Lancaster University

Threshold-free statistical methods for the analysis of continuous health outcomes, with applications to malaria serology

by

Irene Kyomuhangi

This thesis is submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

in the

Faculty of Health and Medicine,

Lancaster University
Medical School

August 2021

Declaration

I, Irene Kyomuhangi, declare that this thesis titled, "*Threshold-free statistical methods for the analysis of continuous health outcomes, with applications to malaria serology*" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____ Date: _____

Abstract

Continuous measurements of health outcome data are often dichotomized into binary (i.e. positive/negative) data for diagnosis and subsequent statistical analysis. The disadvantages of dichotomizing continuous data for statistical inference are well established in the literature, yet this practice is commonplace in health research.

In this thesis, we investigate the impact of dichotomization of data when the aim of analysis is to determine disease prevalence and risk, and propose solutions to some of the main challenges introduced by dichotomization in the context of global heath research.

First, using model-based geostatistics, we show how dichotomization reduces the predictive performance of geostatistical models through loss of information and by reducing the reliability of parameter estimates. We demonstrate this using a simulation study, as well as mapping prevalence and risk of anaemia in Ethiopia, and stunting in Ghana.

We then explore the limitations dichotomization introduces to estimation of malaria transmission in serology models, and propose a novel flexible and unified modelling framework which uses continuous antibody measurements instead of dichotomized data to estimate transmission intensity. Using Western Kenya, we demonstrate the properties of this new approach.

Finally, we address the use of thresholds for dichotomization of continuous antibody measurements when the goal is to estimate malaria seroprevalence. We utilize the principles of the unified modelling framework to develop a threshold-free approach to estimating seroprevalence. Using the same Western Kenyan data-set, we show how this new approach improves model fit and provides more consistent estimates than traditional methods.

Together, these investigations demonstrate the significant impact dichotomization of continuous data has on statistical inference across different areas of health research, and that this practice should be avoided where possible.

Acknowledgements

First, I would like to thank my primary supervisor Dr. Emanuele Giorgi, who has been an absolute pleasure to work with. I am grateful for his mentorship, support, enthusiasm and confidence throughout my PhD. I am also grateful for the support and contribution of my second supervisor Dr Thomas Keegan, as well as support from Dr. Luigi Sedda, Prof. Peter Diggle and Prof. Jo Knight, who at various points offered advice and direction in my research.

I would like to express my gratitude to all the collaborators who made this work possible. Prof. Niel Hens and his group at Hasselt University , as well as Prof. Chris Drakeley and his group at LSHTM for hosting me on research visits, and sharing data and expertise. Thanks to Malaria Consortium, KEMRI and DHS for sharing data, and all co-authors for their contributions to the studies presented within this thesis. Special thanks to all the in-country research teams who did the extensive and important collection of the data included in this research.

My sincere thanks goes to all the members of CHICAS for their professional help and camaraderie throughout my PhD. Special thanks to Remy, Rachel, Lisa, Poppy, Fran, Olatunji, Erick, Max, Ben, Claudio and Barry. Additionally, I would like to thank colleagues and friends from DHR: Maddy, Annie, Clare, Mario, and Jack for the chatty coffee breaks.

I am very grateful for the financial support from my PhD funders, the Commonwealth Scholarship Commission, which enabled me to undertake this research.

Finally, I would like to thank all my family and friends who enrich my life and continue to support my work. Special thanks to my parents, my siblings, and my partner for their unwavering belief in me.

Contents

List of Figures

List of Tables

List of Papers

Paper 1 ***Understanding the effects of dichotomization of continuous outcomes on geostatistical inference.***

Authors: Irene Kyomuhangi, Tarekegn A. Abeku, Matthew J. Kirby, Gezahegn Tesfaye, Emanuele Giorgi.

Published in: *Spatial Statistics*;

Contribution: Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Visualization, Writing-original draft, Writing-review and editing.

Paper 2 ***A unified and flexible modelling framework for the analysis of malaria serology data***

Authors: Irene Kyomuhangi, Emanuele Giorgi

Published in: *Epidemiology and Infection*;

Contribution: Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Visualization, Writing-original draft, Writing-review and editing.

Paper 3 ***A threshold-free approach with age-dependency for estimating malaria seroprevalence***

Authors: Irene Kyomuhangi, Emanuele Giorgi

Submitted to *Malaria Journal*;

Contribution: Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Visualization, Writing-original draft, Writing-review and editing.

For Clemence and Fabian

Chapter 1

Introduction

Across the world, data on health outcomes are routinely collected for monitoring and surveillance of disease prevalence and risk. In the global south, for example, data are collected on a variety of disease indicators for malaria, anaemia, diarrhoeal diseases, HIV/AIDS, and Neglected Tropical Diseases (NTDs), among others.

There are a variety of ways these data can be managed and analysed in order to make statistical inference on disease prevalence and risk, as well as monitor how these parameters change over time. A key goal of analysis is to make informed data-driven decisions on intervention strategy and health policy.

In this thesis we explore research questions around improving the estimation and prediction of disease prevalence and risk, with a focus on applications in the global south. Specifically, we challenge the common practice of dichotomizing continuous health outcome data into binary data, i.e. positive/negative data, for statistical inference.

First, we explore how dichotomization affects the performance of geostatistical models in mapping disease prevalence and risk, using anaemia and stunting data. We then address dichotomization in the analysis of malaria serology data, and present a novel modelling framework which uses continuous antibody measurements, rather than dichotomized data, to estimate malaria transmission intensity. Finally, we provide a new threshold-free approach for estimating malaria seroprevalence.

1.1 Model-Based Geostatistics and mapping disease prevalence

Maps are important tools in the understanding and visualization of infectious disease prevalence and risk on a spatial scale. Specifically, for infectious diseases where location of individuals is correlated to disease, maps underpin monitoring and surveillance of these diseases, and often provide an evidence base for the design and implementation of prevention and control programs. Model-based Geostatistics (MBG) is a branch of spatial statistics which allow us to model spatial variation in disease prevalence, taking into account various sources of uncertainty.

1.1.1 The standard geostatistical model for mapping disease prevalence

Let Y_i be the number of individuals who test positive for a disease of interest out of a total number of m_i individuals (i.e. m is the sample size), sampled at locations $x_i : i = 1, ...n$ (with n being the total number of locations), which represent distinct geographical coordinates of sampled villages or households.

The standard geostatistical model for this data is a generalized linear model with a binomial distribution, logistic link function, and a linear predictor [1, 2]. The linear predictor consists of explanatory variables and an additional Gaussian spatial random process $S = S(x) : x \in \mathbb{R}^2$ on which prevalence is statistically dependent.

If $p(x)$ represents the prevalence at location x, the sampling distribution of the resulting data is denoted as $Y_i|S(x_i), m_i, p(x_i) \sim Bin(m_i, p(x_i))$ for $i = 1, ...n$. The link for $p(x_i)$ at different locations of x in a logistic regression model is defined as

$$log\left(\frac{p(x_i)}{1-p(x_i)}\right) = \alpha + \beta^T d(x_i) + S(x_i) + Z_i \tag{1.1}$$

where $d(x_i)$ is a vector of location-specific covariates such as social economic status and interventions at household level, and β is the corresponding vector of regression coefficients for these covariates. $S(x_i)$ is a spatial random effect used to account for spatial correlation between observations, while Z_i is the 'nugget effect' or 'noise'. The nugget effect represents the unstructured residual variation which can either be small-range spatial variation, or within-household variations like genetic variation of individuals. The goal of most geostatistical analysis is to predict S at an unobserved location x.

We model $S(x)$ as a stationary and isotropic Gaussian process with mean zero,

variance σ^2, and an exponentially decaying correlation function expressed as

$$\rho(u) = e^{-u/\phi} \tag{1.2}$$

where u is the Euclidean distance between points x and x', and ϕ is the scale of spatial correlation.

The model in (1.1) can be expanded to include individual level information such that the outcome variable Y_i becomes Y_{ij} representing the outcome for the $j-$th individual in the $i-$th household. The sampling model for the resulting data is then denoted as $Y_{ij} \sim Bernoulli(p_{ij})$ where p_{ij} is the probability of a positive test for the individual i in household j. The link for p_{ij}, at different locations of x in a logistic regression model takes the form

$$log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha + \beta^T d(x_i) + \gamma^T e_{ij} + S(x_i) + Z_{ij} \tag{1.3}$$

where we distinguish between covariates, $d(x_i)$, that express properties of the locations and covariates, e_{ij}, for individual traits e.g. age and gender.

1.1.2 Dichotomization of data in geostatistical inference

The standard geostatistical model is based on binary (ie. positive/negative) disease outcome data. Equation (1.1) takes aggregated binary data, while (1.3) takes individual-level binary data at a location x. This reflects the reality that in practice, continuous health outcome data are often dichotomized into binary data for diagnosis and subsequent statistical analysis. The disadvantages of this approach include loss of information, which impacts the ability to reliably recover regression relationships, and reduces the precision of parameter estimates, [3–10]. Despite these drawbacks, dichotomization of data is common practice in health research.

In the first paper (chapter 2), we investigate how the dichotomization of continuous outcomes affects the performance of geostatistical models both in parameter estimation and prediction of disease risk. To address this question, we carry out a comparative analysis between geostatistical models where the outcome variable is continuous, versus models where the derived binary outcome is analysed.

We compare two possible approaches for disease mapping: the first defines a geostatistical model for a continuous measurement, and uses this to map disease prevalence and predict the probability of exceeding a threshold relevant to public health policy. The second approach first dichotomizes the continuous measurement, based on a diagnostic threshold, into a binary outcome, and then develops a geostatistical logistic regression based on this binary outcome.

The specific research questions we address are: 1) how does spatial correlation affect the loss of information and estimation of regression relationships, 2) can dichotomization lead to substantially different and more uncertain spatial predictions in disease prevalence, and 3) how can we use linear geostatistical models to map disease prevalence and thus avoid the drawbacks of dichotomization?

1.2 Models for analysing malaria serology data

In the second paper (chapters 3), we address the use of dichotomization in the analysis of malaria serology data. Serology (i.e. antibody) data are an increasingly important tool in malaria surveillance. Current models for analysing malaria serology data are limited by the need to dichotomize continuous antibody measurements, as well as strict assumptions about malaria transmission dynamics. In this paper, we propose a novel unified modelling framework that eliminates the need for dichotomization, and combines existing models while addressing their limitations. We also propose an alternative empirical approach to analysing malaria serology data which relies entirely on the data, rather than biological assumptions inherent to existing models.

1.2.1 The role of antibodies in protection against clinical malaria

Malaria is a mosquito-borne disease caused by *Plasmodium* parasites, and carried by infected *Anopheles* mosquitoes. The majority of malaria cases worldwide are attributed to the *Plasmodium falciparum*, (*Pf*) species, however there are regional variations in the prevalence of *Plasmodium* species including *P.vivax P.malariae*, *P.ovale* and *P.knowlesi* [11, 12].

Once in the human host, the malaria parasites undergo several transformations and migrations as part of their life cycle [13, 14]. Clinical manifestations of malaria occur during the blood stage of the parasites, and antibodies which target blood stage malaria parasites are important in the development of immunity to the disease. These antibodies occur in a broad range of specificities, and their functions include blocking important parasite processes such as adhesion to and invasion of cells, as well as cooperation with immune cells to tag and clear the parasites [14, 15].

In general, antibody responses to malaria antigens are characterised by the following properties: a) they confer 'non-sterile' immunity (i.e. incomplete protection), such that despite the presence of antimalarial antibodies, individuals remain susceptible to re-infections, b) antibody levels are boosted upon re-infection, c) in malaria endemic settings, antibody levels generally increase with age, and d)

malaria antibodies decay in the absence of re-infection [14, 16–20].

1.2.2 The role of antibodies in malaria surveillance

Traditional methods of measuring malaria transmission rely on the detection of the *Plasmodium* parasite in humans and mosquito populations. Parasite prevalence (PrP) is determined by the proportion of infected individuals at time of data collection [20, 21], while the gold-standard measurement of transmission, the entomological inoculation rate (EIR), is determined by the frequency at which individuals are bitten by infectious mosquitoes [22]. Both of these measures are affected by several factors, including: a) seasonal variations in transmission, b) low densities of parasites in both humans and mosquitoes which result in low probability of sampling infected people and mosquitoes particularly in low transmission areas, and c) deriving EIR, is labour-intensive and expensive [20–25].

Antibodies provide an alternative approach to measuring transmission intensity. Because antibodies persist after infection, they: a) provide information on cumulative exposure to malaria parasites over time, b) are more resistant to the effects of seasonality in transmission, and c) allow estimation of transmission intensity with more feasible sample sizes even in low transmission settings [21, 25–29].

1.2.3 Analysing malaria antibody data

Based on the profile of malaria antibodies described in section 1.2.1, three types of models for the analysis of malaria serology data have been proposed so far: Reversible catalytic models (RCMs), Antibody Acquisition Models (AAMs), and the Antibody Density Model (ADM).

These models provide different measures of transmission intensity. In the RCM, transmission is defined as the rate at which individuals convert from seronegative to seropositive upon exposure to malaria parasites; in the AAM, transmission intensity is defined as the rate at which antibodies are boosted upon exposure to parasites; while in the AAM, transmission intensity is defined as the rate of exposure to an infectious mosquito bite.

Reversible catalytic models (RCMs)

RCMs, are most commonly used. These models rely on first dichotomizing continuous antibody measurements in order to define seropositive (S^+) or seronegative (S^-) status. Following dichotomization, the resulting S^+ and S^- outcomes are modelled using the RCM, which operates under the assumption that individuals are born S^-, they can 'seroconvert' to S^+ upon exposure to malaria, and in the

absence of exposure, 'serorevert' to S^-. The seroconversion rate (λ) is related to underlying transmission intensity while the seroreversion rate (ω) represents antibody decay in the absence of malaria infection [20, 25, 26, 30–32]. Figure 1.1 illustrates this mechanistic model.

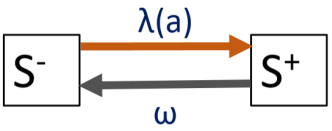

Figure 1.1: A representation of the reversible catalytic model (RCM) where individuals transition between seronegative (S^-) and seropositive (S^+) states through the seroconversion rate, $\lambda(a)$ and the seroreversion rate, ω

.

Dichotomization in the RCM requires thresholds to define S^+ status. These thresholds are usually estimated using 1) a negative control group, which usually consists of serum form non-exposed individuals (e.g. Europeans who never travelled to endemic areas) [25, 33], or 2) a mixture model which assumes a latent bi-modal distribution of S^- and S^+ populations in the study sample itself [20]. The former method is less prevalent due to potential underlying differences between immune responses in these individuals and those living in endemic countries [21].

For the finite mixture model, assuming independent and identically distributed realizations for a sample of n individuals, we write the density function of Y_i as

$$f(y_i) = \prod_{i=1}^{n} \left[(1-p) f_{S^-}(y_i; \mu_{S^-}, \sigma^2_{S^-}) + p f_{S^+}(y_i; \mu_{S^+}, \sigma^2_{S^+})\right] \tag{1.4}$$

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^+. By using the point estimates of the mean, $\hat{\mu}_{S^-}$ and standard deviation, σ_{S-}, of the seronegative distribution S^-, the seropositive threshold is often set to $\hat{\mu}_{S^-} + 3\sigma_{S^-}$ [20, 32, 34, 35], while some studies have used $\hat{\mu}_{S^-} + 2\sigma_{S^-}$ and $\hat{\mu}_{S^-} + 5\sigma_{S^-}$ [36–39]. Alternatively, thresholds can be defined based on the predictive probability of being S^+ resulting from the fitted mixture distribution [20]. After classification, the S^+ and S^- outcome variable is input for the RCM.

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 at age a, i.e $p(a)$, are expressed by the following differential equation

$$\frac{dp}{da} = \lambda(a)(1 - p(a)) - \omega p(a). \tag{1.5}$$

Three transmission profiles are often proposed to model $\lambda(a)$: constant transmission; a sharp stepwise drop in transmission which may occur when transmission is suddenly interrupted due to, say, the introduction of an intervention; and a linear drop in transmission when the reduction is more gradual [20, 30, 32, 40, 41]. These are strong assumptions on the temporal dynamics of transmission, and their validity is often questionable. Additionally, RCMs do not sufficiently account for antibody boosting due to repeated exposure to malaria parasites. A more recent study by Varela et al. [42] proposes an extension to the RCM where the number of times that λ changed in the past, is also estimated from the data.

In order to circumvent the problem of dichotomization, and account for antibody boosting, the AAM, and the ADM have been proposed.

Antibody Acquisition Models (AAMs)

The AAM, rather than the dichotomize the data, makes use of the full continuous antibody measurement to obtain an alternative measure of transmission [20, 30, 31, 40]. The AAM relies on the assumption that the boosting rate (γ), i.e the rate at which antibodies are acquired upon exposure to parasites, can be used as a proxy for the underlying transmission intensity, and in the absence of exposure, antibodies decay at rate r. These dynamics are illustrated in figure 1.2.

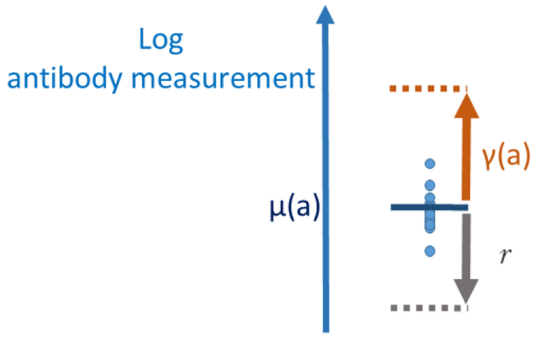

Figure 1.2: A representation of the antibody acquisition model (AAM) where antibody levels are boosted at rate $\gamma(a)$ upon exposure, and decay at rate r in the absence of exposure.

Assuming that antibody levels of individuals at age a follow a log-Gaussian distribution with mean $\mu(a)$ and variance σ^2 [20, 30, 40], let $\mu(a)$ denote the average antibody values (on the log scale) in the general population of individuals of age a. According to the AAM, the temporal dyamics that regulate $\mu(a)$ can be expressed by the following differential equation

$$\frac{d\mu}{da} = \gamma(a) - r\mu(a). \tag{1.6}$$

This equation can be used to infer changes in average antibody levels as a function of age a. $\gamma(a)$ is often modelled according to the same three transmission profiles described for $\lambda(a)$ in the RCM.

Antibody Density Model (ADM)

Similar to the AAM, the ADM makes use of antibody boosting to derive a measure of transmission intensity [21]. The model assumes that the rate at which an individual is bitten by an infectious mosquito, $\tilde{\lambda}$, induces an antibody boost in the individual, and that in the absence of exposure, antibody levels decay at rate $\tilde{\rho}$. Implementation of this model requires discretizing the antibody antibody measurements into compartments where individuals antibodies are boosted or decay. In this framework, individuals move into higher antibody compartments at rate $\tilde{\lambda}$ as their antibodies are boosted by exposure to parasites, and inversely, they move into lower antibody compartments at rate $\tilde{\rho}$ as their antibodies decay in the absence of exposure. In the ADM, $\tilde{\lambda}$ is the measure of underlying transmission intensity. Let y_i denote compartment i, with width Δ, where $1 \leq i \leq N$ and N is the total number of compartments. According to the ADM, the proportion k, of the population in each compartment, y_i, at time t, is defined by the following differential equation

$$\frac{dy_i}{dt} = \tilde{\lambda} \sum_{\substack{j<i \\ i \neq 1}} k_{ij} y_j + \frac{\tilde{\rho}}{\Delta} y_{i+1} - \tilde{\lambda} \sum_{\substack{h>i \\ i \neq N}} k_{hi} y_i - \frac{\tilde{\rho}}{\Delta} y_i, \tag{1.7}$$

where h, i, j index the antibody compartments. The dynamics described in this equation are illustrated in figure 1.3.

The RCM, AAM and ADM have strengths and limitations which are summarised in table 1.1. In the second paper (chapter 3) of this thesis, we address many of these limitations and propose a novel unified mechanistic model which combines the properties of the RCM, mixture model and AAM. This novel framework also provides additional flexibility in how we estimate malaria transmission intensity. The key features of this new framework are 1) the use of continuous antibody measurements, rather than dichotomized data 2) age-dependency of the antibody levels and mixture distribution is accounted for, 3) assumptions around malaria transmission profiles are relaxed, 3) added flexibility through linear regression, and 4) joint estimation of the transmission parameters $\lambda(a)$ and $\gamma(a)$.

Furthermore, as an alternative to the unified mechanistic model, we also present an alternative empirical approach to modelling serology data where analysis is based entirely on the data, rather than biological assumptions inherent to existing

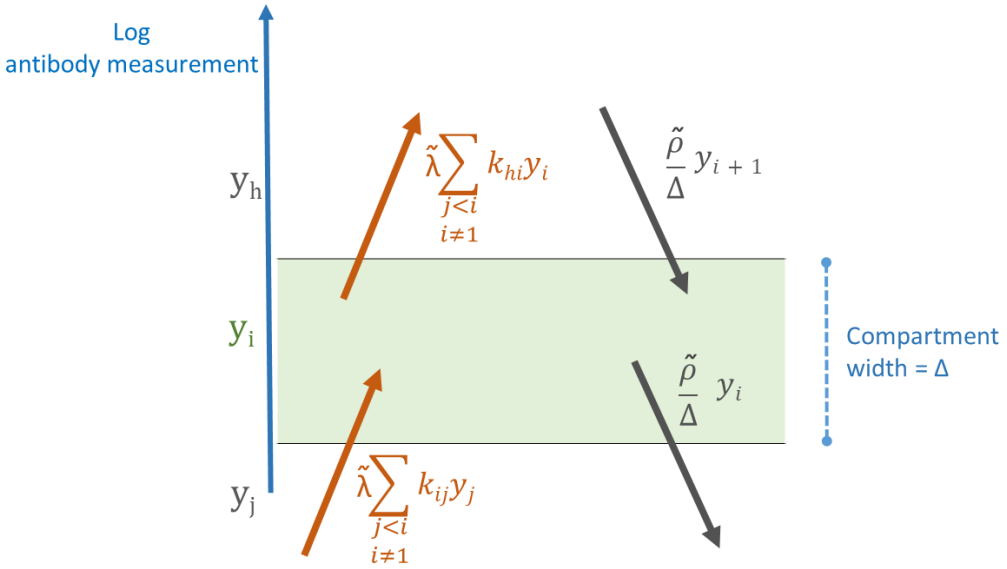

Figure 1.3: A representation of the antibody density model (ADM) where individuals transition between antibody level compartments of width Δ. The transitions are driven by the rate of exposure $\tilde{\lambda}$, which induces a boost of antibodies to higher compartments, as well as antibody decay ρ, which reverts individuals to lower compartments.

models. The empirical model can be used to validate the assumptions of the unified mechanistic model.

Finally, in the third paper (chapter 4), we use the empirical model introduced in the second paper to develop a threshold-free approach for estimating seroprevalence. While the goal of the RCM, AAM and ADM is to investigate historical changes in transmission intensity, seroprevalence itself is a useful metric as it provides as snap-shot of malaria exposure detected at a specific time point and location. As previously described, in order to estimate seroprevalence, individuals are first classified as S^- or S^+ based on a suitable threshold. However, the choice of the threshold is arbitrary, and the same threshold is used across all data, ignoring age dependency of antibody levels.

Therefore, we present a threshold-free approach to estimating seroprevalence which 1) accounts for age dependency of antibody levels and the mixture distribution, 2) eliminates the need for a threshold, and 3) propagates the uncertainty in the seroprevalence estimate. We demonstrate how to propagate the uncertainty around these seroprevalence estimates in further analysis, using the RCM as an example. Note that in this new approach individuals are assigned 'seropositive' or 'seronegative' based on both their antibody level for their age, whereas in classical analysis, this classification is based solely on the individual's antibody level.

The malaria serology models proposed in this thesis aim to provide a more statisti-

Table 1.1: Summary of strengths and limitations of current malaria serology models

	Model		
	RCM	AAM	ADM
Parameters	seroconversion rate (λ) seroreversion rate (ρ)	boosting rate (γ) decay rate (r)	exposure rate ($\tilde{\lambda}$) decay rate ($\tilde{\rho}$)
Strengths	· effective in low transmission settings where seropositivity correlates to exposure · Data from multiple antigens can be combined	· Uses full continuous antibody measurement · Accounts for boosting due to re-exposure · effective even in high transmission settings · Better precision in estimation of parameters than the RCM	· Accounts for boosting due to re-exposure · Better precision in estimation of parameters than the RCM
Limitations	· The need to dichotomize continuous antibody measurements · Strong assumptions around the temporal dynamics of transmission · Difficulty in accounting for boosting due to re-exposure · Not effective in high transmission areas where many individuals are seropositive	· Strong assumptions around the temporal dynamics of transmission · Assumes log normal distribution	· The need to discretize continuous antibody data · Computationally intensive

cally sound likelihood-based approach to analysing malaria serology data without the use of thresholds. These models assume a mixture distribution of antibody data - however, in cases where the antibody distribution is not a mixture, or where there is poor separation between the components of the mixture, the utility of these models may be limited.

Chapter 2

Paper 1: Understanding the effects of dichotomization of continuous outcomes on geostatistical inference

Irene Kyomuhangi[1], Tarekegn A. Abeku[2], Matthew J. Kirby[2,3], Gezahegn Tesfaye[2,4], Emanuele Giorgi[1]

1 CHICAS, Lancaster Medical School, Lancaster University, Lancaster, UK
2 Malaria Consortium, UK
3 London School of Hygiene and Tropical Medicine, UK
4 PATH/MACEPA, Ethiopia

Published in: *Spatial Statistics*

Summary

Diagnosis is often based on the exceedance or not of continuous health indicators of a predefined cut-off value, so as to classify patients into positives and negatives for the disease under investigation. In this paper, we investigate the effects of dichotomization of spatially-referenced continuous outcome variables on geostatistical inference. Although this issue has been extensively studied in other fields, dichotomization is still a common practice in epidemiological studies. Furthermore, the effects of this practice in the context of prevalence mapping have not been fully understood. Here, we demonstrate how spatial correlation affects the loss of information due to dichotomization, how linear geostatistical models can be used to map disease prevalence and thus avoid dichotomization, and finally, how dichotomization affects our predictive inference on prevalence. To pursue these objectives, we develop a metric, based on the composite likelihood, which can be used to quantify the potential loss of information after dichotomization without requiring the fitting of Binomial geostatistical models. Through a simulation study and two applications on disease mapping in Africa, we show that, as thresholds used for dichotomization move further away from the mean of the underlying process, the performance of binomial geostatistical models deteriorates substantially. We also find that dichotomization can lead to the loss of fine scale features of disease prevalence and increased uncertainty in the parameter estimates, especially in the presence of a large noise to signal ratio. These findings strongly support the conclusions from previous studies that dichotomization should be always avoided whenever feasible.

Keywords: binary data; dichotomization; disease mapping; linear geostatistical model; model-based geostatistics; prevalence.

2.1 Introduction

Continuous measurements of disease indicators - e.g. concentration of antibodies in a blood sample - are used in many branches of health research to aid diagnosis and treatment of patients, as well as monitoring and surveillance of diseases in populations. Diagnosis is often based on the exceedance or not of a cut-off value by the continuous indicator, to identify positives and negatives for the disease of interest [43]. In some cases, for instance in anaemia epidemiology, multiple cut-offs are also used to further categorize patients into groups, such as mild, moderate and severe [44]. The rationale for such groupings is to aid and simplify both interpretation and presentation of the results [5, 6, 45], while in clinical settings the groupings are used for targeted treatment. As a result, statistical analysis is often carried out on the categorical outcome obtained through the discretization of the continuous measurement.

The disadvantages and loss of information yielded by this practice have been investigated in several studies. Fedorov et al. [3] showed that dichotomization of continuous outcome variables can lead to a loss of 36% in terms of the Fisher's information for the population average. As a result of this, the statistical power required to estimate regression relationships between a health outcome and risk factors is also diminished [4]. Furthermore, in cases where the relationship is non-linear or non-monotonic, dichotomization or categorization into few classes may make that undetectable [5, 6]. All these issues are further exacerbated when the choice of specific cut-offs is inconsistent or, in some cases, even arbitrary [7–9]. For example, cut-offs may vary within and across studies due to differences in the sample populations from which they are derived or due to changes in how they are defined according to clinical practice and operational policy.

In this paper, we investigate the effects of dichotomization of spatially-referenced continuous outcome variables on geostatistical inference. Model-based geostatistics (MBG) [1] is a likelihood-based paradigm that allows us to carry out spatially continuous predictive inference on an outcome of interest using a spatially discrete set of data. Over the last two decades, MBG has been increasingly used to map the prevalence of infectious diseases [46], especially in low-resource settings where disease registries are non-existent or geographically incomplete. In this context, there have been global efforts to increase the use of rapid diagnostic tests for diseases such as malaria and HIV [47–50], which are typically recorded as binary by labelling the tested individuals as either positive or negative. In other cases, instead, dichotomization is first carried out on a continuous disease indicator variable and a geostatistical model is then developed on the binary outcome. For

example, in Zimmerman et al. [51], a continuous score quantifying the deviation from normal growth in children is dichotomized in order map stunting prevalence; following a similar approach, Magalhaes et al. [52] fit a binomial geostatistical model to dichotomized continuous haemoglobin densities so as to map anaemia prevalence in West Africa.

The effects of dichotomization on geostatistical inference have not been fully understood and, to the best of our knowledge, this is the first study that attempts to pursue this objective in the context of prevalence mapping. More specifically, in this paper, we provide answers to the following questions: 1) How does spatial correlation affect the loss of information and the estimation of regression relationships? 2) Can dichotomization lead to substantially different and more uncertain spatial predictions in disease prevalence? 3) How can we use linear geostatistical models to map disease prevalence and thus avoid the drawbacks of dichotomization?

The structure of the paper is as follows. In section 2.2 we describe the geostatistical modelling framework for disease prevalence mapping, and outline the differences and links between geostatistical models based on binary and continuous outcomes. In section 2.3, we first explore the information loss due to dichotomization in terms of the Fisher's information for two observations. We then develop a metric which can be used to assess the loss of information for the estimation of the regression coefficients of any geostatistical model. We also carry out a simulation study to extend our investigation to the estimation of the covariance parameters and spatial predictions for prevalence. In section 2.4 we illustrate two applications on the mapping of anaemia and stunting prevalence in Africa. Section 2.5 is a concluding discussion.

In what follows, fitting of geostatistical models and geostatistical prediction have been carried out using the Monte Carlo maximum likelihood method implemented in the `PrevMap` package [53] available from the Comprehensive R Network archive (`cran.r-project.org`. In this paper, maximum likelihood estimation is facilitated by unconstrained and box-constrained optimization using PORT routines through the nlminb function in R.

2.2 The link between geostatistical models for continuous and binary outcomes

Consider data from a cross-sectional survey where information on a continuous health outcome, the random variable Y_{ij}, is collected through examination of n_i

individuals residing at location x_i for $j = 1, \ldots, n_i$ and $i = 1, \ldots, m$. We then assume that conditionally on a spatial Gaussian process $S = \{S(x) : x \in \mathbb{R}^2\}$, the Y_{ij} are random Gaussian variables with mean $\mu_{ij} + S(x_i)$ and variance τ^2. From the linear properties of Gaussian distributions, we can write the model in the following compact form

$$Y_{ij} = \mu_{ij} + S(x_i) + Z_{ij}, \tag{2.1}$$

where Z_{ij} are i.i.d. Gaussian variables with mean zero and variance τ^2, representing unexplained individual-level variation and the mean component μ_{ij} is modelled as a linear regression taking the form

$$\mu_{ij} = \alpha + \beta^\top d(x_i) + \gamma^\top e_{ij}$$

where we distinguish between covariates, $d(x_i)$, that express properties of the locations and covariates, e_{ij}, for individual traits e.g. age and gender.

We model $S(x)$ as a stationary and isotropic Gaussian process with mean zero, variance σ^2 and correlation function $\text{Cor}\{S(x), S(x')\} = \rho(u)$ where $u = ||x - x'||$ denotes the Euclidean distance between x and x'. In the remainder of this paper, we shall define $\rho(\cdot)$ to be an exponentially decaying function with scale parameter ϕ, i.e. $\rho(u) = \exp\{-u/\phi\}$.

Based on a predefined threshold c, whose exceedance or not defines the disease status of an individual, we define the binary outcome $\tilde{Y}_{ij}$ as

$$\tilde{Y}_{ij} = \begin{cases} 1 & \text{if } Y_{ij} < c \\ 0 & \text{if } Y_{ij} \geq c \end{cases}, \tag{2.2}$$

with $\tilde{Y}_{ij} = 1$ indicating a positive case for the disease under investigation and $\tilde{Y}_{ij} = 0$ for a negative case. Note that in some diseases, the threshold is dependent on individual characteristics like age or sex, as demonstrated in the 'Applications' section of this paper.

From the model of the continuous outcome in (2.1), it follows that the model for

$\tilde{Y}_{ij}$ is given by

$$\begin{aligned} P\left(\tilde{Y}_{ij}=1\middle|S(x_i)\right) &= P\left(Y_{ij}<c|S(x_i)\right) \\ &= P\left(\frac{Y_{ij}-\mu_{ij}-S(x_i)}{\tau}<\frac{c-\mu_{ij}-S(x_i)}{\tau}\middle|S(x_i)\right) \\ &= \Phi\left(\frac{c-\mu_{ij}-S(x_i)}{\tau}\right)=p_{ij}, \end{aligned} \tag{2.3}$$

where $\Phi(\cdot)$ is the cumulative density function of a standard Gaussian variable. Hence, the resulting model for $\tilde{Y}_{ij}$ is a Binomial geostatistical model with probit link function and linear predictor

$$\eta_{ij}=\Phi^{-1}(p_{ij})=\tilde{\mu}_{ij}+\tilde{S}(x_i) \tag{2.4}$$

where $\tilde{\mu}_{ij}=-\mu_{ij}/\tau$ and $\tilde{S}(x_i)=-S(x_i)/\tau$.

The functional relationships that link the parameters of the geostatistical model for $\tilde{Y}_{ij}$ with those of the model for Y_{ij} are the following

$$\begin{cases} \tilde{\alpha}=(c-\alpha)/\tau \\ \tilde{\beta}=-\beta/\tau \\ \tilde{\gamma}=-\gamma/\tau \\ \tilde{\sigma}^2=\sigma^2/\tau^2 \end{cases}. \tag{2.5}$$

The above equations can thus be used to obtain the parameter estimates for a geostatistical model for $\tilde{Y}_{ij}$ by transforming the parameter estimates from the geostatistical model for Y_{ij}. Note that it is not possible, instead, to map the estimates from the model for $\tilde{Y}_{ij}$ into those for Y_{ij} because the parameter τ^2 cannot be estimated from binary data. The unstructured component Z_{ij} is in fact integrated out in (2.4) and, as shown in (2.5), all the parameters on the left hand-side are expressed as a ratio between τ and all other parameters in the model for Y_{ij}. Finally, ϕ is not included in (2.5), since the scale of the spatial correlation of $\tilde{S}(x)$ is the same as that of $S(x)$.

2.3 Quantifying the effects of dichotomization

In Section 2.3.1, we first study the loss of information due to dichotomization for the estimation of the mean of the process using an intercept-only model, when all other parameters are known. In Section 2.3.2, we carry out a simulation study to the more common case when all parameters are unknown. Here we study the effect

on dichotomization on parameter estimation. In both sections, we shall restrict our attention to the scenario of a single observation per location, hence we set $n_i = 1$ for all i and drop the j subscript.

2.3.1 Unknown regression coefficients and known covariance parameters

2.3.1.1 Special case of m=2 for an intercept-only model

The objective in this section is to quantify the loss of information in terms of the expected Fisher information (EFI) with respect to $\tilde{\alpha}$, the parameter which regulates the mean level of disease prevalence. Here, we restrict our attention to the simpler case of two observations at two locations, hence $m = 2$ and $n_1 = n_2 = 1$, for an intercept-only model. As it will be shown in the applications of Section 2.4, this simpler scenario provides useful insights on the effects of dichotomization which are consistently observed in the case of more than two observations. A more general scenario, however, shall also be considered in the next section.

We re-express the linear geostatistical model for a continuous outcome Y_i as

$$Y_i = \alpha + S(x_i) + Z_i, \text{ for } i = 1, \ldots, m \tag{2.6}$$

where $S(x_i)$ is a stationary and isotropic Gaussian process with the same properties as defined in equation (2.1).

Let $\Sigma_Y = \Sigma + \tau^2 I$ be the covariance matrix of the vector $Y = (Y_1, \ldots, Y_m)$, with Σ and I denoting the spatial covariance matrix with (i, j)-th entry $\sigma^2 \exp\{-\|x_i - x_j\|/\phi\}$ and an m by m identity matrix, respectively.

In order to quantify the loss of information that arises from the dichotomization of the Y_i, we first re-parametrize the linear model in (2.6) with respect to the prevalence parameters as defined in (2.5); note that $\alpha = c - \tau\tilde{\alpha}$. We then obtain the EFI for $\tilde{\alpha}$ under the linear model, given by

$$I_Y(\tilde{\alpha}) = \tau^2 \mathbb{1}^T \Sigma_Y^{-1} \mathbb{1}, \tag{2.7}$$

where $\mathbb{1}$ is a vector with all entries equal to 1.

In the case of the dichotomized outcome $\tilde{Y}_i$, the computation of the EFI is further complicated by the fact that the log-likelihood function is not available in closed form. More specifically, this is given by the marginal distribution of the outcome

$\tilde{Y} = (\tilde{Y}_1, \ldots, \tilde{Y}_m)$, i.e.

$$\log L(\tilde{\alpha}) = \log\left(\int_{\mathbb{R}^m} f(\tilde{s}) f(\tilde{y}|\tilde{s};\tilde{\alpha})\, d\tilde{s}\right), \tag{2.8}$$

where $f(\tilde{s})$ is the density of a multivariate Gaussian distribution with mean zero and covariance matrix $\tilde{\Sigma} = \Sigma/\tau^2$, whilst

$$\begin{aligned} f(\tilde{y}|\tilde{s};\tilde{\alpha}) &= \prod_{i=1}^{m} f(\tilde{y}_i|\tilde{s}_i;\tilde{\alpha}) \\ &= \exp\left\{\sum_{i=1}^{m}\left[\tilde{y}_i \log\left(\frac{p_i}{1-p_i}\right) + (1-\tilde{y}_i)\log\{1-p_i\}\right]\right\} \\ &= \exp\{g(\tilde{y}|\tilde{s};\tilde{\alpha})\} \end{aligned} \tag{2.9}$$

where $\Phi^{-1}(p_i) = \tilde{\alpha} + \tilde{S}(x_i)$. To obtain the EFI for $\tilde{\alpha}$, we first take the second derivative of (2.8) with respect to $\tilde{\alpha}$ to give

$$\begin{aligned} \frac{d^2 \log L(\tilde{\alpha})}{d^2\tilde{\alpha}} &= L^{-1}(\tilde{\alpha}) \int_{\mathbb{R}^m} f(\tilde{s}) f(\tilde{y}|\tilde{s};\tilde{\alpha}) \left[\left(\frac{dg(\tilde{y}|\tilde{s};\tilde{\alpha})}{d\tilde{\alpha}}\right)^2 + \right. \\ &\quad \left. \frac{d^2 g(\tilde{y}|\tilde{s};\tilde{\alpha})}{d^2\tilde{\alpha}}\right] d\tilde{s} + \left(\frac{d\log L(\tilde{\alpha})}{d\tilde{\alpha}}\right)^2, \end{aligned} \tag{2.10}$$

where

$$\frac{d\log L(\tilde{\alpha})}{d\tilde{\alpha}} = L^{-1}(\tilde{\alpha}) \int_{\mathbb{R}^m} f(\tilde{s}) f(\tilde{y}|\tilde{s};\tilde{\alpha}) \frac{dg(\tilde{y}|\tilde{s};\tilde{\alpha})}{d\tilde{\alpha}}\, d\tilde{s}.$$

Finally, we average over the distribution of $\tilde{Y}$

$$I_{\tilde{Y}}(\tilde{\alpha}) = E_{\tilde{Y}}\left[-\frac{d^2 \log L(\tilde{\alpha})}{d^2\tilde{\alpha}}\right].$$

Since the above quantity is not available in closed form we compute $I_{\tilde{Y}}(\tilde{\alpha})$ using Monte Carlo methods.

To quantify the loss of information, we then use the following metric $R(\tilde{\alpha}) = 1 - I_{\tilde{Y}}(\tilde{\alpha})/I_Y(\tilde{\alpha})$. In the special case of $S(x) = 0$ for all x, $R(\tilde{\alpha})$ reduces to

$$R(\tilde{\alpha}) = 1 - \left[\frac{[\Phi''(\tilde{\alpha})][1-\Phi(\tilde{\alpha})] - [\Phi'(\tilde{\alpha})]^2}{1-\Phi(\tilde{\alpha})} - \frac{[\Phi''(\tilde{\alpha})][\Phi(\tilde{\alpha})] - [\Phi'(\tilde{\alpha})]^2}{\Phi(\tilde{\alpha})}\right] \tag{2.11}$$

where $\Phi'(\cdot)$ and $\Phi''(\cdot)$ are the first and second derivative of $\Phi(\cdot)$, respectively. Fedorov et al. [3] have shown that $I_{\tilde{Y}}(\tilde{\alpha}) \leq I_Y(\alpha)$, and that the lower limit of (2.11) is about 36%. Also, note that (2.11) is not dependent on m.

To compute the integrals which define $-d^2 \log L(\tilde{\alpha})/d^2\tilde{\alpha}$, we use a quadrature approach based on Quasi Monte Carlo methods. Finally, for the computation of the

expectation in $I_{\tilde{Y}}(\tilde{\alpha})$, we use 10,000 samples and vary the spatial correlation between the two observations over $\rho \in \{i/10; i = 1, \ldots, 7\}$. Figure 2.1 shows different curves of $R(\tilde{\alpha})$, as a percentage, by setting $\sigma^2 = 1$ and letting τ^2 vary over the set $\{0.5, 1, 2\}$. Notice that these curves are symmetric with respect to 0, although they are shown only for positive values of $\tilde{\alpha}$. Across all three panels of Figure 2.1, we observe that increasing values of ρ lead to a reduction in the loss of information, although such reduction becomes smaller when the data are more noisy, i.e. when τ^2 also increases. Most notably, the largest loss of information is observed for values of prevalence close to 100% and 0% corresponding to large positive and negative values for $\tilde{\alpha}$, respectively. The variance τ^2 of the unstructured residuals Z_i also plays a very important role as shown by the dramatic increase in $R(\tilde{\alpha})$ for $\tau^2 = 2$, with all curves placed above $R(\tilde{\alpha}) = 0.65$.

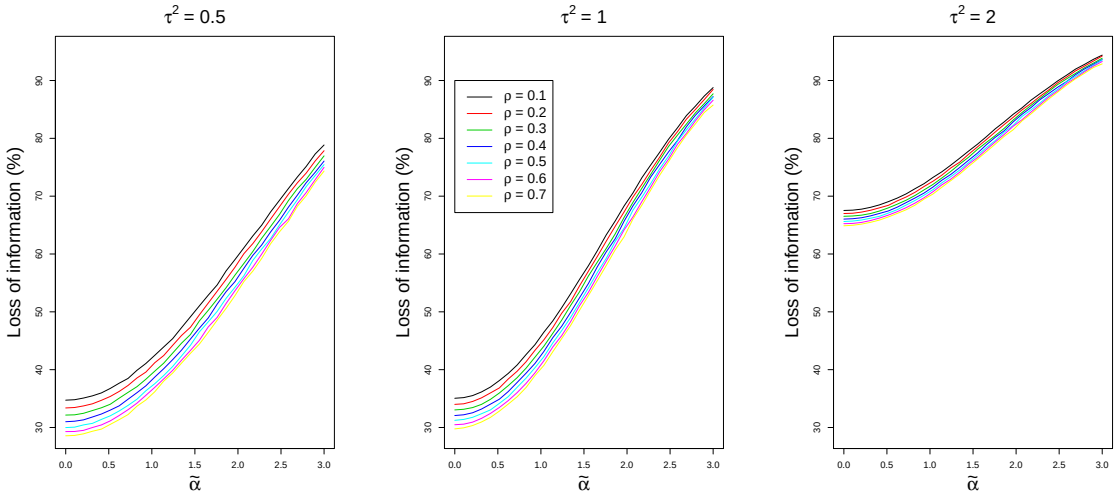

Figure 2.1: Curves for $R(\tilde{\alpha})$, shown as a percentage, by fixing $\sigma^2 = 1$ and varying $\tau^2 \in \{0.5, 1, 2\}$ and the spatial correlation between two observations $\rho \in \{i/10; i = 1, \ldots, 7\}$.

2.3.1.2 General case m > 2

For the general case of more than two locations (i.e. $m > 2$), the effects of dichotomization will also be dependent on the spatial arrangement of the sampled locations. In this section, we develop a metric that allows to quantify the potential loss of information due to dichotomization of continuous outcomes with respect the estimation of the regression coefficients $\theta = (\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma})$ of any geostatistical model as defined by (2.1).

In order to lower the computational burden, we first approximate the likelihood function of both the continuous and dichotomized data using a composite likelihood approach [54]. More specifically, we consider all possible pairs of $\tilde{Y}_{ij}$ and treat each of these as independent bivariate distributions. Let $\tilde{Y}_h$ and $\tilde{Y}_k$ denote

the vectors of binary outcomes associated with locations x_h and x_k and which are obtained through dichotomization of Y_h and Y_k, respectively. We then write

$$L(\theta) \approx L_{CL}(\theta) = \prod_{h=1}^{m-1} \prod_{k=h+1}^{m} f\left(\tilde{y}_h, \tilde{y}_k; \tilde{\theta}\right). \tag{2.12}$$

In the above equation the bivariate probability functions $f(\tilde{y}_h, \tilde{y}_k; \tilde{\theta})$ are expressed by the following integral in two dimensions

$$f(\tilde{y}_h, \tilde{y}_k; \tilde{\theta}) = \int_{R^2} f(\tilde{s}) f\left(\tilde{y}_h, \tilde{y}_k | \tilde{s}; \tilde{\theta}\right) \, d\tilde{s}$$

where $f\left(\tilde{y}_h, \tilde{y}_k | \tilde{s}; \tilde{\theta}\right)$ consists of a product of $n_h + n_k$ probability functions for the binary observations in y_h and y_k.

Let $\hat{\theta}_{LM}$ denote the maximum likelihood estimates of θ obtained from the linear model using the system of equations in (2.5). In order to understand how more or less dispersed the composite likelihood becomes after dichotomization, we proceed as follows. We first compute the second derivative of the composite log-likelihood at $\hat{\theta}_{LM}$, i.e.

$$H_{\tilde{Y}}(\hat{\theta}_{LM}) = \left[\frac{\partial^2 \log L_{CL}(\theta)}{\partial\theta_i \partial\theta_j}\right]_{\theta=\hat{\theta}_{LM}}.$$

For a continuous outcome Y, we have

$$H_Y(\theta) = -\tau^2 \sum_{h=1}^{m-1} \sum_{k=h+1}^{m} D_h^\top \Sigma_{hk}^{-1} D_k$$

where Σ_{hk} is covariance matrix between Y_h and Y_k, and D_h and D_k are the design matrices associated with locations x_h and x_k, respectively. To quantify the change in the dispersion of the composite likelihood around $\hat{\theta}_{LM}$, we finally compute

$$\begin{aligned} CLD(\hat{\theta}_{LM}) &= \log\{\det[-H_{\tilde{Y}}^{-1}(\hat{\theta}_{LM})]\} - \log\{\det[-H_Y^{-1}(\hat{\theta}_{LM})]\} \\ &= \log\{\det[-H_Y(\hat{\theta}_{LM})]\} - \log\{\det[-H_{\tilde{Y}}(\hat{\theta}_{LM})]\} \end{aligned}$$

where $\det(A)$ denotes the determinant of a square matrix A. Large values of $CLD(\hat{\theta}_{LM})$ indicate a more dispersed composite likelihood around $\hat{\theta}_{LM}$ for the binary data $\tilde{Y}$, which we interpret as evidence of loss of information due to dichotomization. Also, note that computation of CLD can be carried out without fitting any Binomial geostatistical model.

In the applications of Sections 2.4.1 and 2.4.2, we compute the CLD metric, by plugging-in the maximum likelihood estimates for the covariance parameters from the linear geostatistical models.

2.3.2 Simulation study

The objective of this simulation is to quantify the effects of dichotomization on parameter estimation and spatial predictions of prevalence. To this end we consider the linear model for Y_i as specified in (2.6) and its dichotomization using a threshold c to give $\tilde{Y}_i = 1$ if $Y_i < c$ and 0 otherwise.

In the simulation, we set $\alpha = 0$ and $\sigma^2 = 1$. We consider several scenarios obtained through all possible combinations of values for the model parameters defined in Table 2.1. For a given scenario, we simulate 1,000 data-sets of both continuous and dichotomized outcomes and fit their corresponding geostatistical models. We use a regular grid covering the unit square with spacing of 1/14 (i.e. a regular grid of 15 x 15 equally spaced points), to give a sample size of $m = 225$. For each of the fitted models, we extract the estimates for $\tilde{\alpha}$, $\tilde{\sigma}^2$ and ϕ, and predict prevalence $p_i = \Phi\left(\tilde{\alpha} + \tilde{S}(x_i)\right)$ at each of the grid points. We summarize the results using the bias and mean square error (MSE) and, for the prevalence predictions, average these two indices over the grid locations.

Tables 2.2 and 2.3 report the results for the model parameters and spatial predictions for prevalence, respectively. Overall, bias and MSE for $\tilde{\alpha}$ and ϕ are consistently smaller in the model fitted to the continuous data (C) than for that fitted to the binary data (B). In the case of $\tilde{\sigma}^2$, instead, the performance of both models is strongly affected by the scale of the spatial correlation ϕ: for $\phi = 0.1$ the model B outperforms model C in terms of bias and MSE, whilst the opposite is observed for $\phi = 0.2$. A possible explanation for this may be due to the fact that in the linear geostatistical model, higher spatial correlation helps to better separate the individual contributions of the signal component $S(x_i)$ and the noise component Z_i to the total variation in the outcome Y_i, thus improving the estimation of $\tilde{\sigma}^2 = \sigma^2/\tau^2$. In the case of the binary data, instead, $\tilde{S}(x_i)$ is the only source of over-dispersion and, as a result of this, a higher spatial correlation leads to a larger number of concordant binary outcomes and, therefore, to a poorer estimate of the variance of $\tilde{S}(x_i)$. Also, we notice that the estimation of $\tilde{\sigma}^2$ and ϕ does not appear to be affected by the threshold c, unlike $\tilde{\alpha}$. Finally, the results for the spatial predictions of prevalence show that the performance of model C is unaffected by changes in c and τ^2, while for $\phi = 0.2$ the predictions have slightly lower MSE than for $\phi = 0.1$. Model B, instead, delivers predictions with higher bias for increasing c which can be partly explained by the positive increase in the bias in the estimates of $\tilde{\alpha}$ for increasing c.

We have also conducted further simulations under the same scenarios defined in Table 2.1 but for a larger sample size $m = 450$, by placing additional points on a

regular grid adjacent to the unit square so as to cover the rectangle $[0, 2] \times [0, 1]$. The results, reported in the supplementary material (Tables S1 and S2), lead to the same conclusions drawn for $m = 225$.

Table 2.1: True values for τ^2, ϕ and c used in the simulation study.

	Symbol	Variations
Variance of the nugget effect	τ^2	0.5, 1, 2
True scale of spatial correlation	ϕ	0.1, 0.2
Cut-off	c	0, 0.2, 0.4

Table 2.2: Bias and mean square error (in brackets) for $\tilde{\alpha}$, $\tilde{\sigma}^2$ and the estimate $\hat{\phi}$ obtained from the geostatistical models fitted the binary (B) and continuous (C) outcomes.

			c=0		c=0.2		c=0.4	
Parameter	τ^2	ϕ	B	C	B	C	B	C
$\tilde{\alpha}$	0.5	0.1	0.009 (0.168)	0.005 (0.113)	0.063 (0.312)	0.017 (0.115)	0.157 (0.263)	0.042 (0.139)
	1	0.1	-0.009 (0.126)	-0.008 (0.068)	0.060 (0.409)	0.004 (0.064)	0.153 (0.119)	0.041 (0.080)
	2	0.1	-0.006 (0.079)	-0.003 (0.036)	0.077 (0.493)	0.022 (0.040)	0.148 (0.169)	0.050 (0.054)
	0.5	0.2	-0.025 (0.624)	-0.013 (0.296)	0.156 (0.648)	0.048 (0.282)	0.238 (0.722)	0.028 (0.303)
	1	0.2	-0.018 (0.332)	-0.007 (0.150)	0.093 (0.585)	0.008 (0.137)	0.215 (0.323)	0.025 (0.145)
	2	0.2	-0.007 (0.185)	-0.004 (0.080)	0.106 (0.566)	0.023 (0.080)	0.164 (0.272)	0.022 (0.088)
$\tilde{\sigma}^2$	0.5	0.1	-0.011 (1.296)	0.788 (5.843)	-0.033 (1.324)	0.604 (5.170)	0.022 (1.594)	0.734 (5.741)
	1	0.1	0.234 (0.787)	0.822 (5.953)	0.190 (0.688)	0.750 (5.255)	0.224 (0.701)	0.741 (5.942)
	2	0.1	0.211 (0.326)	0.672 (5.047)	0.217 (0.364)	0.600 (4.961)	0.204 (0.324)	0.653 (4.833)
	0.5	0.2	1.641 (8.894)	0.527 (3.091)	1.784 (12.672)	0.574 (3.268)	1.566 (8.464)	0.515 (3.401)
	1	0.2	1.162 (3.755)	0.399 (1.712)	1.064 (3.243)	0.372 (1.619)	1.048 (2.949)	0.365 (1.671)
	2	0.2	0.548 (0.871)	0.254 (1.104)	0.575 (1.074)	0.304 (1.829)	0.534 (0.910)	0.341 (1.707)
$\hat{\phi}$	0.5	0.1	0.088 (0.016)	0.007 (0.002)	0.084 (0.015)	0.009 (0.002)	0.085 (0.015)	0.007 (0.002)
	1	0.1	0.071 (0.014)	0.004 (0.003)	0.072 (0.014)	0.006 (0.004)	0.074 (0.015)	0.006 (0.003)
	2	0.1	0.060 (0.017)	0.010 (0.007)	0.063 (0.022)	0.009 (0.007)	0.056 (0.014)	0.010 (0.006)
	0.5	0.2	0.076 (0.029)	-0.017 (0.011)	0.083 (0.034)	-0.020 (0.008)	0.076 (0.030)	-0.021 (0.009)
	1	0.2	0.068 (0.030)	-0.014 (0.016)	0.058 (0.027)	-0.028 (0.010)	0.061 (0.034)	-0.023 (0.013)
	2	0.2	0.032 (0.025)	-0.024 (0.019)	0.037 (0.031)	-0.024 (0.016)	0.028 (0.019)	-0.032 (0.015)

Table 2.3: Bias and mean square error (in brackets), averaged over a 1/14 by 14 regular grid covering the unit square (hence, $m = 225$), for the spatial predictions of prevalence obtained from the geostatistical models fitted to the binary (B) and continuous (C) outcomes.

		c=0		c=0.2		c=0.4	
τ^2	ϕ	B	C	B	C	B	C
0.5	0.1	0.001 (0.060)	0.001 (0.039)	0.018 (0.059)	0.001 (0.038)	0.034 (0.058)	0.001 (0.036)
1	0.1	-0.001 (0.051)	0.001 (0.038)	0.018 (0.051)	0.001 (0.038)	0.038 (0.050)	0.001 (0.036)
2	0.1	-0.001 (0.040)	0.001 (0.033)	0.020 (0.040)	-0.001 (0.033)	0.038 (0.040)	-0.002 (0.032)
0.5	0.2	-0.001 (0.042)	0.001 (0.030)	0.013 (0.042)	-0.001 (0.030)	0.025 (0.041)	-0.001 (0.029)
1	0.2	0.001 (0.037)	0.001 (0.028)	0.014 (0.037)	-0.001 (0.028)	0.030 (0.036)	-0.001 (0.027)
2	0.2	-0.001 (0.031)	0.001 (0.024)	0.019 (0.031)	-0.001 (0.024)	0.034 (0.031)	-0.002 (0.024)

2.4 Applications

2.4.1 Mapping anaemia prevalence in Ethiopia

In this section, we analyse data collected from the Beyond Garki project [1]. This project consisted of cross-sectional surveys which were conducted in selected study sites in Ethiopia and Uganda to monitor changes in malaria risk in the context of interventions that had been implemented. The study sites were defined as a 'health centre and the catchment population in selected villages around it'. Here we subset the data for the Hembecho site, in Ethiopia, collected during the 2012 survey, where a random sample of households were selected from a list of enumerated households from all villages within a radius of 2 to 6 kilometers of the health facility. Among the data obtained in this survey were continuous measurements of haemoglobin density (g/dL), taken from blood samples of individuals living in the households. These measurements were then used to determine the anaemia status of individuals. Further details of the study design and data collection can be found in Abeku et al. [55].

In this analysis, the objective is to identify areas where the anaemia prevalence is highly likely to exceed a 20% threshold for 20 year old women. Hence, we map and compare exceedance probabilities from the geostatistical models for the continuous and binary outcomes. The chosen threshold for anaemia prevalence has clinical, operational and public health significance for policy decisions, with the World Health Organisation (WHO) classifying 20% anaemia prevalence as 'moderate public health significance'[44]. Finally, the rationale for carrying out predictions for 20 year old women is that one of the key WHO Global Nutrition Targets for 2025 is a 50% reduction of anaemia in women of reproductive age [44, 56], which is defined as 15-49 years [57].

The data-set contains information on 1712 individuals distributed over 457 households, with an average of 3.7 individual in each household. The continuous outcome variable, Y_{ij}, is the log-transformed haemoglobin density for the j-th individual at the i-th household. To account for the non-linear relationship between the log-transformed anaemia density and age, as shown in Figure 2.2, we use a linear spline with knots at 15 and 30 years. Our proposed linear model for Y_{ij} is thus expressed as

$$Y_{ij} = \alpha + \sum_{h=1}^{3} \beta_h b_h(a_{ij}) + \beta_4 d_{ij} + S(x_i) + Z_{ij}, \tag{2.13}$$

[1] www.malariaconsortium.org/beyondgarki

where: a_{ij} is the age, in years, of an individual; d_{ij} is a binary indicator of the sex of an individual, with "female" as reference category; $b_h(\cdot)$ are the base functions of the linear spline defined as $b_1(a) = a$, $b_2(a) = max\{0, a - 15\}$ and $b_3(a) = max\{0, a - 30\}$.

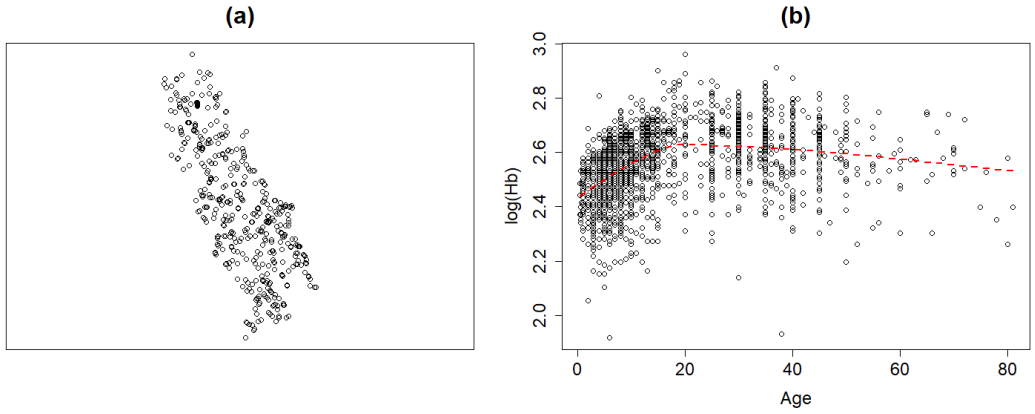

Figure 2.2: (a) locations of the households in the survey; (b) scatter plot of the log-transformed haemoglobin density against age, in years. The dashed red line in the in panel (b) is a least square fit of the linear spline defined in the main text of Section 2.4.1.

Dichotomization of Y_{ij} results in the binary outcome variable Y_{ij}^* indicating anaemia status, where $Y_{ij}^* = 1$ denoting a positive case and $Y_{ij}^* = 0$ a negative case for severe anaemia. In order to classify an individual as positives or negatives, thresholds for severe anaemia have been applied using individual-level information on age, sex and pregnancy status as defined in Table 2.4. As result of this, we then modify equation (2.3) as

$$p_{ij} = \Phi\left(\frac{c_{ij} - \mu_{ij} - S(x_i)}{\tau}\right)$$

where c_{ij} is the logarithm of the threshold values which must now be incorporated as an additional covariate into the linear predictor, i.e.

$$\Phi^{-1}(p_{ij}) = \tilde{\alpha} + \sum_{h=1}^{3} \tilde{\beta}_h b_h(a_{ij}) + \tilde{\beta}_4 d_{ij} + \tilde{\beta}_5 c_{ij} + \tilde{S}(x_i)$$

where $\tilde{\beta}_5 = 1/\tau$.

Table 2.5 reports the maximum likelihood estimates and 95% confidence intervals of the model parameters for the binary and continuous outcomes. The linear geostatistical model gives an estimate for τ^2, the variance of Z_{ij}, of about 1.133×10^{-2} (95% CI: 1.050×10^{-2}, 1.222×10^{-2}) and for σ^2 of about 1.558×10^{-3} (95% CI: 0.954×10^{-3}, 2.422×10^{-3}), yielding an estimated noise to signal ratio τ^2/σ^2 of about 7.3. The CLD metric (Section 2.3.1.2) is 771.235 which indicates a larger dispersion of the composite likelihood for the binary data than for the continuous

Table 2.4: Thresholds of haemoglobin densities (g/dL) for anaemia diagnosis [44].

	Anaemia		
Age or Sex group	Mild	Moderate	Severe
Children (Age 6-59 months)	10.0-10.9	7.0-9.9	< 7.0
Children (Age 5-11 yrs)	11.0-11.4	8.0-10.9	< 8.0
Children (Age 12-15yrs)	11.0-11.9	8.0-10.9	< 8.0
Pregnant women (Age $>$ 15yrs)	10.0-10.9	7.0-9.9	< 7.0
Non-pregnant women (Age $>$ 15 yrs)	11.0-11.9	8.0-10.9	< 8.0
Men (Age $>$ 15 yrs)	11.0-12.9	8.0-10.9	< 8.0

data. We observe that the estimates of the parameters are all comparable except for $\tilde{\sigma}^2$ $(= \sigma^2/\tau^2)$, as indicated by the non-overlapping confidence intervals from the two models. More importantly, we observe that the confidence intervals for the regressions coefficients are narrower for the linear model.

Figure 2.3 shows the resulting anaemia prevalence predictions for 20 year old women from the two models. While the overall pattern of predicted prevalence is similar between the models (see Figures 2.3(a) and 2.3(b)), there are non-negligible differences ranging from -8.23% to 7.85% prevalence (Figure 2.3(c)). Similarly, the maps of the exceedance probability qualitatively show similar spatial patterns (figures 2.3(d) and 2.3(e)). However, we identify small areas where the differences range from -38.6% to 31.20% (Figure 2.3(f)).

Table 2.5: Maximum likelihood estimates with associated 95% confidence intervals (CI) for the geostatistical models fitted to the anaemia data.

	Binomial model		Linear model	
Term	Estimate	95% CI	Estimate	95% CI
$\tilde{\beta}_1$	0.079	(-0.220, 0.378)	-0.278	(-0.377, -0.179)
$\tilde{\beta}_2$	-0.066	(-0.124, -0.008)	-0.115	(-0.131, -0.100)
$\tilde{\beta}_3$	0.052	(-0.025, 0.129)	0.097	(0.072, 0.122)
$\tilde{\beta}_4$	0.071	(0.021, 0.121)	0.050	(0.031, 0.069)
$\tilde{\sigma}^2$	0.527	(0.395,0.705)	0.138	(0.082, 0.218)
ϕ	0.325	(0.201,0.396)	0.250	(0.093, 0.549)

2.4.2 Mapping stunting prevalence in Ghana

The data analysed in this section are from the 2014 Demographic and Health Survey[2] (DHS) conducted in Ghana. DHS are nationally representative household surveys conducted about every 5 years, and provide data on health and population indicators for monitoring and impact evaluation across Africa. The DHS surveys follow a stratified two-stage cluster design where in the first stage, enumeration areas are selected from previous population census files, followed by a second stage

[2] `dhsprogram.com`

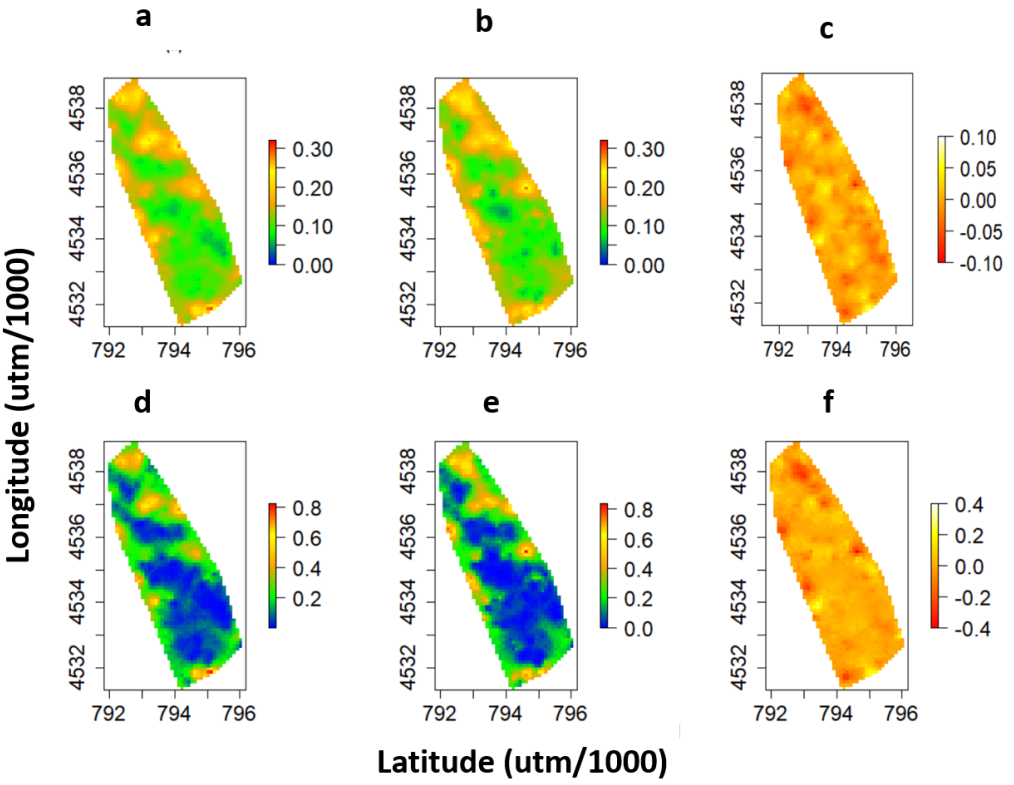

Figure 2.3: Predicted anaemia prevalence for a 20 year old woman. Upper panels: prevalence surfaces from the binomial model (a) and the linear models (b), and the difference between the first and the second (c). Lower panels: exceedance probabilities for a 20% prevalence threshold obtained from the binomial (d) and the linear models (e), and their difference (f). This study area is within 2–6 kms radius of a local health facility

where, for each selected enumeration area, samples of households are sampled from updated lists of households to generate the so called sampling clusters. The GPS locations of a sampling cluster is then assigned to each of the individuals falling within that cluster.

Among the health indicators collected in this survey are anthropometric measurements, which are used to calculate the height-for-age Z-score (HAZ). HAZ are standardized scores which indicate the standard deviation from the mean of children's heights based on the WHO growth standards [58, 59] and are comparable across ages and sex. HAZ values below -2 are taken as an indication of stunted growth.

One of the key WHO Global Nutrition Targets for 2025 is a 40% reduction in the number of children under-5 who are stunted [56, 60]. Additionally, a stunting prevalence above 40% is considered a high public health significance [61]. Accordingly, we aim to map the exceedance probability of 40% stunting prevalence for

a 2 year old who falls in the lowest wealth index category and whose mother has poor education.

The data include information on children under 5 years old, with a total of 2671 sampled children and 410 clusters, giving an average of 6.5 children per cluster. The continuous outcome variable, Y_{ij}, is the HAZ for child j in cluster i. Figure 2.4 (a) shows the empirical relationship between HAZ and age in years. Using a similar approach of the previous analysis, we capture the non-linear relationship with a linear spline having knots at 1 and 2 years. Hence, the resulting linear geostatistical model is

$$Y_{ij} = \alpha + \sum_{h=1}^{3} \beta_h b_h(a_{ij}) + \beta_4 d_{ij} + \beta_5 e_{ij} + S(x_i) + Z_{ij}, \tag{2.14}$$

where: a_{ij} is the age of a child; the basis functions of the linear splines are $b_1(a) = a$, $b_2(a) = \max\{0, a-1\}$ and $b_2(a) = \max\{0, a-2\}$; d_{ij} is a score of maternal education, taking integer values from 1="Poorly educated" to 3="Highly educated"; e_{ij} is a wealth index of the household, taking integer values from 1="Poor" to 3="Rich".

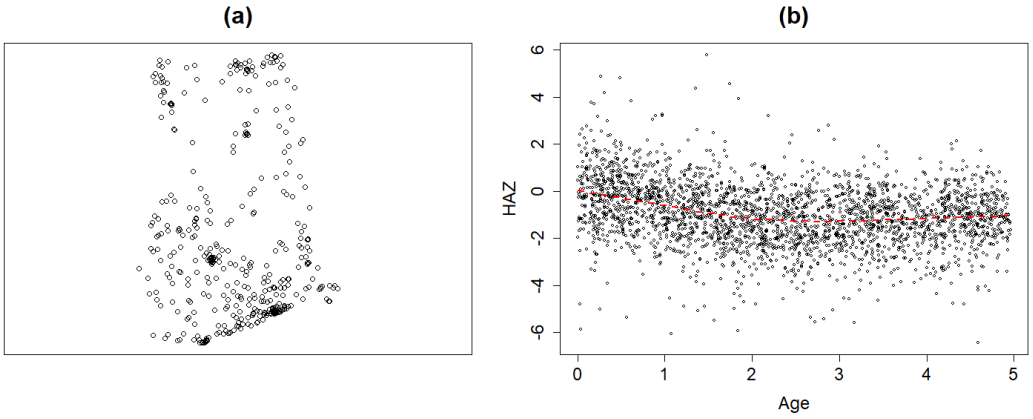

Figure 2.4: Figure (a) shows the spatial distribution of households included in the analysis, while (b) shows the relationship between HAZ and age. The red dashed line in panel (b) corresponds to a least square fit of the linear spline defined in (2.14).

The maximum likelihood estimates and associated 95% confidence intervals are shown in Table 2.6. We also report that the estimates for σ^2 and τ^2 from the linear geostatistical model are 0.071 (95% CI: 0.037, 0.126) and 1.396 (95% CI: 1.318, 1.477), respectively. Hence, the estimated ratio τ^2/σ^2 is about 20, indicating that the data are substantially more noisy than those analysed in the previous section. This is also reflected in the CLD metric yielding a value of 9667.012 which is substantially larger than that reported for the anaemia analysis. Following from the results of Section 2.3, this suggests that the effects of dichotomization on

Table 2.6: Maximum likelihood estimates with associated 95% confidence intervals (CI) for the geostatistical models fitted to the data on childhood malnutrition.

	Binomial model		Linear model	
Term	Estimate	95% CI	Estimate	95% CI
$\tilde{\beta}_1$	0.772	(-0.002, 1.545)	0.554	(0.338, 0.771)
$\tilde{\beta}_2$	0.774	(-0.331, 1.879)	0.168	(-0.168, 0.502)
$\tilde{\beta}_3$	-1.942	(-2.463, -1.421)	-0.830	(-1.024, -0.639)
$\tilde{\beta}_4$	-0.721	(-1.174, -0.269)	-0.139	(-0.239, -0.039)
$\tilde{\beta}_5$	-0.444	(-0.666, -0.221)	-0.259	(-0.332, -0.186)
$\tilde{\sigma}^2$	0.256	(0.102, 0.528)	0.051	(0.026, 0.091)
ϕ	157.301	(59.706, 341.451)	51.899	(14.657, 136.232)

geostatistical inference will also be stronger. Note that this may also be affected by the the spatial scale of Ghana, which is much larger than for the Ethiopia site. We observe that the estimates of the regression coefficients are concordant in sign but the size of the effects of the covariates are different as indicated by the non-overlapping confidence intervals; as in the previous section, we observe that the confidence intervals for the regression coefficients from the linear model are all narrower. The estimated $\tilde{\sigma}^2$ and ϕ are also substantially different, with the linear geostatistical model providing lower estimates and narrower confidence intervals for both parameters.

The differences in the parameter estimates are also reflected in Figure 2.5 which shows the predicted surfaces of stunting prevalence and the exceedance probabilities from the two models. These predictions are for a 2 year old who falls in the lowest wealth index category, and whose mother has poor education. Qualitatively, both models identify high and low levels of prevalence in the same areas. However, the differences in the predicted prevalence between the binomial model (Figure 2.5(a)) and the continuous model (Figure 2.5(b)), range from -9.38% to 19.98% prevalence (Figure 2.5(c)). Most notably, the binomial model presents much smoother maps than those from the linear model. For example, the binomial model identifies a single large hot-pot in the eastern part of Ghana, as being highly likely to exceed 40%. The linear model, instead, shows three neighbouring hot-spots in the same area. The differences in exceedance probabilities between the two models range from -51.50% to 64.90% (Figure 2.5(f))

2.5 Discussion

Understanding of the effects of dichotomization of continuous outcomes is especially important in medical research where cut-offs are used for diagnosis. These can be derived using different approaches: empirical approaches, where thresh-

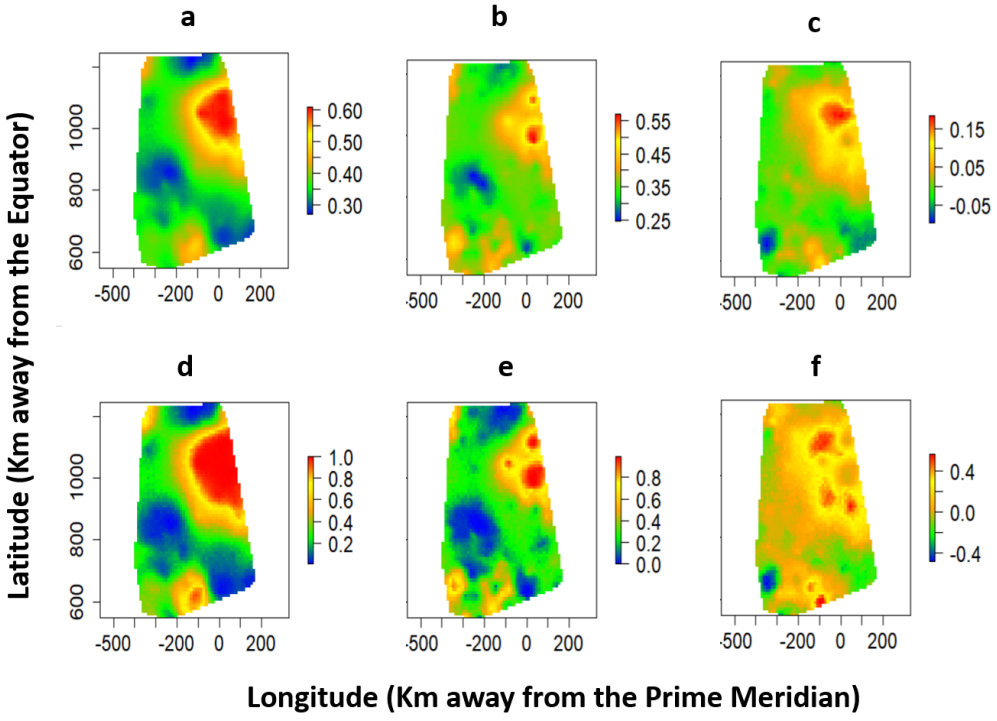

Figure 2.5: Predicted stunting prevalence for a 2 year old who falls in the lowest wealth index category and whose mother has poor education. Upper panels: prevalence surfaces from the binomial model (a) and the linear models (b), and the difference between the first and the second (c). Lower panels: exceedance probabilities for a 40% prevalence threshold obtained from the binomial (d) and the linear models (e), and their difference (f).

olds are obtained through summary statistics of measurements taken from healthy individuals; clinical approaches, which utilize a specified threshold above which symptoms and complications become more frequent; prognostic approaches, where a threshold is defined based on clinical results which may be symptom-less but carry an adverse prognosis; or, finally, operational approaches, where a threshold may be based on management and/or operational guidelines [43]. In this study, we have investigated the effects of dichotomization in the context of geostatistical modelling of disease prevalence data through a simulation study and two applications, and have proposed a likelihood-based metric to quantify the potential loss information arising from this practice. All of these provided evidence that dichotomization of the data can lead to substantial loss of information for both parameter estimation and spatial prediction. We found that spatial correlation may alleviate the effects of dichotomization for parameter estimation but the increase in uncertainty and bias still remained substantially larger than those of the linear model. In particular, one of the key factors that more strongly affects the loss of accuracy and precision is the distance of the threshold from the mean of the underlying process. As such distance increases, both the bias and the MSE in

the estimation of the mean component and in the spatial predictions of prevalence also increase. Another important factor is the magnitude of the noise variance τ^2 relative to the signal variance σ^2. This was especially evident in the application of Section 2.4.1, where a τ^2 about 20 times larger than σ^2 led to the loss of the fine scale features in the spatial pattern of disease prevalence.

As shown in the application of Section 2.4.1, when thresholds vary across individuals, these should be accounted for in the model for binary data, an aspect that has been ignored in previous studies of anaemia mapping. Also, this could be especially problematic if some of the covariates on which the the cut-offs are based are missing (e.g. the pregnancy status of a woman). An additional problem that arises in the context of anaemia epidemiology, is that the cut-offs described in Table 2.4 are based on guidelines from 1992 and 2001 [44] which may be subject to amendment as scientific research or clinical practices evolve.

We have only considered the case of a Gaussian distribution for the unstructured component Z_i. Assuming a symmetric distribution for Z_i implies that, on average, misclassifications of individuals as false positives and false negatives balance out after dichotomization. However, if Z_i followed a skewed distribution, this could introduce additional bias in the geostatistical model for binary data as more individuals could be misclassified as either false positives or false negatives. Hence, we expect that under these scenarios the negative effects of dichotomization on geostatistical inference would be even stronger than those shown in this study.

It is important to note that in our study we compared the performance of binary and linear geostatistical models for cut-offs that are dependent on the scale of the continuous measurement. In other cases, the Y_i may follow a mixture distribution with a probability mass in zero. For example, malaria parasite density may exhibit this feature if a large proportion of the general population has not been infected and is thus clear of parasites. In this case dichotomization of the continuous outcome as $Y_i^* = 1$ if $Y_i > 0$ and $Y_i^* = 0$ otherwise $Y_i = 0$, would not lead to any loss of information. Similarly, where the distinction between positive and negative is minimal, for example where there is a negligible overlap between two latent populations in a mixture distribution (e.g, seronegative and seropositive in the case of malaria serology analysis), then loss of information may not be a problem.

A final remark relates to the computational burden of binary and linear geostatistical models. The likelihood function of the latter can be, most of the times, expressed in closed form, while the former requires numerical procedures based on analytical or Monte Carlo approximations of the likelihood function in order to be fitted. Hence, the increase in the computational burden is a further reason to

avoid dichotomization of the data.

2.6 Conclusion

In the context of geostatistical inference, dichotomization of continuous outcomes can lead to a substantial loss of efficiency for both parameter estimation and spatial prediction. Such loss is further compounded as cut-offs used for dichotomization are further away from the mean. In addition, dichotomization can also result in the loss of fine scale features of disease prevalence, especially in the presence of a large noise to signal ratio. The findings in this study strongly support the conclusions drawn from previous studies that, whenever feasible, dichotomization should be avoided by developing models for the continuous measurements which can then be used to estimate prevalence.

CRediT authorship contribution statement

Irene Kyomuhangi: Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Visualization, Writing-original draft.
Tarekegn A. Abeku, **Matthew J. Kirby** and **Gezahegn Tesfaye**: Resources, Writing-review and editing
Emanuele Giorgi: Conceptualization, Methodology, Formal analysis, Supervision, Writing-review and editing.

Acknowledgements

We thank Dr. Luigi Sedda (Lancaster University) for his comments on the manuscript, and Dr Claudio Fronterrè (Lancaster University) for useful discussions on the computational aspects of the study.

We thank the study participants of the Beyond Garki and DHS projects, the staff of Malaria Consortium and the DHS program who were involved in the data collection, as well as the funders of the surveys presented.

Irene Kyomuhangi is a Commonwealth Scholar, funded by the UK government. .

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Chapter 3

Paper 2: A unified and flexible modelling framework for the analysis of malaria serology data

Irene Kyomuhangi[1], Emanuele Giorgi[1]

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

Published in: *Epidemiology and Infection*

Summary

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.

Keywords: malaria serology; reversible catalytic model; antibody acquisition model; mixture model; malaria antibody; seroprevalence.

3.1 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 [11]. Even with the decline of malaria prevalence in endemic countries [62], there are still challenges which require robust mechanisms for monitoring malaria transmission and evaluation of elimination efforts [11].

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 African, while *Plasmodium vivax (Pv)* dominates in the Americas and South East Asia [11]. Parasite prevalence (PrP) is determined by the proportion of infected individuals at time of data collection [20, 21], while the entomological inoculation rate (EIR) is the rate at which individuals are bitten by infectious mosquitoes [22]. Both of these measures may very 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 [20, 21]. Additionally, the collection of entomological data is labour-intensive, expensive, and excludes the recruitment of children, due to ethical considerations [23–25].

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 [21, 25, 26, 28].

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 [16, 17]. 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 [18, 20]. 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. [20, 25, 26]. 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 seropostives and seronegatives [20, 63].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 close to 50% [64, 65]. Bollaerts et al. [64] and Hens et al. [65] 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. [65].

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 [20, 30, 40]. Antibody aquisiton 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 [40].

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 [20, 30, 32, 40]. 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, reversible catalytic models, and antibody acquisition models in order to reliably estimate malaria transmission intensity. We also show that the additional flexibility brought by this novel model allows 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 3.2 provides an overview of current models for malaria serology analysis. Section 3.3 introduces a unified mechanistic model and outlines an alternative empirical approach that can be used to analyse malaria serology data. In section 3.4 we apply this new framework to cross-sectional antibody data from Western Kenya, and section 3.5 is a discussion of the results. Finally section 3.6 provides a summary and conclusion.

3.2 Existing models

3.2.1 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 [20, 66]. 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 realizations for a sample of n individuals, we write the density function of Y_i as

$$f(y_i) = \prod_{i=1}^{n} \left[(1-p) f_{S^-}(y_i; \mu_{S^-}, \sigma^2_{S^-}) + p f_{S^+}(y_i; \mu_{S^+}, \sigma^2_{S^+}) \right] \tag{3.1}$$

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 [20, 32, 34, 64, 65]. 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.

$$\begin{aligned} P(C_i^* = S^+ | y_i) &= \frac{p f_{S^+}(y_i; \theta_{S^+})}{(1-p) f_{S^-}(y_i; \theta_{S^-}) + p f_{S^+}(y_i; \theta_{S^+})} \\ P(C_i^* = S^- | y_i) &= 1 - P(C_i^* = S^+ | Y_i = y_i) \end{aligned} \tag{3.2}$$

where $\theta_{S^-} = (\mu_{S^-}, \sigma^2_{S^-})$ and $\theta_{S^+} = (\mu_{S^+}, \sigma^2_{S^+})$.

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

$$C_i = \begin{cases} S^- & \text{if } P(C_i^* = S^-|Y_i = y_i) \leq c^- \\ I & \text{if } c^- < P(C_i^* = S^-|Y_i = y_i) < c^+ \\ S^+ & \text{if } P(C_i^* = S^+|Y_i = y_i) \geq c^+ \end{cases}, \tag{3.3}$$

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 [64, 65, 67].

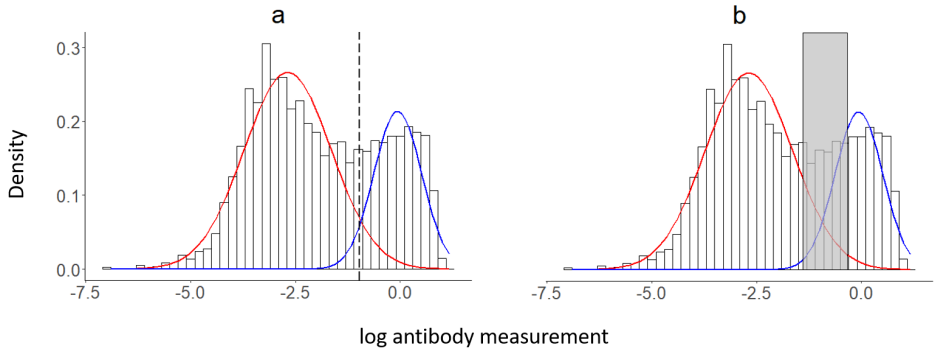

Figure 3.1: An illustration of the mixture model showing the bi-modal distributions for the S^- (red) and S^+ (blue) populations. The dotted line in Figure (a) shows the seropositivity threshold $\mu_{S^-} + 3\sigma_{S^-}$, above which individuals are classified as S^+. The grey rectangle in Figure (b) shows the inconclusive cases as defined by equation 3.3. 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 3.4.

In malaria serology, most studies favour the first threshold-based approach that does not introduce the classification for inconclusive cases [30, 32, 36, 40, 68, 69]. 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. [20].

However, both of these threshold-based approaches are prone to misclassification, which can create bias in estimating epidemiological parameters [20, 64, 65]. 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. [20] 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

$$f(y_i) = \prod_{i=1}^{n} \left[\sum_{k=1}^{K} p_k f_k(y_i; \theta_k) \right]. \tag{3.4}$$

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

3.2.2 Reversible catalytic models

Following the dichotomization of the continuous antibody measurements through the application of a mixture model, the resulting S^+ and S^- outcomes are modelled using a reversible catalytic model (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 3.2a. 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 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

$$\frac{dp}{da} = \lambda(a)(1 - p(a)) - \omega p(a). \tag{3.5}$$

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 [25], while ω is typically fixed and assumed to be constant [20]. However, some authors Bosomprah [32] and Akpogheneta et al. [70] suggest that ω may be age-dependent. Sepúlveda et al. [20] 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 (3.5) gives the following solution

$$p(a) = \frac{\lambda}{\lambda + \omega}\Big(1 - e^{-(\lambda+\omega)a}\Big). \tag{3.6}$$

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 [20, 25]. 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 [20, 34, 66].

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 (3.5) cannot be solved analytically and numerical procedures must instead be used.

In the study by Yman et al. [40], 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 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. [42] propose a model where the number of times that λ changed in the past, 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 [32] 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. [20]. 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 3.2b. The SIM also faces challenges with interpretation of results where initial exposure

and boosting between the multiple seropositive classes may be conflated [20, 21].

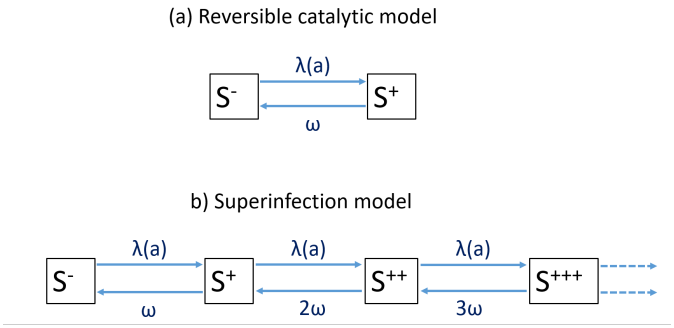

Figure 3.2: (a) is a representation of the reversible catalytic model (RCM) where individuals transition between seronegative (S^-) and seropositive (S^+) states through the SCR, $\lambda(a)$ and the SRR, ω. (b) is a representation of the superinfection model (SIM) where individuals can have their antibodies 'boosted' through increasing seropositive ($S^{+\cdots}$) states depending on the cumulative exposure to malaria parasites.

3.2.3 Antibody acquisition models

An alternative modelling approach to estimate transmission intensity is to use antibody acquisition models (AAMs) [30, 40]. Unlike RCMs, AMMs use the full antibody measurements without requiring any dichotomization 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 [20, 30, 31, 40]. 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

$$\frac{d\mu}{da} = \gamma(a) - r\mu(a). \tag{3.7}$$

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 (3.7) 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 [20, 30, 40].

Similar to the way seroconversion rates have been modelled in RCMs (section 3.2.2), 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 (3.7)

$$\mu(a) = \frac{\gamma}{r}\left(1 - e^{-ra}\right). \tag{3.8}$$

Similarly to RCMs, extensions of the AAM assumes either a step-change or linear reduction in the boosting rate γ ; see Sepúlveda et al. [20], Weber et al. [30], and Yman et al. [40] 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. [40] finds that the AAMs provide a more consistent fit to age-dependent antibody data compared to RCM fit to age-dependent seroprevalence data, particularly in the estimation of the change point (i.e when there was a change in transmission). 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 [40], or where an antigen is highly immunogenic, leading to high seropositivity to its antibody in the population [30].

3.3 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 3.2.1 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 (3.5) 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

$$p(a) = \sum_{h=1}^{a} \frac{\lambda(h)}{\prod_{k=h}^{a}(1 + \lambda(h-k) + \omega)} \tag{3.9}$$

This allows us to specify any function for $\lambda(a)$ without being constrained to three

options described in section 3.2.2. 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 (3.7) as

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

or, equivalently,

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

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

$$\mu(a) = \sum_{h=1}^{a} \gamma(h) \left(\frac{1}{1+r}\right)^{a-h+1} \tag{3.10}$$

Similar to the interpretation of (3.9), 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 [14, 31, 71].

To model the temporal changes in $\lambda(h)$ and $\gamma(h)$, in 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

$$\begin{aligned} \lambda(h) &= \exp\{l_0 + l_1(a-h)\} \\ \gamma(h) &= \exp\{g_0 + g_1(a-h)\} \end{aligned} \tag{3.11}$$

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 (Figure S1 of the supplementary material further illustrates the mechanism of this approach).

Assuming $\mu(a_i)$ in (3.10) to be the mean level of antibodies in the S^- population, the density function of the resulting mixture model using the 'direct' approach is

$$f(y_i) = \prod_{i=1}^{n} \Big[(1 - p(a_i)) f_{S^-}(y_i; \mu(a_i), \sigma^2_{S^-}) + p(a_i) f_{S^+}(y_i; \delta\mu(a_i), \sigma^2_{S^+})\Big] \tag{3.12}$$

where $\delta > 1$ is a multiplicative factor accounting for the higher mean levels of antibodies in the S^+ population. Note that the seronegative distribution is also modelled as age-dependent to account for potential residual antibody levels due to previous infections. In the 'direct' approach, we utilize the underlying structure of the mixture distribution in order to estimate transmission parameters in the unified mechanistic model, thus avoiding dichotomization 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 3.3(a).

When analysing cross-sectional data, estimation of the model in (3.12) can be problematic because of the large number of parameters to estimate. In 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). The AIC is defined as $2p - 2\log(\hat{L})$, where p is the number of parameters in the model and $\hat{L}$ is the value of the likelihood function evaluated at the maximum likelihood estimate. The AIC is used to quantify the goodness of fit of a model to the data while penalizing models that contain a larger number of parameters. The AIC can be used to compare models that are not nested, i.e. models that are not contained within each other. A lower AIC usually indicates a better fit to the data.

Another simplification that we introduce in the maximization of the likelihood function is to fix the seroreversion rate ω. In practice, we found that using numerical optimization with a continuous ω was unstable as a result of a very flat likelihood surfaces. Maximization of the likelihood estimation is carried through unconstrained optimization using PORT routines as implemented in the "nlminb" function in R

3.3.1 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,

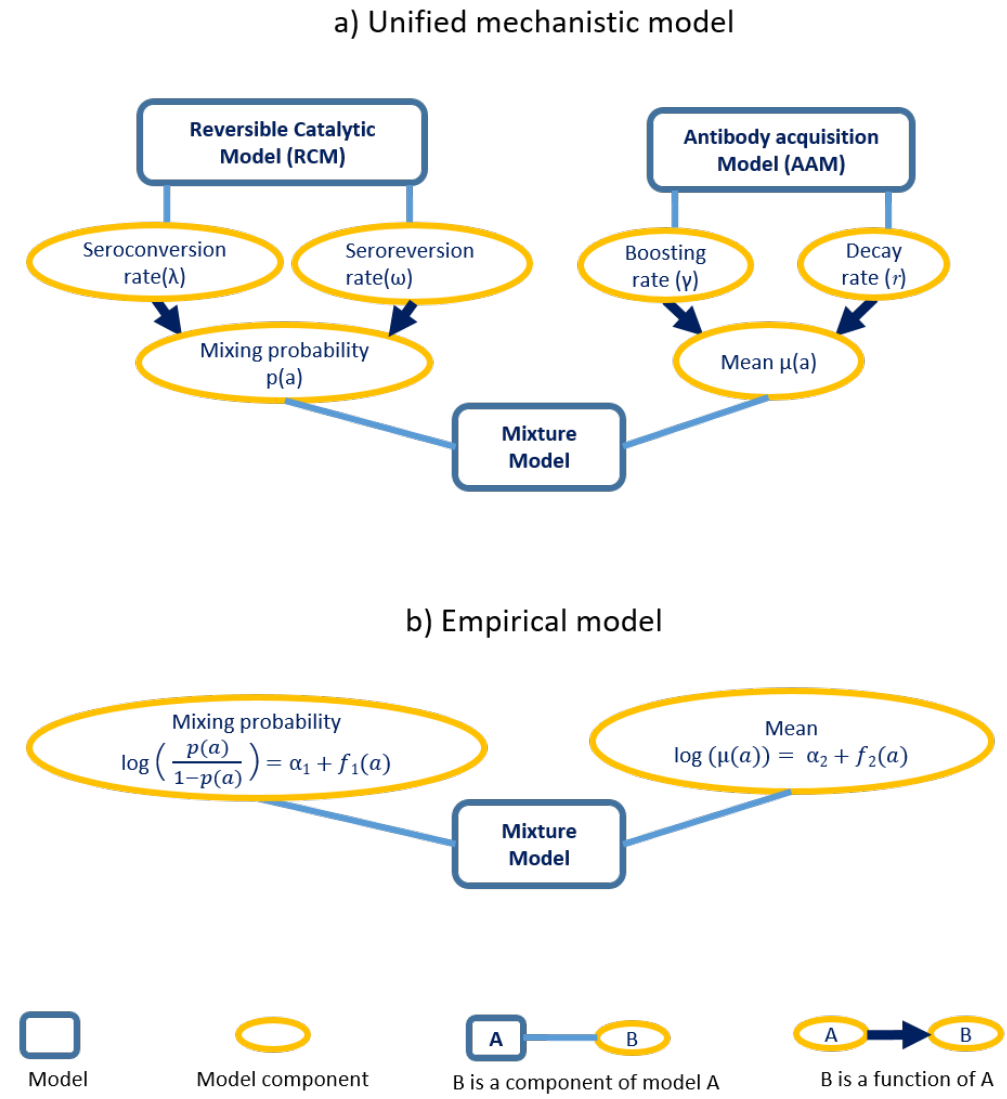

Figure 3.3: (a) is a representation of the unified mechanistic model, showing how the reversible catalytic model and antibody acquisition model are incorporated into the mixture model for antibody data. (b) is a representation of the empirical model used to model age-dependence in the mixing probabilities and mean antibody level.

we model the age-dependency in $\lambda(a)$ and $p(a)$ using a log-linear and logit-linear regression as

$$\begin{aligned} \log\left\{\frac{p(a)}{1-p(a)}\right\} &= \alpha_1 + f_1(a) \\ \log\{\mu(a)\} &= \alpha_2 + f_2(a) \end{aligned} \tag{3.13}$$

where $f_1(a)$ and $f_2(a)$ are functions which can be specified with the aid of simple graphical tools, such as scatter plots. The resulting structure of the empirical model is summarized in Figure 3.3b, and we give examples of this in the application of section 3.4.

3.4 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 [72]. Transmission was characterized as seasonal, following peaks in rainfall, typically between March-June and October-November [72, 73].

Most malaria was attributed to *Pf*, with predominant vector species being *Anopheles gambiae s.s., A. arabiensis,* and *A. funestus* [74, 75]. Malaria control interventions at the time included distribution of long-lasting Insecticide-treated nets (LLINs) which had been ongoing for many years, and Indoor Residual Spraying (IRS) which started in 2009 [75]. Further details of the study design and data collection can be found in Bousema et al. [72, 75]

In the study, finger prick blood was collected from all participants on filter paper and used to detect total Immunoglobulin G (IgG) antibodies against the blood-stage *Pf* antigen apical membrane antigen 1 (*Pf*AMA1) using the Enzyme-linked immunoassay (ELISA). 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 6-9 months[26, 40, 76]. 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 [40]. The

noisier distribution in older individuals may result from an accumulation of factors which increase the variation in antibody responses to malaria in the long-term, for example individuals' nutrition history, migration between regions of different endemicity, varying histories of intervention use, e.t.c.

The data-set consists of $n = 9549$ children. Figure 3.4 shows the age and OD distributions of the individuals included in the analyses.

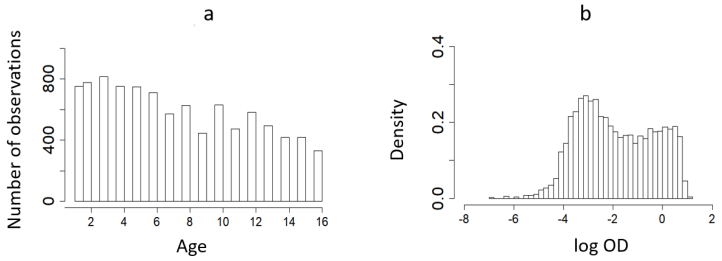

Figure 3.4: Descriptive plots of the age distribution (a) and the log OD distribution (b) of individuals aged 1-16, who are included in the *Pf*AMA1 antibody analysis.

We fit both unified mechanistic and empirical models to the *Pf*AMA1 antibody data using the maximum likelihood method of estimation. Maximum likelihood estimation is facilitated by unconstrained and box-constrained optimization using PORT routines through the nlminb function in R. The full reproducible code is available on GitHub (see 'Data availability'.

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 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.

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 [25, 26, 70]. In what follows, we present results for the best performing value for ω, i.e. $\omega = 0.01$.

To summarize, 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 (3.9) and (3.11); boosting rate γ and decay rate r from (3.10); and the mixture distribution parameters δ, $\sigma^2_{S^-}$, and $\sigma^2_{S^+}$ from (3.12).

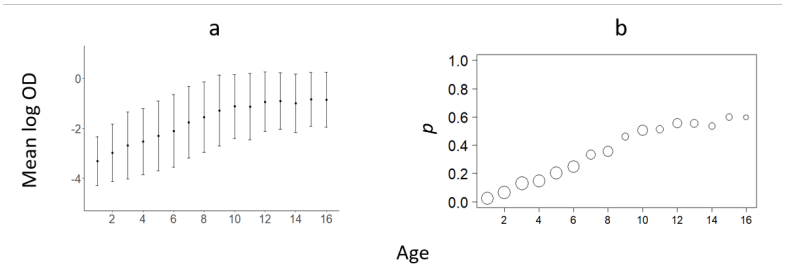

Figure 3.5: Exploratory analysis of the Rachuonyo South District *Pf*AMA1 antibody data. (a) shows the geometric mean OD across age while (b) shows the proportions of S^+ individuals, p, as defined by (3.1), using the seropositivity threshold (i.e. $\mu_{S^-} + 3\sigma_{S^-}$). The circle sizes in (b) are proportional to the sample size in each age group.

For the empirical model, $\mu(a)$ and the mixing probability are modelled according to (3.13), and are informed by Figure 3.5. We apply a linear spline with a knot at age 10, based on the empirical trend for $\mu(a)$ observed in Figure 3.5a, to give

$$\mu(a) = \exp\{\beta_1 + \beta_2 a + \beta_3 (a - 10) I(a > 10)\}, \tag{3.14}$$

where $I(a > 10)$ is an indicator function that takes value 1 if $a > 10$, and 0 otherwise. Based on Figure 3.5b, we introduce the log-transformed age as a logit-linear predictor for $p(a)$, such that

$$p(a) = \frac{\exp\{\tilde{\beta}_0 + \tilde{\beta}_1 \log a\}}{1 + \exp\{\tilde{\beta}_0 + \tilde{\beta}_1 \log a\}} \tag{3.15}$$

Thus, the model parameters to estimate for the empirical model are: the regression coefficients β_1, β_2, and β_3 in (3.14), and $\tilde{\beta}_0$ and $\tilde{\beta}_1$ in (3.15); 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 3.6 shows a bi-modal antibody distribution between ages 5 to 10, 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 (Figure 3.6), we notice that, while

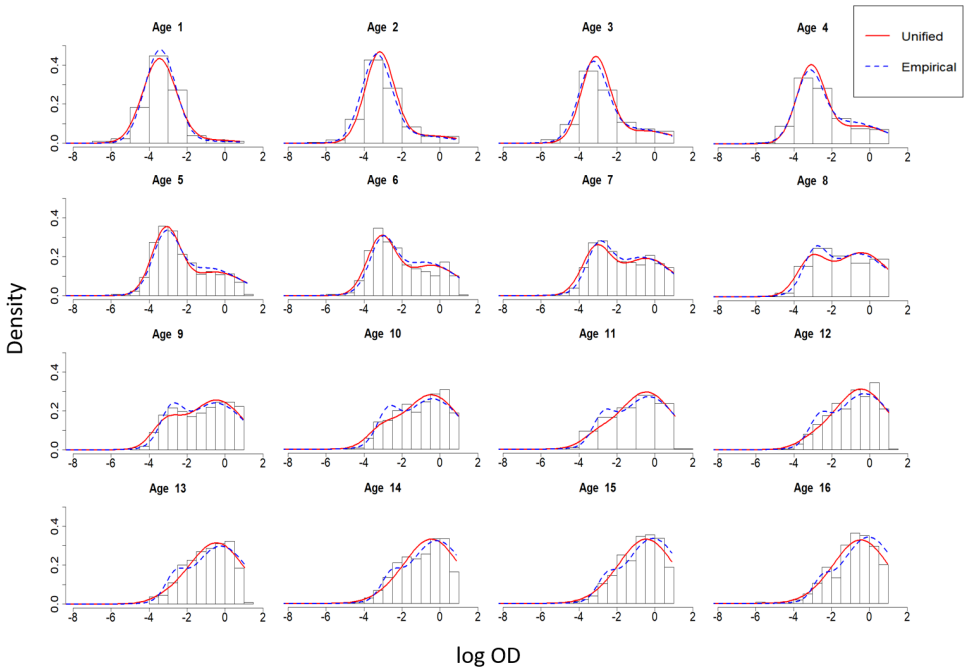

Figure 3.6: Age-dependent mixture distributions of *Pf*AMA1 antibodies for individuals 1 to 16 years of age in Rachuonyo South District. The red line indicates distributions derived from the unified mechanistic model, while the blue dotted line indicates distributions derived from the alternative empirical model.

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 3.1).

With regards to $\lambda(h)$, Figure 3.7 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 3.6. 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.

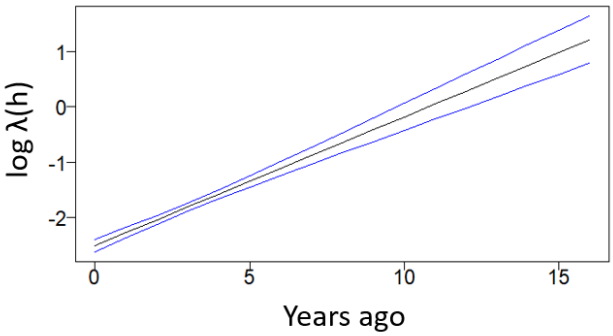

Figure 3.7: Changes in λ over historical time as derived from the unified mechanistic model fitted to *Pf*AMA1 antibody data. The blue lines indicate 95% CIs. 'Years ago' corresponds to $(a - h)$ as described in (3.11).

Table 3.1: 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. The Akaike Information Criterion (AIC) is also reported.

equation	parameter	UFM	EM
eq 3.9 and 3.11	l_0	-2.696 (-2.627, -2.397)	
	l_1	0.246 (0.202, 0.264)	
eq 3.10	γ	-1.5 (-1.687, -1.291)	
	r	3.806 (3.122, 4.754)	
eq 3.12	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)
eq 3.14	β_1		-3.141 (-3.191, -3.087)
	β_2		0.052 (0.046, 0.058)
	β_3		-0.021 (-0.032, -0.005)
eq 3.15	$\tilde{\beta}_0$		-3.031 (-3.194, -2.69)
	$\tilde{\beta}_1$		2.005 (1.915, 2.188)
	AIC	29791.910	29711.460

3.5 Discussion

We have introduced a unified mechanistic model which 1) avoids the dichotomization 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 the reversible catalytic model (RCM) and antibody acquisition model (AAM), respectively. The disadvantages of dichotomizing continuous data into binary data, a common practice in the standard use of RCMs, are well established. Dichotomiza-

tion can lead to the loss of information which affects the ability to reliably recover regression relationships and the precision of parameter estimates [3–6, 77]. 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. Therefore where the interest is in estimating $\lambda(a)$ and $\gamma(a)$, the unified mechanistic model is preferable, however when the interest is simply in describing the age dependency of the mixture model, or estimating seroprevalence, then the empirical model is preferable as it is more parsimonious. 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, as shown in Figure 3.6 were small for most ages.

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 $PfMSP1_{19}$ showed limited 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 the population is seropositive [21, 40]. Additionally, the seroreversion rate ω may also difficult to estimate in this scenario and, for this reason, is often fixed[20]. This is one of the main limitations in reversible catalytic models, which also applies to the unified mechanistic model. Generally, to alleviate the problem of over-parametrization, 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 flex-

ibility 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.

3.6 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 dichotomization of the continuous antibody measurements, a common practice in the analysis of malaria serology data. More importantly, the unified framework allows us to critically assess and evaluate assumptions on the heterogeneity of biological indicators of malaria transmission using a principled likelihood-based framework.

CRediT authorship contribution statement

Irene Kyomuhangi: Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Visualization, Writing-original draft, Writing-review and editing.
Emanuele Giorgi: Conceptualization, Methodology, Formal analysis, Supervision, Writing-review and editing.

Acknowledgements

IK is a Commonwealth Scholar, funded by the UK government. EG acknowledges support from the Academy of Medical Sciences thorough a Springboard Award (SBF0041009). 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. Special thanks to Prof. Niel Hens, Dr. Toman Neyens, Dr. Lindsey Wu, Prof. Chris Drakeley, and Dr. Gillian Stresman for useful discussions on this work.

Declaration of Competing Interest

We declare we have no competing interest.

Data availability

R scripts for implementation of the unified mechanistic and empirical models are available on request from the authors,and available on GitHub (https://github.com/kyomuhai/Kyomuhangi-and-Giorgi_unified-mechanistic-model.git).

Chapter 4

Paper 3: A threshold-free approach with age-dependency for estimating malaria seroprevalence

Irene Kyomuhangi[1], Emanuele Giorgi[1]

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

Submitted to: *Malaria Journal*

Summary

In malaria serology analysis, the standard approach to obtain seroprevalence, i.e the proportion of seropositive individuals in a population, is based on a threshold which is used to classify individuals as seropositive or seronegative. We argue that the choice of this threshold is often arbitrary and is based on methods that ignore the age-dependency of the antibody distribution. Using cross-sectional antibody data from the Western Kenyan Highlands, we introduce a novel approach that has three main advantages over the current threshold-based approach: it avoids the use of thresholds; it accounts for the age dependency of malaria antibodies; and it allows us to propagate the uncertainty from the classification of individuals into seropositive and seronegative when estimating seroprevalence. The reversible catalytic model is used as an example for illustrating how to propagate this uncertainty into the parameter estimates of the model. We find that accounting for age-dependency leads to a better fit to the data than the standard approach which uses a single threshold across all ages. Additionally, we also find that the proposed threshold-free approach is more robust against the selection of different age-groups when estimating seroprevalence. The novel threshold-free approach presented in this paper provides a statistically principled and more objective approach to estimating malaria seroprevalence. The introduced statistical framework also provides a means to compare results across studies which may use different age ranges for the estimation of seroprevalence.

Keywords: malaria serology; geostatistical model; reversible catalytic model; antibody acquisition model; unified mechanistic model.

4.1 Introduction

Thanks to increased diagnostic capacity, preventative measures and a scale-up of interventions, there has been an overall decrease in malaria burden worldwide [78, 79]. However, malaria still remains a significant global public health threat in sub-Saharan Africa, where *Plasmodium falciparum (Pf)* is the predominant parasite. A total 229 million cases and 409,000 deaths have been estimated globally in 2019 [80]. Additionally, the decrease in malaria is heterogeneous across regions, countries and communities [79–83], posing additional challenges to malaria elimination efforts. These challenges require robust surveillance mechanisms which can adapt to the changing epidemiology, enabling a more targeted approach to intervention strategies [81, 84].

To estimate malaria exposure and transmission, analysis of human serology data has emerged as a viable alternative approach to disease risk metrics that are based on the detection of malaria parasites in humans and mosquito populations [20, 25, 40]. Because of the persistence of antibodies after infection, their concentration is less influenced by the seasonality of transmission and can be used as an indicator of the cumulative exposure to malaria. Additionally, antibodies, unlike the *Plasmodium* parasite, can be easily detected even in low transmission areas [21, 25, 26, 28].

Analysis of seroprevalence - i.e the proportion of 'seropositive' individuals - is often carried out using reversible catalytic models (RCM). These models allow for the estimation of seroconversion rates which quantify the transmission intensity and correspond to the rate at which individuals convert from seronegative to seropositive through exposure to malaria parasites over time [20, 25]. Alternatively, continuous antibody measurements can be used in antibody acquisition models to estimate boosting rates, another measure of transmission intensity, which represents the rate at which antibodies are boosted upon exposure to parasites [20, 40, 85]. Such indicators of transmission intensity can be used to inform decisions on intervention strategies by identifying hot-spots of transmission where individuals are likely to exceed a specified degree of exposure [73, 86].

To estimate seroprevalence, classification of individuals as seropositive or seronegative is required. The most commonly used approach is to identify a suitable threshold of antibody density beyond which individuals are classified as seropositive, and below as seronegative [20, 25, 26]. To this end, mixture distributions are first fitted to the antibody density data, assuming that continuous antibody measurements consists of two latent distributions, one for the seronegative and one for the seropositive populations. By using the point estimates of the mean,

μ_{S^-} and standard deviation, σ_{S-}, of the seronegative distribution S^-, the seropositivity threshold is often set to $\mu + 3\sigma$ [20, 32, 34, 35], while other studies have instead used $\mu + 2\sigma$ [36–38]. An alternative to this approach is to define thresholds based on the predictive probability of being seropositive resulting from the fitted mixture distribution [20].

The major drawback of threshold-based approaches is that the choice of the threshold is arbitrary and it is unclear to what extent this affects the results of the statistical analysis of serological data, as biased estimates of seroprevalence can in fact arise from the misclassification of individuals as seronegative or seropositive [64]. Additionally, in the case of the probability thresholds, individuals whose probability of belonging to either the seronegative or seropositive groups is close to 50% are often classified as 'intermediate', and are therefore excluded from analysis [20, 64]. Furthermore, the uncertainty around the estimated thresholds and probabilities used for the classification of individuals, is ignored.

In addition to these drawbacks, classical analysis of malaria serology data does not account for the age dependency of antibody distribution when calculating thresholds. Typically in mixture models, a threshold is obtained by assuming a constant mixing probability across all ages [85]. This assumption is questionable since, in malaria endemic settings, it is well known that antibody levels are in fact age-dependent [87, 88] and thus the likeihood of being seropositive is expected to increase with age. A 2011 study by Ster [89] incorporated age-dependency for varicella zoster virus serology mixture models, however to our knowledge, this principle has not been applied to malaria serology data

To address these issues, Kyomuhangi et al. [85] proposed a unified modelling framework for the analysis of malaria serology data that uses the continuous antibody measurement rather than thresholds to estimate transmission parameters. However, as acknowledged by the authors, this modelling framework requires a larger amount of data than is usually available in serological studies to reliably estimate all the model parameters, thus limiting its applicability.

In this paper, we propose a novel modelling approach for the analysis of serological data that retains the same properties of the approach proposed in Kyomuhangi et al. [85], but is also more parsimonious. More specifically, our novel approach satisfies the following requirements: 1) it accounts for age dependency of antibody levels; 2) it avoids the use of any threshold; and 3) it allows us to account for and propagate the uncertainty in the classification of seropositive and seronegative individuals. Using cross-sectional antibody data from Western Kenya, we demonstrate 1) the properties of this new methodology for estimating malaria seroprevalence,

and 2) how to incorporate the uncertainty around the resulting seroprevalence estimates, using the standard RCM as an example. In the discussion, we explain how the principles used to develop this novel approach can be extended to more complex analysis of serological data.

Methods

Existing methods for estimating seroprevalence

Here, we outline the most commonly used methods in the analysis of malaria serology data, to classify individuals as seropositive and seronegative, using a two-component mixture distribution.

Let Y_i denote the log-transformed antibody measurement for the i-th individual in a sample, S^- denote the seronegative classification, and S^+ denote the seropositive classification. Assuming independent and identically distributed realizations for a sample of n individuals, and μ to be the mean level of antibodies in the S^- distribution, the density function of $Y = (Y_1, \ldots, Y_n)$ is

$$f(y) = \prod_{i=1}^{n} \left[(1-p) f_{S^-}(y_i; \mu, \sigma^2_{S^-}) + p f_{S^+}(y_i; \delta\mu, \sigma^2_{S^+}) \right] \tag{4.1}$$

where f_{S^-} is a univariate log-normal distribution with mean μ and variance $\sigma^2_{S^-}$ for the S^- population, and analogously for S^+, with $\delta > 1$ being a multiplicative factor accounting for higher mean antibodies in the S^+ distribution. 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. Based on the seropositivity threshold κ, the classification of individuals, say C_i, into S^+ and S^- is defined as

$$C_i = \begin{cases} S^- & \text{if } Y_i < \kappa \\ S^+ & \text{if } Y_i \geq \kappa \end{cases}. \tag{4.2}$$

In our application, we shall use $\kappa = \mu_{S^-} + 3\sigma_{S^-}$, as this is used in most other statistical analyses of malaria serology data.

Proposed method for estimating seroprevalence

We propose a novel modelling framework that overcomes the limits of the approach described in the previous section, by incorporating age-dependency into the mixture distribution in (4.1), and by propagating the uncertainty in the classification

of individuals into S^+ and S^- using a Monte Carlo approach.

In this framework, age dependency is introduced into (4.1) using linear regression, as described in Kyomuhangi et al. [85]. Assuming $\mu(a_i)$ to be the mean level of antibodies in the S^- distribution for a given age a_i, (4.1) becomes

$$f(y) = \prod_{i=1}^{n} \left[(1 - p(a_i)) f_{S^-}(y_i; \mu(a_i), \sigma^2_{S^-}) + p(a_i) f_{S^+}(y_i; \delta\mu(a_i), \sigma^2_{S^+}) \right] \tag{4.3}$$

where $p(a_i)$ is the probability of being S^+ at age a. Note that the seronegative distribution is also modelled as age-dependent to account for potential residual antibody levels due to previous infection. The age dependencies in $p(a)$ and $\mu(a)$ are modeled using logit linear and log linear regression respectively such that

$$\begin{aligned} \log\left\{ \frac{p(a)}{1 - p(a)} \right\} &= \alpha_1 + g_1(a) \\ \log\{\mu(a)\} &= \alpha_2 + g_2(a) \end{aligned} \tag{4.4}$$

where $g_2(a)$ is a function of age that can be specified through empirical inspection of the data. In the case of $g_1(a)$, identifying a suitable specification may be more problematic because of the need to dichotomize the data. However, because it is well established that $p(a)$ increases for increasing a, a pragmatic approach would be, for example, to specify a logit-linear regression on a as illustrated later in this paper. Note that predictor for these models can take other functional forms such as polynomials and smoothing splines to increase their flexibility.

Using the resulting mixture distribution, we compute the predictive probability of belonging to the S^+ distribution for each sampled individual, by conditioning on the observed antibody measurement $Y_i = y_i$ and age a_i, to give

$$\begin{aligned} P(C_i^* = S^+ | y_i, a_i) &= \frac{p(a_i) f_{S^+}(y_i; \theta_{S^+})}{(1 - p(a_i)) f_{S^-}(y_i; \theta_{S^-}) + p(a_i) f_{S^+}(y_i; \theta_{S^+})} \\ P(C_i^* = S^- | y_i, a_i) &= 1 - P(C_i^* = S^+ | Y_i = y_i, a_i) \end{aligned} \tag{4.5}$$

where $\theta_{S^-} = (\mu(a_i), \sigma^2_{S^-})$ and $\theta_{S^+} = (\delta\mu(a_i), \sigma^2_{S^+})$. Based on the above expressions, when then simulate 10,000 classifications C_i^* for a every single sampled individual. The resulting 10,000 data-sets generated from this process are then fed into the second stage of the analysis, which we explain in the next section.

We point out two main advantages of this modelling approach. The first is that it avoids the use of a threshold κ as in (4.2) and uses the generated samples C_i to propagate the uncertainty of the classification into S^+ and S^-. The second is that the empirical approach used to account the age-dependency combines information

across all ages as described in (4.4), and is therefore more efficient than fitting separate mixtures distribution for each age.

The structure of this modeling framework is illustrated in Figure 4.1.

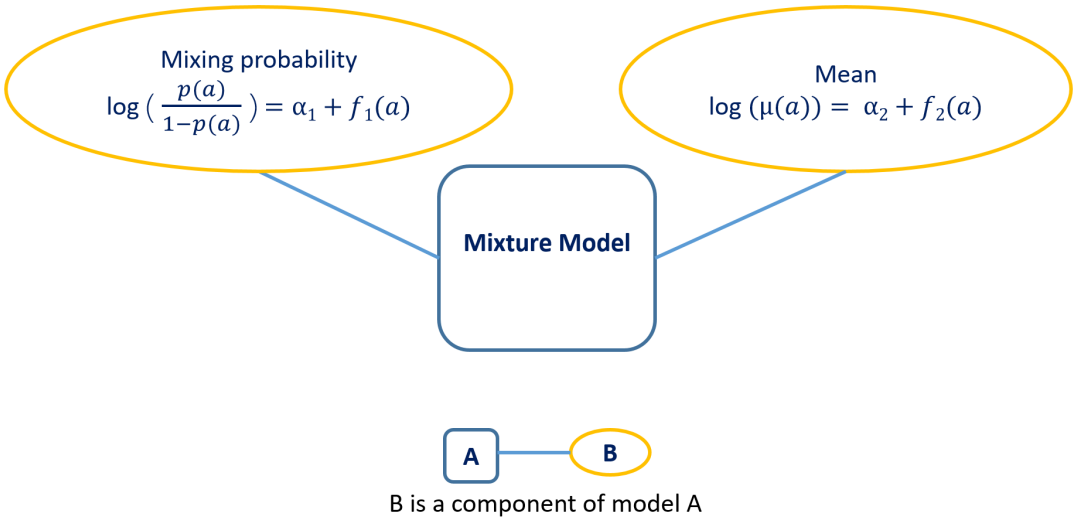

Figure 4.1: An illustration of the empirical model introduced in Kyomuhangi et al. [85].This model is used to describe the antibody mixture distribution as indicated in equations (4.3) and (4.4)

The reversible catalytic model

The RCM assumes that individuals are born S^- and, after becoming S^+ upon exposure to malaria, can revert to S^- in the absence of exposure. Since antibody data are believed to represent the cumulative exposure of individuals during their lifetime, an individual's age prior to the sample collection is used as 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, i.e. $p(a)$, are expressed by the following differential equation

$$\frac{dp}{da} = \lambda(a)(1 - p(a)) - \omega p(a). \tag{4.6}$$

The seroconversion rate $\lambda(a)$ can be modelled using a variety of approaches, the simplest of which assumes constant transmission, i.e. $\lambda(a) = \lambda$ for all a. Due to poor identifiability of the seroreversion rate ω, this is typically fixed and assumed to be constant across ages [20, 40, 42, 85]. Assuming a constant λ and ω in (4.6)

gives the following solution

$$p(a) = \frac{\lambda}{\lambda + \omega}\Big(1 - e^{-(\lambda+\omega)a}\Big). \tag{4.7}$$

More flexible models could also be used to account for the temporal variation in λ, including a step-wise reduction or linear reduction in transmission [20, 42]. Additionally, other specifications of the RCM, for example the superinfection RCM [32] could be applied in the proposed approach. However, in this paper, we restrict our attention to the RCM defined in the above equation for simplicity, as we compare existing methods and the proposed approach described in the previous sections.

In order to propagate the uncertainty in classification of individuals as S^+ and S^-, for the purpose of estimating parameters of the RCM, we maximize the likelihood of a Binomial distribution with probability $p(a)$, as indicated in (4.7), for each of the 10,000 data-sets for the outcome C_i as described in the previous sections. This gives 10,000 different estimates for λ, which we summarize by taking their mean and 2.5% and 97.5% quantiles.

The estimation of the model parameters is conducted using the maximum likelihood estimation method. Let z_i denote the binary variable indicating seropositivity ($z_i = 1$) or seronegativity ($z_i = 0$) for the i-th individual; the likelihood function for the RCM in (4.7) is then

$$\prod_{i=1}^{n} p(a_i)^{z_i}(1 - p(a_i))^{1-z_i}$$

Data

We analyse data from cross-sectional survey which was conducted in Rachuonyo South District (34.75 to 34.95°E, 0.41 to 0.52°S), in the western Kenyan highlands (1400 m to 1600 m altitude), in 2011 over a 100 km^2 area. This survey was the baseline for a cluster-randomized controlled trial whose aim was to determine the community effect of interventions targeted at malaria prevalence hotspots. Further details of the study protocol can be found in Bousema et al. [72] At the time of the survey, malaria transmission in this area was described as low but highly heterogeneous, and seasonal, following peaks in rainfall, typically between March-June and October-November [72, 73].

The majority of malaria cases were attributed to *Pf*, with *Anopheles gambiae s.s., A. arabiensis,* and *A. funestus* being the predominant vector species. Malaria control interventions at the time included distribution of Insecticide-treated nets

which had been ongoing for many years, and Indoor Residual Spraying which started in 2009[74, 75].

To generate the serology data, finger prick blood samples were collected from participants on filter paper and used to detect total Immunoglobulin G (IgG) antibodies against the blood-stage *Pf* antigens, apical membrane antigen 1 (*Pf*AMA1) and merozoite surface protein-1_{19} (*Pf*MSP1_{19}). Standard Enzyme-linked immunoassay (ELISA) methods [26, 90] were used to obtain Optical density (OD) values. Further details of the study design and data collection can be found in Bousema et al. [72].

We first restrict analysis to individuals between 1 and 16 years. Additional analysis on 1-20 year olds, 1-30 year olds, and 1-50 year olds is presented in the supplementary material. We split the data this way in order to investigate the effect of selecting different age-groups on the performance of M1 and M2. In what follows, we shall first focus in the 1-16 year old age group.

The data-set consists of $n = 9549$ children for the *Pf*AMA1 analysis and $n = 9576$ for the *Pf*MSP1_{19} analysis. Figure 4.2 shows the age and OD distributions of the individuals included in the analyses.

Specifications of the model components

In this analysis, we compare two modelling approaches in the estimation of seroconversion rates, for both *Pf*AMA1 and *Pf*MSP1_{19}.

The first, which we refer to as M1, is the classic threshold-based approach as defined in (4.1), which considers seropositivity according to (4.2). After dichotomization of the antibody measurements, the RCM, as described by (4.6), is fitted using the maximum likelihood method.

The second modelling approach, which we refer to as M2, is the proposed threshold-free approach described in the previous sections. For this analysis, the age-dependency of the mixture models for the two antigens is modelled using a an empirical approach. Based on the Figure 4.3(a) for *Pf*AMA1, we use a linear spline with a knot at the age of 10 years, formally expressed as

$$\mu(a) = \exp\{\beta_0 + \beta_1 a + \beta_2 (a - 10) I(a > 10)\}, \tag{4.8}$$

where $I(a > 10)$ is an indicator function that takes value 1 if $a > 10$, and 0 otherwise. For *Pf*MSP1_{19}, based on the trend observed in Figure 4.3(b), we use

log-linear model, given by

$$\mu(a) = \exp\{\beta_0 + \beta_1 a\}. \tag{4.9}$$

To account for the age dependency in $p(a)$, we introduce age as a logit-linear predictor of $p(a)$, i.e.

$$p(a) = \frac{\exp\{\tilde{\beta}_0 + \tilde{\beta}_1 a\}}{1 + \exp\{\tilde{\beta}_0 + \tilde{\beta}_1 a\}}. \tag{4.10}$$

Note that M1 is recovered when all the regression parameters except β_0 and $\tilde{\beta}_0$ in (4.8), (4.9) and (4.10) are set to 0. Therefore for M1, we will report the estimates for β_0 and $\tilde{\beta}_0$ only.

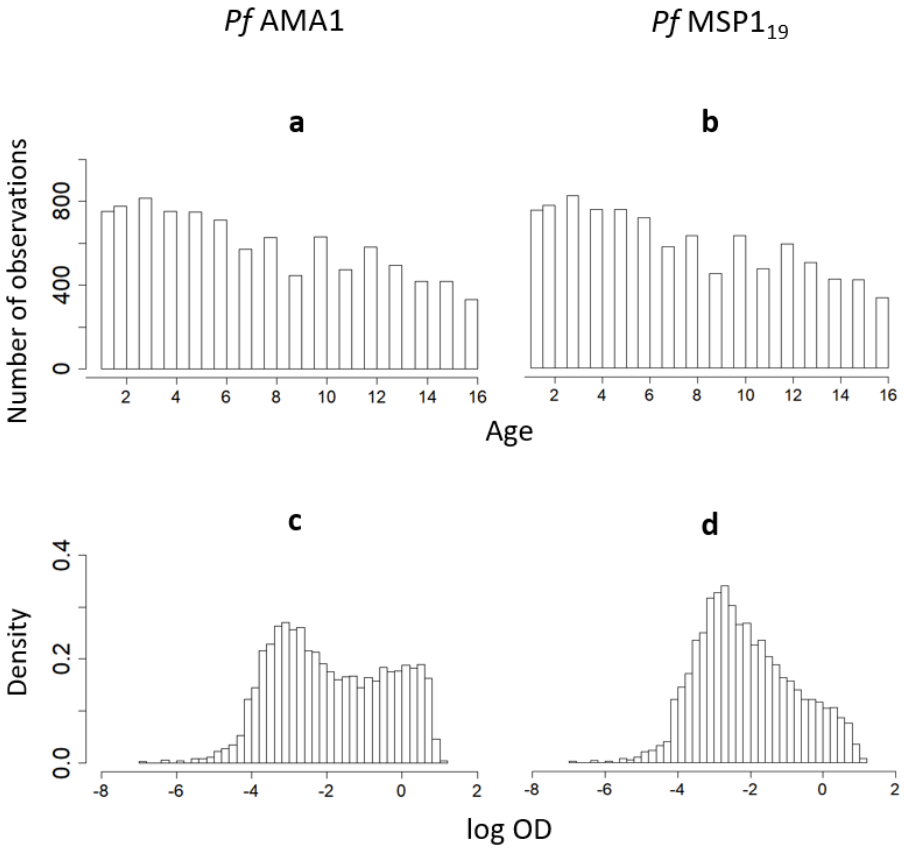

Figure 4.2: Descriptive plots of *Pf*AMA1 and *Pf*MSP1$_{19}$ antibodies for individuals between ages 1 and 16. The top row shows the age distribution, the bottom row shows the log OD distribution of individuals included in the analysis

For both M1 and M2, due to the truncated nature of the antibody distributions, we use truncated log-normal distributions for both antigens. The upper limit, say $y_{max}(a_i)$ of the truncation is estimated for each age group as the maximum

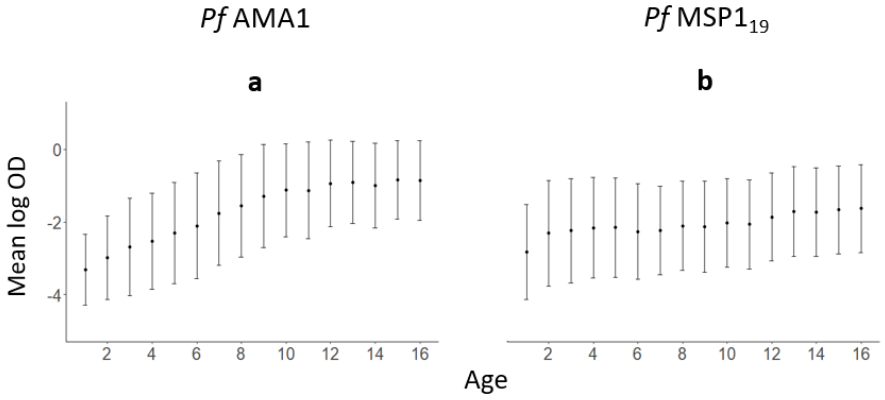

Figure 4.3: Exploratory analysis of *Pf*AMA1 and *Pf*MSP1$_{19}$ antibodies for individuals between ages 1 and 16. The figure shows the geometric mean OD by age, with associated error bars

observed value of OD. Hence, the likelihood function in (4.3) now becomes

$$f(y) = \prod_{i=1}^{n} \frac{[(1-p(a_i))f_{S^-}(y_i;\mu(a_i),\sigma^2_{S^-}) + p(a_i)f_{S^+}(y_i;\delta\mu(a_i),\sigma^2_{S^+})]}{[(1-p(a_i))F_{S^-}(y_{max};\mu(a_i),\sigma^2_{S^-}) + p(a_i)F_{S^+}(y_{max};\delta\mu(a_i),\sigma^2_{S^+})]}, \tag{4.11}$$

where F_{S^+} and F_{S^-} are the cumulative distribution functions of seropositive and seronegative probability distributions, respectively.

Finally, for the RCM, we considered a range of values from 0.01 to 1 for ω hence assuming that seroreversion events for individuals would occur between 1 and 100 years [25, 26, 70, 86]. Profile likelihood analysis indicated flat likelihood surfaces (see Figure S2), therefore we let ω take three values, namely 0.01, 0.5 and 1 to represent low, medium and high seroconversion rate respectively. In what follows, we present results for the best performing value of ω for each antigen, i.e. $\omega = 0.01$ for *Pf*AMA1 and $\omega = 1$ for *Pf*MSP1$_{19}$. Note that these values are not the maximum likelihood estimates for ω, but rather the best performing values out of the three choices stated above.

A summary of model parameters to estimate in this analysis is provided in Table 4.1. In order to compare how well M1 and M2 fit the data, we calculate the Akaike information criterion (AIC), defined as $2p - 2\log(\hat{L})$, where p is the number of parameters in the model and $\hat{L}$ is the value of the likelihood function evaluated at the maximum likelihood estimate. The AIC is used to quantify the goodness of fit of a model to the data while penalizing models that contain a larger number of parameters. The AIC can be used to compare models that are not nested, i.e. models that are not contained within each other. A lower AIC usually indicates a better fit to the data. All statistical analyses are conducted in the R version

4.1.1 (2021-08-10) [91] software environment, and maximization of the likelihood estimation is carried through unconstrained optimization using PORT routines as implemented in the "nlminb" function in R. The full reproducible code is available on GitHub (see 'Availability of data and material'.

Table 4.1: Model specification for the analysis

model	equations	age-dependency	threshold	parameters to estimate
M1	(4.1), (4.8), (4.9), (4.10), (4.7)	No	Yes	δ, $\sigma^2_{S^-}$, $\sigma^2_{S^+}$, β_0, $\tilde{\beta}_0$, λ
M2	(4.3), (4.8), (4.9), (4.10), (4.7)	Yes	No	δ, $\sigma^2_{S^-}$, $\sigma^2_{S^+}$, β_0, β_1, β_2, $\tilde{\beta}_0$, $\tilde{\beta}_1$, λ

Results

A comparison of AIC in Table 4.2 shows a lower value for M2 than M1 for both antigens (29669.940 versus 33354.100 for *Pf*AMA1, and 31162.920 versus 31886.310 for $Pf\text{MSP1}_{19}$), indicating that the age-dependent mixture model in M2 is a better fit to the data compared to M1, which assumes a single mixture distribution across all ages. This age dependency is illustrated in Figures 4.5 and 4.6, which show an increase in mean antibody levels and the mixture distribution with age. Of note, the increase is much more prominent for *Pf*AMA1, than for $Pf\text{MSP1}_{19}$.

Additionally, in both M1 and M2, the separation between the two components of the mixture distribution is more prominent in *Pf*AMA1 (Figure 4.5) than in $Pf\text{MSP1}_{19}$ (Figure 4.6) where there is poor separation of the S^+ and S^- distributions. In the M2 *Pf*AMA1 analysis, the bi-modal distribution is more pronounced between the ages of 5 to 10 years, and less so in younger and older individuals. Figures 4.5 and 4.6 also indicate that age modulates the seropositivity threshold.

Figure 4.7 shows the difference in seroprevalence estimation between M1 and M2, with overall higher estimates across age in the latter model. For both antigens, we observe that the uncertainty resulting from M2, as quantified by the 95% confidence intervals (CIs), in the seroprevalence estimates of the RCM is considerably larger than M1. This is because the M2 estimates are obtained by incorporating the uncertainty in the seropositivity classification, while M1 ignores this uncertainty, resulting in very narrow confidence intervals for M1.

Figure 4.7 also shows that the RCM fitted using M2, provide a good interpolation of the seroprevalence for $Pf\text{MSP1}_{19}$ but less so for the *Pf*AMA1. Although most of the seroprevalence points fall within the 95% confidence interval, it is evident that, as we approach 15 years of age, where the observed seroprevalence is not contained within the 95% intervals, the model underestimates seroprevalence. This is made more clear by visualizing the the y-axis of the plot in Figure 4.7 on the logit-scale

Table 4.2: Maximum likelihood estimates with associated 95% CIs (within brackets) for M1 and M2, fitted to *Pf*AMA1 and *Pf*MSP1$_{19}$ antibody data. The Akaike Information Criterion (AIC) is also reported for the mixture models.

		parameter	M1	M2
*Pf*AMA1	Mixture Model	β_0	-2.338 (-2.428, -2.249)	-3.164 (-3.217, -3.111)
		β_1		0.052 (0.045, 0.058)
		β_2		-0.037 (-0.052, -0.023)
		$\tilde{\beta}_0$	-0.565 (-0.671, -0.460)	-2.085 (-2.281, -1.890)
		$\tilde{\beta}_1$		0.401 (0.371, 0.432)
		δ	11.706 (10.778, 12.722)	30.613 (26.224, 35.764)
		$\sigma^2_{S^-}$	0.014 (0.011, 0.019)	$1.665 \cdot 10^{-3}$ ($1.383 \cdot 10^{-3}$, $2.003 \cdot 10^{-3}$)
		$\sigma^2_{S^+}$	0.884 (0.716, 1.092)	43.521 (25.898, 73.138)
		AIC	33354.100	29669.940
	RCM	λ	0.022 (0.020, 0.023)	0.175 (0.109, 0.286)
*Pf*MSP1$_{19}$	Mixture model	β_0	-2.165 (-2.2656, -2.064)	-2.915 (-2.989, -2.841)
		β_1		0.031 (0.028, 0.034)
		$\tilde{\beta}_0$	-1.220 (-1.429, -1.010)	0.081 (-0.114, 0.277)
		$\tilde{\beta}_1$		0.038 (0.022, 0.054)
		δ	9.256 (8.624, 9.941)	11.698 (10.385, 13.193)
		$\sigma^2_{S^-}$	0.021 (0.015, 0.028)	$2.770 \cdot 10^{-3}$ ($2.081 \cdot 10^{-3}$, $3.687 \cdot 10^{-3}$)
		$\sigma^2_{S^+}$	0.994 (0.735, 1.346)	5.340 (3.387, 8.420)
		AIC	31886.310	31162.920
	RCM	λ	0.060 (0.055, 0.066)	1.459 (0.760, 2.675)

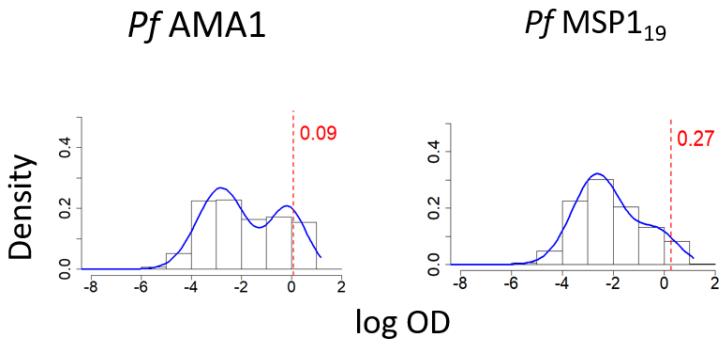

Figure 4.4: Mixture distributions of *Pf*AMA1 and *Pf*MSP1$_{19}$ antibodies for individuals between ages 1 and 16 using M1. These mixture distributions are derived from equation (4.1), and all the data of individuals aged 1-16 are analysed together. The red dotted lines illustrate the seropositivity thresholds ($\mu_{S^-} + 3\sigma_{S^-}$), above which individuals are be classified as S^+ in traditional analysis.

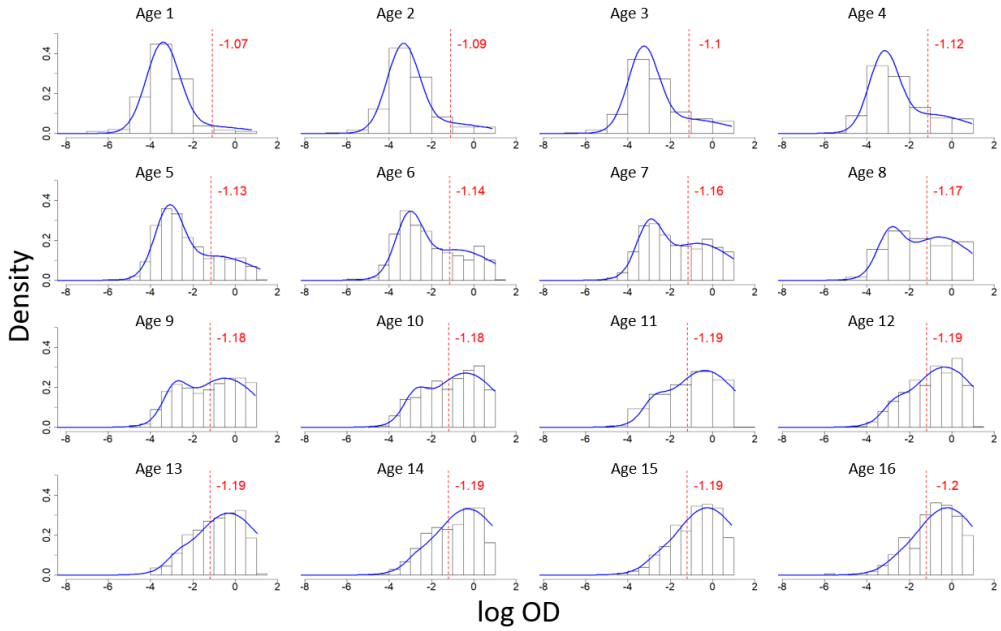

Figure 4.5: Age-dependent mixture distributions of *Pf*AMA1 antibodies for individuals between ages 1 and 16 using M2. The blue line shows fitted distributions derived from equations (4.3), (4.8) and (4.10). The red dotted lines illustrate the seropositivity thresholds ($\mu_{S^-} + 3\sigma_{S^-}$), above which individuals would be classified as S^+ in M1. Note that the red dotted lines are for illustration only - M2 does not use thresholds

(see supplementary Figure S3). This indicates that, in the case of *Pf*AMA1, the assumptions of the standard RCM may not be fully supported by the data, which is undetected by the standard threshold-based model M1.

The distributions of λ estimates derived from M2 for both antigens are shown in Figure 4.8. For *Pf*AMA1, λ is 0.175 (0.109, 0.286), while for *Pf*MSP1$_{19}$, this is 1.459 (0.760, 2.675). Note that these estimates represent the mean, 2.5% and 97.5% quantiles from the Monte Carlo distributions of the maximum likelihood estimates for λ.

Finally, supplementary Figures S4 and S5 show that M2 is consistent in the estimation of both seroprevalence and λ, even when different age groups are considered in analysis, unlike M1. Figure S6 also shows the additional variation in seroprevalence estimates for M1 when different seropositivity thresholds are used. Note the marked decrease in seroprevalence estimates as the threshold increases (see supplementary material).

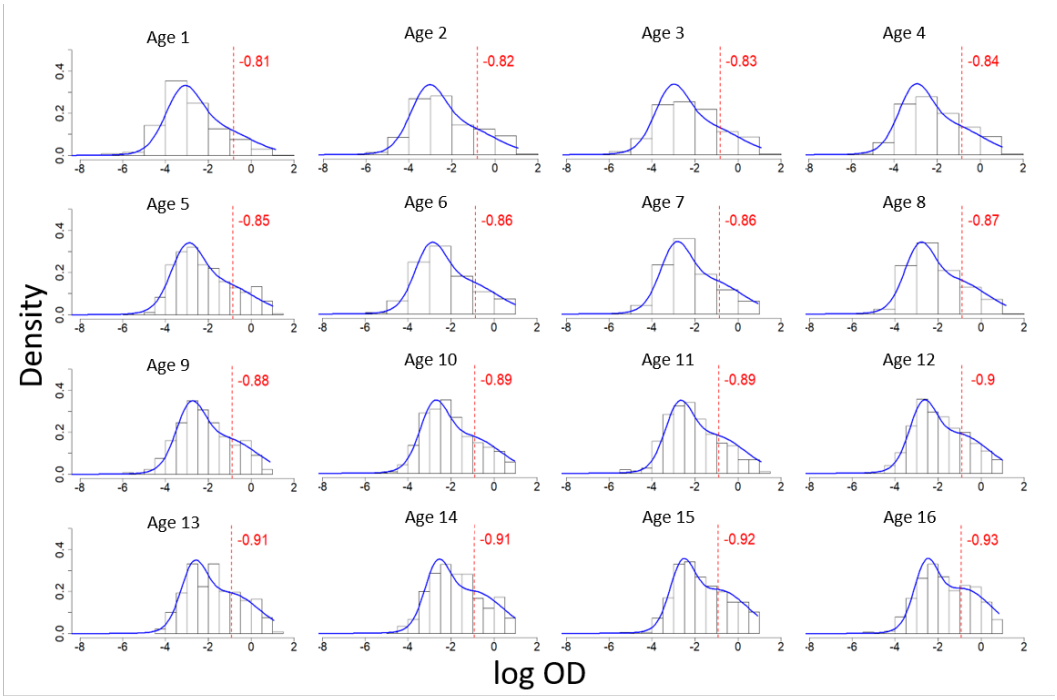

Figure 4.6: Age-dependent mixture distributions of *Pf*MSP1$_{19}$ antibodies for individuals between ages 1 and 16 using M2. The blue line shows fitted distributions derived from equations (4.3), (4.9) and (4.10). The red dotted lines show the seropositivity thresholds ($\mu_{S^-} + 3\sigma_{S^-}$), above which individuals would be classified as S^+ in M1. Note that the red dotted lines are for illustration only - M2 does not use thresholds

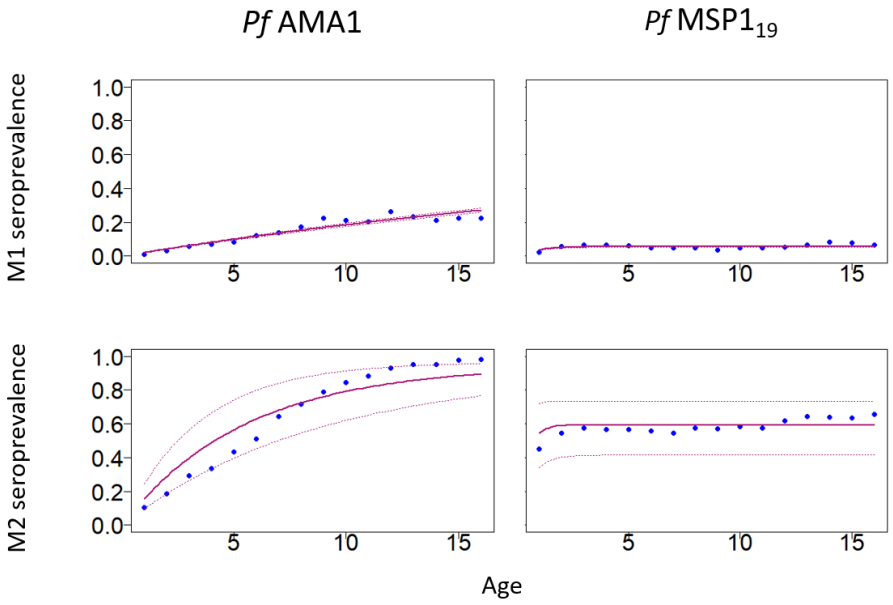

Figure 4.7: *Pf*AMA1 and *Pf*MSP1$_{19}$ seroprevalence estimates from M1, and seroprevalence distributions from M2, for individuals between ages 1 and 16. The top row shows M1 seroprevalence point estimates (blue dots), as well as the fitted seroprevalence curve (purple curve) and 95% CIs (purple dotted curves) from the RCM. The bottom row shows the mean of the seroprevalence distribution derived from M2 (blue dots), as well as the fitted seroprevalence curve (purple curve) and 95% CIs (purple dotted curves) from the RCM.

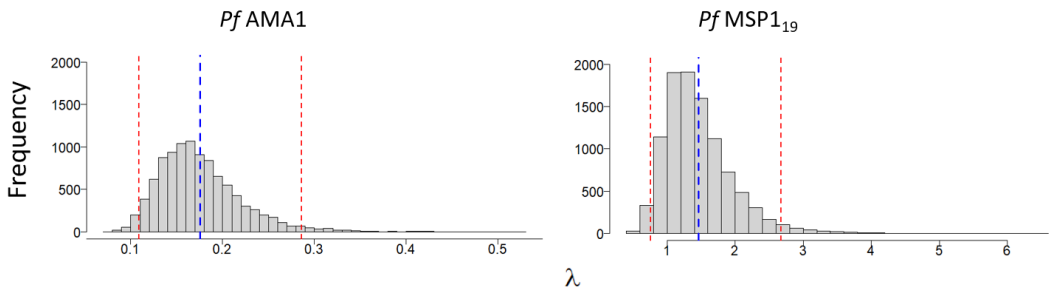

Figure 4.8: Distributions of the seroconversion rate λ derived from M2 for *Pf*AMA1 and *Pf*MSP1$_{19}$. The mean and 95% CIs for λ are indicated by blue and red dotted lines respectively. For *Pf*AMA1, these are 0.175 (0.109, 0.286), while for *Pf*MSP1$_{19}$, they are 1.459 (0.760, 2.675)

Discussion

In this paper we have presented a threshold-free method for estimating seroprevalence that incorporates the age dependency of malaria antibodies in the classification of individuals into seropositive and seronegative. Additionally, we have demonstrated how the uncertainty of this classification can be accounted for in a the RCM. We also point out that this approach can be applied to other types of analyses that require the use of models different from the RCM. For example, if the goal of the study is to map seroprevalence data within a study area, the simulated classifications (previously denoted by C_i) could be used as the input of a geostatistical model whose results are then summarized in a similar fashion as presented for the RCM in this paper.

In the application of our modelling framework to the RCM, seroprevalence is modelled into two different stages, using two different approaches: first, in a mixture distribution, using a logit-linear regression; and secondly, in an RCM, using equation (4.7). This raises the question of a mathematical inconsistency since both equations cannot be simultaneously true. We point out that this issue also applies to previous work which uses threshold-based RCMs [25, 26, 34, 86], whereby the threshold is first generated using a constant mixing probability, which would correspond to an intercept-only logit-linear regression in our case, and is then modelled using equation (4.7). To avoid this issue, one solution would be to replace the logit-linear regression on age for seroprevalence, with equation (4.7), hence embedding the assumptions of the RCM directly into the mixture distribution. However, our preference remains with the approach illustrated in this paper for the following reasons. First, the use of a logit-linear regression on age in the mixture distributions allows us to develop an empirical approach that is more flexible than an RCM and can better capture the variations of the antibody distributions across

age. Secondly, the use of the RCM-based equation (4.7) for seroprevalence also in the mixture distributions would yield a circular argument, whereby the outcome to be modelled with the RCM would be already generated under an RCM, thus making any validation of the RCM assumptions a vain exercise. As shown in our case-study with western Kenya data, our approach can in fact better detect the inadequacy of the RCM than the current threshold-based approach.

The results in this paper show clear age-dependency in the mean antibody levels, the mixture distribution, and the threshold. The differences between *Pf*AMA1 and *Pf*MSP1$_{19}$ indicate that the magnitude of this dependency is likely dependent on the type of antigen and the dynamics of the immune response to it. Notably, results provide evidence that different combinations of age-groups in analysis lead to different seropositivity thresholds and, therefore different seroprevalence estimates. This inconsistency has significant implications for control programs which rely on these results to direct intervention strategies. A key advantage of the threshold-free approach is that it is unaffected by the age limits considered for the analysis.

Furthermore, different definitions of the seropositivity threshold (i.e. between 2 and 5 standard deviations of the mean of the seronegative distribution) are an additional source of inconsistency in current literature. This makes the comparability of results reported across malaria serology studies more difficult. Avoiding the use of an arbitrary threshold, as described in this paper, provides a statistically rigorous solution to this problem and facilitates the comparison of results across studies.

The limitations of dichotomizing continuous measurements into positive and negative for statistical analysis are well established in the literature, and include loss of information which affects the ability to reliably recover regression relationships, as well as reducing the the precision of parameter estimates [3–5]. However when the scientific interest is in estimating seroprevalence - as we set out to do in this paper - rather than modeling the dynamics that affect mean antibody antibody levels, dichotomization may be appropriate. This is because the approach presented in this paper results in a more parsimonious model than the unified mechanistic model presented in Kyomuhangi et al. [85], allowing for a more efficient estimation of parameters that only modulate seroprevalence.

Depending on the degree of overlap between the seronegative and seropositive populations in the sample, mixture models can be difficult to estimate. The *Pf*MSP1$_{19}$ analysis illustrates this key limitation. Due to the poor separation of the seronegative and seropositive populations, the estimate for λ shows a large value, which is

inconsistent with both epidemiological data from the study site. This poor separation could be a biological feature of the antibody response to $Pf\text{MSP1}_{19}$, or due to poor dynamic range of the serological assay that generated the data. Similarly, in areas of high transmission where the majority of the population is seropositive [21, 40], or in elimination settings where there are very few seropositive cases, estimating the model parameters may be difficult. In these scenarios, if prior knowledge on some of the components of the model is available, Bayesian methods of inference can be used to alleviate estimation issues though the specification of suitable prior distributions. Additionally, to deal with skewness of the antibody distributions which can still persists after taking the logarithmic transformation, a mixture of skew-Normal distributions can be used in the mixture model to model the left asymmetry of the seropositive population.

When fitting the RCM, the seroreversion rate may also be difficult to estimate, hence ω is usually fixed [20]. In this paper, we considered the simplest form of the RCM, which assumes constant transmission and ignores possible changes in transmission due to, for example interventions in the recent past. While the resulting seroprevalence curves from the RCM do not fit the data very well in Figure 4.7, the majority of seroprevalence points fall within the 95% CIs of the seroprevalence curves. Several studies have proposed modifications which relax this assumption of constant transmission [20, 34, 42, 66], and each of these can be fitted by using the Monte Carlo approach proposed in this paper to propagate the uncertainty in the classification of seropositive individuals.

Conclusion

We have proposed a new threshold-free method for estimating malaria seroprevalence which accounts for age dependency of antibodies through regression, and incorporates uncertainty around the estimates in subsequent analysis of the data. This method is more robust to varying conditions of analysis and provides more consistent estimates than the traditional threshold-based approach.

Ethics approval and consent to participate

Ethical approval for collecting the data included in the paper was granted by the London School of Hygiene and Tropical Medicine (LSHTM-5721) and the Kenya Medical Research Institute (SSC-1802). All methods were performed in accordance with good research practices and written informed consent was obtained from all participants, and, if appropriate, their parents or guardians.

Consent for publication

Not applicable

Availability of data and material

The data set included in this paper is not publicly available but may be requested from Prof Chris Drakeley at The London School of Hygiene and Tropical Medicine. The R script to run both M1 and M2 is available from the authors upon request, and accessible on Github (https://github.com/kyomuhai/Kyomuhangi-and-Giorgi_-thresholdfree). Supplementary material is available as part of this submission.

Competing interests

The authors declare that they have no competing interests.

Funding

IK is a Commonwealth Scholar, whose PhD is funded by the UK government. EG acknowledges support from the Academy of Medical Sciences thorough a Springboard Award (SBF0041009). Both funders had no role in the design of the study, the collection, analysis, and interpretation of data, or in writing the manuscript.

Author's contributions

IK: conceptualization, methodology, formal analysis, data curation, data visualization, writing-original draft.
EG: conceptualization, methodology, formal analysis, supervision, writing-review and editing.

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, as well as Prof. Chris Drakeley's group at LSHTM for sharing the data.

Chapter 5

Future work and conclusions

Each of the three papers presented in this thesis contains a detailed discussion, therefore this chapter focuses on potential implications of the findings on future research, and how these analyses can be expanded.

5.1 Paper 1

In paper 1 we conduct a comparative analysis between geostatistical models which use continuous measurements of disease outcomes (i.e. a linear model), versus those which use dichotomized data (i.e. a binomial model model), and measure their performance in both parameter estimation and prediction of disease risk.

The results, which show how dichotomization leads to loss of information and reduced predictive performance of geostatistical models, are in alignment with similar comparisons in other fields of research. The practice of dichotomizing data in disease mapping has potential ramifications for public health policy decisions. For instance, where intervention strategies are developed based on exceedence probabilities, greater uncertainty around estimates can translate into ineffective public health and intervention policy.

Further comparative analysis on the performance of linear vs binomial geostatistical models could be conducted by looking at different performance measures, for example sensitivity and specificity in detecting hotspots, or considering other categorizations of continuous data (i.e. multiple categories, rather than simply positive/negative).

5.2 Paper 2

In paper 2, we explore malaria serology models, where the prevailing method for analysing serology data depends on dichotomization of continuous antibody measurements. We propose a novel unified mechanistic model which 1) eliminates the need to dichotomize continuous antibody measurements into seropositive and seronegative data, 2) relaxes assumptions about malaria transmission dynamics, 3) adds flexibility in how transmission intensity can be estimated using regression analysis, 4) incorporates age-dependency of the antibody distribution, and 5) allows for joint estimation of malaria transmission intensity from both the reversible catalytic and antibody acquisition models.

While the unified mechanistic model relaxes assumptions on the temporal dynamics of $\lambda(a)$ and $\gamma(a)$, the model still assumes a monotone change in these transmission parameters, which may not capture all the variation these parameters over time. In order to go beyond the assumption of monotonicity, longitudinal data is required. This would likely be more reliable than cross-sectional data which, in this context, can only be analysed under certain assumptions.

Of note, results from an additional analysis of another antigen, $Pf\text{MSP1}_{19}$ showed limited indication of a bi-modal distribution and age-dependency, making estimation of parameters from the unified mechanistic model unfeasible. Application of the model to a wider range of anti-malaria antibodies could clarify which of the many antigens currently under investigation would be more useful than others in estimation of transmission parameters using this model.

Additionally, the results from this analysis indicated that a unified mechanistic model that assumes a time-varying $\lambda(a)$ and a constant γ provides a better fit to the data than a model where the reverse assumptions is made, suggesting that $\lambda(a)$ may be a more suitable measure for surveillance purposes when the focus of analysis is historical changes in transmission.

The unified mechanistic model provides opportunity to estimate both $\lambda(a)$ and $\gamma(a)$, therefore the choice of metric could be informed by the properties of the antibody response and distribution. This is particularly significant as further research emerges about which specific antibodies correlate with protection, and as multiplex immunoassays, which facilitate the analysis of multiple antigens simultaneously, become more common.

Other studies have described the study area as having low but highly heterogeneous transmission [72, 73], and while the heterogeneity of transmission in this area is not assessed in paper 2, the flexibility of the unified mechanistic model allows for

an extension incorporating geostatistical analysis. This would make it suitable for analysis of spatial variation in transmission intensity.

Results from the paper 2 analysis indicate a reduction in $\lambda(a)$ over time, which is consistent with previous findings for this study site [72, 73]. Explanation for why transmission may have increased would require knowledge of additional factors such presence and coverage of interventions, climate, as well as mosquito habitat, characteristics and behaviour, among others. Nevertheless, the advantage of the unified mechanistic model is that it provides flexibility to detect increases or decreases in transmission, unlike current models which are bound by more rigid assumptions on transmission profiles. Serology models alone can not explain why changes have occurred, but rather, they provide important information on how transmission parameters change over time. The unified mechanistic model improves on how we derive this information from serology data.

5.3 Paper 3

In paper 3 we utilize the principles of the empirical model to estimate seroprevalence without thresholds. While the goal of most serology models, including the unified mechanistic model, is to investigate historical changes in transmission, seroprevalence itself gives us a snapshot of malaria exposure at a given time and location.

The threshold-free approach proposed in paper 3 provides three key solutions in seroprevalence estimation: it accounts for the age dependency of malaria antibodies; allows us to propagate the uncertainty around classification of individuals as seropositive and seronegative; and avoids the use of any threshold.

Given the age-dependency evidenced in both papers 2 and 3, additional consideration of the age distribution of sample populations may be needed when designing future studies. For example, many of the study areas included in this thesis have a left skewed age distribution where younger individuals are much more prevalent than older individuals in the sample. Depending on the goal of analysis, this should be taken into account in the study design when recruiting study participants.

An application of this threshold-free approach in malaria research would be mapping of seroprevalence using geostatistical methods, where there is an urgent need to identify hot-spots of disease burden [73] using a variety of malaria indicators, particularly in places of low or highly heterogeneous transmission.

5.4 Conclusion

In this thesis, we set out to explore the impact of thresholds and dichotomization on statistical inference, with the aim of developing methods that improve estimation and prediction of disease prevalence and risk.

The studies presented in this thesis demonstrate the significant impact using thresholds and dichotomization of continuous health outcome data can have on statistical inference. We show how dichotomization leads to loss of information, reduces the reliability of parameter estimates, and reduces the predictive performance of geostatistical models. We also demonstrate how the use of thresholds in malaria serology models reduces the reliability of seroprevalence estimates.

Importantly, we present alternative statistical models which make full use of the continuous health outcome measurements, rather than dichotomize the data. In geostatistical analysis, we demonstrate the advantage of using a linear geostatistical model compared to a binomial geostatistical model. For malaria serology analysis, we develop a unified modelling framework which uses continuous antibody measurements and combines existing models while addressing their current limitations.

Acknowledging that in some cases where dichotomization of data is necessary and appropriate, for instance when the scientific interest is in estimating malaria seroprevalence, we propose a novel threshold-free approach for estimating seroprevalence.

Finally, in this thesis we present models for analysing continuous stunting, anaemia, and malaria serology data, however the statistical principles underpinning these proposed approaches can be applied to a variety of diseases, in order to develop models which use continuous measurements of the disease indicator.

When developing these models, consideration should be given to key questions including the research context, model complexity and computational intensity of the new methods. For example, in cases where the disease outcome is determined by presence or absence of an indicator rather than the scale of the continuous measurement for the indicator, dichotomization of this data into positive and negative would not lead to any loss of information. In this context a model using continuous measurements would not be suitable. Additionally, in the case of the unified mechanistic model proposed in paper 2, the complexity of this model necessitates a large data-set and advanced computational capacity in order to estimate parameters reliably. This reduces the applicability of this method, particularly in resource-limited settings where this analysis is most relevant.

Together, the studies presented in this thesis strongly support the conclusion that, whenever feasible, dichotomization should be avoided by developing models for the continuous measurements which can then be used to estimate model parameters.

Supplementary material

S.1 Paper 1 Supplementary material

S.1.1 Simulation study results for sample size m=450

Table S1: Bias and mean square error (in brackets) for $\tilde{\alpha}$, $\tilde{\sigma}^2$ and the estimate $\hat{\phi}$ obtained from the geostatistical models fitted the binary (B) and continuous (C) outcomes. The following are results when the number of observations, $n = 450$

			c=0		c=0.2		c=0.4	
Parameter	τ^2	ϕ	B	C	B	C	B	C
$\tilde{\alpha}$	0.5	0.1	-0.013 (0.090)	-0.014 (0.056)	0.089 (0.216)	0.017 (0.068)	0.173 (0.180)	0.045 (0.085)
	1	0.1	-0.007 (0.06)	0.001 (0.032)	0.081 (0.325)	0.019 (0.039)	0.159 (0.054)	0.034 (0.040)
	2	0.1	-0.009 (0.031)	-0.004 (0.017)	0.073 (0.448)	0.021 (0.019)	0.120 (0.128)	0.027 (0.022)
	0.5	0.2	-0.030 (0.292)	-0.029 (0.151)	0.100 (0.382)	0.013 (0.151)	0.228 (0.417)	0.023 (0.162)
	1	0.2	-0.014 (0.169)	-0.012 (0.077)	0.093 (0.401)	0.008 (0.071)	0.175 (0.167)	0.009 (0.079)
	2	0.2	-0.006 (0.093)	-0.001 (0.040)	0.079 (0.497)	0.011 (0.041)	0.156 (0.169)	0.015 (0.040)
$\tilde{\sigma}^2$	0.5	0.1	0.603 (0.554)	0.735 (4.688)	0.624 (0.579)	0.747 (4.983)	0.570 (0.508)	0.780 (4.932)
	1	0.1	0.289 (0.137)	0.543 (3.119)	0.297 (0.136)	0.679 (3.828)	0.274 (0.122)	0.508 (3.276)
	2	0.1	0.125 (0.029)	0.417 (2.781)	0.112 (0.027)	0.462 (2.806)	0.107 (0.026)	0.446 (2.966)
	0.5	0.2	1.096 (2.059)	0.289 (1.197)	1.120 (2.086)	0.387 (1.408)	1.078 (2.015)	0.230 (1.055)
	1	0.2	0.566 (0.548)	0.131 (0.334)	0.565 (0.559)	0.133 (0.267)	0.533 (0.498)	0.105 (0.218)
	2	0.2	0.252 (0.126)	0.100 (0.519)	0.243 (0.119)	0.116 (0.471)	0.228 (0.114)	0.075 (0.157)
ϕ	0.5	0.1	0.031 (0.002)	0.001 (0.001)	0.033 (0.002)	0.002 (0.001)	0.031 (0.002)	0.001 (0.001)
	1	0.1	0.031 (0.002)	0.001 (0.002)	0.032 (0.002)	-0.002 (0.001)	0.03 (0.002)	0.001 (0.002)
	2	0.1	0.028 (0.002)	0.002 (0.002)	0.026 (0.002)	0.001 (0.002)	0.024 (0.002)	0.001 (0.003)
	0.5	0.2	0.088 (0.018)	-0.017 (0.004)	0.087 (0.017)	-0.017 (0.005)	0.088 (0.018)	-0.011 (0.005)
	1	0.2	0.095 (0.021)	-0.012 (0.007)	0.093 (0.019)	-0.014 (0.006)	0.090 (0.018)	-0.014 (0.006)
	2	0.2	0.088 (0.020)	-0.011 (0.010)	0.083 (0.018)	-0.015 (0.009)	0.078 (0.017)	-0.016 (0.009)

Table S2: Bias and mean square error (in brackets), averaged over a 1/14 by 1/14 regular grid in $[0, 2] \times [0, 1]$ (hence, $m = 450$), for the spatial predictions of prevalence obtained from the geostatistical models fitted to the binary (B) and continuous (C) outcomes.

		c=0		c=0.2		c=0.4	
τ^2	ϕ	B	C	B	C	B	C
0.5	0.1	0.001 (0.053)	0.001 (0.038)	0.013 (0.053)	0.001 (0.037)	0.029 (0.051)	0.001 (0.036)
1	0.1	-0.001 (0.047)	0.001 (0.037)	0.019 (0.047)	-0.001 (0.037)	0.034 (0.046)	-0.001 (0.035)
2	0.1	-0.002 (0.037)	-0.001 (0.032)	0.020 (0.037)	0.001 (0.032)	0.037 (0.036)	-0.001 (0.031)
0.5	0.2	-0.001 (0.041)	0.001 (0.029)	0.012 (0.041)	0.001 (0.029)	0.026 (0.039)	-0.001 (0.027)
1	0.2	0.001 (0.036)	0.001 (0.027)	0.017 (0.036)	-0.001 (0.027)	0.032 (0.035)	-0.001 (0.025)
2	0.200	-0.001 (0.029)	0.001 (0.022)	0.020 (0.029)	-0.001 (0.022)	0.037 (0.029)	-0.001 (0.021)

S.2 Paper 2 Supplementary material

S.2.1 Additional illustration of the mechanisms underlying the unified mechanistic model

Contribution of individuals of different ages to the estimation of transmission parameters λ and γ

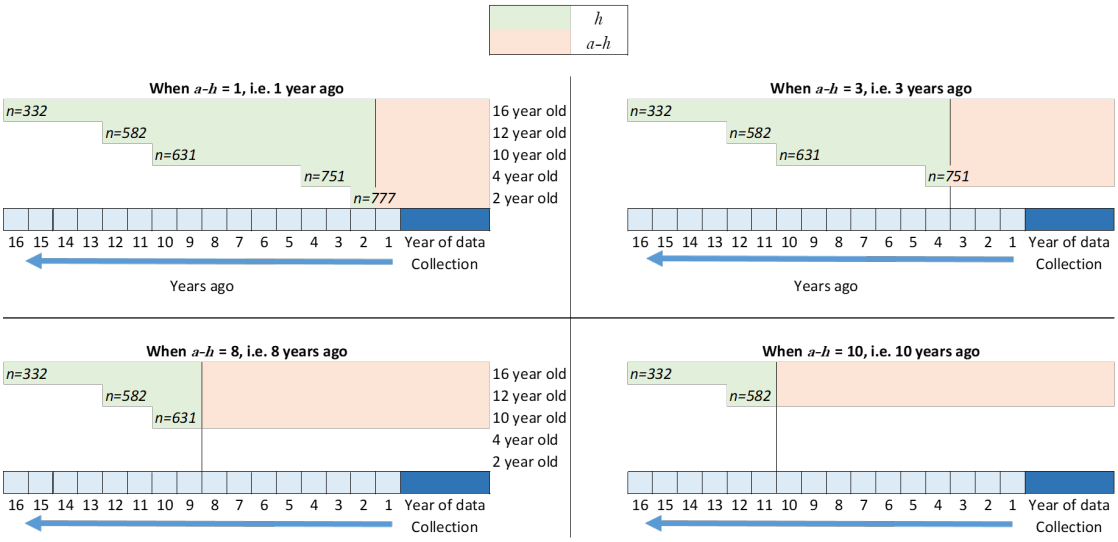

Figure S1: An illustration of how individuals of different ages contribute to estimation of λ and γ for *Pf*AMA1 through historical time. Data is taken from section 3.4. $a-h$ and h, as defined by equation 3.11, represent X-years ago, and the age of the individual X-years ago, respectively. For example in the top right panel, all individuals above 1 year will contribute to the estimation of γ one year ago, however in the bottom right panel, only individuals above 10 years will contribute to the estimation of γ 10 years ago. Note that individuals who contribute to the estimation of γ do so equally, regardless of how old they were at the time, i.e. regardless of the value of h. Also note that the further back in time we estimate γ, the fewer the number individuals, n, contribute to the estimate.

Model formulations for the unified mechanistic model

Time discretization of the RCM

Let $p(a)$ be the proportion of seropositive (S^+) individuals at age a. Given that individuals seroconvert from seronegative S^- to S^+ at rate $\lambda(a)$, and serorevert from S^+ to S^- at rate ω, the standard expression of the temporal dynamics in the RCM:

$$\frac{dp}{da} = \lambda(a)(1-p(a)) - \omega p(a)$$

We then approximate this as

$$\frac{dp}{da} \approx p(a) - p(a-1)$$

Therefore,

$$\begin{aligned}
p(a) - p(a-1) &= \lambda(a)(1-p(a)) - \omega(a) \\
p(a) - p(a-1) &= \lambda(a) - (\lambda(a)p(a)) - \omega p(a) \\
p(a) + (\lambda(a)p(a)) + \omega p(a) &= \lambda(a) + p(a-1) \\
p(a)(1+\omega+\lambda(a)) &= \lambda(a) + p(a-1) \\
p(a) &= \frac{1}{1+\omega+\lambda(a)}(\lambda(a) + p(a-1))
\end{aligned}$$

Assuming $\lambda(0) = 0$,

it follows that $p(0) = 0$

It then follows that,

$$\begin{aligned}
p(1) &= \frac{1}{1+\omega+\lambda_1}(\lambda_1 + 0) \\
&= \frac{\lambda_1}{1+\omega+\lambda_1} \\
p(2) &= \frac{1}{1+\omega+\lambda_2}\Big(\lambda_2 + \frac{\lambda_1}{1+\omega+\lambda_1)}\Big) \\
p(3) &= \frac{1}{1+\omega+\lambda_3}\left(\lambda_3 + \frac{1}{1+\omega+\lambda_2}\left(\lambda_2 + \frac{\lambda_1}{1+\omega+\lambda_1}\right)\right)
\end{aligned}$$

And more generally,

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

Time discretization of the AAM

Let $\mu(a)$ be geometric mean antibody level of individuals at age a. Assuming anti-malaria antibodies of individuals are boosted at rate $\gamma(a)$ upon exposure, and decay at rate r in the absence of exposure, the standard expression of the temporal dynamics in the AAM:

$$\frac{d\mu}{da} = \gamma(a) - r\mu_a$$

We then apply the approximation

$$\frac{d\mu}{da} \approx \mu_a - \mu_{a-1}$$

which leads to

$$\begin{aligned} \mu(a) - \mu_{a-1} &= \gamma(a) - r\mu_a \\ \mu_a &= \frac{1}{1+r}\Big(\gamma(a) + \mu_{a-1}\Big) \end{aligned}$$

Assuming $\gamma(0) = 0$, we have $\mu(0) = 0$

It then follows that,

$$\begin{aligned} \mu(1) &= \frac{1}{1+r}\gamma \\ \mu(2) &= \frac{1}{1+r}\left(\gamma + (\frac{1}{1+r}\gamma_1)\right) \\ &= \frac{1}{1+r}\gamma_2 + \left(\frac{1}{1+r}\right)^2\gamma_1 \\ \mu(3) &= \frac{1}{1+r}\left(\gamma_3 + \frac{1}{1+r}\gamma_2 + \left(\frac{1}{1+r}\right)^2\gamma_1\right) \\ &= \frac{1}{1+r}\gamma_3 + \left(\frac{1}{1+r}\right)^2\gamma_2 + \left(\frac{1}{1+r}\right)^3\gamma_1\Big) \end{aligned}$$

And more generally,

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

S.2.2 Implementation of the unified mechanistic model in Section 3.4

Akaike Information Criterion (AIC) comparisons for the implementation of the unified mechanistic model

Table S3: Preliminary analysis of Western Kenya data, comparing the AIC for the empirical model (EM) and unified mechanistic models (UFM) with time-varying λ & constant γ, constant λ & time-varying γ, and different values of ω.

Model	ω	AIC
EM	–	29711.460
UFM, constant λ, time-varying γ	Continuous	30166.680
UFM, time-varying λ, constant γ	Continuous	29801.920
	0.01	29791.910
	0.5	29800.680
	1	29799.920

S.3 Paper 3 Supplementary material

S.3.1 Profile likelihood for different values of ω

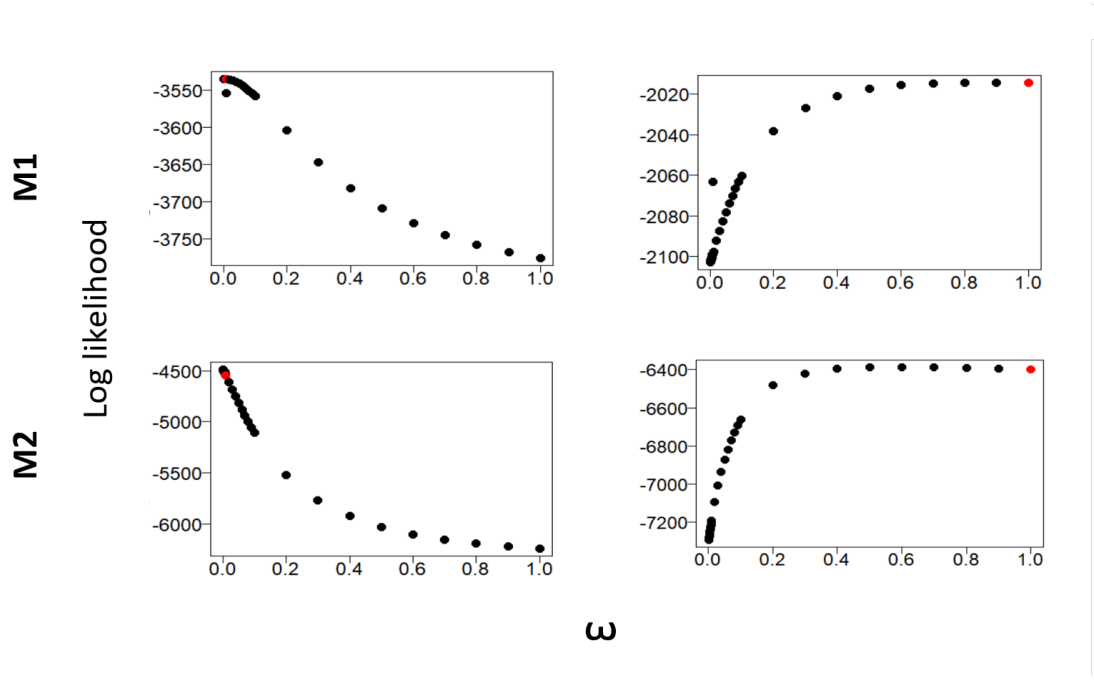

Figure S2: Profile likelihood analysis for different values of ω in the *Pf*AMA1 and *Pf*MSP1$_{19}$ analyses

S.3.2 Logit-transformed prevalence estimates from M2

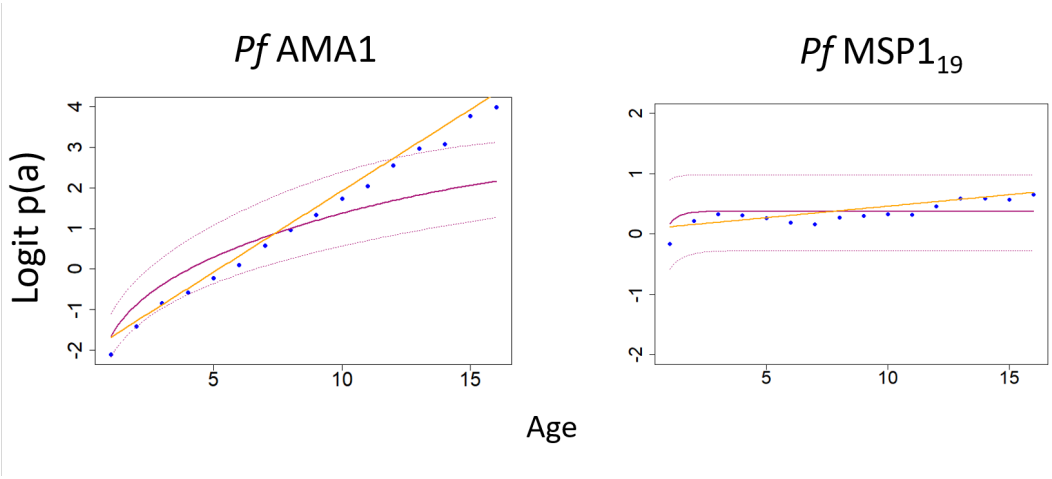

Figure S3: Logit-transformed prevalence estimates from M2. The mean of the seroprevalence distribution is indicated by blue dots; the purple solid and dotted curves represent the fitted seroprevalence and 95% CIs, respectively, from the RCM; and the orange line indicates the fitted seroprevalence estimate from the age-dependent mixture model, as defined by equation (10)

S.3.3 M1 and M2 analysis using data from different age groups

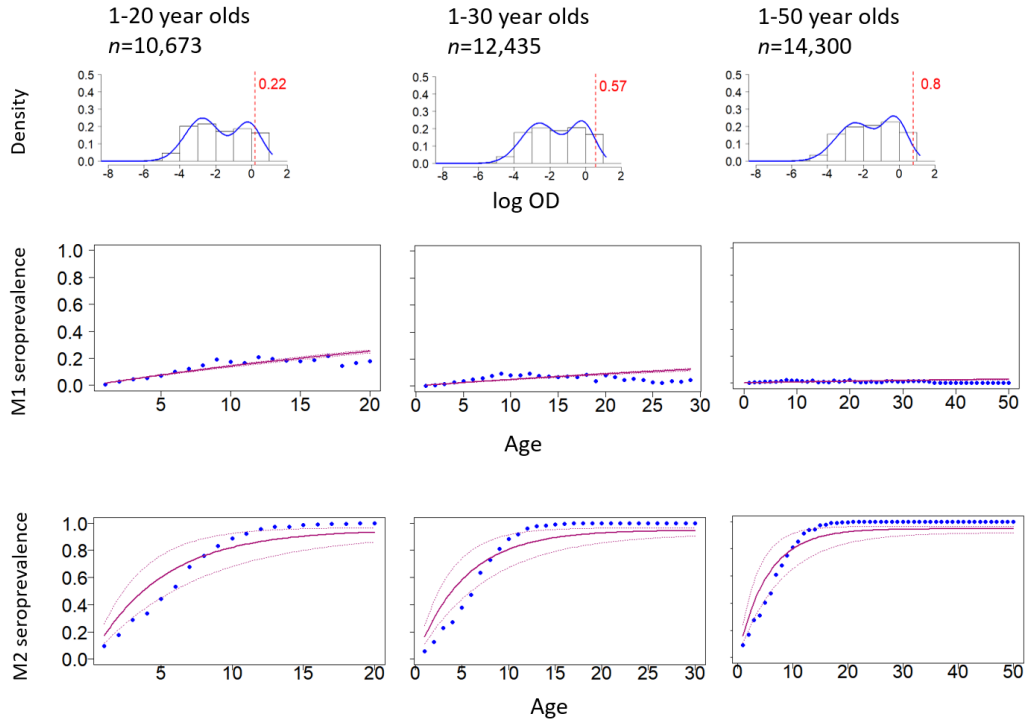

Figure S4: Analysis of *Pf*AMA1 mixture distributions using different age-groups in both M1 and M2. The mixture distributions in the top row are derived from M1 (see equation (1)), and show the seropositivity thresholds (red dotted lines represent $\mu_{S^-} + 3\sigma_{S^-}$) when different age groups are used in analysis. The middle row shows M1 seroprevalence point estimates (blue dots), as well as the fitted seroprevalence curve (purple curve) and 95% CIs (purple dotted curves) from the RCM. The bottom row shows the mean of the seroprevalence distribution derived from M2 (blue dots), as well as the fitted seroprevalence curve (purple curve) and 95% CIs (purple dotted curves) from the RCM.

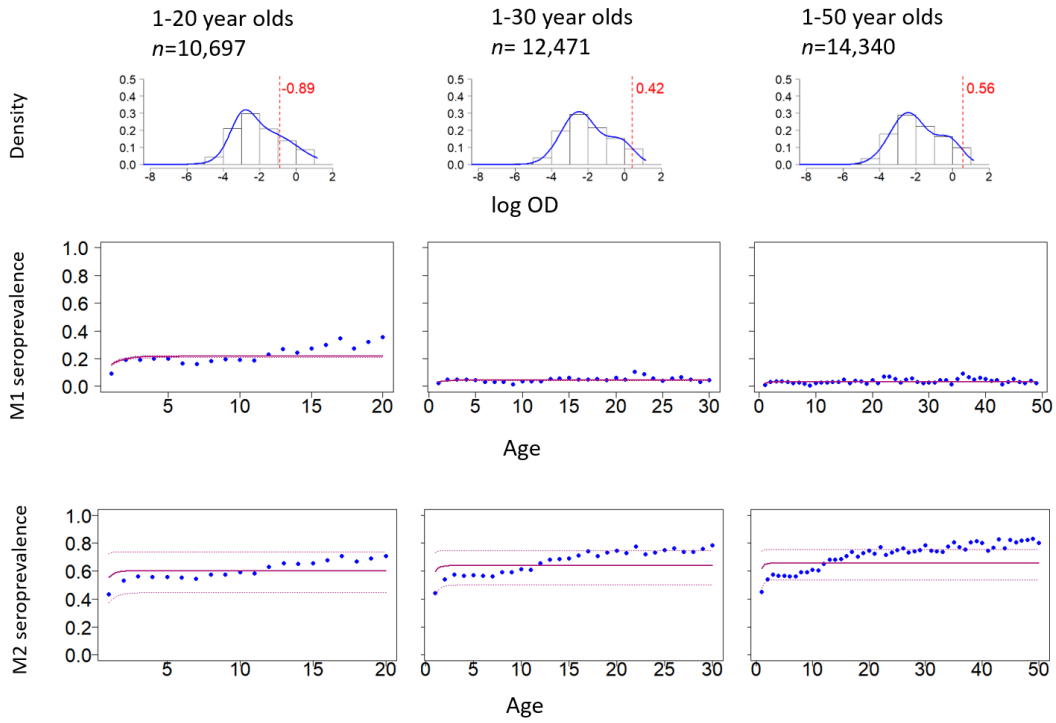

Figure S5: Analysis of *Pf*MSP1$_{19}$ mixture distributions using different age-groups in both M1 and M2. The mixture distributions in the top row are derived from M1 (see equation (1)), and show the seropositivity thresholds (red dotted lines represent $\mu_{S^-} + 3\sigma_{S^-}$ thresholds) when different age groups are used in analysis. The middle row shows M1 seroprevalence estimates (blue dots), as well as the fitted seroprevalence curve (purple curve) and 95% CIs (purple dotted curves) from the RCM. The bottom row shows the mean of the seroprevalence distribution derived from M2 (blue dots), as well as the fitted seroprevalence curve (purple curve) and 95% CIs (purple dotted curves) from the RCM.

S.3.4 M1 analysis using different seropositivity thresholds for individuals aged 1-16 years

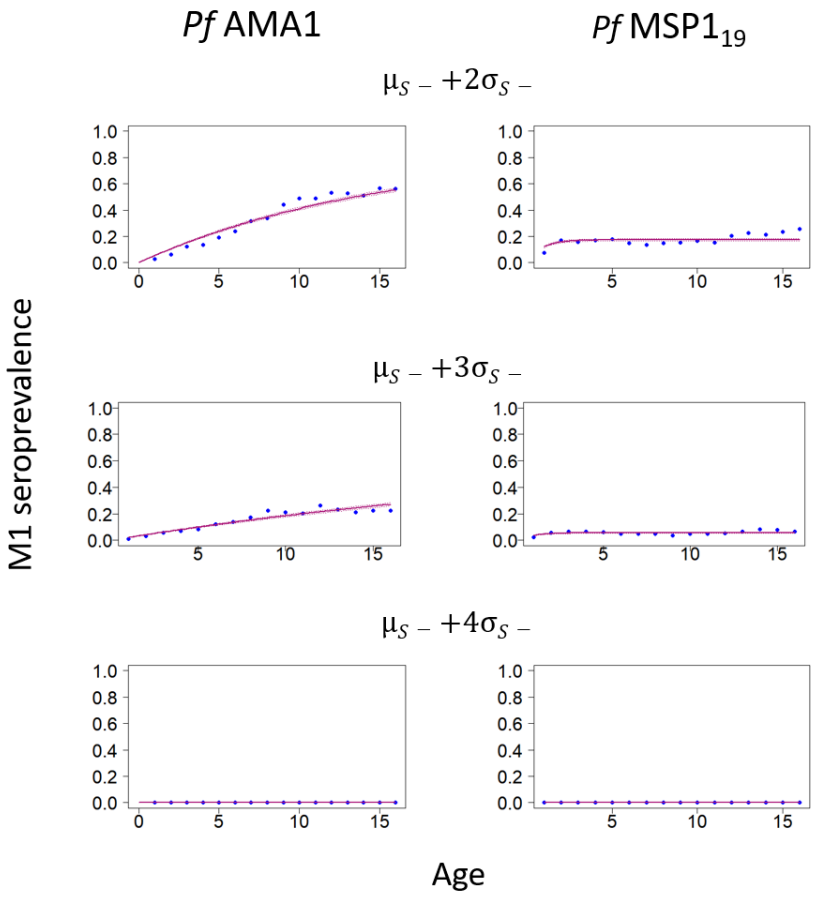

Figure S6: The top row shows M1 seroprevalence point estimates (blue dots) where seropositivity is defined as $\mu_{S^-} + 2\sigma_{S^-}$, as well as the fitted seroprevalence curve (purple curve) and 95% CIs (purple dotted curves) from the RCM. The middle row shows where seropositivity is defined as $\mu_{S^-} + 3\sigma_{S^-}$, and the bottom row shows where seropositivity is defined as $\mu_{S^-} + 4\sigma_{S^-}$

Bibliography

[1] Peter J Diggle, JA Tawn, and RA Moyeed. "Model-based geostatistics". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 47.3 (1998), pp. 299–350.

[2] Peter J Diggle and Emanuele Giorgi. *Model-based geostatistics for global public health: methods and applications*. CRC Press, 2019.

[3] Valerii Fedorov, Frank Mannino, and Rongmei Zhang. "Consequences of dichotomization". In: *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry* 8.1 (2009), pp. 50–61.

[4] Douglas G Altman and Patrick Royston. "The cost of dichotomising continuous variables". In: *Bmj* 332.7549 (2006), p. 1080.

[5] Patrick Royston, Douglas G Altman, and Willi Sauerbrei. "Dichotomizing continuous predictors in multiple regression: a bad idea". In: *Statistics in medicine* 25.1 (2006), pp. 127–141.

[6] Caroline Bennette and Andrew Vickers. "Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents". In: *BMC medical research methodology* 12.1 (2012), p. 21.

[7] Frank Harrell. *Problems caused by categorising continuous variables*. http://biostat.mc.vanderbilt.edu/wiki/Main/CatContinuous. Accessed: 2019-8-8. June 2004.

[8] Petra Buettner, Claus Garbe, and Irene Guggenmoos-Holzmann. "Problems in defining cutoff points of continuous prognostic factors: example of tumor thickness in primary cutaneous melanoma". In: *Journal of clinical epidemiology* 50.11 (1997), pp. 1201–1210.

[9] Robert C MacCallum, Shaobo Zhang, Kristopher J Preacher, and Derek D Rucker. "On the practice of dichotomization of quantitative variables." In: *Psychological methods* 7.1 (2002), p. 19.

[10] Peter C Austin and Lawrence J Brunner. "Inflation of the type I error rate when a continuous confounding variable is categorized in logistic regression analyses". In: *Statistics in medicine* 23.7 (2004), pp. 1159–1178.

[11] World Health Organization et al. "World malaria report 2019". In: (2019).

[12] Beatrice Autino, Alice Noris, Rosario Russo, and Francesco Castelli. "Epidemiology of malaria in endemic areas". In: *Mediterranean journal of hematology and infectious diseases* 4.1 (2012).

[13] *CDC Malaria Life Cycle*. https://www.cdc.gov/malaria/about/biology/index.html. Accessed: 2020-11-15. July 2020.

[14] Denise L Doolan, Carlota Dobaño, and J Kevin Baird. "Acquired immunity to malaria". In: *Clinical microbiology reviews* 22.1 (2009), pp. 13–36.

[15] Andrew Teo, Gaoqian Feng, Graham V Brown, James G Beeson, and Stephen J Rogerson. "Functional antibodies and protection against blood-stage malaria". In: *Trends in parasitology* 32.11 (2016), pp. 887–898.

[16] Carole A Long and Fidel Zavala. "Immune responses in malaria". In: *Cold Spring Harbor perspectives in medicine* 7.8 (2017), a025577.

[17] Ian A Cockburn and Robert A Seder. "Malaria prevention: from immunological concepts to effective vaccines and protective antibodies". In: *Nature immunology* 19.11 (2018), pp. 1199–1211.

[18] OJ Akpogheneta, S Dunyo, M Pinder, and DJ Conway. "Boosting antibody responses to Plasmodium falciparum merozoite antigens in children with highly seasonal exposure to infection". In: *Parasite immunology* 32.4 (2010), pp. 296–304.

[19] Danielle I Stanisic, Freya JI Fowkes, Melanie Koinari, Sarah Javati, Enmoore Lin, Benson Kiniboro, Jack S Richards, Leanne J Robinson, Louis Schofield, James W Kazura, et al. "Acquisition of antibodies against Plasmodium falciparum merozoites and malaria immunity in young children and the influence of age, force of infection, and magnitude of response". In: *Infection and immunity* 83.2 (2015), pp. 646–660.

[20] Nuno Sepúlveda, Gillian Stresman, Michael T White, and Chris J Drakeley. "Current mathematical models for analyzing anti-malarial antibody data with an eye to malaria elimination and eradication". In: *Journal of immunology research* 2015 (2015).

[21] Emilie Pothin, Neil M Ferguson, Chris J Drakeley, and Azra C Ghani. "Estimating malaria transmission intensity from Plasmodium falciparum serological data using antibody density models". In: *Malaria journal* 15.1 (2016), p. 79.

[22] Maxwell Kilama, David L Smith, Robert Hutchinson, Ruth Kigozi, Adoke Yeka, Geoff Lavoy, Moses R Kamya, Sarah G Staedke, Martin J Donnelly, Chris Drakeley, et al. "Estimating the annual entomological inoculation rate for Plasmodium falciparum transmitted by Anopheles gambiae sl using three sampling methods in three sites in Uganda". In: *Malaria journal* 13.1 (2014), p. 111.

[23] T Smith, G Killeen, C Lengeler, and M Tanner. "Relationships between the outcome of Plasmodium falciparum infection and the intensity of transmission in Africa". In: *The American journal of tropical medicine and hygiene* 71.2_suppl (2004), pp. 80–86.

[24] Lucy S Tusting, Teun Bousema, David L Smith, and Chris Drakeley. "Measuring changes in Plasmodium falciparum transmission: precision, accuracy and costs of metrics". In: *Advances in parasitology*. Vol. 84. Elsevier, 2014, pp. 151–208.

[25] Patrick Corran, Paul Coleman, Eleanor Riley, and Chris Drakeley. "Serology: a robust indicator of malaria transmission intensity?" In: *Trends in parasitology* 23.12 (2007), pp. 575–582.

[26] CJ Drakeley, PH Corran, PG Coleman, JE Tongren, SLR McDonald, I Carneiro, R Malima, J Lusingu, A Manjurano, WMM Nkya, et al. "Estimating medium-and long-term trends in malaria transmission by using serological markers of malaria exposure". In: *Proceedings of the National Academy of Sciences* 102.14 (2005), pp. 5108–5113.

[27] Laveta Stewart, Roly Gosling, Jamie Griffin, Samwel Gesase, Joseph Campo, Ramadan Hashim, Paul Masika, Jacklin Mosha, Teun Bousema, Seif Shekalaghe, et al. "Rapid assessment of malaria transmission using age-specific seroconversion rates". In: *PloS one* 4.6 (2009), e6083.

[28] Teun Bousema, Randa M Youssef, Jackie Cook, Jonathan Cox, Victor A Alegana, Jamal Amran, Abdisalan M Noor, Robert W Snow, and Chris Drakeley. "Serologic markers for detecting malaria in areas of low endemicity, Somalia, 2008". In: *Emerging infectious diseases* 16.3 (2010), p. 392.

[29] Maristela G Cunha, Eliane S Silva, Nuno Sepúlveda, Sheyla PT Costa, Tiago C Saboia, João F Guerreiro, Marinete M Póvoa, Patrick H Corran, Eleanor Riley, and Chris J Drakeley. "Serologically defined variations in malaria endemicity in Pará state, Brazil". In: *PloS one* 9.11 (2014), e113357.

[30] Grace E Weber, Michael T White, Anna Babakhanyan, Peter Odada Sumba, John Vulule, Dylan Ely, Chandy John, Evelina Angov, David Lanar, Sheetij Dutta, et al. "Sero-catalytic and antibody acquisition models to estimate differing malaria transmission intensities in Western Kenya". In: *Scientific reports* 7.1 (2017), p. 16821.

[31] Michael T White, Jamie T Griffin, Onome Akpogheneta, David J Conway, Kwadwo A Koram, Eleanor M Riley, and Azra C Ghani. "Dynamics of the

antibody response to Plasmodium falciparum infection in African children". In: *The Journal of infectious diseases* 210.7 (2014), pp. 1115–1122.

[32] Samuel Bosomprah. "A mathematical model of seropositivity to malaria antigen, allowing seropositivity to be prolonged by exposure". In: *Malaria journal* 13.1 (2014), p. 12.

[33] Tamaki Kobayashi, Aarti Jain, Li Liang, Joshua M Obiero, Harry Hamapumbu, Jennifer C Stevenson, Philip E Thuma, James Lupiya, Mike Chaponda, Modest Mulenga, et al. "Distinct antibody signatures associated with different malaria transmission intensities in Zambia and Zimbabwe". In: *Msphere* 4.2 (2019).

[34] Jackie Cook, Immo Kleinschmidt, Christopher Schwabe, Gloria Nseng, Teun Bousema, Patrick H Corran, Eleanor M Riley, and Chris J Drakeley. "Serological markers suggest heterogeneity of effectiveness of malaria control interventions on Bioko Island, equatorial Guinea". In: *PloS one* 6.9 (2011), e25137.

[35] Ryan A Simmons, Leonard Mboera, Marie Lynn Miranda, Alison Morris, Gillian Stresman, Elizabeth L Turner, Randall Kramer, Chris Drakeley, and Wendy P O'Meara. "A longitudinal cohort study of malaria exposure and changing serostatus in a malaria endemic area of rural Tanzania". In: *Malaria journal* 16.1 (2017), pp. 1–13.

[36] Michael E von Fricken, Thomas A Weppelmann, Brandon Lam, Will T Eaton, Laura Schick, Roseline Masse, Madsen V Beau De Rochars, Alexandre Existe, Joseph Larkin, and Bernard A Okech. "Age-specific malaria seroprevalence rates: a cross-sectional analysis of malaria transmission in the Ouest and Sud-Est departments of Haiti". In: *Malaria journal* 13.1 (2014), p. 361.

[37] Joseph Okebe, Muna Affara, Simon Correa, Abdul Khalie Muhammad, Davis Nwakanma, Chris Drakeley, and Umberto D'Alessandro. "School-based countrywide seroprevalence survey reveals spatial heterogeneity in malaria transmission in the Gambia". In: *PloS one* 9.10 (2014), e110926.

[38] Michelle K Muthui, Alice Kamau, Teun Bousema, Andrew M Blagborough, Philip Bejon, and Melissa C Kapulu. "Immune Responses to Gametocyte Antigens in a Malaria Endemic Population—The African falciparum Context: A Systematic Review and Meta-Analysis". In: *Frontiers in immunology* 10 (2019), p. 2480.

[39] Eric Rogier, Ryan Wiegand, Delynn Moss, Jeff Priest, Evelina Angov, Sheetij Dutta, Ito Journel, Samuel E Jean, Kimberly Mace, Michelle Chang, et al. "Multiple comparisons analysis of serological data from an area of low Plasmodium falciparum transmission". In: *Malaria journal* 14.1 (2015), pp. 1–12.

[40] Victor Yman, Michael T White, Josea Rono, Bruno Arcà, Faith H Osier, Marita Troye-Blomberg, Stéphanie Boström, Raffaele Ronca, Ingegerd Rooth, and Anna Färnert. "Antibody acquisition models: a new tool for serological surveillance of malaria transmission intensity". In: *Scientific reports* 6 (2016), p. 19472.

[41] Rajika Lasanthi Dewasurendra, Janaka Nandana Dias, Nuno Sepulveda, Geethika Sharmini Abayaweera Gunawardena, Naduviladath Chandrasekharan, Chris Drakeley, and Nadira Dharshani Karunaweera. "Effectiveness of a serological tool to predict malaria transmission intensity in an elimination setting". In: *BMC infectious diseases* 17.1 (2017), p. 49.

[42] Marie-Louise Varela, David Koffi, Michael White, Makhtar Niang, Babacar Mbengue, Fatoumata Diene Sarr, André Offianan Touré, and Ronald Perraut. "Practical example of multiple antibody screening for evaluation of malaria control strategies". In: *Malaria journal* 19.1 (2020), pp. 1–12.

[43] David Coggon, David Barker, and Geoffrey Rose. *Epidemiology for the Uninitiated*. John Wiley & Sons, 2009.

[44] World Health Organisation. *Haemoglobin concentrations for the diagnosis of anaemia and assessment of severity*. Tech. rep. World Health Organisation, 2011.

[45] Giuseppe Del Priore, Peyman Zandieh, and Men-Jean Lee. "Treatment of continuous data as categoric variables in obstetrics and gynecology". In: *Obstetrics & Gynecology* 89.3 (1997), pp. 351–354.

[46] Peter J Diggle and Emanuele Giorgi. "Model-based geostatistics for prevalence mapping in low-resource settings". In: *Journal of the American Statistical Association* 111.515 (2016), pp. 1096–1120.

[47] Sandra Incardona, Elisa Serra-Casas, Nora Champouillon, Christian Nsanzabana, Jane Cunningham, and Iveth J González. "Global survey of malaria rapid diagnostic test (RDT) sales, procurement and lot verification practices: assessing the use of the WHO–FIND Malaria RDT Evaluation Programme (2011–2014)". In: *Malaria journal* 16.1 (2017), p. 196.

[48] Jinkou Zhao, Marcel Lama, Eline Korenromp, Patrick Aylward, Estifanos Shargie, Scott Filler, Ryuichi Komatsu, and Rifat Atun. "Adoption of rapid diagnostic tests for the diagnosis of malaria, a preliminary analysis of the Global Fund program data, 2005 to 2010". In: *PloS one* 7.8 (2012), e43549.

[49] GCM Kalla, E Voundi Voundi, R Guiadem, F Angwafor Iii, L Bélec, and F-X Mbopi-Keou. "Mass campaigns for HIV, HBV (HBsAg) and HCV screening by multiplex rapid diagnostic test in sub-Saharan Africa using mobile units: the game changer". In: *International Journal of Infectious Diseases* 79 (2019), p. 107.

[50] World Health Organization. "Consolidated Guidelines on HIV Testing Services: 5Cs: consent, confidentiality, counselling, correct results and connection 2015". In: (2015).

[51] Aaron Zimmerman, Anoushka I. Millear, Rebecca W. Stubbs, Chloe Shields, Brandon Pickering, Lucas Earl, Nicholas Graetz, Damaris K. Kinyoki, Sarah E. Ray, Samir Bhatt, et al. "Mapping child growth failure in Africa between 2000 and 2015". In: *Nature* 555 (Mar. 2018), pp. 41–47. DOI: 10.1038/nature25760.

[52] Ricardo J Soares Magalhaes and Archie CA Clements. "Mapping the risk of anaemia in preschool-age children: the contribution of malnutrition, malaria, and helminth infections in West Africa". In: *PLoS medicine* 8.6 (2011), e1000438.

[53] Emanuele Giorgi and Peter J Diggle. "PrevMap: an R package for prevalence mapping". In: *J Stat Softw* 78.8 (2017), pp. 1–29.

[54] Cristiano Varin, Nancy Reid, and David Firth. "An overview of composite likelihood methods". In: *Statist. Sinica* (2011), pp. 5–42.

[55] Tarekegn A Abeku, Michelle EH Helinski, Matthew J Kirby, Takele Kefyalew, Tessema Awano, Esey Batisso, Gezahegn Tesfaye, James Ssekitooleko, Sarala Nicholas, Laura Erdmanis, et al. "Monitoring changes in malaria epidemiology and effectiveness of interventions in Ethiopia and Uganda: Beyond Garki Project baseline survey". In: *Malaria journal* 14.1 (2015), p. 337.

[56] World Health Organisation. *Global nutrition targets 2025: Policy brief series*. Tech. rep. World Health Organisation, 2014.

[57] *Infertility definitions and terminology*. https://www.who.int/teams/sexual-and-reproductive-health-and-research/key-areas-of-work/fertility-care/infertility-definitions-and-terminology. Accessed: 2019-8-5.

[58] World Health Organization. "WHO child growth standards: length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age: methods and development". In: (2006).

[59] WHO Multicentre Growth Reference Study Group. "WHO Child Growth Standards based on length/height, weight and age." In: *Acta paediatrica (Oslo, Norway: 1992). Supplement* 450 (2006), p. 76.

[60] World Health Organisation. *Global nutrition targets 2025: Stunting policy brief*. Tech. rep. World Health Organisation, 2014.

[61] World Health Organization. *Nutrition Landscape Information System (NLIS) country profile indicators: interpretation guide*. Tech. rep. World Health Organisation, 2010.

[62] *The Malaria Atlas Project*. https://malariaatlas.org/. Accessed: 2019-12-11.

[63] Dominic P Kwiatkowski. "How malaria has affected the human genome and what human genetics can teach us about malaria". In: *The American Journal of Human Genetics* 77.2 (2005), pp. 171–192.

[64] K Bollaerts, M Aerts, Z Shkedy, C Faes, Y Van der Stede, Ph Beutels, and Niel Hens. "Estimating the population prevalence and force of infection directly from antibody titres". In: *Statistical Modelling* 12.5 (2012), pp. 441–462.

[65] Niel Hens, Ziv Shkedy, Marc Aerts, Christel Faes, Pierre Van Damme, and Philippe Beutels. *Modeling infectious disease parameters based on serological and social contact data: a modern statistical perspective*. Vol. 63. Springer Science & Business Media, 2012.

[66] Jackie Cook, Heidi Reid, Jennifer Iavro, Melissa Kuwahata, George Taleo, Archie Clements, James McCarthy, Andrew Vallely, and Chris Drakeley. "Using serological measures to monitor changes in malaria transmission in Vanuatu". In: *Malaria journal* 9.1 (2010), p. 169.

[67] Emanuele Del Fava, Grazina Rimseliene, Elmira Flem, Birgitte Freiesleben De Blasio, Gianpaolo Scalia Tomba, and Piero Manfredi. "Estimating age-specific immunity and force of infection of varicella zoster virus in Norway using mixture models". In: *PloS one* 11.9 (2016).

[68] Benjamin F Arnold, Mark J van der Laan, Alan E Hubbard, Cathy Steel, Joseph Kubofcik, Katy L Hamlin, Delynn M Moss, Thomas B Nutman, Jeffrey W Priest, and Patrick J Lammie. "Measuring changes in transmission of neglected tropical diseases, malaria, and enteric pathogens from quantitative antibody levels". In: *PLoS neglected tropical diseases* 11.5 (2017), e0005616.

[69] Ruth A Ashton, Takele Kefyalew, Alison Rand, Heven Sime, Ashenafi Assefa, Addis Mekasha, Wasihun Edosa, Gezahegn Tesfaye, Jorge Cano, Hiwot Teka, et al. "Geostatistical modeling of malaria endemicity using serological indicators of exposure collected through school surveys". In: *The American journal of tropical medicine and hygiene* 93.1 (2015), pp. 168–177.

[70] Onome J Akpogheneta, Nancy O Duah, Kevin KA Tetteh, Samuel Dunyo, David E Lanar, Margaret Pinder, and David J Conway. "Duration of naturally acquired antibody responses to blood-stage Plasmodium falciparum is age dependent and antigen specific". In: *Infection and immunity* 76.4 (2008), pp. 1748–1755.

[71] Ann M Moormann. "How might infant and paediatric immune responses influence malaria vaccine efficacy?" In: *Parasite immunology* 31.9 (2009), pp. 547–559.

[72] Teun Bousema, Jennifer Stevenson, Amrish Baidjoe, Gillian Stresman, Jamie T Griffin, Immo Kleinschmidt, Edmond J Remarque, John Vulule, Nabie Bayoh, Kayla Laserson, et al. "The impact of hotspot-targeted interventions on malaria transmission: study protocol for a cluster-randomized controlled trial". In: *Trials* 14.1 (2013), pp. 1–12.

[73] Gillian H Stresman, Emanuele Giorgi, Amrish Baidjoe, Phil Knight, Wycliffe Odongo, Chrispin Owaga, Shehu Shagari, Euniah Makori, Jennifer Stevenson, Chris Drakeley, et al. "Impact of metric and sample size on determining malaria hotspot boundaries". In: *Scientific reports* 7 (2017), p. 45849.

[74] Erin M Stuckey, Jennifer C Stevenson, Mary K Cooke, Chrispin Owaga, Elizabeth Marube, George Oando, Diggory Hardy, Chris Drakeley, Thomas A Smith, Jonathan Cox, et al. "Simulation of malaria epidemiology and control in the highlands of western Kenya". In: *Malaria journal* 11.1 (2012), p. 357.

[75] Teun Bousema, Gillian Stresman, Amrish Y Baidjoe, John Bradley, Philip Knight, William Stone, Victor Osoti, Euniah Makori, Chrispin Owaga, Wycliffe Odongo, et al. "The impact of hotspot-targeted interventions on malaria transmission in Rachuonyo South District in the Western Kenyan Highlands: a cluster-randomized controlled trial". In: *PLoS medicine* 13.4 (2016), e1001993.

[76] Katherine R Dobbs and Arlene E Dent. "Plasmodium malaria and anti-malarial antibodies in the first year of life". In: *Parasitology* 143.2 (2016), pp. 129–138.

[77] Irene Kyomuhangi, Tarekegn A Abeku, Matthew J Kirby, Gezahegn Tesfaye, and Emanuele Giorgi. "Understanding the effects of dichotomization of continuous outcomes on geostatistical inference". In: *Spatial Statistics* (2020), p. 100424.

[78] World Health Organisation. *World Malaria Report 2018*. Tech. rep. World Health Organisation, 2018.

[79] World Health Organisation. *World Malaria Report 2019*. Tech. rep. World Health Organisation, 2019.

[80] World Health Organisation. *World Malaria Report 2020*. Tech. rep. World Health Organisation, 2020.

[81] Irene N Nkumama, Wendy P O'Meara, and Faith HA Osier. "Changes in malaria epidemiology in Africa and new challenges for elimination". In: *Trends in parasitology* 33.2 (2017), pp. 128–140.

[82] Pierre De Beaudrap, Carolyn Nabasumba, Francesco Grandesso, Eleanor Turyakira, Birgit Schramm, Yap Boum, and Jean-François Etard. "Heterogeneous decrease in malaria prevalence in children over a six-year period in south-western Uganda". In: *Malaria journal* 10.1 (2011), p. 132.

[83] Luis Fernando Chaves, Masahiro Hashizume, Akiko Satake, and Noboru Minakawa. "Regime shifts and heterogeneous trends in malaria time series from Western Kenya Highlands". In: *Parasitology* 139.1 (2012), pp. 14–25.

[84] Bruno Moonen, Justin M Cohen, Robert W Snow, Laurence Slutsker, Chris Drakeley, David L Smith, Rabindra R Abeyasinghe, Mario Henry Rodriguez, Rajendra Maharaj, Marcel Tanner, et al. "Operational strategies to achieve and maintain malaria elimination". In: *The Lancet* 376.9752 (2010), pp. 1592–1603.

[85] Irene Kyomuhangi and Emanuele Giorgi. "A unified and flexible modelling framework for the analysis of malaria serology data". In: *Epidemiology & Infection* (2021), pp. 1–18.

[86] Jennifer C Stevenson, Gillian H Stresman, Amrish Baidjoe, Albert Okoth, Robin Oriango, Chrispin Owaga, Elizabeth Marube, Teun Bousema, Jonathan Cox, and Chris Drakeley. "Use of different transmission metrics to describe malaria epidemiology in the highlands of western Kenya". In: *Malaria journal* 14.1 (2015), p. 418.

[87] Michael White and James Watson. "Malaria: age, exposure and immunity". In: *Elife* 7 (2018), e40150.

[88] I Rodriguez-Barraquer, E Arinaitwe, P Jagannathan, MR Kamya, PJ Rosenthal, J Rek, G Dorsey, J Nankabirwa, SG Staedke, M Kilama, et al. *Quantification of anti-parasite and anti-disease immunity to malaria as a function of age and exposure. Elife 7*. 2018.

[89] Irina Chis Ster. "Inference for serological surveys investigating past exposures to infections resulting in long-lasting immunity–an approach using finite mixture models with concomitant information". In: *Journal of Applied Statistics* 39.11 (2012), pp. 2523–2542.

[90] Patrick H Corran, Jackie Cook, Caroline Lynch, Heleen Leendertse, Alphaxard Manjurano, Jamie Griffin, Jonathan Cox, Tarekegn Abeku, Teun Bousema, Azra C Ghani, et al. "Dried blood spots as a source of anti-malarial antibodies for epidemiological studies". In: *Malaria journal* 7.1 (2008), p. 195.

[91] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2013. URL: http://www.R-project.org/.