N *et al.*** (2012) *Modeling Infectious Disease Parameters Based on Serological and Social Contact Data: A Modern Statistical Perspective*, vol. **63**. Springer Science & Business Media 3, pp. 79–85.
17. **Yman V *et al.*** (2016) Antibody acquisition models: a new tool for serological surveillance of malaria transmission intensity. *Scientific Reports* **6**, 19472.
18. **Weber GE *et al.*** (2017) Sero-catalytic and antibody acquisition models to estimate differing malaria transmission intensities in Western Kenya. *Scientific Reports* **7**, 16821.
19. **Bosomprah S** (2014) A mathematical model of seropositivity to malaria antigen, allowing seropositivity to be prolonged by exposure. *Malaria Journal* **13**, 12.
20. **Cook J *et al.*** (2010) Using serological measures to monitor changes in malaria transmission in Vanuatu. *Malaria Journal* **9**, 169.
21. **Cook J *et al.*** (2011) Serological markers suggest heterogeneity of effectiveness of malaria control interventions on Bioko Island, equatorial Guinea. *PLoS ONE* **6**, e25137.
22. **Del Fava E *et al.*** (2016) Estimating age-specific immunity and force of infection of varicella zoster virus in Norway using mixture models. *PLoS ONE* **11**(9), pe0163636.
23. **von Fricken ME *et al.*** (2014) Age-specific malaria seroprevalence rates: a cross-sectional analysis of malaria transmission in the Ouest and Sud-Est departments of Haiti. *Malaria Journal* **13**, 361.

24. **Arnold BF *et al.*** (2017) Measuring changes in transmission of neglected tropical diseases, malaria, and enteric pathogens from quantitative antibody levels. *PLoS Neglected Tropical Diseases* **11**, e0005616.
25. **Ashton RA *et al.*** (2015) Geostatistical modeling of malaria endemicity using serological indicators of exposure collected through school surveys. *The American Journal of Tropical Medicine and Hygiene* **93**, 168–177.
26. **Akpogheneta OJ *et al.*** (2008) Duration of naturally acquired antibody responses to blood-stage *Plasmodium falciparum* is age dependent and antigen specific. *Infection and Immunity* **76**, 1748–1755.
27. **Varela M-L *et al.*** (2020) Practical example of multiple antibody screening for evaluation of malaria control strategies. *Malaria Journal* **19**, 1–12.
28. **White MT *et al.*** (2014) Dynamics of the antibody response to *Plasmodium falciparum* infection in African children. *The Journal of Infectious Diseases* **210**, 1115–1122.
29. **Moormann AM** (2009) How might infant and paediatric immune responses influence malaria vaccine efficacy? *Parasite Immunology* **31**, 547–559.
30. **Doolan DL *et al.*** (2009) Acquired immunity to malaria. *Clinical Microbiology Reviews* **22**, 13–36.
31. **Bousema T *et al.*** (2013) The impact of hotspot-targeted interventions on malaria transmission: study protocol for a cluster-randomized controlled trial. *Trials* **14**, 1–12.
32. **Stresman GH *et al.*** (2017) Impact of metric and sample size on determining malaria hotspot boundaries. *Scientific Reports* **7**, 45849.
33. **Stuckey EM *et al.*** (2012) Simulation of malaria epidemiology and control in the highlands of western Kenya. *Malaria Journal* **11**, 357.
34. **Bousema T *et al.*** (2016) The impact of hotspot-targeted interventions on malaria transmission in Rachuonyo South District in the Western Kenyan Highlands: a cluster-randomized controlled trial. *PLoS Medicine* **13**, e1001993.
35. **Dobbs KR *et al.*** (2016) Plasmodium malaria and antimalarial antibodies in the first year of life. *Parasitology* **143**, 129–138.
36. **Fedorov V *et al.*** (2009) Consequences of dichotomization. *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry* **8**, 50–61.
37. **Altman DG *et al.*** (2006) The cost of dichotomising continuous variables. *BMJ* **332**, 1080.
38. **Royston P *et al.*** (2006) Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine* **25**, 127–141.
39. **Bennette C *et al.*** (2012) Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents. *BMC Medical Research Methodology* **12**, 21.
40. **Kyomuhangi I *et al.*** (2020) Understanding the effects of dichotomization of continuous outcomes on geostatistical inference. *Spatial Statistics*, 100424.