Estimating Infectious Disease Parameters for the Transmission of Malaria in Ugandan Children

Levicatus Mugenyi

Promoters: Prof. Niel Hens
Dr. Sarah Staedke
Co-promoter: Dr. Steven Abrams

Acknowledgements

Firstly, I would like to express my sincere gratitude to my promoters Prof. Niel Hens, Dr. Sarah Staedke and my co-promoter Dr. Steven Abrams for the continuous support of my PhD study and related research, for their patience, motivation, immense knowledge and encouragement. Their guidance helped me in all the time of research and writing of this thesis.

Besides my promoters, I would like to thank my doctoral committee member, Prof. Marc Aerts for his insightful comments, questions and suggestions. Also, I thank the Jury members for their valuable time invested in reading my thesis and for their feedback that made the thesis a success.

I gratefully acknowledge the financial support by a grant of the Vlaamse Interuniversitaire Raad (VLIR) without which, this work could not even take its first step. I appreciate the same support for my Masters in Biostatistics which also formed a good foundation for this PhD.

I acknowledge the support by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under Award Number U19AI089674 that enabled the PRISM study whose data form a backbone of this thesis. Also, I thank the Makerere University Centre for Health and Population research which manages the Iganga-Mayuge health and demographic surveillance site funded under the Makerere University and Sweden International Development Agency (SIDA) Research Collaboration (Project Number 372) whose data form part of the thesis. The support from the PRISM team including Moses Kamya, Grant Dorsey and Sarah Staedke is highly appreciated.

Also, I want to express my gratitude to the University of Hasselt and Censtat, for giving me the opportunity to interact in an academic environment. In particular, the administrators like Martine Machiels and Marc Theolen have not only made a comfortable environment for my stay in Belgium, but they have also made the administrative tasks regarding my PhD so easy for me. I want to thank all my colleagues at Censtat for the friendly environment over the last four years.

I acknowledge the Infectious Diseases Research Collaboration (IDRC), my local institution in Uganda for a wonderful research environment that enabled me come up with the research idea for my PhD. In a special way, I thank the executive director of IDRC, Prof. Moses Kamya for his fatherly support. I am also grateful to the administration and finance department of IDRC headed by Catherine Tugaineyo for managing my local PhD research funds. I cannot forget to say thank you to all my colleagues at IDRC, whose friendly chats and outings did not only relieve me of my PhD stress but also made IDRC a comfortable home.

Last but not the least, I would like to thank my family: my wife Grace Mugenyi and children (Celine, Trevor, Bibiana, Malaika and Graciela) for their moral and spiritual support throughout my PhD research. I thank my father Mzee Peter Kyaluzi (RIP) and mother Venny Nakabuye for laying a foundation for my education.

Above all, I thank the Almighty God for life and wisdom that has enabled me reach this milestone.

And to everybody I've forgotten to acknowledge, thank you very much!

Levicatus Mugenyi
Diepenbeek, 12 April 2018

List of Publications

The materials presented in the dissertation are based on the following publications:

Mugenyi L, Abrams S, Hens N. Estimating age-time-dependent malaria force of infection accounting for unobserved heterogeneity. Epidemiology and Infection. 2017:1-18.

Mugenyi L, Herzog SA, Hens N, Abrams S. Modelling longitudinal binary outcomes with outcome-dependent observation times with an application to a malaria cohort study. **Submitted**

Mugenyi L, Staedke S, Abrams S, Galiwango E, Wabwire F, Kajungu D, Hens N. Determinants of malaria-related death among children in Iganga and Mayuge districts, Uganda, in 2008-2012. **In preparation**

List of other publications not covered in this dissertation:

Staedke SG, Maiteki-Sebuguzi C, DiLiberto DD, Webb EL, **Mugenyi L**, Mbabazi E, Gonahasa S, Kigozi SP, Willey BA, Dorsey G, Kamya MR, Chandler CI. The Impact of an Intervention to Improve Malaria Care in Public Health Centers on Health Indicators of Children in Tororo, Uganda (PRIME): A Cluster-Randomized Trial. American Journal of Tropical Medicine and Hygiene. 2016;95(2):358-67.

Asua V, Tukwasibwe S, Conrad M, Walakira A, Nankabirwa JI, **Mugenyi L**, Kamya MR, Nsobya SL, Rosenthal PJ. Plasmodium Species Infecting Children Presenting with Malaria in Uganda. American Journal of Tropical Medicine and Hygiene. 2017;97(3):753-7.

Tukwasibwe S, **Mugenyi L**, Mbogo GW, Nankoberanyi S, Maiteki-Sebuguzi C, Joloba ML, Nsobya SL, Staedke SG, Rosenthal PJ. Differential prevalence of transporter polymorphisms in symptomatic and asymptomatic falciparum malaria infections in Uganda. Journal of Infectious Diseases. 2014;210(1):154-7.

Contents

List of Tables

List of Figures

List of abbreviations

Here, we give a list of the abbreviations used in the dissertation.

ACT	:	Artemisinin-based combination therapy
AIC	:	Akaike information criteria
AL	:	Artemether-lumefantrine
AR	:	Attack rate
BIC	:	Bayesian information criteria
CDC	:	Centers for Disease Control
CIF	:	Cumulative incidence function
CMI	:	Clinical malaria incidence
CR	:	Competing risks
EIR	:	Entomological innoculation rate
FOI	:	Force of infection
GLM	:	Generalized linear model
GLMM	:	Generalized linear mixed model
HBR	:	Human biting rate
HF	:	Health facility
HIV/AIDS	:	Human immunodeficiency virus/Acquired immunodeficiency syndrome
HR	:	Hazard ratio
IDRC	:	Infectious Diseases Research Collaboration
IMHDSS	:	Iganga-Mayuge Health and Demographic Surveillance Site
KM	:	Kaplan-Meier
MOH	:	Ministry of health
MSE	:	Mean squared error
NIH	:	National Institute of Health
ODEs	:	Ordinary differential equations
ODS	:	Outcome-dependent sampling
OR	:	Odds ratio

PCR	:	Polymerase chain reaction
PH	:	Proportional hazard
PR	:	Parasite rate
PRISM	:	Program for Resistance, Immunology, Surveillance and Modelling of malaria
RDT	:	Rapid diagnostic test
SI	:	Susceptible-Infected
SIR	:	Susceptible-Infected-Recovered
SIRS	:	Susceptible-Infected-Recovered-Susceptible
SIR(T)S	:	Susceptible-Infected-Recovered(Treatment)-Susceptible
SIS	:	Susceptible-Infected-Susceptible
SR	:	Sporozoite rate
UMSP	:	Uganda Malaria Surveillance Project
VA	:	Verbal autopsy
VLIR	:	Vlaamse Interuniversitaire Raad
WHO	:	World Health Organization

Chapter 1

General introduction

This chapter presents a general introduction to the thesis. It starts by giving a definition of an infectious disease followed by the epidemiology of malaria infection, the diagnostics of malaria infection and the key parameters that describe the dynamics of malaria transmission. Later, a review of the literature on malaria is given followed by the problem statement, rationale for the thesis and research objectives. The chapter ends by giving an overview of the thesis.

1.1 What is an infectious disease?

An infectious disease, also known as transmissible disease or communicable disease, is an illness that results from an infection. The term infection refers to the invasion of an organism's body tissues by disease-causing agents, their multiplication, and the reaction of host tissues to these organisms and the toxins they produce. Infectious diseases kill millions of people worldwide, which is more than any other single cause. These diseases are caused by different kinds of germs, which include bacteria, viruses, fungi, protozoa and parasites. For an infectious disease to multiply, the infecting organism should be able to leave an existing reservoir and cause infection elsewhere; this is referred to as infectious disease transmission. There are many potential transmission routes including droplet contact, faecal-oral, sexual, oral, direct contact, vertical, and vector-borne transmission, among others. An example of an infectious disease that results from a vector-borne transmission route is malaria. The work presented in this dissertation focuses on malaria.

1.2 Epidemiology of malaria infection

Malaria is a mosquito-borne disease. This potentially lethal infection causes a wide range of symptoms such as high fever, chills, and flu-like illness symptoms. It can also cause death. Malaria parasites are microorganisms belonging to the genus *Plasmodium*. More specifically, there are five major species of *Plasmodium* parasites causing malaria, namely: *P. falciparum*, *P. malariae*, *P. ovale*, *P. vivax* and *P. knowlesi* [73, 75, 118]. *P. falciparum* is the most severe and prevalent species in sub-Saharan Africa, and it is responsible for the majority of malaria deaths globally [122]. It is characterized by various clinical features, including fever, chills, headache, muscle aches and weakness, vomiting, cough, diarrhoea and abdominal pain [60, 75, 119]. Other symptoms related to organ failure may supervene, such as acute renal failure, pulmonary oedema, generalized convulsions, and circulatory collapse, followed by coma and death [75]. The initial symptoms, which may be mild, may not be easy to recognize as being due to malaria [122].

In humans, the parasites grow and multiply first in the liver cells and then in the red blood cells. In the blood, successive broods of parasites grow inside and destroy the red blood cells, releasing daughter parasites ("merozoites") that continue the cycle by invading other red blood cells. Blood stage parasites are those that cause the symptoms of malaria. See the life cycle for the malaria parasite in Figure 1.1. When certain forms of blood stage parasites ("gametocytes") are picked up by a female *Anopheles* mosquito during a blood meal, they start another, different cycle of growth and multiplication in the mosquito. After 10–18 days, the parasites are found (as "sporozoites") in the mosquito's salivary glands. When the *Anopheles* mosquito takes a blood meal from another human, sporozoites are injected with the mosquito's saliva and start another human infection when they parasitize liver cells. Thus, the mosquito carries the disease from one human to another (acting as a "vector"). Unlike the human host, the mosquito vector does not suffer from the presence of the parasites [14].

1.3 Malaria diagnosis

Malaria infection should be properly detected and treated in time to avoid further spread of the disease. Delays in detection and inappropriate treatment of malaria increase morbidity and mortality [41, 108]. Malaria can be detected and confirmed in the laboratory by one or by a combination of the following diagnostic tools

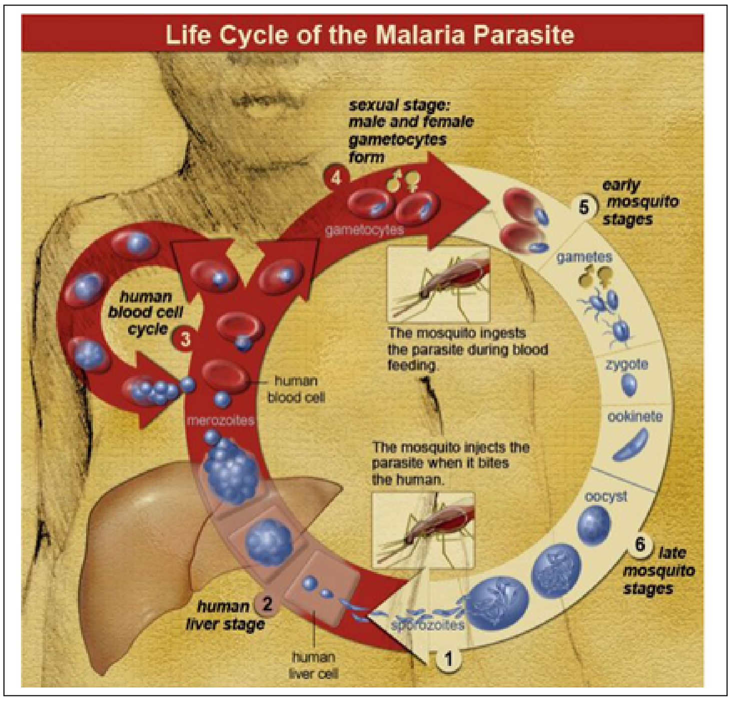
Life Cycle of the Malaria Parasite
sexual stage: male and female gametocytes form
4
gametocytes
5
early mosquito stages
gametes
human blood cell cycle
3
The mosquito ingests the parasite during blood feeding.
zygote
human blood cell
merozoites
ookinete
The mosquito injects the parasite when it bites the human.
oocyst
6
late mosquito stages
human liver stage
2
sporozoites
human liver cell
1

Figure 1.1: Life cycle of the malaria parasite showing the various stages of the parasite in a human host (red arrow) and mosquito vector (white arrow). The cycle begins when an infected mosquito injects parasites by biting a human.
Credited: National Institute of Allergy and Infectious Diseases (NIAID).
Source: https://www.niaid.nih.gov/diseases-conditions/infectious-diseases

[16, 29, 78], namely: antigen detection, parasite detection, molecular tests and serology. Microscopic diagnosis, which remains a gold standard [29, 39, 124], involves identifying malaria parasites using a drop of the patient's blood, spread out as a "blood smear" on a microscope slide and examined under a microscope. The microscopic diagnostic tool was used in the Program for Resistance, Immunology, Surveillance and Modelling of malaria (PRISM) study for which the data have been used largely throughout this thesis. Figure 1.2 shows the microscopic examination for the malaria parasite. Antigen detection test kits, commonly referred to as Rapid Diagnostic Tests (RDTs), detect antigens derived from malaria parasites and offer the opportunity for feasible diagnostic capacity in resource-limited areas because they are easy, fast, inexpensive, and less subjective than microscopy, and they require minimal personnel training [13, 29, 58]. However, RDTs have several limitations. RDTs cannot quantitate parasite density or differentiate non-*falciparum* species; they are less

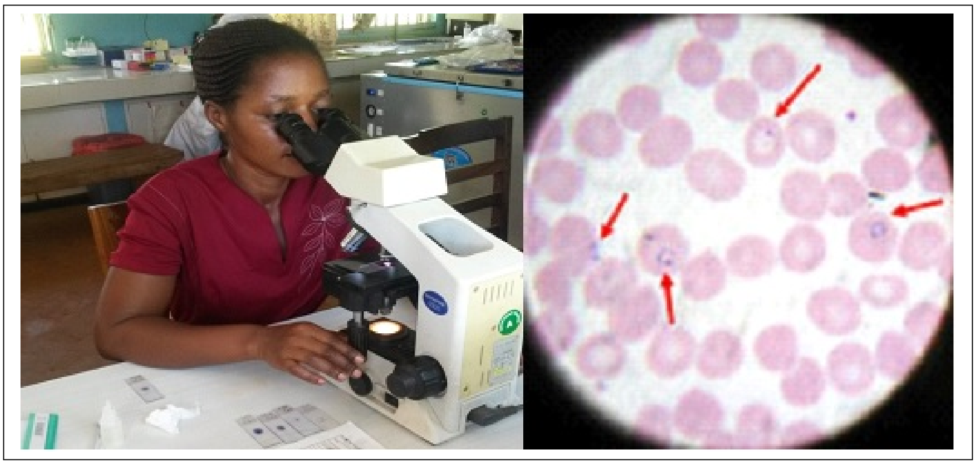

Figure 1.2: Left panel: A laboratory technician at one of the PRISM study sites reading a blood smear under a microscope (field work photo). Right panel: A blood smear from a patient with malaria; microscopic examination shows *Plasmodium falciparum* parasites (arrows) infecting some of the patient's red blood cells (CDC photo).

specific than microscopy and should not be used to evaluate treatment success [13, 59].

Molecular diagnosis involves detecting parasite nucleic acids using polymerase chain reaction (PCR). Although this technique is more sensitive than microscopy [39, 78], it is of limited utility for the diagnosis of acutely ill patients in the standard healthcare setting. PCR results are often not available quickly enough to be of value in establishing the diagnosis of malaria infection [13], limiting the use of PCR outside of research studies. PCR is most useful for confirming the species of malarial parasites or for confirming a negative result after the diagnosis has been established by either microscopy or RDT [39]. The other molecular diagnostic tool that is increasingly being used to diagnose submicroscopic parasitemia is loop-mediated isothermal amplification (LAMP). It is a very sensitive, easy and time-efficient method that uses a single tube technique for the amplification of the deoxyribonucleic acid (DNA). For details on the LAMP technique, the reader is referred to Katrak *et al.* [45]. Serology, on the other hand, detects antibodies against malaria parasites, using either indirect immunofluorescence (IFA) [4, 27] or enzyme-linked immunosorbent assay (ELISA) [52, 70]. Serology does not detect current infection but rather measures past exposure. Figure 1.3 shows an RDT test kit demonstrating a positive test result (left panel) and an IFA showing the presence of malaria parasites in a blood serum sample (right panel).

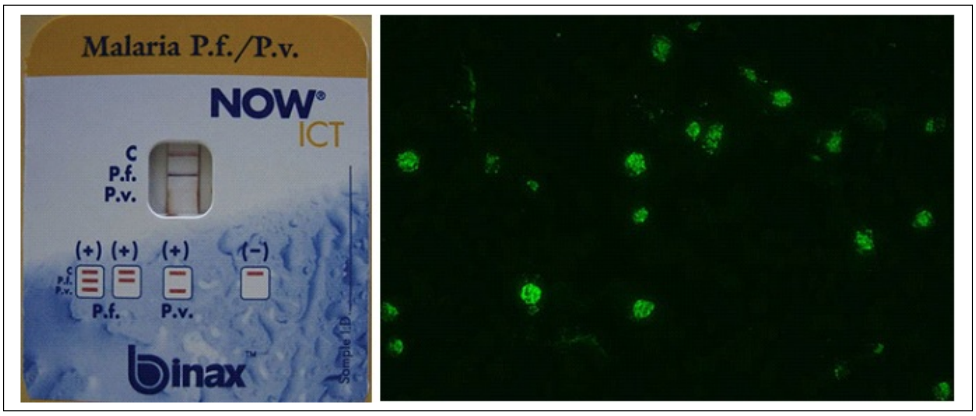

Figure 1.3: Left panel: RDT picture demonstrating a positive test for *Plasmodium falciparum* (Howden BP *et al.* [38]). Right panel: Indirect fluorescent antibody (IFA) test. The fluorescence indicates that the patient serum being tested contains antibodies that are reacting with the antigen preparation (here, *Plasmodium falciparum* parasites) (CDC photo).

1.4 Key malaria transmission parameters

The burden of malaria can be quantified using measures of transmission, including entomological inoculation rate (EIR), force of infection (FOI), clinical incidence, and parasite prevalence. EIR is defined as the number of infectious bites per person per unit time [71], and FOI is defined as the number of infections per person per unit time [97]. Mueller *et al.* [64] defined the FOI for malaria as the number of *Plasmodium* infections acquired over time and devised a way of measuring it molecularly. The FOI counts all incident (that is, new) human malaria infections in some time interval regardless of clinical symptoms, and whether or not a person is already infected [97]. In theory, there should be a close correspondence between EIR and the FOI in children who have not developed immunity. In practice, however, there is a discrepancy between the two. The efficiency of transmission can be estimated by taking the ratio of the two measures, i.e., the ratio of the EIR to the FOI, the number of infectious bites required to cause an infection. The lower the number, the higher the efficiency of transmission. Most studies have shown that malaria transmission is highly inefficient [97]. For example, using annual EIRs of 300, 32 and 2.8 [42] and FOIs of 0.320, 0.108 and 0.152 for children aged <1, which were assumed symptomatic at the previous visit [65], for Nagongera, Kihihi and Walukuba, respectively, the ratios between the EIRs and FOIs range from 18.4 to 937.5. While the clinical incidence of malaria looks at the number of clinical episodes of malaria (defined by symptoms, typically fever; hence symptomatic infections) experienced over a given time period, the FOI looks at both symptomatic and asymptomatic infections acquired over time. The prevalence of infection or parasite rate (PR) refers to the proportion of people who are patently infected with malaria parasites. It is used as a measure of how common malaria is within a population at a point in time. The EIR differs from the FOI since it estimates the number of bites by infectious mosquitoes (other than the number of infections) per person per unit time [49]; notably, not every infectious bite causes an infection, and the bites are distributed unevenly so that not every bite lands on a unique individual [97]. Intuitively, it is expected that a decrease in the mosquito population would lower the FOI and, consequently, decrease the disease incidence for malaria [22]. On the other hand, when the incidence is approximately constant for the duration of the disease, prevalence = incidence × duration [48]. In high intensity settings, where a person can become superinfected (i.e., infected with many different parasites), the duration of infection includes patent infection with any one of the parasites.

The relationship between the EIR (which itself is a product of the proportion of infectious mosquitos, called the sporozoite rate (SR), and the number of vectors attempting to feed on a human each day, called the human biting rate (HBR)), FOI, PR and clinical malaria incidence (CMI) can be illustrated as in Figure 1.4 according to Smith & McKenzie [99]. This relationship implies that a decrease in the mosquito population leads to a decrease in the HBR, which leads to a reduction in the EIR. A decrease in the EIR (given the transmission efficiency, the probability that a bite by an infectious mosquito results in an infection) in turn leads to a decrease in the FOI, in the PR and, consequently, in CMI. Generally, if the human-host immunity is increased due to previous exposure to infections or due to maturation of innate immunity and gradual specific immune development and/or a specific protection (i.e., skin thickness), then the probability that an infectious mosquito bite will result in an infection will diminish. In other words, an immune human host will require a larger number of infectious bites (EIR) to get infected, thereby implying a reduced FOI, PR and CMI. Consequently, the number of infectious humans that can infect mosquitoes will go down, resulting in a reduced number of infectious mosquitoes, which in the end can lead to a reduction in malaria disease transmission and malaria cases.

1.5 Review of literature

Despite increased efforts to eliminate malaria worldwide resulting in reductions in malaria incidence and mortality by 21% and 58% between 2010 and 2015,

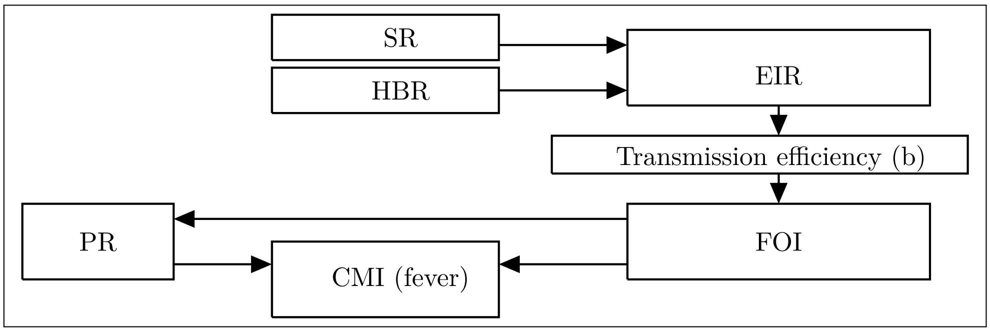

Figure 1.4: Schematic diagram showing the common malaria transmission parameters and the relationships between them.

respectively, malaria remains a major health problem among children, killing a child every 2 minutes globally [122]. For example, 212 million cases of malaria and 429,000 malaria deaths were recorded worldwide in 2015, of which 92% were recorded in the WHO Africa Region [122]. Although there have been substantial reductions in malaria burden since 2010, the trend indicates an increasing burden between 2014 and 2016, with malaria cases increasing by 5 million in 2016 compared to those in2015; moreover, the number of deaths remained largely the same [123]. In Uganda, where the data forming the backbone of this thesis were collected, like in the rest of the sub-Saharan Africa, malaria is still the leading cause of morbidity and mortality among children with 6,100 total malaria deaths registered in 2015 [122]. According to the Uganda's Health Management Information System data from 2014, malaria accounts for 33% and 30% of the outpatient visits and hospital admissions, respectively [3]. Malaria is responsible for approximately 22.6% of the total number of deaths among the under 5 inpatient admissions according to the Uganda Ministry of Health (MOH) Financial Year Report 2014/2015 [110].

One of the challenges probably hindering the elimination of malaria is that the disease is linked to poverty [26, 125], which is a characteristic shared by many people and households in Africa. For example, in Uganda, approximately 20% of the population, i.e., 8 million people, have been reported to live below the poverty line, meaning that they earn less than 57 dollars per month (i.e., less than 1.9 dollars per day) [37]. The other important challenge is that data on malaria-related deaths are still sparse in many African countries like Uganda, because most deaths are either unregistered or registered without specifying the cause by the national mortality registries [12]. Even though the Health Management Information Systems (HMIS) in Uganda systematically collect data on inpatient malaria admissions and deaths, these data are often incomplete, delayed, or inaccurate, restraining the utility of their use to estimate malaria mortality [105], which is an important gap we wish to narrow in this dissertation.

Recently, the number of infectious mosquito bites per person per year, or the entomological inoculation rate (EIR), has been estimated to range from 2.8 to 310 bites for areas of low-to-high transmission intensities in Uganda[42]. In the same paper, Kamya *et al.* [42] estimated the malaria parasite prevalence to be 7.4%, 9.3%, and 28.7% for areas with low, medium and high intensities, respectively. In the past, Felger *et al.* [24] reported that the malaria FOI was moderately age-dependent with children less than two years acquiring less new *P. falciparum* clones than children of older ages. They further stated that the FOI was significantly correlated to the incidence of episodes.

However, it was not clearly stated by the aforementioned authors whether and how they accounted for several other risk factors when estimating the above parameters, which could have introduced bias into the estimates. Modelling is a natural choice for estimating these parameters to adjust for all the risk factors. We, therefore, intended to address this gap in this dissertation.

A large amount of work has been done thus far to model the parameters of infectious diseases using mathematical and statistical models. A mathematical model for malaria transmission was first published in 1908 by Ronald Ross [85]. His model describes changes in the proportion of infected humans or infectious mosquitoes during an epidemic. In his paper published in 1916, Ronald Ross recommended that both the mathematical and statistical modelling frameworks need to be used to correctly understand infectious diseases transmission [84]. The work by Ross was not firmly established until 1950 following the work by George Macdonald which was based on Ross's concept [96]. The Ross-Macdonald theory has since played a central role in the development of research on mosquito-borne pathogen transmission and the development of strategies for mosquito-borne disease prevention [96].

Using the Ross-Macdonald concept, several mathematical models have been proposed to estimate malaria transmission parameters, e.g., see Smith *et al.* [97, 98] and Keeling and Rohani [47]. For example, Smith *et al.*, [97] proposed a nonlinear relationship between the EIR and FOI given by equation (1.1).

$$FOI = \frac{\log(1 + \alpha bEt)}{\alpha t}, \tag{1.1}$$

where b is the transmission efficiency, E is the EIR, and t is the observation time (exposure time). In an earlier work, Smith *et al.* [98] showed that the parasite rate (PR), assuming heterogeneous biting with gamma-distributed biting rates (with a mean of one and a variance of $1/k$) is given by the equation

$$PR = 1 - \left(1 + \frac{bE}{rk}\right)^{-k}, \tag{1.2}$$

where PR, b and E are as defined above, and $1/r$ is the expected time to clear each infection. Keeling and Rohani [47] give several mathematical models covering a whole scope of infectious diseases; in particular, they provide a compartmental model and a system of differential equations linking the transmission dynamics between the mosquito vector and the human host.

On the other hand, the statistical modelling of infectious disease parameters has been the standard methodology for analysing data for many decades now. For example, 84 years have passed since Muench formulated the first catalytic model to estimate the force of infection (unspecific to infectious agent) from current status data in 1934, after which several authors addressed the estimation of this parameter by more advanced statistical methods [35]. A historical overview discussing the relevance of Muench's work and a wide array of newer methods with illustrations on pre-vaccination serological survey data of two airborne infections, rubella and parvovirus B19, was given by Hens *et al.* [35]. Other authors, e.g., Shkedy *et al.* [93] used fractional polynomials to model the age-dependent FOI from seroprevalence data. In a different paper, Shkedy *et al.* [92] used a hepatitis A dataset from Bulgaria to illustrate the use of local polynomials as a nonparametric method to estimate both the prevalence and FOI. Hens *et al.* [36] gave an overview of methods to estimate both the prevalence and FOI from serological and incidence data. Elsewhere, Mueller *et al.* [64] showed that adjusting for individual differences in molecular FOI completely explained spatial variation, age trends, and the effect of insecticide-treated nets (ITN) use.

Ideally, the two distinct approaches to data, which have been called a *priori* (i.e., mathematical modelling) and a *posteriori* (i.e., statistical modelling), could be used to understand infectious disease transmission. Mathematical models are based on facts about the population dynamics, and any lack of knowledge about the underlying mechanisms can make it difficult to apply this approach, which is a disadvantage [95]. On the other hand, statistical modelling analyses available observations and works backward to the underlying cause [36], thus making assumptions about the process that generates the observations, which is a possible drawback. Because of these concerns, neither modelling approach is superior, and the two approaches can be used to complement each other, which is one of the intentions of this dissertation.

Indeed, different infectious diseases are spread differently, and many of the existing models developed for other infectious diseases do not directly apply to malaria. A key idea is heterogeneity or frailty, which is defined as intrinsic variability among individuals in the risk of infection. In particular, malaria is said to be highly heterogeneous [96, 100]. However, the above two models in (1.1) and (1.2) may not efficiently estimate the burden of malaria as they do not indicate how several risk factors and heterogeneity may be accounted for. Due to their flexibility in accounting for several risk factors and heterogeneity, statistical models can provide a more efficient way of estimating the burden of malaria in Uganda and the rest of the world. Therefore, there is a clear need to develop appropriate statistical models for malaria, i.e., the aim of this project. There is also a need to compare and harmonize the results of different methodologies. Here, we develop and harmonize statistical and mechanistic models of the dominant epidemiological and entomological metrics of malaria. For example, the work by Mugenyi *et al.* [65], which forms part of this thesis, harmonizes statistical and mechanistic models when estimating malaria parasite prevalence and FOI while accounting for both observed and unobserved heterogeneity. These authors document that the FOI significantly varies with age and is estimated to be highest among children aged 5–10 years in areas of high and medium malaria transmission and to be highest in children aged below 1 year in a low-transmission setting [65]. Mugenyi *et al.* [65] further show that heterogeneity in malaria infection is greater between households than within households, and it increases with decreasing risk of malaria infection.

The other important scenario that if ignored can possibly hinder the proper understanding, estimation and assessment of malaria control strategies is outcome-dependent sampling (ODS). Malaria follow-up studies involve routine and clinical visits where in the latter case, the infection triggers outcome assessment, leading to ODS [106]. It has been documented that ordinary methods used to analyse longitudinal data ignore ODS and can lead to biased estimates [65, 86, 106] with a consequence of making an incorrect assessment and evaluation of the impact of malaria control strategies. Even though several authors developed methods to accommodate ODS in various settings (e.g., see [83, 106]), literature on how to account for ODS with data coming from both routine and clinical malarial visits does not exist. We therefore intended to narrow this gap.

1.6 Problem statement

Uganda is a country with many factors hindering its development, including diseases (mainly malaria and HIV/AIDS) and poverty, which are closely linked. Malarial disease has impacted humans for thousands of years, and it has continued to do so even though methods for preventing and curing malaria are now available [60, 75]. One of the challenges in properly understanding the dynamics of malaria and in estimating the transmission parameters is that the disease is highly heterogenous [65, 96, 100]. Estimating and understanding how these parameters vary with risk factors while addressing unobserved heterogeneity will allow for the advancement of knowledge regarding efforts for improving malaria control. Additionally, since limited information is available on modelling (statistical modelling in particular) infectious disease parameters, especially in Uganda, the results from this project will form a great foundation for further research in the field of malaria infection. In particular, the estimates for the force of infection will be useful in describing the rate at which susceptible persons acquire an infection that later leads to malaria. The FOI that is estimated by modelling will act as a key parameter (marker) in estimating the burden of malaria and the effectiveness and cost-effectiveness of malaria control.

1.7 Rationale for the thesis

To properly understand the burden of malaria and to facilitate financial funding towards its elimination, data should be available not only on its prevalence and incidence but also on the related mortality. Data on malaria-related mortality are still insufficient in Uganda, a gap this thesis intended to minimize by estimating age-specific mortality rates and the determinants. Although Ronald Ross [84] recommended the use of both the mathematical and the statistical modelling frameworks when estimating infectious diseases parameters, limited work, particularly for malaria infection, has been done to estimate transmission parameters while linking the two frameworks, a gap that is addressed in this dissertation. For example, we derived a methodology to estimate the malaria FOI linking the susceptible-infected-susceptible (SIS) (mathematical) and generalized linear mixed model (GLMM) (statistical) models. Estimates for the malaria FOI are not only lacking, but this parameter has an advantage over incidence and prevalence because it takes susceptibility into account and because it can be used to compare the rate of transmission between different groups of the population for the same infectious disease. On the other hand, it has been noted that the failure to account for the fact that clinical observations are triggered by the longitudinal outcome, a scenario referred to as ODS, may lead to biased estimates, which may lead to the inaccurate measurement of outcomes. This scenario has not been accounted for in the context of malaria infection, a gap this dissertation aimed to address. To do this, a joint model has been developed to include both the longitudinal outcome (parasitemia) and clinical observation times when estimating the age-dependent malaria FOI and parasite prevalence.

1.8 Research objectives

In this doctoral thesis, I was mainly interested in addressing three major gaps: 1) the insufficient information about malaria-specific mortality and determinants; 2) the limited work linking mathematical and statistical modelling frameworks, particularly when estimating transmission parameters for malaria infection; and 3) the limited work on accounting for outcome-dependent sampling (ODS), mainly in the context of malaria infection. Predominantly, we developed models to estimate indicators of malaria burden, including malaria-related mortality, FOI, parasite prevalence and parasite clearance rates. We also suggest several mathematical models describing the various dynamics in malaria transmission and relate these models to the statistical modelling framework. Emphasis has been placed on estimating these parameters while accounting for both observed and unobserved heterogeneity. Particularly, 1) we analyse the determinants of malaria-related death among Ugandan children dying before the age of 15 years; 2) we derived an expression to link the two modelling frameworks and used it to estimate age-time dependent malaria FOI and parasite prevalence, accounting for both observed and unobserved heterogeneity; and 3) we developed a novel methodology to model longitudinal binary outcomes with outcome-dependent observation times with application to a malaria cohort study (PRISM study, see Section 3.1).

1.9 Overview of the dissertation

The thesis is organized as follows. Chapter 2 presents various mathematical models for infectious diseases focusing on malaria, whereas Chapter 3 gives a description of the two data sources used in this dissertation. Chapter 4 presents estimates for the malaria-related mortality and the determinants among children who died between the ages of 29 days and 14 years. Though Chapter 4 presents a stand-alone project using malaria-related mortality data, the results presented in this chapter act as an ice breaker to the malaria problem in Uganda. In Chapter 5, estimates for age-time dependent FOI and parasite prevalence accounting for both observed and unobserved heterogeneity in the acquisition of malaria infection are presented. In this chapter, we show how mathematical and statistical models can be connected and how estimates from the latter can be substituted in the former, thus refining malaria spread models. In clinical study designs, participants often make unscheduled visits triggered by the study outcome such as malaria symptomatic infections, creating a dependency between the observation time process and the process that generates the outcome of interest. If this dependency is not accounted for, there is potential for obtaining biased estimates. Methods to account for this dependency are presented and applied to a malaria cohort study in Chapter 6. Chapters 5 and 6 are linked in the sense that we rely on the same parasitemia data collected routinely and clinically from a cohort of Ugandan children aged 0.5–10 years from 3 regions of varying transmission intensities. In particular, Chapters 5 and 6 account for both observed and unobserved heterogeneity in malaria infection. The dissertation ends with Chapter 7 presenting a general discussion and recommendations for future research.

Chapter 2

Mathematical models

This chapter presents possible transmission models for infectious diseases focusing on malaria. The chapter starts by giving a definition of a mathematical model and later uses schematic diagrams and systems of ordinary differential equations (ODEs) to illustrate and describe dynamics for the various mathematical transmission models. .

2.1 Definition

A mathematical model is a conceptual tool that uses the language of mathematics to explain how an object (system or objects) will behave, and it is used to express quantitative relationships. These models enable us to predict the population-level epidemic dynamics from individual-level knowledge of epidemiological factors and/or the impact of external interventions, like vaccination [47]. A mathematical model translates the infection stages into compartments, also known as classes, describing how individuals move from one compartment to another at a given rate either because they are susceptible, infected or recovered. These models are sometimes referred to as compartmental models. A mathematical model can be classified as either deterministic or probabilistic (stochastic). Unlike the deterministic model, the probabilistic model includes elements of randomness.

Mathematical/compartmental models provide a simplification of the disease dynamics in real-life. Therefore, one needs to understand the disease under investigation in the sense that the infection profile for the disease at hand should be fairly known (and potentially simplified). First, one needs to clearly understand the characterization of a disease at hand. That is, the infection profile for the disease should be fairly known and simplified as appropriate. This profile starts from the point (time 0) when the susceptible hosts encounter an infectious agent (e.g., individual, mosquito vector, animal or plant) and get infected, after which they become infectious, diseased (symptomatic), recovered and later may become immune or susceptible again. Keeling and Rohani [47] give a simplified infection profile in Figure 1.2 of their book. Second, a model that explains the dynamics of the transmission of an infectious disease needs to be simple but not too simple. That is, tangible assumptions should be made to simplify the model while retaining key ingredients. Though all models, even the most complex ones, are essential "wrong", some are useful [10, 47]. Third, formulating a model for a particular problem is a trade-off between three important and often conflicting elements: *accuracy*, *transparency*, and *flexibility.* Accuracy refers to the ability to reproduce the observed data and reliably predict the future dynamics, transparency comes from being able to understand (either analytically or numerically) how the various model components influence the dynamics and interact, and flexibility measures the ease with which the model can be adapted to new situations [47]. Finally, models have two distinct roles, *prediction* and *understanding* which are related to the model properties of accuracy and transparency. Usually, predictive models require a high level of accuracy, whereas models meant to improve our understanding of the problem require transparency [47].

Throughout this section, capital letters will represent the compartments or classes and small letters will represent the proportions of individuals in the respective compartments, i.e., S, I, and R will represent the compartments for the susceptible, infected, and recovered, respectively. Consequently, s, i, and r will represent the corresponding proportions. The parameters, λ and γ will denote the FOI and the recovery or parasite clearance rate, respectively. Time will be represented by t with (t) and subscript t representing the continuous and discrete time frameworks, respectively. The ordinary differential equations (ODEs) will be based on proportions. We represent the differentials with respect to t; $\frac{ds(t)}{dt}$, $\frac{di(t)}{dt}$ and $\frac{dr(t)}{dt}$ with $s'(t)$, $i'(t)$ and $r'(t)$, respectively. For simplicity, we ignore the vital dynamics (demography), i.e., birth, death and migration.

2.2 SIR model without demography

Although the SIR model is not ideal for malaria transmission (because it assumes permanent immunity after recovery from an infection, which is untenable for malaria infection), we consider it here because it is a basic epidemiological model [36]. This model describes the movement of individuals from the susceptible class (S) to the infected class (I) and then to the recovered class (R), i.e., susceptible-infected-recovered (SIR). For the SIR model, we assume permanent immunity after recovery from an infection, and reinfection is thus not possible. Therefore, all infections observed in time period [0, t] are new, with at most one infection per infected individual. Here, (under SIR dynamics), the number of new infections is equal to the number of infected individuals. However, this does not hold for models where there is no permanent immunity (e.g., susceptible-infected-susceptible, SIS model) as we shall see in the following sections. The SIR model is graphically shown in Figure 2.1. Here, $\lambda(t)$ represents the instantaneous rate at which individuals leave the S-class and move into the I-class, also known as the FOI; $\gamma(t)$ denotes the rate at which individuals leave the I-class and move into the R-class, also known as the recovery rate; $s(t)$ is the proportion of susceptible individuals; $i(t)$ is the proportion infected and $r(t)$ represents the proportion recovered at time t. The flow of individuals in the deterministic SIR model can be described using the following system of three ODEs:

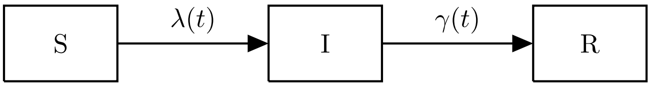

Figure 2.1: Schematic diagram of the SIR compartmental model illustrating the dynamics in transmission of diseases that provide permanent immunity after infection (e.g., measles and chickenpox).

$$\begin{aligned} s'(t) &= -\lambda(t)s(t), \\ i'(t) &= \lambda(t)s(t) - \gamma(t)i(t), \\ r'(t) &= \lambda(t)i(t), \end{aligned} \tag{2.1}$$

with initial condition $s(t) + i(t) + r(t) = 1$.

2.3 SIS model without demography

This model describes the movement of individuals from the S-class to the I-class and then back to the S-class. This model assumes that individuals regain susceptibility upon recovery, hence no permanent immunity. A graphical representation of this model is given in Figure 2.2. By assuming the same parameters as those in the SIR, a

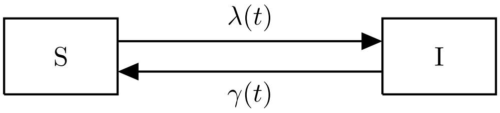

Figure 2.2: Schematic diagram of the SIS compartmental model illustrating the dynamics in transmission of diseases that do not provide immunity after infection.

system of ODEs describing transitions in the compartmental SIS model is as follows:

$$
\begin{aligned}
s'(t) &= -\lambda(t)s(t) + \gamma(t)i(t), \\
i'(t) &= \lambda(t)s(t) - \gamma(t)i(t),
\end{aligned}
\tag{2.2}
$$

where $s(t) + i(t) = 1$.

2.4 SIRS model without demography

Unlike the SIR model, the SIRS model assumes that individuals stay in the R-class for a given period, after which they move back into the S-class at some rate $\sigma(t)$. More specifically, we can re-write the SIRS model as SIR(T)S to imply that temporary recovery is due to prior treatment (T). The SIRS compartmental model is graphically shown in Figure 2.3. Without loss of generality, we represent the rate at which individuals leave the R-class at time t by $\sigma(t)$. The system in (2.3) describes the SIRS model.

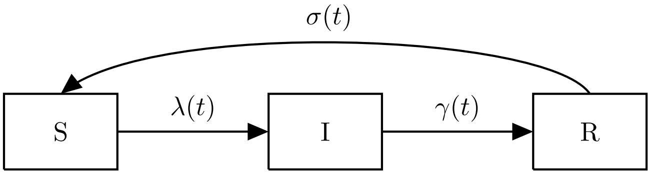

Figure 2.3: Schematic diagram of the SIRS compartmental model illustrating the dynamics in transmission of diseases that provide short-lived immunity.

$$
\begin{aligned}
s'(t) &= -\lambda(t)s(t) + \sigma(t)r(t), \\
i'(t) &= \lambda(t)s(t) - \gamma(t)i(t), \\
r'(t) &= \gamma(t)i(t) - \sigma(t)r(t),
\end{aligned}
\tag{2.3}
$$

where $s(t) + i(t) + r(t) = 1$.

Chapter 3

Data sources

This chapter presents the two sources of Ugandan data used in this dissertation. The first data set comes from the Program for Resistance, Immunology, Surveillance and Modelling of malaria (PRISM) study and the second comes from the Iganga-Mayuge Health and Demographic Surveillance Site (IMHDSS). Details regarding these data sources are given in Sections 3.1 and 3.2.

3.1 PRISM data

The PRISM study, which started in August 2011 in Uganda, collected data from entomological studies, cohort studies, and community & school surveys. This study was conducted under the Infectious Diseases Research Collaboration (IDRC)-Uganda and funded by the National Institutes of Health (NIH). The study was conducted in three sub-counties located in Uganda: the Nagongera sub-county, located in the Tororo district in eastern Uganda; the Walukuba sub-county, located in the Jinja district in eastern Uganda; and the Kihihi sub-county, located in the Kanungu district in south-western Uganda. Nagongera and Kihihi are rural areas, and Walukuba is peri-urban. These regions are characterized by distinct transmission intensities. The EIRs were previously estimated to be 310, 32 and 2.8 infectious bites per unit year for Nagongera, Kihihi, and Walukuba, respectively [42]. A map showing these study sites is given in Figure 3.1.

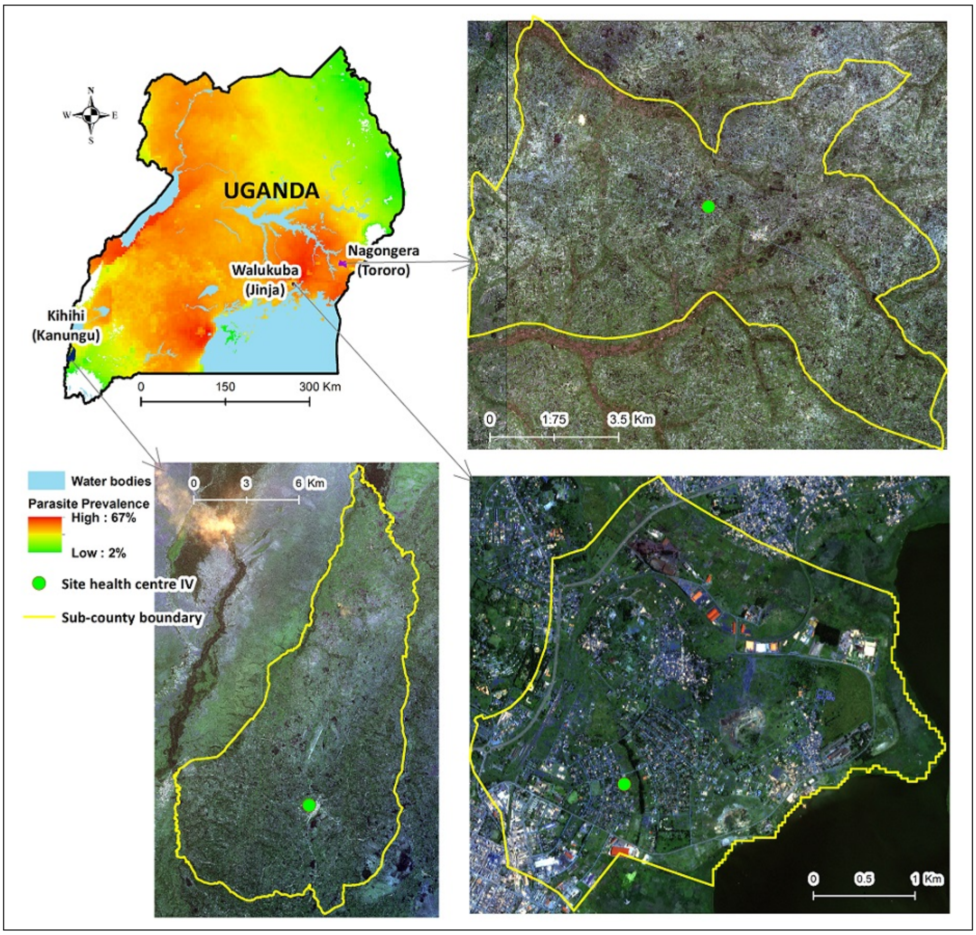

Figure 3.1: A map showing the three study regions for the PRISM study in Uganda. The colour gradient on the map of Uganda (top left) from green to red corresponds to an increasing transmission intensity of malaria. The green dot in each region's boundary corresponds to the location of the health centre IV clinic where the study participants were being tested and treated (map by Simon Peter Kigozi).

Data from only the cohort study were used in this dissertation. The study participants were recruited from 300 randomly selected households (100 per site) located within the catchment areas. The screening visit and all subsequent study visits took place at a designated study clinic in the health centre IV facilities located in each catchment area. A study clinician assessed the eligibility of potential of the study participants using the following selection criteria. The inclusion criteria included (1) the documented age between 6 months to less than 10 years, (2) resident in the sub-county at the household selected for recruitment, (3) no intention to move out of the sub-county for the next two years, (4) agreement to come to the study clinic at the UMSP health centre sentinel site for any febrile illness, (5) agreement to avoid antimalarial medications administered outside the study, and (6) provision of written informed consent. The exclusion criteria included (1) the presence of a chronic medical condition requiring specialized primary health care and (2) enrolment in another research study. Blood samples were collected at enrolment at each sick (clinical) visit with suspected malaria and routinely every 3 months, and the samples were tested for the presence of *Plasmodium* parasites. All children born into a household during the follow-up were recruited into the cohort at the time they were 6 months of age. The major reasons for doing clinic visits were to measure the overall malaria incidence, including both asymptomatic and symptomatic infections, and to provide the recruited children with general medical care for all other illnesses. If a subject had a history of and/or documented fever (temperature$\geq$ 38.0°C), a blood sample was obtained and tested for malaria. Patients who did not have fever and were not suspected to have malaria received standard-of-care treatment as per local treatment guidelines. Details of this study can be found in Kamya *et al.* [42].

3.2 IMHDSS data

Data on all deaths and their causes were collected from the Iganga-Mayuge Health and Demographic Surveillance Site (IMHDSS) in the Iganga and Mayuge districts in the East Central region of Uganda. A map showing the IMHDSS study area is shown in Figure 3.2. According to the Uganda Malaria Indicator Survey (2014-15), this region has the highest prevalence of anaemia and malaria by microscopy, estimated at 8% and 37%, respectively. The IMHDSS has conducted 17 individual and household-level data update cycles since its inception in 2004 ($http://www.indepth-network.org/Profiles/iganga\_mayuge\_hdss\_2013.pdf$). The IMHDSS is managed by the Makerere University Centre for Health and Population research (MUCHAP), which is aimed at strengthening the platform for institutional research and research training and ensuring a robust research agenda. These data were collected using standardized VA questionnaires designed by the INDEPTH network (www.indepth-network.org) with detailed information regarding signs and symptoms that led to the deceased deaths. These data were collected in 65 villages with a population of approximately 85,000 people by conducting interviews of close relatives or caretakers who were present during the period from illness until death. Unlike post-mortem analyses in clinical research, VA serves the same purpose, and it is an indirect way of ascertaining the cause of death. The WHO categorizes these questionnaires depending on the age of the deceased, i.e., a neonatal tool for 0-28-day-old infants, a child tool for those aged 29 days to 14 years, and an adult tool for persons aged 15 years and above [69]. The data used in this thesis were collected using the child tool. These data were examined afterwards by 2 to 3 physicians, who assigned a cause of death depending on the reported signs and symptoms prior to death.

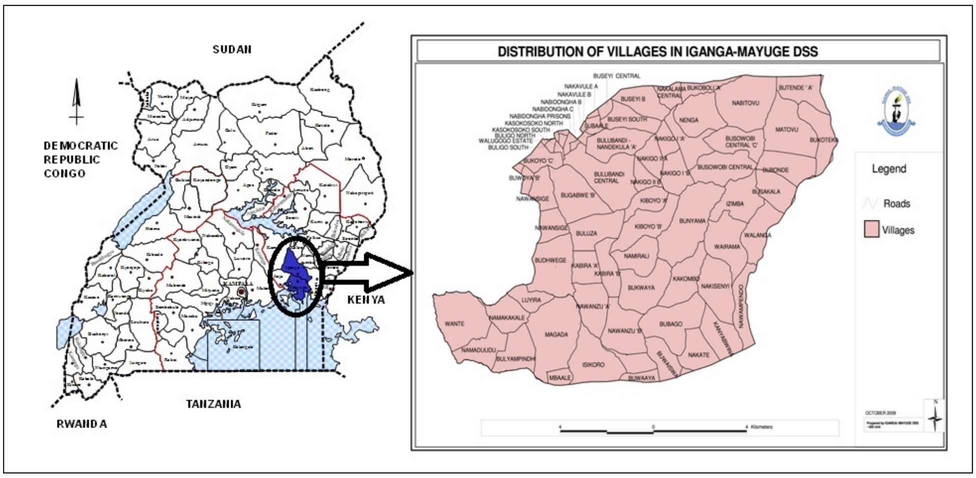

Figure 3.2: A map showing the Iganga-Mayuge Health and Demographic Surveillance Site (IMHDSS) in Iganga and Mayuge districts in the East Central region of Uganda.

Chapter 4

Malaria-related mortality and determinants

4.1 Summary

Malaria remains a leading cause of disease and death among children in sub-Saharan Africa. In low-resource countries, many deaths are not captured by routine registration systems, leading to underreporting of mortality. To better understand the malaria burden in Uganda, we analysed mortality data extracted from a verbal autopsy (VA) study conducted in the Iganga and Mayuge districts from 2008 to 2012. Competing risks models were used to estimate the hazard and determinants of malaria-related deaths among children (aged 29 days to 14 years) who died. Of 781 deaths, 396 (50.7%) were attributed to malaria; nearly all malaria deaths (93.7%) occurred in children under five. Of children who died of malaria, 169 (42.7%) died in a hospital or health facility, 103 (26.0%) died en route, while 124 (31.3%) died at home or elsewhere. In these children who died of malaria before 15 years of age, the hazard of dying was higher in those with fever (adjusted hazard ratio, HRa= 3.72, 95% confidence interval [CI]: 2.81 − 4.92), suggesting that children with fever were younger when they died compared to those without fever. The hazard of dying due to malaria was lower in children who died at home (HRa = 0.61, 95% CI: 0.47 – 0.78), suggesting that those who died at home were older than those who died in a hospital or health facility. Among children who died before the age of 15 years in Uganda, VA data suggest that malaria was the leading cause of death, and that over half of children died outside of health centers. Strengthening malaria surveillance at health facilities

and within communities to accurately capture data on malaria mortality is essential. Educating caregivers about the symptoms of malaria and importance of seeking care promptly, to ensure appropriate diagnosis and effective treatment of malaria, should continue to be a priority.

4.2 Introduction

Despite recent efforts to scale-up coverage of malaria control interventions, and the renewed focus on elimination, malaria remains a major global health problem [122]. In Africa, where approximately 91% of malaria deaths occur, the number of malaria deaths has decreased by 37% since 2010. However, between 2015 and 2016, there was no significant change in the malaria mortality rate in Africa suggesting progress may have stalled [2]. In Uganda, malaria is a leading cause of morbidity and mortality with 28.7 million suspected malaria cases and 5,635 malaria deaths recorded in 2016 [123]. Among children under-five admitted to a hospital in Uganda during 2014/2015, an estimated 22.6% of deaths were attributed to malaria [110]. However, estimating malaria mortality overall in Uganda is still a challenge because most deaths occur at home and are not registered, or are recorded in national registries without specifying the cause of death [12]. Indeed, although information on inpatient malaria admissions and deaths are systematically collected in the Health Management Information Systems (HMIS) in Uganda, these data are often incomplete, delayed, or inaccurate, limiting the utility of HMIS data on malaria-specific mortality [105]. Reliable measures of the burden of malaria mortality in Uganda are needed to more accurately quantify the burden of malaria and to evaluate the impact of control interventions.

Verbal autopsy (VA) is an indirect method of determining cause of death based on an interview with the caretakers of a deceased individual, which has been widely used to collect information on cause-specific mortality where medical information on deaths is incomplete [12, 63, 104]. Information about specific signs and symptoms, and circumstances preceding the terminal event, are used to ascertain the most likely cause or causes of death. Although the sensitivity of VA for determining malaria-specific mortality may be low in some settings, and methods used to interpret VA data can vary substantially between sites [7, 102], VA remains the only available method for determining cause-specific mortality data in regions where vital registration systems are deficient [7]. A study of the validity of VA methods for determining malaria deaths in Uganda concluded that VA had an acceptable level of diagnostic accuracy for determining malaria deaths at the population level in high and medium transmission areas, albeit not in an area with low transmission [62]. To further explore all-cause and malaria-specific mortality, we applied a competing risk analysis to VA data collected from the Iganga and Mayuge districts in eastern Uganda. Specifically, we estimated the malaria-specific mortality hazard and related factors in a population of children aged 29 days to 14 years who died between 2008 and 2012.

4.3 Methods

4.3.1 Data source

In this chapter, we use VA data collected by the Iganga-Mayuge Health and Demographic Surveillance Site (IMHDSS). Data on all causes of death for children aged 29 days up to 14 years were extracted from the VA database. Other extracted variables included: age at death, gender, relationship with next of kin, place of death, presence of comorbidities, symptoms and signs noted, the duration of the final illness that led to death, and treatment and health service use for the final illness. These data were analyzed using STATA (version 12.0, College Station, TX) and R (R Core Team (2015), URL: https://www.R-project.org/). The R code is given in Appendix A.

4.3.2 Statistical analysis

Survival analysis methods were used with the child's age at death taken as the survival time variable. In survival analysis, standard Kaplan-Meier (KM) estimates of survival curves [43] and the Cox proportional hazards (PH) model [19] are commonly used. These methods deal with only one event type, for example death, and right-censor all other observations where the event of interest is not experienced by the end of the study period [30, 76, 81]. The main assumptions of these methods are that individuals with right-censored observations will experience the event of interest eventually if followed long enough in time and that censoring is independent implying that censored individuals should be representative for those still at risk at a specific time point [25, 76]. If these assumptions are intangible or if the occurrence of other events prevents the occurrence of the event of interest, then a competing risk analysis should be preferred [25, 90]. Estimating the marginal survival function using KM, hence, censoring the competing events, can be interpreted as probabilities to survive up to a specific time in a hypothetical but unrealistic world in which no other competing risks exist [28]. In case the independent censoring assumption is violated, the KM approach overestimates the event probability yielding misleading results. Here, causes of death including malnutrition, anemia, pneumonia, diarrhea, measles, tetanus, road accident and HIV/AIDS are considered as competing against malaria-related mortality. Therefore, we apply competing risks analysis techniques to estimate the malaria-specific mortality hazard and study its determinants in the presence of all other causes of death.

4.3.3 Competing risks analysis

In recent years, competing risks (CR) analysis became an important method in survival analysis, particularly in situations where interest is in a specific cause of failure in the presence of other competing causes, which alter the probability of experiencing the event of interest [25, 76, 89]. In this case, instead of the 1-KM estimate, from here onwards referred to as the marginal Cumulative Incidence Function (marginal CIF), the cause-specific CIF, representing the event probability for one of the competing events in the presence of all others, is estimated [25]. It has been shown that for dependent competing risks, the marginal CIF always overestimates the cause-specific CIF [76]. A similar observation was made based on our data (see Figure 4.3). The cause-specific CIF partitions the probability of failure into probabilities corresponding to each competing event [5, 76, 89]. To assess statistical significance of a prognostic factor in the competing risks analysis, Gray's test [32] was used instead of the log-rank and Wilcoxon tests which are commonly used in standard survival analyses [89].

In order to determine factors associated with a particular event of interest in the presence of one or more competing risks, several regression models have been proposed, including the cause-specific hazards model. For details on these methods, see, e.g., Fine and Gray [25], Klein and Anderson [51], Logan *et al.* [56], and Pintilie [27]. In particular, Fine and Gray [25] noted that the effect of a covariate on a cause-specific hazard function of a particular event type may be very different from its effect on the corresponding CIF. This led to the so-called Fine and Gray proportional hazards model for the subdistribution of a competing risk [25], which is recommended for determining the covariate effects in a competing risks setting [76]. Therefore, in this chapter, the model proposed by Fine and Gray [25], is considered which estimates the so-called subdistribution hazard of an event following a specific cause in the presence of competing causes. In their approach, Fine and Gray adopted a semiparametric PH model for the subdistribution hazard of cause r for a subject with covariate vector $\mathbf{X}$

as follows [90],

$$\lambda_r(t|\mathbf{X}) = \lambda_{ro}(t)\text{exp}(\beta_r^T\mathbf{X}), \tag{4.1}$$

where $\lambda_{r0}(t)$ is the subdistribution hazard of cause r, β_r is a vector of coefficients associated with covariates $\mathbf{X}$. Similar to the estimation of model parameters in the semiparametric Cox PH model, the estimation follows a partial likelihood approach. Since this is not a full likelihood approach, the test statistics that are asymptotically distributed as chi-square (χ^2), such as the likelihood ratio test, become invalid and can therefore not be used for model selection. In a similar way, the commonly used Akaike information criterion (AIC) and the Bayesian information criterion (BIC) cannot be used directly. However, differences in AIC and BIC values for models with respect to the smallest value obtained for model (1) in a set of candidate models (including the null model), are valid to be used for model comparison [90]. For example, using $AIC = -2l + 2p$ where l denotes the log-likelihood and p the number of model parameters, then AIC difference for model i with respect to the smallest value of AIC for a set of candidate models is defined as $\Delta AIC_i = AIC_i - min(AIC)$. By rule of thumb, it's argued that $\Delta AIC_i > 10$ provides very strong evidence against the candidate model as compared to model (1), whereas $0 < \Delta AIC_i < 2$ suggests that the candidate model has substantial support and should be used to make inference [44, 90]. Therefore, ΔAIC_i was used for model selection. In the Appendix A, we present the fit statistics and describe the steps followed to arrive at a better fit for which the results are presented as adjusted estimates in Table 4.2 . In the same appendix, we give a list of candidate models and their fit statistics in Table A.1.

4.4 Results

Of 781 children contributing to the VA data, 404 (51.7%) were males and 490 (62.7%) were under the care of their biological mothers. The median age at death (all deaths) was 1.2 years (interquartile range, IQR: 0.8 – 3 years) with 31.2% dying before their first birthday, and only 13.6% dying after having reached the age of 5 years (i.e., between 5 and 14 years of age). The age at death distribution encompassing deaths due to all causes is given in in Figure 4.1 (A). Of the total number of deaths, 50.7% (396/781) was attributed to malaria infection followed by malnutrition (12.9%) according to the VA physicians who reviewed the questionnaires. These proportions are referred to as cause-specific mortality fractions. See Figure 4.1 (B) for details.

Out of the 396 malaria deaths, 54.0% were males, 64.4% were under the care of their

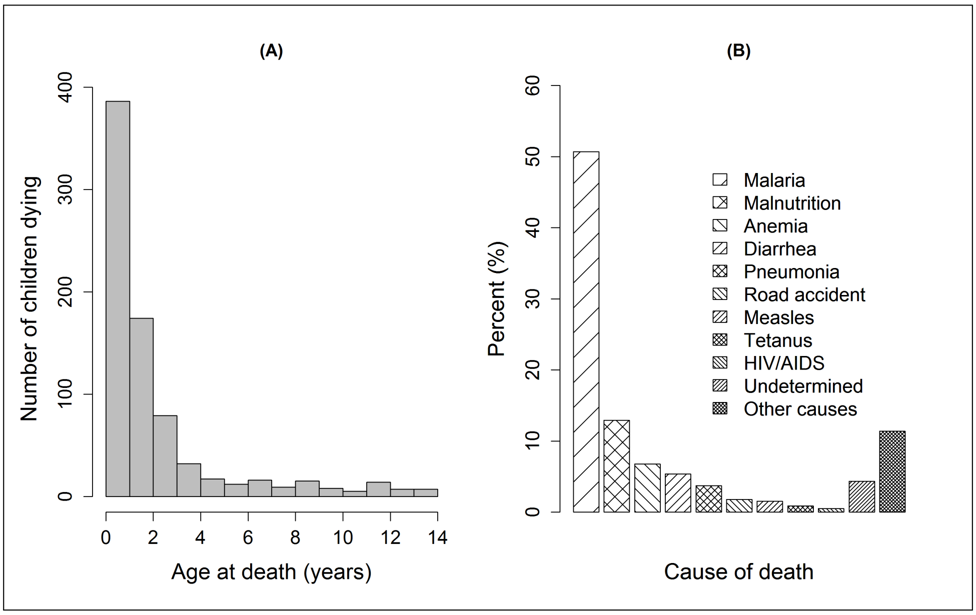

Figure 4.1: All-cause mortality among Ugandan children aged 29 days to 14 years. (A) Frequency distribution of age at time of death. (B) Cause-specific mortality fractions.

mothers with 57.3% dying outside a hospital or health facility. About 78% of the malaria deaths occurred within the first three years of life with 53.3% dying when aged one year and below. Of those dying within their first year of life, 79.1% were aged between six months and one year. The median age at death due to malaria was 1 year (IQR: 0.75 – 2 years). The other summary statistics are given in Table 4.1.

Of the children dying from malaria, 21.7% had convulsions, 79.8% had fever (54.3% of the fever cases were considered severe by the caretaker), 42.9% experienced vomiting, 16.2% had diarrhea, 6.8% had headache and 6.6% had comorbidities. Sickle-cell anemia was the most frequent comorbidity (2.0%) followed by asthma (1.5%), malnutrition (1%), heart attack (0.8%), congenital malformation (0.8%), tuberculosis (0.5%), and epilepsy (0.3%). None of the children dying from malaria had HIV/AIDS, cancer or diabetes. We refer to Table 4.2 for the other summary statistics.

A plot for the estimated probabilities of dying at a given age for all causes of death for this study population is given in Figure 4.2, which shows that malaria was the dominant cause of death among the children dying before the age of 15 years.

Table 4.1: Baseline characteristics of children dying from malaria at the ages of 29 days to 14 years, in the two study sites in Uganda (Iganga and Mayuge), period 2008 to 2012.

		Malaria deaths (N=396)
		n (%)
Sex	Female	182 (46.0)
Age at death	<1 year	131 (33.1)
	1-2 years	178 (45.0)
	3-4 years	62 (15.7)
	5-14 years	25 (6.3)
Caretaker	Mother	255 (64.4)
	Father	72 (18.2)
	Other	69 (17.4)
Place where death occurred	Hospital/facility	169 (42.7)
	Home	96 (24.2)
	On way to hospital/health facility	103 (26.0)
	Other	28 (7.1)

The marginal and cause-specific CIF estimates for the probability of dying from malaria by age are presented in Figure 4.3. The marginal CIF (long-dashed line) represents the probability of dying in the absence of any other competing cause, and conditional on dying before the age of 15. As indicated before, Figure 4.3 clearly shows that the marginal CIF is larger than the cause-specific CIF (solid line). The cause-specific estimator shows that about 40% of the children who died from malaria before the age of 15 years and in the presence of other competing risks, died within the first two years of life whereas malaria-specific mortality remains almost constant at 50% after five years of age. Graphical representations of the factor level malaria-specific CIF curves are given in Figure 4.4. The CIF was higher on the way to hospital or health facility and lower at home compared to the CIF at the hospital or health facility, the CIF was higher for presence of convulsions, and lower for children with comorbidities. All these differences were significant (Gray's test: p-value < 0.001).

Next, the determinants associated with malaria-specific deaths resulting from the Fine and Gray model are presented. Here forth, the malaria-specific mortality

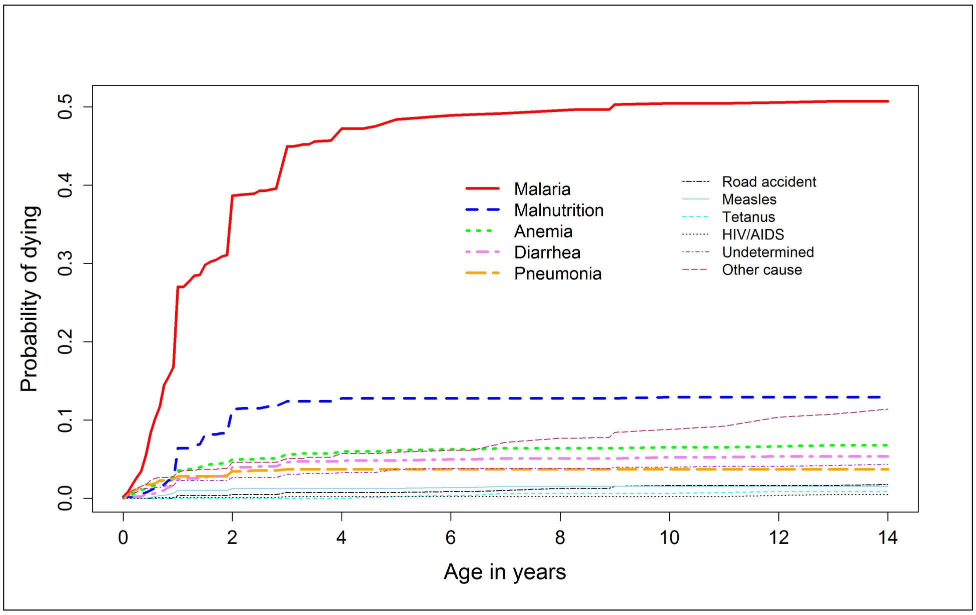

Figure 4.2: Estimated cumulative incidence curves showing age-specific probabilities of dying from all causes among Ugandan children aged 29 days to 14 years.

subdistribution hazard is referred to as the hazard, and the abbreviations, HRc and HRa are used to represent crude and adjusted (derived from the final fit) estimates for the hazard ratios, respectively. Table 4.2 shows the number of malaria deaths for different factor levels, the crude/unadjusted and the adjusted estimates for the hazard. Factors that had significant crude estimates were subjected to a multivariable analysis leading to adjusted estimates.

The subdistribution hazard was significantly lower at home [HRa = 0.61 (95% CI: 0.47 – 0.78), p-value < 0.001] compared to the hospital or health facility (HF), and lower among children with comorbidities [HRa = 0.40 (95% CI: 0.28 – 0.58), p-value < 0.001]. The hazard was significantly higher among children with fever [HRa = 3.72 (95% CI: 2.81 – 4.92), p-value < 0.001] but lower among those presenting with a headache [HRa = 0.60 (95% CI: 0.45 – 0.81), p-value = 0.001]. Children who had a longer illness duration, had a lower hazard [HRa = 0.73 (95% CI: 0.67 – 0.80), p-value < 0.001]. Whereas the crude estimate indicated a higher hazard associated with convulsions [HRc = 1.60 (95% CI: 1.28 – 2.00), p-value < 0.001], the adjusted estimate was no longer significant [HRa = 1.23 (95% CI: 0.96 – 1.58)]. Treating a child within 24 hours (treatment promptness) was found not to be associated with

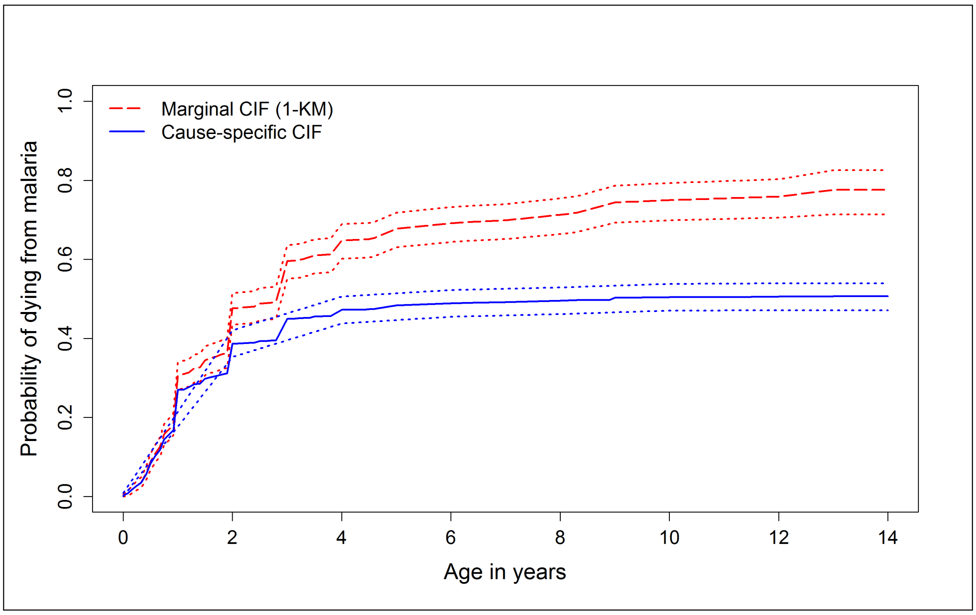

Figure 4.3: Estimated cumulative incidence curve for malaria-related death among Ugandan children with 95% pointwise confidence intervals (dotted lines).

the event time of interest ([HRc = 1.10 (95% CI: 0.71 – 1.71)] if treated at home within 24 hours, and [HRc = 1.12 (95% CI: 0.84 – 1.50)] if treated elsewhere within 24 hours). See Table 4.2 for more details.

4.5 Discussion and conclusion

Accurate estimates of malaria mortality are essential for understanding malaria burden, assessing the impact of interventions, and targeting resources efficiently. Given the gaps in HMIS malaria surveillance data, we analysed data from a VA study to evaluate malaria-specific mortality and its determinants among children who died in between 29 days and 14 years of age in Iganga and Mayuge. A WHO standard tool for age group 29 days to 14 years was used to determine causes of death. Our results suggest that malaria is still the leading cause of death in this population with over a half of these children dying when aged one year and below, and of these, more than three-quarters died when aged between six months and one year. The high mortality within the age group 0.5–1 years, is in line with the work by Riley *et al.* [82] that immunity wanes before 6 months of age. Also, more than half of the malaria deaths occurred outside of a hospital or health facility. In these

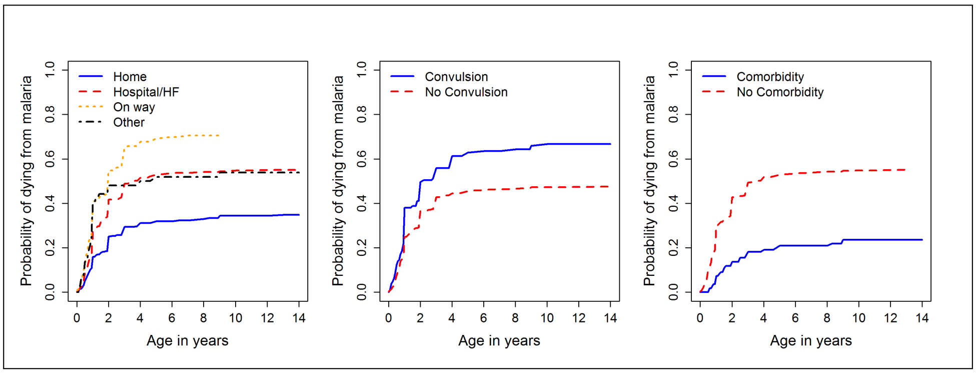

Figure 4.4: Estimated cumulative incidence curves comparing group probabilities of dying from malaria. Left: By place where death occurred (Gray's test P<0.001). Middle: By presence of convulsions (Gray's test P<0.001). Right: By presence of comorbidities (Gray's test P<0.001).

children who died of malaria before 15 years of age, those with fever were younger when they died compared to those without fever, while those who died at home were older than those who died in a hospital or health facility. Our results also suggest that children who had comorbidities, a longer duration of illness, or complained of headache, were older when they died than their respective comparators. These findings are likely reflective of treatment seeking behaviors and variation in the clinical presentation of malaria due to age and naturally acquired immunity. Programmes aiming to ensure universal coverage of diagnostic testing and antimalarial treatment, and those promoting prompt effective treatment of malaria, as advocated by the WHO's Global Malaria Programme, should continue to be a priority [120, 121].

The results presented here do not differ from those reported by Mpimbaza *et al.* [62] who estimated the malaria mortality fraction at 47.8% based on medical hospitalization records and 35.8% using VA questionnaires in the Tororo district in Uganda within a similar setting. However, the estimate of about 50% of malaria deaths among the under 5 years more than doubles the national estimate of 22.6% reported by the MOH (2014). The difference can be attributed to the fact that most malaria deaths in regions like Uganda occur outside the hospitals or the health facilities as noted by Streatfield (2014) and as found in this paper. But, it is also worthwhile noting that there was a recorded decrease in malaria deaths of 35% among the under 5 years old between 2010 to 2015 (WHO, 2016) which could at least partly explain the difference between our estimate and that of the MOH (2014).

The results on determinants of malaria-specific mortality imply that among children dying before the age of 15, the disease has potential to kill children faster especially if the illness involves fever. This is the reason by which the WHO promotes prompt treatment of fever (within 24 hours) [91]. The lower hazard of dying from malaria at home in this population can be attributed to the several initiatives like community health workers (CHWs) that reach households in Uganda performing malaria diagnostic and treatment [40, 68]. The finding for the decrease in the hazard after the age of 5 years could be that these children likely will get acquired immunity due to past illnesses and/or increasing age as discussed by Doolan *et al.* [21]. We noted changes between crude and adjusted estimates. For example, the effect of convulsions changed by 18.1% when fever was introduced into the model. Therefore, further research is needed to investigate the causal pathway of convulsions and fever.

The results presented in this chapter are restricted to children who died, moreover from only 2 of the 112 districts in Uganda, thus limiting the ability to generalize our results to a wider population. Also, the VA data used for the analysis were based on physician review of records with the existing concerns regarding this method in terms of repeatability and reliability of the collected data.

In conclusion, the work in this chapter seems to agree with the claim that malaria is the leading cause of death among children in Uganda, with more than half of the malaria-related deaths occurring outside hospitals or health facilities and which are possibly missing in the national registries. However, since our results are limited to only a small part of the country, it's possible that elsewhere (e.g., in areas with low malaria intensity), more children could be dying from other causes than malaria. Children whose last illness involved fever died at a younger age compared to those who died without fever. We recommend the strengthening of the malaria surveillance at health facilities and within communities to accurately capture data on malaria mortality. In addition, caregivers should continue to receive education about the symptoms of malaria and importance of seeking care promptly, to ensure appropriate diagnosis and effective treatment of malaria.

Table 4.2: Crude/unadjusted and adjusted hazard ratios obtained using the Fine & Gray subdistribution hazard model.

Factor°	Malaria deaths (N=396) n (%)	Crude estimates		Adjusted estimates	
		HR_c (95% CI)	p-value	HR_a (95% CI)	p-value
Place of death					
Hospital/HF	169 (42.68)	Reference		Reference	
Home	96 (24.24)	0.55 (0.43–0.70)	<.001	0.61 (0.47–0.78)	<.001
On the way	103 (26.01)	1.49 (1.18–1.88)	<.001	1.22 (0.94–1.59)	0.130
Unspecified	28 (7.07)	1.06 (0.70–1.61)	0.780	1.35 (0.90–2.02)	0.150
Convulsions					
No	310 (78.28)	Reference		Reference	
Yes	86 (21.72)	1.60 (1.28–2.00)	<.001	1.23 (0.96–1.58)	0.110
Comorbidities[†]					
Non-comorbid	310 (78.28)	Reference		Reference	
Comorbid	86 (21.72)	0.34 (0.23–0.49)	<.001	0.40 (0.28–0.58)	<.001
Illness duration,					
mean (SD)[ϕ]	1.33 (1.15)	0.73 (0.68–0.79)	<.001	0.73 (0.67–0.80)	<.001
Fever					
No	80 (20.20)	Reference		Reference	
Yes	316 (79.80)	3.07 (2.41–3.91)	<.001	3.72 (2.81–4.92)	<.001
Headache					
No	369 (93.18)	Reference		Reference	
Yes	27 (6.82)	0.69 (0.50–0.96)	<.001	0.60 (0.45–0.81)	0.001
Hospitalization history					
No	161 (40.66)	Reference			
Yes	155 (39.14)	0.77 (0.62–0.95)	0.017		
Don't know	80 (20.20)	0.74 (0.57–0.96)	0.025		
Vomiting					
No	226 (57.07)	Reference			
Yes	170 (42.93)	1.10 (0.91–1.34)	0.310		
Diarrhea					
No	332 (83.84)	Reference			
Yes	64 (16.16)	0.54 (0.41–0.70)	<.001		
Source of care					
Hospital	146 (36.87)	Reference			
Home	83 (20.96)	0.93 (0.72–1.21)	0.580		
Health center	65 (16.41)	1.06 (0.81–1.40)	0.670		
Private clinic	71 (17.93)	1.09 (0.82–1.44)	0.550		
Drug shop	17 (4.29)	1.00 (0.62–1.60)	0.990		
Trad. healer	5 (1.26)	0.77 (0.30–1.99)	0.590		
Other	9 (2.27)	1.33 (0.69–2.57)	0.400		
Sex					
Male	214 (54.04)	Reference			
Female	182 (45.96)	0.91 (0.75–1.10)	0.310		
Caretaker					
Mother	255 (64.39)	Reference			
Father	72 (18.18)	1.05 (0.82–1.35)	0.710		
Other	69 (17.42)	0.76 (0.59–0.98)	0.037		

[†] *sickle-cell anemia, asthma, malnutrition, heart attack, congenital malformation, TB and epilepsy;* [ϕ] *log transformed;* [°] *treatment promptness was not included in the table because it was making it messy, moreover only crude estimates were available*

Chapter 5

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

5.1 Summary

Despite well-recognized heterogeneity in malaria transmission, key parameters such as the force of infection (FOI) are generally estimated ignoring the intrinsic variability in individual infection risks. Given the potential impact of heterogeneity on the estimation of the FOI, we estimate this quantity accounting for both observed and unobserved heterogeneity. We used cohort data of children aged 0.5–10 years evaluated for the presence of malaria parasites at three sites in Uganda (see Section 3.1). Assuming a Susceptible-Infected-Susceptible model, we show how the FOI relates to the point prevalence, enabling the estimation of the FOI by modeling the prevalence using a generalized linear mixed model. We derive bounds for varying parasite clearance distributions. The resulting FOI varies significantly with age and is estimated to be highest among children aged 5–10 years in areas of high and medium malaria transmission and highest in children aged below 1 year in a low transmission setting. Heterogeneity is greater between than within households and it increases with decreasing risk of malaria infection. This suggests that next to the individual's

age, heterogeneity in malaria FOI may be attributed to household conditions. When estimating the FOI, accounting for both observed and unobserved heterogeneity in malaria acquisition is important for refining malaria spread models.

5.2 Introduction

Estimating the burden of malaria and evaluating the impact of control strategies, requires reliable estimates of transmission intensities [17]. Measures of malaria transmission intensity include the entomological inoculation rate (EIR), parasite prevalence and force of infection (FOI) [17, 42, 50, 71, 97, 98]. The EIR is defined as the number of infectious bites per person per unit time [71, 72] whereas the FOI is defined as the number of infections per person per unit time [97] or the per capita rate at which a susceptible individual acquires infection [18, 35]. The malaria FOI counts all incident (that is, new) human malaria infections in a specified time interval regardless of clinical symptoms, and recurrent infections [97]. The EIR and FOI are related but differ; the EIR considers the number of infective bites delivered by the mosquito vector, whereas the FOI focuses on the infections acquired by the human host. In theory, there should be a close relationship between the EIR and the FOI, especially in children with less developed immunity. In practice, however, there is a discrepancy between the two because not every infectious bite results in an infection due to various factors [99]. The efficiency of transmission can be estimated by taking the ratio of the two measures, i.e., the ratio of the EIR to the FOI, the number of infectious bites required to cause an infection [99]. A smaller ratio of the EIR to the FOI implies higher transmission efficiency. Most studies have shown that malaria transmission is highly inefficient [97]. Whereas more recently malaria FOI has been estimated from serological data [17, 113] by detecting past exposure to malaria infection, here we focus on estimating malaria FOI from parasitemia data [8, 87, 94].

Despite well-recognized heterogeneity in malaria transmission [96, 100], the FOI is often estimated ignoring intrinsic variability in the individual risk of malaria infection. Heterogeneity in malaria infection arises due to variability in risk factors, including environmental, vector, and host-related factors [117]. Taking these sources of heterogeneity into account [100, 117] in population-based epidemiological studies has been shown to be important [18].

Ronald Ross first published a mathematical model for malaria transmission in 1908 [85, 96]. This model was only firmly established in 1950 by the work of George

Macdonald who used Ross's idea [96]. The "Ross-Macdonald" model describes a simplified set of concepts that serves as a basis for studying mosquito-borne pathogen transmission [96]. Using this concept, mathematical methods to estimate the FOI in relation to the EIR have been proposed by, e.g., Smith *et al.* [97, 98], Keeling and Rohani [47] and Aguas *et al.* [1]. Some of the parameters involved in these models are often unknown and should be estimated from data [36]. A solution proposed by Ross in 1916 is to iterate between two modelling frameworks, that is, mathematical and statistical models [36, 84]. The major difference in these two is that the mathematical models (*priori*) are based on differential equations describing the biological mechanism and causal pathway of transmission, whereas the statistical models (*posteriori*) start by the statistical analysis of observations and work backwards to the underlying cause [36]. These two frameworks complement each other and, here, we provide an explicit link between them.

In this paper, we use the well-known generalized linear mixed model (GLMM) framework [see, e.g., [61]] to estimate the point prevalence accounting for both observed and unobserved heterogeneity and show how the FOI can be obtained from the point prevalence based on a mathematical Susceptible-Infected-Susceptible model. We derive an expression and easy-to-calculate bounds of the FOI for varying parasite clearance distributions. Our results can be used to refine mathematical malaria transmission models.

5.3 Methods

5.3.1 Data source

The results in this paper are based on cohort data from children aged 0.5 to 10 years in three regions in Uganda; Nagongera sub-county, Tororo district; Kihihi sub-county, Kanungu district; and Walukuba sub-county, Jinja district. Data were routinely collected every 3 months (routine visits) and for non-routine clinical (symptomatic) visits. Individuals were tested for the presence of *Plasmodium* parasites using microscopy from August 2011 to August 2014 (3 years). All symptomatic malaria infections were treated with artemether-lumefantrine (AL) anti-malarial medications. A detailed description of these data are given in Chapter 3, Section 3.1 (also see [42, 65]). Given that for clinical visits the sampling process is outcome-dependent (see discussion), the analysis here is restricted to the planned routine visits yielding unbiased estimates (simulation study shown in Chapter 6). These data were analyzed using R (R Core

Team (2015), URL: https://www.R-project.org/) and SAS (SAS Institute Inc 2013. SAS/ACCESS 9.4) statistical software. See Appendix B, Sections B.5 and B.6 for the R code and the SAS macro, respectively.

5.3.2 The SIS model, point prevalence and FOI

A simplified version of malaria transmission can be described using the so-called Susceptible (S) - Infected (I) - Susceptible (S), or SIS, compartmental transmission model. This mathematical model classifies the population into two compartments, i.e., the susceptible (S) and the infected (I) class, which can be graphically depicted as shown in Chapter 2, Figure 2.2. Whereas the rate $\lambda(t)$ is referred to as the force of infection, γ represents the time-invariant clearance rate at which individuals regain susceptibility after clearing malaria parasites from their blood. With $s(t)$ denoting the proportion of susceptible individuals in the population and $i(t)$ the proportion of infected individuals at calendar time t, i.e., the (point) prevalence, then the set of ODEs describing transitions in the compartmental SIS model without demography is given by the system of equations in (2.2). As individuals are either susceptible to infection or malaria infected (at least in the aforementioned simplified SIS model), we have $s(t) = 1 - i(t)$. Substituting this expression for *s(t)* in a system of ODEs in (2.2) yields:

$$\lambda(t) = \frac{i(t)\gamma + i'(t)}{1 - i(t)} \tag{5.1}$$

where $i'(t)$ is the derivative of the point prevalence with respect to t. The force of infection $\lambda(t)$ can thus be estimated using an estimate for the prevalence $i(t)$ and the clearance rate γ. Relaxing the assumption of an exponentially distributed parasite clearance distribution in the SIS model can be done by dividing the I compartment into J sub-compartments, such that infected individuals move from the first sub-compartment I_1 to the second I_2, and later to the J^{th} sub-compartment I_J during the different phases of clearing malaria parasites. Using identical rates γ for the transitions between these sub-compartments and for moving from I_J back to the S compartment results in an Erlang distribution with shape parameter J and rate γ for the time spent in all of the sub-compartments [20]. It is easily shown that equation (5.1) yields an upper bound for the FOI when compared to the aforementioned Erlang clearance distribution (see Appendix B, Section B.4). A lower bound is readily obtained by taking $\gamma = 0$ in equation (5.1) (SI model - see Appendix B, Section B.4). The FOI is thus bounded by $[\lambda_L(t), \lambda_U(t)] = \left[\frac{i'(t)}{1-i(t)}, \frac{i(t)\gamma+i'(t)}{1-i(t)}\right]$. Estimates for both the exponential assumption (upper bound) as well as the lower bound are presented in this chapter. In order to estimate the prevalence $\pi(t) = i(t)$, we use a generalized linear mixed model to account for individual- and household-specific clustering. This will enable us to explicitly model the observed and unobserved heterogeneity in the acquisition of malaria infection.

5.3.3 Generalized linear mixed model

Generalized linear mixed models (GLMMs) extend the well-known generalized linear models by explicitly taking into account (multiple levels of) clustering of observations [61]. Let Y_{ijk} denote the binary response variable indicating parasitemia in the blood (1 if parasites are present - malaria infected; and 0 if not - malaria uninfected) for the i^{th} individual nested in the j^{th} household at the k^{th} visit. Similarly, let X_{ijk} be a $(p+1)\times 1$ vector containing covariate information on p independent variables, and Z_{ijk} be a $q\times 1$ vector of information associated with q random effects. Given the subject-specific random effects $\boldsymbol{b}_{ij}$ and the covariate information X_{ijk}, the random variables $Y_{ijk}|X_{ijk}$ are assumed to be conditionally independent with conditional mean $\pi(X_{ijk}|\boldsymbol{b}_{ij}) = E(Y_{ijk}|X_{ijk}, \mathbf{b}_{ij}) = P(Y_{ijk} = 1|X_{ijk}, \boldsymbol{b}_{ij})$. The GLMM relates the conditional mean to the covariates X_{ijk} and Z_{ijk} as follows:

$$g[\pi(X_{ijk}|\boldsymbol{b}_{ij})] = g[P(Y_{ijk} = 1|X_{ijk}, \boldsymbol{b}_{ij})] = X_{ijk}^T\beta + Z_{ijk}^T\boldsymbol{b}_{ij}. \tag{5.2}$$

Here, g is a monotonic link function (e.g., logit, cloglog and log); $\eta(X_{ijk}, \boldsymbol{b}_{ij}) = X_{ijk}^T\beta + Z_{ijk}^T\mathbf{b}_{ij}$ is the linear predictor with β a vector of unknown regression parameters for the fixed effects; $\boldsymbol{b}_{ij} \sim N(0, \boldsymbol{D})$ a vector of subject-specific random effects for subject i in household j for which elements are assumed to be mutually independent; and $\boldsymbol{D}$ a $q \times q$ variance-covariance matrix [128]. Using equations (5.1) and (5.2), the FOI can be obtained using different link functions. Table 5.1 presents the prevalence and FOI when selecting either the logit, cloglog or log-link function in the GLMM.

Table 5.1: General structures for the FOI according to different link functions in a GLMM framework.

Link function (g)	Prevalence (π)	FOI (λ)
Logit	$\frac{e^\eta}{1+e^\eta}$	$\gamma e^\eta + \eta' \frac{e^\eta}{1+e^\eta}$
Clog-log	$1 - e^{-e^\eta}$	$\gamma\Big(e^{e^\eta} - 1\Big) + \eta' e^\eta$
Log	$1 - e^{-\eta}$	$\gamma(e^\eta - 1) + \eta'$

η refers to the linear predictor $\eta(X_{ijk}|\mathbf{b}_{ij})$ and η' represents the derivative of the linear predictor with respect to the predictor of interest.

5.3.4 Flexible parametric modeling

In a parametric framework such as the GLMM, fractional polynomials provide a very flexible modelling tool for the linear predictor $\eta(X_{ijk}|\boldsymbol{b}_{ij})$ [36, 23, 92]. In this paper, a GLMM using a fractional polynomial of degree one with regard to age, with power p selected from a grid $(-3, -2, -1, -0.5, 0, 0.5, 1, 2, 3)$ using Akaike's information criterion (AIC), is used [2]. More precisely, we use

$$\eta(X_{ijk}|\boldsymbol{b}_{ij}) = \eta(a_{ijk}, l_{ij}|\mathbf{b}_{ij}) = \beta_0 + \beta_1 age^p_{ijk} + \beta_2 l_{ij} + b_{0i(j)} + b_{1i(j)} age^p_{ijk}, \tag{5.3}$$

where $b_{0i(j)}$ is the nested random intercept and $b_{1i(j)}$ is the nested random slope for age. Nesting is done to explicitly acknowledge that individuals make up households. Furthermore, shifted year of birth: l_{ij}, defined as the child's birth year minus the birth year of the oldest child in the cohort (i.e., baseline year 2001), is used in the model to account for the (calendar) time effect since [calendar time] = [birth year] + [age]. The linear predictor (5.3) can be further extended to include additional covariates.

5.3.5 Age-time dependent force of infection

In equation (5.2), the conditional mean $\pi(X_{ijk}|\boldsymbol{b}_{ij})$ is the point prevalence conditional on the random and fixed effects. In this paper, we use the logit-link function, which enables easy calculation of the intra-cluster correlation coefficient (ICC) through an approximation indicating how much the elements within a cluster are correlated [61, 66, 126]. The age-time dependent FOI, conditional on random effects, is estimated by plugging in the parameter estimates obtained from the final fit in equation (5.1). More specifically, using a logit-link, the conditional age-time dependent FOI is estimated as follows:

$$\hat{\lambda}_{l_{ij}}(a_{ijk}|\boldsymbol{b}_{ij}) = \hat{\gamma}\exp[\hat{\eta}(a_{ijk}, l_{ij}|\boldsymbol{b}_{ij})] + \hat{\eta}'(a_{ijk}, l_{ij}|\boldsymbol{b}_{ij})\hat{\pi}_{l_{ij}}(a_{ijk}, l_{ij}|\boldsymbol{b}_{ij}), \tag{5.4}$$

where $\hat{\gamma}$ is an estimate for the clearance rate and $\hat{\pi}_{l_{ij}}(a_{ijk}, l_{ij}|\mathbf{b}_{ij})$ is the estimated age- and time-dependent conditional prevalence. For the lower boundary of FOI, $\hat{\gamma}\exp[\hat{\eta}(a_{ijk}, l_{ij}|\mathbf{b}_{ij})]$ is omitted in equation (5.4). In the above expression, an estimate for the clearance rate γ is required. Previously, Bekessy *et al.* [8] estimated annual clearance rates of 1.643, 0.584 and 0.986 years^{-1} for children aged less than 1 year, 1–4 years and 5–8 years, respectively. Later, Singer *et al.* [94] estimated these rates as 1.917, 1.425 and 2.364 years^{-1} for ages less than 1 year, 1–4 years and 5–8 years, respectively. Sama *et al.* [87] estimated a constant annual clearance rate of 1.825 years^{-1} by assuming an exponential distribution for infection duration or parasite clearance. Most recently, Bretscher *et al.* [11] studied the parametric distributions of the infection durations using Ghanaian data, and concluded based on AIC that a Weibull distribution gave a better fit to the data followed by a gamma distribution, while an exponential one was performing worst. Here, we use both exponential and Erlang clearance distributions to derive estimates for the malaria FOI obtained based on the aforementioned clearance rates as distributional parameters.

Often, an investigator may wish to observe population averaged estimates. Under the random effects framework, this can be achieved by taking the expectation of the conditional estimates (e.g., the FOI in (5.4)) with regard to the random effects distribution resulting into unconditional or marginal estimates. Using the logit-link function, the unconditional (population) force of infection is given by

$$\begin{aligned}\lambda_{l_{ij}}(a_{ijk}) &= E\big[\lambda_{l_{ij}}(a_{ijk}|\boldsymbol{b}_{ij})\big]\\ &= E\big[\gamma\text{exp}\big(\eta(a_{ijk|\boldsymbol{b}_{ij}})\big) + \eta'(a_{ijk}, l_{ij}|\boldsymbol{b}_{ij}) * \pi_{l_{ij}}(a_{ijk}, l_{ij}|\boldsymbol{b}_{ij})\big].\end{aligned} \tag{5.5}$$

Hence, calculation of the marginalized FOI in (5.5), requires integrating out the random effects, $\boldsymbol{b}_{ij}$ over their fitted distribution. This can be done using numerical integration techniques or based on numerical averaging [61].

5.3.6 Model selection

Model building was done using both AIC [88] and a likelihood ratio test for the random effects based on the appropriate mixture of chi-square distributions [112]. Backward model building was performed starting with the random effects and then the fixed effects. The covariates considered in the model building process included study site, age, time since enrolment, shifted birth year (i.e., shifted birth year = birth year - birth year of the oldest child), previous use of AL treatment, and the infectious status at the previous visit. The covariates ‘time since enrolment’ and ‘shifted birth year’ were generated to represent the calendar time, albeit we preferred the latter one since participants were not enrolled at the same time point.

5.4 Results

Of 989 children, recruited between August 2011 to August 2014, 334 (33.8%), 355 (35.9%) and 300 (30.3%) were from Nagongera, Kihihi and Walukuba, respectively. The baseline parasite prevalence among children aged below 5 years was 38.2%, 12.8% and 9.5% for Nagongera, Kihihi and Walukuba, respectively. The monthly parasite prevalence was higher in Nagongera (range: 26.7% to 68.4%) followed by Kihihi (range: 7.0% to 68.0%) and lastly by Walukuba (range: 0% to 42.9%). Other summary statistics including the median monthly prevalence and interquartile range are presented in Table 5.2.

Table 5.2: Recruited number of children, baseline and monthly parasite prevalence, by study site and age group

.

	Nagongera	Kihihi	Walukuba
< 5 years:			
Number	186	188	190
Baseline prevalence[†] (%)	38.2	12.8	9.5
Monthly prevalence[†] (%), range	27.4 - 54.7	7.0 - 64.7	0 - 32.0
Monthly prevalence[†] (%), median (IQR)	40.3 (34.5 - 47.9)	28.2 (18.0 - 43.8)	6.8 (4.4 - 11.9)
5–10 years:			
Number	148	167	110
Baseline prevalence[†] (%)	58.8	18.0	10.9
Monthly prevalence[†] (%), range	26.7 - 68.4	8.3 - 68.0	0 - 42.9
Monthly prevalence[†] (%), median (IQR)	39.1 (34.8 - 47.1)	28.0 (21.6 - 40.0)	11.1 (6.9 - 17.6)
Total:			
Number	334	355	300
Baseline prevalence[†] (%)	47.3	15.2	10.0
Monthly prevalence[†] (%), range	26.7 - 68.4	7.0 - 68.0	0 - 42.9
Monthly prevalence[†] (%), median (IQR)	39.9 (34.6 - 47.9)	28.1 (21.0 - 42.2)	9.0 (5.6 - 14.6)

[†] Parasite prevalence

The parasite prevalence increases with age particularly for children less than 3 years of age and after 7 years of age a decrease is observed (Figure 5.1, panel A). The prevalence increases with calendar time in Kihihi with increasing variability, while it decreases in Walukuba, and slightly increases in Nagongera (Figure 5.1, panel B). These observations suggest a difference in malaria infection risk between the three study sites. Also, the infection risk seems to vary with age and calendar time and it tends to take different trends between sites indicating a possibility for a site-time interaction effect. The relationship with age seems to be non-linear. These observed effects were taken into consideration when building the GLMM.

The mean structure in our model consists of a fractional polynomial of age with power -1 (selected based on AIC) and the following covariates (based on significance testing at 5% significance level): shifted year of birth; infection status at previous visit and AL use; and study site. Goodness-of-fit of the final model was assessed using the ratio of the generalized Chi-square statistic to its degrees of freedom. A

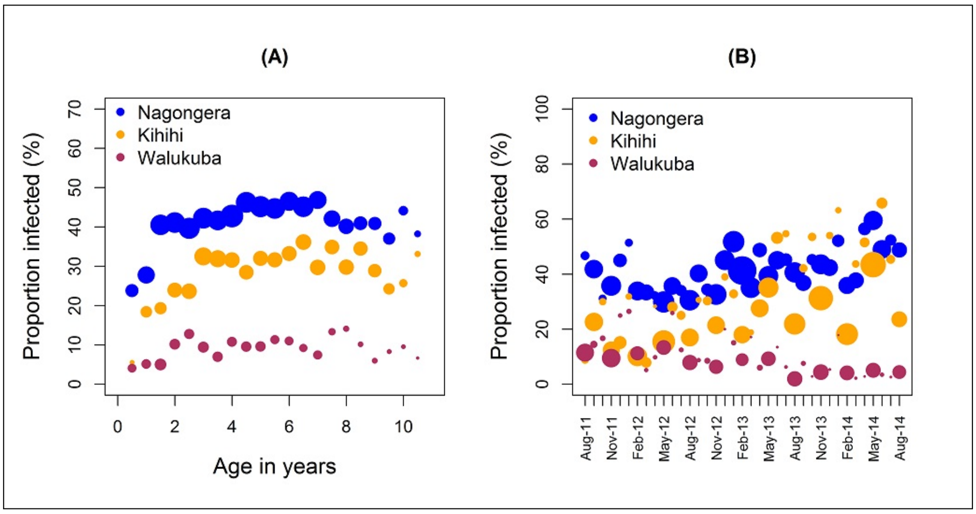

Figure 5.1: Proportion of children infected with malaria parasites (parasitemia) in a cohort followed for 3 years, by study site (Nagongera, Kihihi and Walukuba) in Uganda based on data from August 2011 to August 2014 with the size of the dots proportional to the number of observations. (A) observed parasitemia varying with age; (B) observed parasitemia varying with calendar time.

value of 0.74 was obtained, which is fairly close to 1, indicating that the variability in these data seems to be adequately modelled and little residual over-dispersion remains present [115].

The parameter estimates, standard errors estimates and corresponding test results of the final GLMM fit are shown in Table 5.3. More details about the candidate models can be found in Appendix B (Tables B.1 and B.2) together with the fitted conditional and marginal prevalences for the different AL use categories (Figure B.2). The results in Table 5.3 show an overall significant effect of age and shifted year of birth; the effect of age and shifted year of birth is non-significant and borderline significant, respectively, for Walukuba, whereas the effect of age is significant for Kihihi and Nagongera. Shifted year of birth is significant for Kihihi and non-significant for Nangongera. There is significant heterogeneity in the rate of acquiring malaria infection between households (Walukuba: variance = 2.80; Kihihi: variance = 1.16; Nagongera: variance = 0.21) and between household members (variance = 0.24). The intra-household correlation coefficients are 0.44, 0.25 and 0.06 for Walukuba, Kihihi and Nagongera, indicating moderate, low and very low correlation within households, respectively. The intra-individual correlation coefficients are 0.04, 0.05

and 0.06 for Walukuba, Kihihi and Nagongera, respectively, indicating very low correlation in all sites.

Based on the final model fit and using equations (5.4) and (5.5) both the conditional (given the random effects) and marginal (population averaged) FOIs can be calculated provided that γ can be estimated. However, estimating γ from the same data is not possible due to an identifiability problem: two or more distinct values of γ give rise to the same (log)likelihood (see Figure B.1 in the Appendix B). Therefore, we use γ equal to the annual clearance rates given by Bekessy *et al.* [8] as 1.643, 0.584 and 0.986 years^{-1} for children aged less than 1 year, 1–4 years and 5–10 years, respectively, to calculate the conditional and marginal FOIs. We further conduct a sensitivity analysis by considering different clearance rates ranging from 0 to 3 motivated by the ranges estimated by Bekessy et al. [8], Singer *et al.* [94], Sama *et al.* [87] and Bretscher *et al.* [11] (see Figure 5.4, top row). As discussed before, we also provide lower bounds for the FOI.

Figure 5.2 shows estimates for the marginal FOI together with the corresponding lower bound estimates. We focused on children who were born in the baseline year for graphical reasons. Similar plots were obtained (not shown) for other birth years. Estimates for the lower boundary of the FOI were higher in Nagongera followed by Kihihi and Walukuba. For Nagongera and Walukuba, the lower bound for the FOI was highest for children aged below 1 year and least in those aged 5–10 years, yet. In Kihihi, it is highest among those aged 1–4 years.

Figure 5.2 further shows that in Nagongera and Kihihi, the estimates for the marginal FOI were highest among children aged 5 – 10 years; yet in Walukuba it was highest among those aged below 1 year. The values for the marginal FOI obtained using the upper boundary estimator, stratified by site, age group and the previous infection status and use of AL are given in Table B.3 in the Appendix B. At the extreme, the previously symptomatic children acquire up to 4 infections per year in Nagongera, and 8 infections per year both in Kihihi and Walukuba. Overall, the FOI is highest among the asymptomatic children and smallest among previously symptomatic children across all age groups and sites (Figure 5.2 and Table B.3 in Appendix B). Although Figure 5.2 clearly shows the impact of different distributional assumptions with regard to the clearance time, the lower and upper bound estimates do not fully capture uncertainty around the point estimates. In Table B.4 of the Appendix B, we show the 95% confidence bounds for the age- and time-dependent force of infection.

Table 5.3: Estimates of the fitted GLMM using a fractional polynomial of degree 1 for age and a logit-link function.

Effect	Parameter	log OR (SE)	t-value	P	OR (95% CI)
Intercept	β_0	-3.04 (0.38)	-8.09	<0.001	
Study site (Ref: Walukuba)					
Kihihi	β_1	0.86 (0.43)	2.01	0.045	2.36 (1.02-5.49)
Nagongera	β_2	2.19 (0.40)	5.45	<0.001	8.94 (4.08-19.57)
Infection status at previous visit (Ref: Neg. & No AL)					
Negative + AL	β_3	-0.01 (0.10)	-0.05	0.956	0.99 (0.82-1.21)
Symptomatic	β_4	-0.24 (0.10)	-2.30	0.022	0.78 (0.64-0.97)
Asymptomatic	β_5	1.23 (0.12)	9.94	<0.001	3.43 (2.69-4.37)
Age^{-1}					
Walukuba	β_6	-0.05 (0.83)	-0.06	0.948	0.95 (0.19-4.82)*
Kihihi	β_7	-4.01 (0.87)	-4.62	<0.001	0.02 (0.003-0.10)*
Nagongera	β_8	-1.75 (0.45)	-3.89	0.001	0.17 (0.07-0.42)*
Shifted year of birth†					
Walukuba	β_9	-0.13 (0.06)	-2.00	0.045	0.88 (0.78-1.00)
Kihihi	β_{10}	0.11 (0.04)	2.58	0.010	1.12 (1.13-1.22)
Nagongera	β_{11}	0.04 (0.03)	1.33	0.184	1.04 (0.98-1.10)
Variance components		Variance	Z-value		
Intercepts for subjects	d_{11}	0.24 (0.07)	3.32	<0.001	
Intercepts for households:					
Walukuba	d_{22}	2.80 (0.88)	3.20	0.001	
Kihihi	d_{33}	1.16 (0.28)	4.21	<0.001	
Nagongera	d_{44}	0.21 (0.08)	2.48	0.007	

† birth year - min(birth year)

* note that the OR here should be interpreted at the Age^{-1} level

Figure 5.3 (top row) shows the predicted conditional FOIs for 50 randomly selected individual profiles at each of the three sites based on the lower boundary estimator for the FOI. For graphical purposes, we focused on subjects who were symptomatic at the previous visit and who were born in the baseline year. However, similar plots

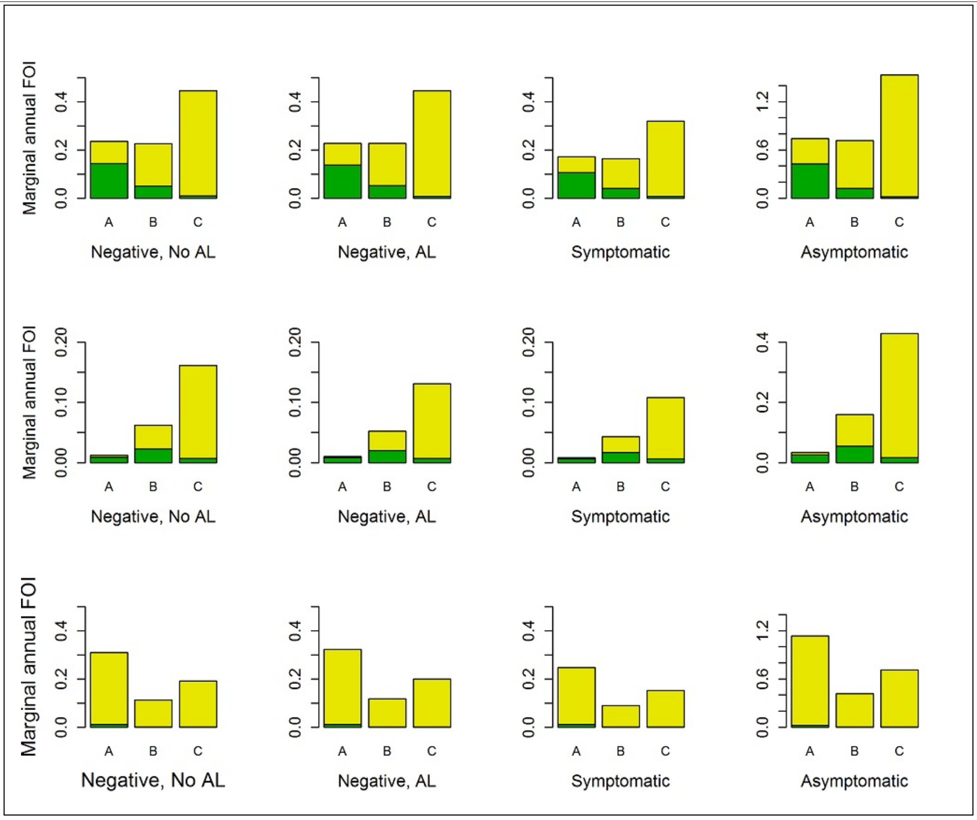

Figure 5.2: The lower bound (green) for the marginal annual FOI and the difference between upper and lower bound (yellow) with full bar showing the upper bound for the FOI, by study site, age group (A: <1 year, B: 1–4 years, and C: 5–10 years) and the infection status at the previous visit and past use of AL (negative and no AL in the past (left column), negative and AL in the past (second left column), symptomatic (second right column) and asymptomatic (right column)) for children assumed to be born in the baseline year (2001). Top row: Nagongera, middle row: Kihihi, bottom row: Walukuba.

are obtained for other levels of the infection status at the previous visit and for different birth years. Figure 5.3 (bottom row) shows the predicted marginal FOIs again based on the lower boundary estimator, by age (continuous scale) and infection status at the previous visit and past AL use. In general, the lower boundary estimator indicates that younger children have the greatest FOI. In all sites, individuals that were asymptomatic at the previous visit have the highest FOI, regardless of age. The depicted conditional FOI curves show that individuals have different profiles, indicating substantial unobserved heterogeneity. The increasing trend in the FOI from 6 months of age is likely attributed to loss of maternal immunity in infants [82].

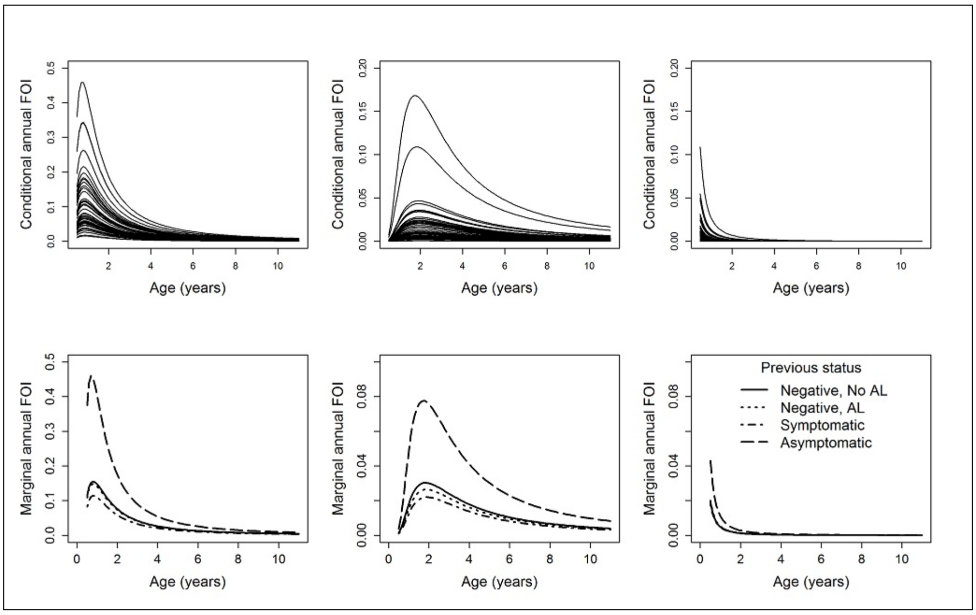

Figure 5.3: Top row: Individual-specific evolutions for the conditional annual FOI obtained using the lower boundary estimator, by study site for children assumed to be symptomatic at the previous visit and who were born in the baseline year (2001). Bottom row: The marginal annual FOI, obtained using the lower boundary estimator, by study site and the infection status at the previous visit and past use of AL (negative and no AL in the past (solid lines), negative and AL in the past (dotted lines), symptomatic (dash-dotted lines) and asymptomatic (long-dashed lines)). Left column: Nagongera, middle column: Kihihi, right column: Walukuba.

Figure 5.4 (top row) shows the marginal FOIs for different clearance rates from 0 up to 3 years^{-1} (y-axis). For graphical purposes, and without loss of generality, we again focused on subjects who were symptomatic at the previous visit and who were born in the baseline year. The colour gradient from green (dark) to brown (light) in Figure 5.4 (top row) corresponds to an increasing FOI. The figure indicates that in Nagongera and Kihihi, children who are below 1 year of age have a lower FOI (green colour) regardless of the presumed clearance rate. Also, in Nagongera and Kihihi, the risk for malaria infection increases with increasing clearance rate, except for the younger children less than 1 to 2 years. In Walukuba, the FOI increases with increasing clearance rate regardless of age.

Figure 5.4 (bottom row) shows how the FOI varied with age group (A, B and C)

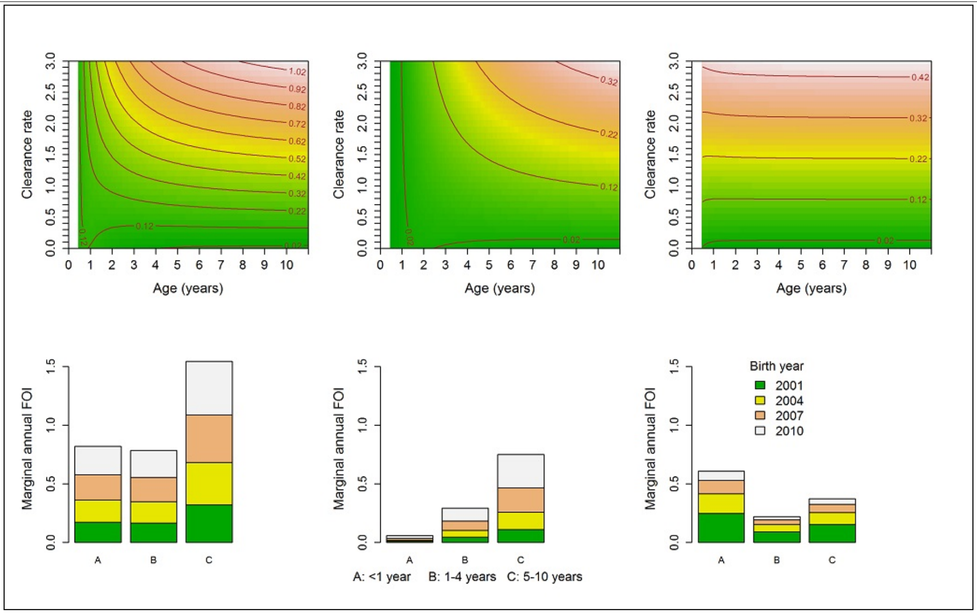

Figure 5.4: Top row: The marginal annual FOI (contour lines) considering different values for the clearance rate ranging from 0 to 3 years^{-1} by study site for individuals assumed to be symptomatic at the previous visit and were born in the baseline year. Bottom row: The marginal annual FOI, obtained using the upper boundary estimator, for individuals assumed to be symptomatic at the previous visit, by study site, birth year (2001, 2004, 2007 and 2010) and by age group (A: <1 year, B: 1–4 years, and C: 5–10 years). Left panel: Nagongera, middle panel: Kihihi, right panel: Walukuba.

and calendar time among subjects assumed to be symptomatic at the previous visit. In Kihihi, the risk of acquiring a new malaria infection is slightly higher for children born in 2010 compared to those born in earlier years across age groups but not for Nagongera and Walukuba. This would be expected since children born at a later year are younger than those born at an earlier year, and hence are at a higher risk of infection.

5.5 Discussion and conclusion

In this chapter, we use data from a cohort study to estimate the malaria FOI among Ugandan children while accounting for observed and unobserved heterogeneity. The results clearly demonstrate the existence of heterogeneity in the acquisition of malaria infections, which is greater between households than between household members. These observations emphasize the claim by Smith *et al.* [96], Smith [100] and White *et al.* [117] that heterogeneity in malaria infection can arise due to several unobserved factors including environmental, vector, and host-related factors. This implies that estimating the malaria transmission parameters assuming homogeneity in the acquisition of infection may yield misleading results.

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

Based on the SIS model, we derived an expression relating the FOI to the prevalence for infectious diseases such as malaria where we cannot assume lifelong immunity. This expression is an extension of the one proposed by Hens *et al.* [36] for a so-called SIR model assuming lifelong immunity after recovery, an assumption, which is untenable for malaria. A compartmental model, which can account for temporally recovery due to prior use of treatment (induced immunity) or due to previous exposure to infection (acquired immunity), that is, Susceptible-Infected-Recovered(Treatment)-Susceptible (SIR(T)S), would potentially offer a better alternative compared to the more restrictive SIS model. However, an SIR(T)S model does not yield a closed-form expression for the point prevalence, and hence, for the force of infection. Nevertheless, the derivations are approximately valid for an SIR(T)S model with short recovery duration (derivations not included here). Consequently, we focused on the SIS model, albeit that we adjusted for the previous infection status and treatment in our model. The standard SIS compartmental model assumes that the clearance rate is exponentially distributed. We derived two estimators for the FOI, which provide a lower and upper boundary for the FOI based on different Erlang distributions for the clearance rate. The lower boundary approximately holds for a scenario in which the clearance rate is small compared to the FOI. Although mathematical models encompassing more complicated and more realistic transmission dynamics for malaria could be considered, we defer their treatment to future research in which we will combine Nonlinear Mixed Model (NLMM) methodology and numerical approaches for the estimation of the model parameters in the presence of unobserved heterogeneity.

The temporal inhomogeneity observed in the data is not in contradiction with the SIS model we used. Heterogeneity, age and temporal aspects are addressed in the GLMM, through the specification of random effects as well as age- and calendar time variables, whereas derivations from the SIS model under endemic equilibrium enable the estimation of the age- and time-dependent force of infection from the estimated age- and time-dependent parasite prevalence. Furthermore, estimation of the reproduction number can be done when focusing on the underlying mechanistic modelling of the FOI. However, we deem this to be beyond the scope of this dissertation. Seasonality is not explicitly modelled here, however, inclusion of a covariate describing the amount of rainfall, due to the absence of a clear distinction between the different seasons, and based on additional information (not part of the PRISM data) would be an interesting topic for further research.

When the clearance rate is considered negligible, the rate at which children get infected is highest among those between 1 and 2 years. When the clearance rate is non-negligible, the infection rate is higher among children older than 5 years in areas with high and medium transmission (e.g., Nagongera and Kihihi) and higher in children below 1 year in areas with low transmission (e.g., Walukuba). In Kihihi, the FOI was least for children aged less than 1 year and it is observed to increase as children grow up from 6 months to 1 year. This could be explained by the fact that children lose maternal immunity in their first year of life [82], which puts them at an increased risk of malaria infection. The higher FOI among children aged 5 years and older could be explained by the fact that these children are often asymptomatic malaria cases and are rarely treated, which makes them reservoirs for infections. This finding concurs with the work by Walldorf *et al.* [114] who reported that children aged 6–15 years were at higher risk of (asymptomatic) infection compared to the younger ones. They concluded that older children represent an underappreciated reservoir of malaria infection and have less exposure to antimalarial interventions.

A higher risk was seen among children in Nagongera compared to those in Kihihi and Walukuba with no significant difference between the latter two sites. This could be explained by the fact that Nagongera is a predominantly rural area with many semi-structured houses and many mosquitoes compared to Walukuba or Kihihi as was noted by Kilama *et al.* [50]. Our results also demonstrated the importance of prior treatment in lowering infection risk due to the post treatment prophylactic effect of longer acting anti-malarials, such as artemether-lumefantrine (AL). For example, children who were previously treated with AL (the symptomatic malaria cases) had a lower risk of getting re-infected compared to those who were asymptomatic or negative at the previous visit.

This study has two major limitations. First, the analysis was based on results of parasite prevalence determined by microscopy, which is less sensitive than molecular methods such as polymerase chain reaction (PCR) or loop-mediated isothermal amplification method (LAMP) [16, 78]. Thus, sub-microscopic infections would not have been detected. This could have resulted into lower estimates of the FOI. In addition, genotyping was not performed to distinguish new and recurrent infections. As a result, the FOI among individuals who were asymptomatic at the previous visit could have been overestimated. Secondly, the unscheduled clinical visits by the symptomatic individuals were triggered by the study outcome (i.e., parasitemia). This creates a dependency between the observation-time and outcome processes. This dependence, if not accounted for, has a potential to introduce bias in the model estimates and hence in the estimation of the FOI. This bias was avoided by dropping clinical visits and by using only routine data, though the infection status and use of treatment during clinical visits was accounted for in the model. This implies that the analysis used less data than was actually available. The latter limitation was later dealt with in Chapter 6 by modelling both the outcome and the observation-time processes concurrently using a joint model [83, 106].

To conclude, we have used longitudinal data from a cohort of Ugandan children to estimate the malaria FOI accounting for both observed and unobserved heterogeneity. First, we show how the FOI relates to parasite prevalence assuming an SIS compartmental model and giving both lower and upper boundaries thereof by relaxing the exponential assumption with regard to the parasite clearance distribution. We estimated the parasite prevalence using a GLMM, whose estimates were used to obtain an estimate for the FOI. The malaria FOI was highest among children aged 1 to 2 years based on the lower boundary estimator, and it was higher among children older than 5 years in areas of high and medium transmission based on the upper boundary estimator. In a low transmission setting, the FOI was highest in children aged below 1 year regardless of the boundary estimator for the FOI. The FOI varied between study sites highest in Nagongera and least in Walukuba. Heterogeneity increases with decreasing FOI and was greater between households than household members. We recommend that estimating the malaria FOI should be done accounting for both observed and unobserved heterogeneity to enable refining existing mathematical models in which the FOI may be unknown.

Chapter 6

Modelling longitudinal binary outcomes with outcome-dependent observation times with an application to a malaria cohort study

6.1 Summary

Malaria follow-up studies typically involve routine visits at pre-scheduled time points and clinical visits upon experiencing malaria-like symptoms. In the latter case, infection triggers outcome assessment, thereby leading to outcome-dependent sampling (ODS). Ordinary methods to analyse such data ignore ODS potentially leading to biased estimates of transmission parameters, hence, inducing an incorrect assessment and evaluation of control strategies. We propose novel methodology to handle ODS using a joint model for the longitudinal binary outcome measured at routine visits and the clinical event times. The methodology is applied to parasitemia data from a cohort of $n = 988$ Ugandan children aged 0.5–10 years in 3 regions (Walukuba - 300,

Kihihi - 355, Nagongera - 333) with varying transmission intensities (entomological inoculation rates of 2.8, 32 and 310 infectious bites per unit year, respectively) collected between 2011–2014. The results indicate that parasite prevalence and force of infection (FOI) increase with age in the high intensity region with highest FOI for 5–10 year olds. For the medium intensity region, the prevalence increases with age and the FOI for the routinely collected data is highest for 5–10 year olds yet for the clinical data, the FOI gradually decreases with increasing age. In the low intensity region, both prevalence and FOI peak at one year of age after which the former remains constant and the latter decreases with age for the clinical observations. In all study sites, the prevalence and FOI are highest among previously asymptomatic children and lowest among their symptomatic counterparts.

6.2 Introduction

Malaria is potentially life-threatening and infections are caused by *Plasmodium* parasites that are transmitted through bites of infected female mosquitoes. In spite of the fact that malaria is a preventable and curable disease for which increased efforts worldwide dramatically reduced malaria incidence (i.e., a reduction of 21% between 2010 and 2015 as reported by WHO), African countries still carry a disproportionately high share of the overall malaria burden. In order to reduce the malaria burden in African countries such as Uganda, a correct assessment and evaluation of the impact of control strategies is essential. Measures of malaria transmission intensity such as the entomological inoculation rate (EIR), the parasite prevalence and the malaria force of infection (FOI) have been used frequently to quantify the impact of various interventions [42, 97]. In general, malaria transmission has been reported to be highly inefficient, meaning that the ratio of EIR to FOI is relatively high [97]. As is the case for other infections, individual- and household-specific heterogeneity in malaria acquisition is often not accounted for in the estimation of the aforementioned epidemiological parameters, albeit that it is well-recognised that variability in environmental and host-related factors, among other sources, has an important effect thereon [65].

Often in clinical trials with follow-up to study (infectious) disease dynamics, study participants are asked to come to the clinic and get examined for malaria infection during scheduled (routine) visits. On top of that, unscheduled (clinical) visits can occur when participants develop symptoms for the disease under consideration, or when they experience symptoms similar to those typically observed for the infection at hand. If infection triggers outcome assessment in between prescheduled follow-up visits, the outcome and observation-time processes are said to be dependent, which in literature is often referred to as outcome-dependent sampling (ODS) (ODS)[106]. Conventional longitudinal methods to analyse repeated measurements for subjects over time assume independence of both processes. Hence, such unscheduled visits, and the ODS they induce, could lead to biased estimation of the epidemiological quantities of interest when it is not appropriately accounted for in the statistical analysis.

Different models have been proposed to address ODS in different experimental settings. For example, Ryu et al. [86] considered studies where the measurement time points are unequally spaced and having a follow-up measurement at any time depends on the history of past visits and outcomes of that individual. These authors discussed limitations of previously proposed models and methods for longitudinal data, such as generalised linear mixed models or generalised estimating equations (GEE), which do not address the association between the outcome and observation time process. Furthermore, these authors proposed a joint model using latent random variables in which the observed follow-up times are described jointly with the longitudinal response data [86]. More recently, Tan [106] considered a joint model with a semi-parametric regression model for the longitudinal outcomes and a recurrent event model for the observation times. Rizopoulos *et al.* [83] stated that an attractive paradigm for the joint modelling of longitudinal and time-to-event processes is the shared parameter framework [127] in which a set of random-effects is assumed to induce the interdependence of the two processes.

Although several authors developed methods to accommodate ODS in various settings, we propose new methodology to cope both with routine and clinical data on malaria infections from a cohort study in Uganda. More specifically, in this chapter focus is on the estimation of the malaria parasite prevalence in three regions of Uganda, accounting for observed and unobserved heterogeneity as done previously [65], while dealing with ODS at the same time. The chapter is organised as follows. Our motivating example is introduced and briefly discussed in Section 6.3. In Section 6.4, we present the general methodology to estimate malaria FOI from parasitaemia data. In Section 6.5, we briefly highlight the impact of ignoring ODS after which our proposed joint model is fitted to the available routine and clinical data on parasite presence in Ugandan children in Section 6.6. Finally, these results are discussed in Section 6.7 together with strengths and limitations of our methodology.

6.3 Motivating example

We consider again longitudinal cohort data from children aged 0.5 to 10 years in three regions in Uganda; Nagongera sub-county, Tororo district; Kihihi sub-county, Kanungu district; and Walukuba sub-county, Jinja district as described in Chapter 3, Section 3.1. The data were collected as part of the Program for Resistance, Immunology, Surveillance and Modelling of malaria (PRISM) study. The aforementioned study regions are characterized by distinct transmission intensities, with the highest intensity reported in Nagongera, followed by Kihihi and with Walukuba having the smallest intensity [65, 42]. The study participants were recruited from 300 randomly selected households (100 per region) located within the catchment areas. In total, $n = 988$ children were followed over time with 300 children in Walukuba, 355 in Kihihi and 333 children in Nagongera. Individuals were routinely tested for the presence of *Plasmodium* parasites using microscopy every three months from August 2011 to August 2014 (3 years). Furthermore, tests were also conducted at unscheduled clinical visits. More detailed information regarding the study design can be found in Kamya *et al.* [42]. Throughout this chapter, the outcome process refers to the occurrence of the longitudinal binary outcome (parasite presence), and the observation-time process relates to the timing of scheduled, i.e., routine, and unscheduled, i.e., clinical visits over the entire follow-up period of the study. The data in this Chapter were analyzed using R (R Core Team (2015), URL: https://www.R-project.org/) and SAS (SAS Institute Inc 2013. SAS/ACCESS 9.4) statistical software. See Appendix C, Sections C.3 and C.4 for the R code and the SAS macro used in this chapter.

6.4 Materials and methods

6.4.1 Malaria dynamics - A simplified transmission model

For the purpose of this paper, we again consider a simplified version of a realistic transmission model to describe malaria infection dynamics. More specifically, following Mugenyi *et al.* [65], a so-called Susceptible (S) - Infected (I) - Susceptible (S), or short SIS, compartmental model dividing the population into two mutually exclusive compartments, i.e., the susceptible (S) and infected(I) class, was used to describe malaria dynamics within the human host. We refer to the discussion of Mugenyi *et al.* [65] for a motivation of the choice of the SIS model and would like to note that the methodology outlined here is more generally applicable in case of other disease dynamics. The schematic diagram depicting the flows between the different states is graphically displayed in Figure 6.1.

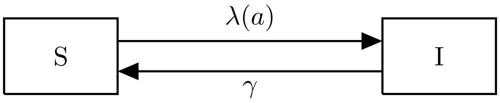

Figure 6.1: A schematic diagram of the SIS compartmental model illustrating the simplified dynamics for malaria transmission: Individuals are born into the susceptible class S and move to the infected state I at rate $\lambda(a)$, after which they become susceptible again at rate γ.

Herein, the force of infection $\lambda(a)$ represents the instantaneous rate at which individuals flow from the susceptible compartment S to the infected state I at age a (i.e., the age-specific rate at which individuals are infected with malaria parasites through effective mosquito bites). Furthermore, γ represents a time- and age-invariant clearance rate at which individuals regain susceptibility after clearing malaria parasites from their blood. Let $s(a)$ denote the proportion of susceptible individuals in the population and $i(a)$ the proportion of infected individuals of age a, i.e., the (point) parasite prevalence, then the set of ordinary differential equations (ODEs), modifying the ODEs in (2.2) (see Chapter 2) replacing t by a is described as follows;

$$\begin{aligned} s'(a) &= -\lambda(a)s(a) + \gamma i(a), \\ i'(a) &= \lambda(a)s(a) - \gamma i(a), \end{aligned} \tag{6.1}$$

Hence, one can easily derive the following expression for the age-dependent force of infection in terms of the point prevalence $i(a)$:

$$\lambda(a) = \frac{i(a)\gamma + i'(a)}{1 - i(a)}, \tag{6.2}$$

using $s(a) + i(a) = 1$. Hence, writing down a model for the point prevalence $i(a)$ will imply a functional form for the underlying force of infection $\lambda(a)$ depending on the clearance rate γ.

6.4.2 Parasite prevalence and routine visits

Consider the binary random variable Y_{ij} representing an indicator for the presence of malaria parasites for individual i at (routine) visit j. Consequently, for scheduled routine visits, $(Y_{ij}|a_{ij}, \boldsymbol{x}_i, \boldsymbol{b}_i) \sim B(1, i(a_{ij}|\boldsymbol{x}_i, \boldsymbol{b}_i))$, where a_{ij} represents the age of individual i at visit j, $\boldsymbol{x}_i$ represents a $(p \times 1)$-vector of covariate information for individual $i = 1, \ldots, n$, and $\boldsymbol{b}_i$ a $(q \times 1)$-vector of individual-specific random effects.

In order to model the parasite prevalence, we formulate a generalized linear mixed model with cloglog-link as follows:

$$\text{cloglog}\left[i(a_{ij}|\boldsymbol{x}_i, \boldsymbol{b}_i)\right] = \eta_{ij} = h(a_{ij}; \boldsymbol{\theta}) + \boldsymbol{\beta}^T \boldsymbol{x}_i + \boldsymbol{b}_i^T \boldsymbol{z}_i, \tag{6.3}$$

where $\boldsymbol{\beta}$ is a vector of unknown regression parameters and $\boldsymbol{z}_i$ an individual-specific $(q \times 1)$ design vector for $\boldsymbol{b}_i$ which is a vector of individual-specific normally distributed random effects, i.e., $\boldsymbol{b}_i \sim N(\boldsymbol{\mu}, \boldsymbol{D})$ thereby addressing the association among repeated measurements over time within the same individual. Here, the variance-covariance matrix $\boldsymbol{D}$ is assumed to have zero elements, except for the variances on the main diagonal. Moreover, $h(a_{ij}; \boldsymbol{\theta})$ is a known function describing the age-effect with parameter vector $\boldsymbol{\theta}$. Note that the calendar time effect can be introduced in the linear predictor by means of the shifted birth year of the ith individual, implying the prevalence, and equivalently the FOI, to depend on both age and calendar time [65]. In Table 6.1, we present some common parametric distributions and their implied functional forms for $h(a_{ij}; \boldsymbol{\theta})$ based on model (6.3) and the corresponding baseline infection risk $\lambda_0(a_{ij}) = h'(a_{ij}; \boldsymbol{\theta}) \exp\left[h(a_{ij}; \boldsymbol{\theta})\right]$ (derived under the assumption of no parasite clearance). In the absence of unscheduled clinical visits ($n_i = n_{i(r)}$, i.e., the number of routine visits for individual i), or under the assumption of independence between the observation time process and the outcome process, we can simply estimate model parameters using maximum likelihood techniques, thereby maximizing a marginal likelihood function with the following individual likelihood contributions:

Table 6.1: Distributional assumptions regarding the underlying age-specific malaria force of infection.

Distribution	$\boldsymbol{\theta}$	$h(a_{ij}; \boldsymbol{\theta})$	$\lambda_0(a_{ij})$
Exponential	$\theta_1 > 0$	$\log(\theta_1 a_{ij})$	θ_1
Weibull	$\theta_1, \theta_2 > 0$	$\log(\theta_1 a_{ij}^{\theta_2})$	$\theta_1 \theta_2 a_{ij}^{\theta_2 - 1}$
Gompertz	$\theta_1 > 0, -\infty < \theta_2 < +\infty$	$\log\left[\frac{\theta_1}{\theta_2}\left(e^{\theta_2 a_{ij}} - 1\right)\right]$	$\theta_1 e^{\theta_2 a_{ij}}$
Log-logistic	$\theta_1, \theta_2 > 0$	$\log\left\{\log\left[1 + (\theta_1 a_{ij})^{\theta_2}\right]\right\}$	$\frac{\theta_1 \theta_2 (\theta_1 a_{ij})^{\theta_2 - 1}}{1 + (\theta_1 a_{ij})^{\theta_2}}$
Fractional polynomial	$\theta_2 < 0$	$\theta_2 a_{ij}^{-1}$	$-\theta_2 a_{ij}^{-2} e^{\theta_2 a_{ij}^{-1}}$

$$\begin{aligned}
L_{1i}(\boldsymbol{\beta}, \boldsymbol{\theta}|\boldsymbol{y}_i, a_{ij}, \boldsymbol{x}_i) &= \int_{\boldsymbol{b}_i} f(\boldsymbol{y}_i|a_{ij}, \boldsymbol{x}_i, \boldsymbol{b}_i) \mathrm{g}(\boldsymbol{b}_i) d\boldsymbol{b}_i \\
&= \int_{\boldsymbol{b}_i} \left\{ \prod_{j=1}^{n_i} f(y_{ij}|a_{ij}, \boldsymbol{x}_i, \boldsymbol{b}_i) \right\} \mathrm{g}(\boldsymbol{b}_i) d\boldsymbol{b}_i,
\end{aligned}$$

with

$$f(y_{ij}|a_{ij}, \boldsymbol{x}_i, \boldsymbol{b}_i) = i(a_{ij}|\boldsymbol{x}_i, \boldsymbol{b}_i)^{y_{ij}} \times [1 - i(a_{ij}|\boldsymbol{x}_i, \boldsymbol{b}_i)]^{(1-y_{ij})},$$
$$\mathrm{g}(\boldsymbol{b}_i) = \frac{1}{\sqrt{|2\pi \boldsymbol{D}|}} e^{-\frac{1}{2}(\boldsymbol{b}_i - \boldsymbol{\mu})^T \boldsymbol{D}^{-1}(\boldsymbol{b}_i - \boldsymbol{\mu})},$$

where y_{ij} is the observed binary outcome for individual i at routine visit $j = 1, \dots, n_i$, and $i(a_{ij}|\boldsymbol{x}_i, \boldsymbol{b}_i)$ is the conditional parasite prevalence. Numerical integration techniques are employed to perform integration over the random effects distribution $\mathrm{g}(\boldsymbol{b}_i)$. In the following subsection, we specifically focus on clinical visits and how to address ODS.

6.4.3 Outcome-dependent sampling and clinical visits

As mentioned before, clinical visits due to symptomatic malaria infections, or malaria-like events giving rise to symptoms similar to those observed for malaria, can not be treated in the same way as described in Section 6.4.2. Let t_{ij} represents the time-at-risk for an individual i for which the jth visit is clinical, and c_{ij} an indicator having value one for an unscheduled clinical visit and 0 for routine data. For the purpose of illustration, we assume that t_{ij} is known, albeit that this is not the case in practice, and statistical ways to deal with this are outlined below. The probability density function for the random variable T_{ij}, suppressing dependence on covariates $\boldsymbol{x}_i$ and $c_{ij} = 1$ for simplicity, is given by:

$$f(t_{ij}|a_{ij}, y_{ij}, \boldsymbol{b}_i) = \left[(1 - \pi_0)\lambda^*(a_{ij} + t_{ij}|\boldsymbol{b}_i) e^{-\int_{a_{ij}}^{a_{ij}+t_{ij}} \lambda^*(u|\boldsymbol{b}_i)du}\right]^{y_{ij}} \times \pi_0^{1-y_{ij}},$$

where $\lambda^*(u|\boldsymbol{b}_i) \equiv \lambda^*(u|\boldsymbol{x}_i, \boldsymbol{b}_i) = e^{\boldsymbol{b}_i^T \boldsymbol{z}_i + \boldsymbol{\zeta}^T \boldsymbol{x}_i} \lambda_0^*(u)$ is the conditional time-varying malaria force of infection under the proportional hazards assumption (with $\boldsymbol{\zeta}$ a vector of model parameters) and π_0 denotes the probability of a malaria-like clinical visit for which no malaria parasites are present in the blood. For the purpose of this manuscript, we will not model the dependence of the probability of having a malaria-like event $\pi_0 = P(Y_{ij} = 0|C_{ij} = 1)$ on the observed covariate information a_{ij} and $\boldsymbol{x}_i$. Different distributional assumptions can be made regarding the time-at-risk distribution, such as, e.g., exponential, Weibull, Gompertz, among others, which also relates to the selected functional form for $h(a_{ij}; \boldsymbol{\theta})$ in the outcome process model (see Section 6.4.2 and Table 6.1). In order to align the models for both processes, the baseline infection risk $\lambda_0^*(u)$ for the observation time process can be of the same type as $\lambda_0(u)$, albeit that distributional parameters, say $\boldsymbol{\vartheta}$, are allowed to be different.

Note that more flexible parametric shapes for $h(a_{ij};\boldsymbol{\theta})$, such as, e.g., using fractional polynomials, could result in non-standard non-negative distributions for the malaria infection times, albeit that unconstrained optimisation could lead to negative FOI estimates. In the statistical analyses, we include parametric fractional polynomials as an alternative to the standard event time distributions.

For outcomes $(\boldsymbol{t}_{i(c)},\ \boldsymbol{y}_{i(c)})$ that are derived from the clinical visits $j = 1, \ldots, n_{i(c)}$, where $n_i = n_{i(r)} + n_{i(c)}$ having $n_{i(r)}$ and $n_{i(c)}$ the number of routine and clinical visits for individual i, respectively, the likelihood function has contributions:

$$L_{2i}(\boldsymbol{\zeta}, \boldsymbol{\vartheta} | \boldsymbol{t}_{i(c)}, \boldsymbol{y}_{i(c)}, \boldsymbol{a}_{i(c)}, \boldsymbol{x}_i) = \int_{\boldsymbol{b}_i} \left\{ \prod_{j=1}^{n_i(c)} f(t_{ij(c)} | a_{ij(c)}, y_{ij(c)}, \boldsymbol{x}_i, \boldsymbol{b}_i) \right\} \mathrm{g}(\boldsymbol{b}_i) d\boldsymbol{b}_i, \quad (6.4)$$

where $\boldsymbol{t}_{i(c)}$ and $\boldsymbol{a}_{i(c)}$ are the vectors of time-at-risk and age values at which the individual becomes at risk for the jth clinical event, respectively. Finally, the likelihood for the joint model including both information on routine and clinical visits is obtained by combining likelihood contributions as described before:

$$\begin{aligned} L_{3i}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\zeta}, \boldsymbol{\vartheta} | \boldsymbol{t}_i, \boldsymbol{y}_i, \boldsymbol{a}_i, \boldsymbol{x}_i, \boldsymbol{c}_i) = \int_{\boldsymbol{b}_i} & \left\{ \prod_{j=1}^{n_i} f(y_{ij} | a_{ij}, \boldsymbol{x}_i, \boldsymbol{b}_i)^{1-c_{ij}} \times \right. \\ & \left. \prod_{j=1}^{n_i(c)} f(t_{ij} | a_{ij}, y_{ij}, \boldsymbol{x}_i, \boldsymbol{b}_i)^{c_{ij}} \right\} \mathrm{g}(\boldsymbol{b}_i) d\boldsymbol{b}_i, \end{aligned} \quad (6.5)$$

at least under the assumption that each malaria event contributes solely to one of the two components (i.e., routine or clinical process) in the likelihood.

As mentioned previously, the time-at-risk for a specific clinical event (i.e., a symptomatic malaria infection) is not precisely known. More specifically, malaria infection times are interval-censored which needs to be taken into account in the statistical analyses through the modification of the likelihood function. For more details on how the interval-censoring has been treated in the analyses, the reader is referred to Appendix C, Section C.2.1.

6.5 Simulation study

In order to study the impact of ignoring ODS, we set up a simulation study which is inspired by the PRISM data under consideration. More specifically, we generate

$M = 1000$ datasets including $n_m \equiv n = 1000$ individuals per simulated dataset ($m = 1, \ldots, M$). Furthermore, we consider a simulation setting in which exponential infection times occur during a follow-up period of 1800 days (≈ 5 years) and with an average duration until acquiring a new infection of about 365 days (1 year: $\lambda_0 = \lambda_0^* = \exp(-5.9) = 0.0027$). Parasite clearance times are exponentially distributed with a mean duration of infectiousness equal to 50 days ($\gamma = 0.02$). Based on the generated infection histories for the individuals, routine and clinical visits are obtained. More specifically, routine visits are scheduled every 90 days and parasite presence is recorded based on the current status at the time of data collection. Varying probabilities for having a symptomatic malaria episode are considered in the simulation whereby symptomatic observations at unscheduled time points were considered as clinical visits (i.e., $P = 20\%, 40\%, 60\%, 80\%, 100\%$). Hence, asymptomatic malaria cases were only included when detected during the routine visits. No malaria-like events were generated such that all clinical visits are due to symptomatic malaria infections (i.e., $\pi_0 = 0$). Individual-specific random intercepts $b_i \sim N(\mu, \sigma_b^2)$, $i = 1, \ldots, n$, with $\mu = -\sigma_b^2/2$ implying a unit mean for the lognormal random terms e^{b_i}, are introduced to induce correlation between repeated measurements for the same subject ($\sigma_b^2 = 0.25$). If a single infection is contributing to both the routine and clinical process (i.e., consecutive observations C^+ and R^+, or vica versa), hence leading to two dependent observations, we drop the second one in Scenario 4. However, without additional information, we cannot determine whether individuals already recovered and got re-infected in between such visits, thereby potentially underestimating the force of infection. We performed a sensitivity analysis given the simulation scenario at hand in order to deduce the time period in which consecutive positive routine and clinical observations can be considered to be the result of a single malaria infection. From this exercise, a period of 35 days is assumed to be optimal (see Appendix C, Figure C.1 for more details thereon). This observation is supported by the literature where 100% recovery rate was reported on day 28 following anti-malaria treatment [57, 67].

Table 6.2: Overview of the different scenarios, corresponding loglikelihood functions to be maximised (see Section 6.4) and parasitaemia data that is included in the analyses.

Scenario	Loglikelihood function	Parasitemia data
1	$ll_1(\boldsymbol{\beta}, \boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{X}) = \sum_{i=1}^{n} \log\left[L_{1j}(\boldsymbol{\beta}, \boldsymbol{\theta} \mid \boldsymbol{y}_i, \boldsymbol{x}_i)\right]$	Routine
2	$ll_1(\boldsymbol{\beta}, \boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{X}) = \sum_{i=1}^{n} \log\left[L_{1j}(\boldsymbol{\beta}, \boldsymbol{\theta} \mid \boldsymbol{y}_i, \boldsymbol{x}_i)\right]$	Routine & clinical†
3	$ll_2(\boldsymbol{\zeta}, \boldsymbol{\vartheta} \mid \boldsymbol{t}, \boldsymbol{y}, \boldsymbol{a}, \boldsymbol{X}) = \sum_{i=1}^{n} \log\left[L_{2i}(\boldsymbol{\zeta}, \boldsymbol{\vartheta} \mid \boldsymbol{t}_i, \boldsymbol{y}_i, \boldsymbol{a}_i, \boldsymbol{x}_i)\right]$	Clinical
4	$ll_3(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\zeta}, \boldsymbol{\vartheta} \mid \boldsymbol{t}, \boldsymbol{y}, \boldsymbol{a}, \boldsymbol{X}) = \sum_{i=1}^{n} \log\left[L_{3i}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\zeta}, \boldsymbol{\vartheta} \mid \boldsymbol{t}_i, \boldsymbol{y}_i, \boldsymbol{a}_i, \boldsymbol{x}_i)\right]$	Routine & clinical

† Scenario does not take ODS into account

6.5.1 Simulation results

Hereunder, we present results from fitting the four scenarios based on the three different likelihoods in Section 6.4 to the simulated data. All models converged for all simulation runs. In Table 6.3, we show the simulation results for the four different scenarios described in Table 6.2 with varying percentages of symptomatic malaria infections. Scenario 2 including both routine and clinical data without accounting for ODS performs worse compared to Scenario 1 in which only routine data is used. Hence, ignoring ODS leads to biased estimates of both the baseline hazard as well as population-averaged hazard functions. Note that Scenario 1 is not influenced by the percentage of symptomatic infections, simply since these clinical infections are not accounted for therein. Our proposed model for the analysis of both clinical and routine parasitaemia data (Scenario 4) outperforms Scenarios 1 and 2 in terms of bias and precision (and consequently MSE) for the baseline hazard function and population-averaged hazard λ_p, at least when P = 60% or higher, and leads in all cases to a reduction in bias. In Scenario 4, we add clinical information to the readily available routine data (i.e., larger sample size), resulting in a lower MSE, bias and empirical variance for the model parameters compared to Scenario 1. The loss of perfomance in Scenario 4 compared to Scenario 3 as P>60% can be explained by the nature of the data since noise is added by combining time-to-event data (which is analysed separately in Scenario 3) with interval-censored data.

Table 6.3: Simulaton results for the different models showing mean estimates for the marginal or population-averaged FOI (λ_p), variance of the random intercepts (σ_b^2), and the corresponding mean squared error (MSE), bias and empirical variance. P represents the percentage of symptomatic infections. †: all data except for positive routine observations following a positive clinical visit, or positive clinical observations following a positive routine visit within a 35 day period. N represents the total number of observations over all individuals averaged over the M datasets.

	Scenario 1 routine data	Scenario 2 all data	Scenario 3 clinical data	Scenario 4 all data†
$P = 20\%$	$n = 20,000$	$n = 21,832$	$n = 1,832$	$n = 21,781$
Population-averaged hazard λ_p				
Mean estimate $\bar{\hat{\lambda}}_p$	0.0034	0.0047	0.0020	0.0028
Mean estimate $\bar{\hat{\sigma}}_b^2$	0.3806	0.4058	0.4727	0.5228
MSE(λ)x10^5	0.0078	0.2460	0.1375	0.0111
Bias(λ)x10^5	23.8252	155.6670	116.2956	9.0705
Var(λ)x10^5	0.0021	0.0037	0.0023	0.0102

Table 3 continued.

	Scenario 1	Scenario 2	Scenario 3	Scenario 4
	routine data	all data	clinical data	all data[†]
$P = 40\%$	$n = 20,000$	$n = 22,678$	$n = 2,678$	$n = 22,576$
Population-averaged hazard λ_p				
Mean estimate $\bar{\bar{\lambda}}_p$	0.0034	0.0060	0.0026	0.0028
Mean estimate $\bar{\bar{\sigma}}_b^2$	0.3806	0.4046	0.3646	0.4402
MSE(λ)x10^5	0.0078	0.8106	0.0270	0.0093
Bias(λ)x10^5	23.8252	283.8542	50.4948	7.0972
Var(λ)x10^5	0.0021	0.0048	0.0016	0.0088
$P = 60\%$	$n = 20,000$	$n = 23,520$	$n = 3,520$	$n = 23,370$
Population-averaged hazard λ_p				
Mean estimate $\bar{\bar{\lambda}}_p$	0.0034	0.0072	0.0029	0.0028
Mean estimate $\bar{\bar{\sigma}}_b^2$	0.3806	0.3908	0.3076	0.3861
MSE(λ)x10^5	0.0078	1.6500	0.0063	0.0081
Bias(λ)x10^5	23.8252	405.4125	22.5147	5.4008
Var(λ)x10^5	0.0021	0.0064	0.0012	0.0078
$P = 80\%$	$n = 20,000$	$n = 24,368$	$n = 4,368$	$n = 24,169$
Population-averaged hazard λ_p				
Mean estimate $\bar{\bar{\lambda}}_p$	0.0034	0.0084	0.0030	0.0028
Mean estimate $\bar{\bar{\sigma}}_b^2$	0.3806	0.3780	0.2713	0.3496
MSE(λ)x10^5	0.0078	2.7799	0.0017	0.0072
Bias(λ)x10^5	23.8252	526.4963	8.4274	4.2648
Var(λ)x10^5	0.0021	0.0064	0.0012	0.0064
$P = 100\%$	$n = 20,000$	$n = 25,218$	$n = 5,218$	$n = 24,971$
Population-averaged hazard λ_p				
Mean estimate $\bar{\bar{\lambda}}_p$	0.0034	0.0072	0.0029	0.0028
Mean estimate $\bar{\bar{\sigma}}_b^2$	0.3831	0.3659	0.2445	0.3265
MSE(λ)x10^5	0.0083	4.1944	0.0007	0.0065
Bias(λ)x10^5	23.8252	646.9211	1.0401	3.5431
Var(λ)x10^5	0.0021	0.0064	0.0012	0.0064

6.6 Data application

In this section, we apply the proposed joint model in Scenario 4 to the observed Ugandan malaria parasitaemia data presented in Section 6.3. The covariates considered in the model building process included study site, age, shifted birth year (i.e., shifted birth year = birth year - birth year of the oldest child), previous use of Artemether-Lumefantrine (AL) treatment, and the infectious status at the previous visit. The covariate 'shifted birth year' was generated to represent the calendar time (see also [65] for details concerning this modelling strategy). Let S_i represent the study site (1 = Walukuba, 2 = Kihihi, 3 = Nagongera), a_{ij} the child's age in years, l_{ij} the shifted birth year, P_{ij} the previous infection status and use of AL (1 = Negative & no AL, 2 = Negative + AL, 3 = Symptomatic, 4 = Asymptomatic) for individual i at visit j. Different parametric distributional assumptions regarding the infection times are explored (i.e., leading to various functional forms for $h(a_{ij};\boldsymbol{\theta})$, and equivalently, for the underlying malaria force of infection) thereby allowing for different distributional parameters $\boldsymbol{\theta}$ and $\boldsymbol{\vartheta}$ for the outcome and infection time process, respectively. Since malaria transmission intensity differs between the three sites (see, e.g., [42, 65]), site-stratified analyses were performed, and model comparison was done based on AIC and BIC in order to select the most appropriate functiontal form for $h(a_{ij};\boldsymbol{\theta})$. Table C.2 in Appendix C provides the site-specific fit statistics for the different models.

In Table 6.4, we show the parameter and standard error estimates (between brackets) for the joint model under Scenario 4, thereby having Gompertz baseline hazard functions $\lambda_0(a)$ and $\lambda_0^*(a)$ for the three study sites (see Table C.2 in Appendix C for more details on the AIC- and BIC-values for the candidate models). A significant effect of shifted year of birth has been observed for Kihihi and Nagongera in both processes, and not for the low transmission intensity site Walukuba. The infection status at previous visit was included only for the outcome process resulting in an overall significant effect at all sites (p-value <0.001). In total, 35%, 43% and 62% of the observed visits were classified as clinical visits in Walukuba, Kihihi and Nagongera, respectively. Of those observed clinical visits, 87%, 48% and 54% are malaria-like clinical visits implying that no evidence of malaria infection was found in children coming to the clinic due to malaria-like symptoms. The estimated values for π_0 are equal to 89% (95% confidence interval (CI): 87% – 91%), 58% (95% CI: 56% – 60%) and 65% (95% CI: 63% – 67%) for Walukuba, Kihihi and Nagongera, respectively, which are quite in line with the observed empirical probabilities.

Figure 6.2 depicts the estimated marginal prevalence by age for children assumed to be born in the baseline year (2001) which were symptomatic (top row) or asymptomatic (bottow row) at the previous visit, and by study site (left to right: Nagongera, Kihihi and Walukuba). The curves are drawn for Scenario 2 (solid blue line) and Scenario 4 (dashed red line). In general, the parasite prevalence increases with increasing age in areas with high (Nagongera) and medium (Kihihi) transmission intensity, though the prevalence is fairly constant for Scenario 4 in the latter case. In Walukuba, the prevalence first increases to a plateau from 6 months up to 2 years after the prevalence remains constant. From the graphs, it is clear that small differences exist between the two scenarios in terms of the estimated marginal prevalence.

In Figure 6.3, we show the estimated marginal force of infection for the outcome (routine) process based on expression (6.2). We consider annual parasite clearance rates (γ) of 1.643, 0.584 and 0.986 years^{-1} for children aged less than 1 year, 1–4 years and 5–10 years, respectively [8]. On top of that, the marginal FOI estimated from the time-to-event process is shown in the bottom row. The marginal FOI for the outcome process increases with increasing age at least for Nagongera and Kihihi, and it is highest among children in age group 5–10 years or those that were previously asymptomatic (gray bars) and least in their symptomatic counterparts (brown bars) in all study areas. For the time process, the marginal FOI in Nagongera is close to zero and constant with time at risk, at least for children aged 1 year when becoming at risk. For children at a higher age, the FOI tends to increase more steeply with increasing time at risk and age. However, the FOI for the time process is highest among children aged about one year in medium (Kihihi) and low (Walukuba) transmission intensities, after which it decreases gradually with increasing time at risk for children of all ages. More specifically, when children are older, the infection risk is smaller as compared to their younger counterparts given the specific time at risk.

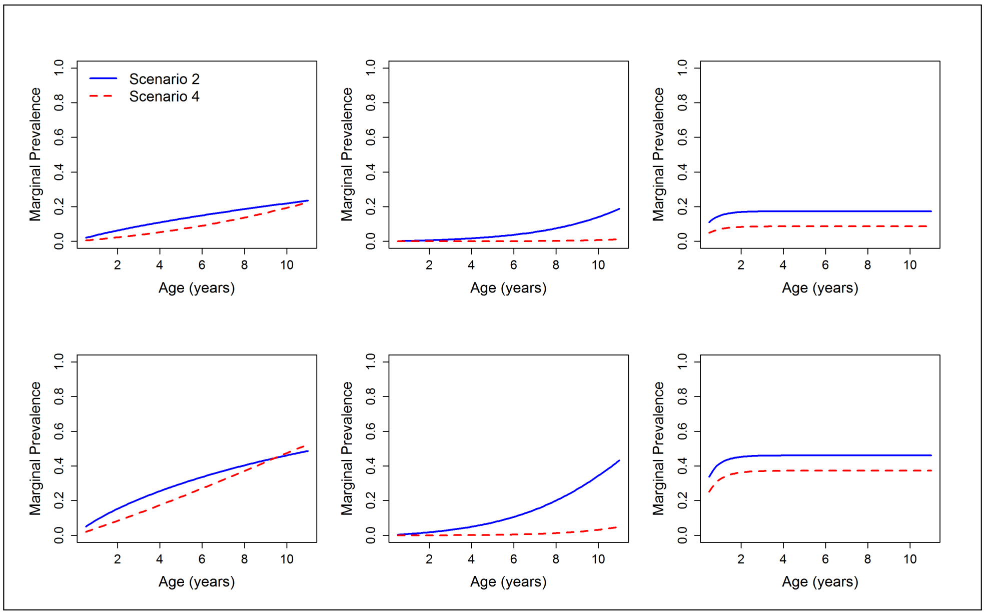

Figure 6.2: Estimated marginal prevalence for children assumed to be born in the baseline year (2001) by age, study site and symptomatic (top row) or asymptomatic (bottom row) at the previous visit. Left to right column: Nagongera, Kihihi and Walukuba.

Table 6.4: Application to PRISM data: results showing parameter and standard error (s.e.) estimates from the joint model (scenario 4) assuming Gompertz-distributed infection times for Walukuba, Kihihi and Nagongera.

Effect		Parameter	Estimate (s.e.)	t-value	p-value
Walukuba (Gompertz):					
Infection status at the previous visit (Ref = Negative & No AL treatment in past):	Negative + AL	β_1	0.12 (0.26)	0.47	0.639
	Symptomatic	β_2	-0.66 (0.46)	-1.43	0.153
	Asymptomatic	β_3	1.33 (0.26)	5.13	< 0.001
Shifted year of birth		β_4	-0.09 (0.05)	-1.93	0.054
Age		θ_1	0.16 (0.23)	0.71	0.480
		θ_2	-1.54 (1.96)	-0.78	0.434
Shifted year of birth^t		ζ	-0.23 (0.17)	-1.37	0.171
Age^t		ϑ_1	36.68 (63.44)	0.58	0.564
		ϑ_2	-0.28 (0.14)	-2.01	0.046
Probability of a malaria-like clinical visit		π_0	0.89 (0.01)	102.18	< 0.001
Variance for random intercepts for subjects		d_{11}	0.25 (0.18)	1.38	0.167
Variance for random intercepts for households		d_{22}	1.22 (0.37)	3.32	0.001
Kihihi (Gompertz):					
Infection status at the previous visit (Ref = Negative & No AL treatment in past):	Negative + AL	β_1	-0.30 (0.06)	-5.29	< 0.001
	Symptomatic	β_2	-1.08 (0.14)	-7.94	< 0.001
	Asymptomatic	β_3	0.65 (0.12)	5.37	< 0.001
Shifted year of birth		β_4	0.49 (0.04)	13.03	< 0.001
Age		θ_1	$4e$-6 ($2e$-6)	2.61	0.009
		θ_2	0.56 (0.04)	13.57	< 0.001
Shifted year of birth^t		ζ	-0.25 (0.04)	-6.94	< 0.001
Age^t		ϑ_1	18.06 (6.98)	2.59	< 0.001
		ϑ_2	-0.26 (0.03)	-7.91	< 0.001
Probability of a malaria-like clinical visit		π_0	0.58 (0.01)	53.58	< 0.001
Variance for random intercepts for subjects		d_{11}	0.27 (0.05)	9.83	< 0.001
Variance for random intercepts for households		d_{22}	4.28 (0.35)	12.30	< 0.001
Nagongera (Gompertz):					
Infection status at the previous visit (Ref = Negative & No AL treatment in past):	Negative + AL	β_1	-0.46 (0.12)	-3.80	< 0.001
	Symptomatic	β_2	-1.24 (0.13)	-9.35	< 0.001
	Asymptomatic	β_3	0.15 (0.13)	1.22	0.222
Shifted year of birth		β_4	0.19 (0.08)	2.44	0.015
Age		θ_1	0.02 (0.02)	1.23	0.219
		θ_2	0.17 (0.11)	1.58	0.115
Shifted year of birth^t		ζ	0.92 (0.05)	18.31	< 0.001
Age^t		ϑ_1	$6e$-4 ($3e$-4)	1.88	0.061
		ϑ_2	1.03 (0.05)	19.64	< 0.001
Probability of a malaria-like clinical visit		π_0	0.65 (0.01)	83.81	< 0.001
Variance for random intercepts for subjects		d_{11}	0.94 (0.15)	6.42	< 0.001
Variance for random intercepts for households		d_{22}	0.37 (0.15)	2.43	0.016

^{t} Time-to-event model effects

6.7 Discussion and conclusion

In this chapter, we have proposed novel methodology to account for outcome-dependent sampling (ODS) when estimating malaria transmission parameters such as, for example, the parasite prevalence and the force of infection (FOI) in case of longitudinal cohort data with routine (scheduled) and clinical (unscheduled) visits. A simulation study, inspired by parasitemia data from a cohort of Ugandan

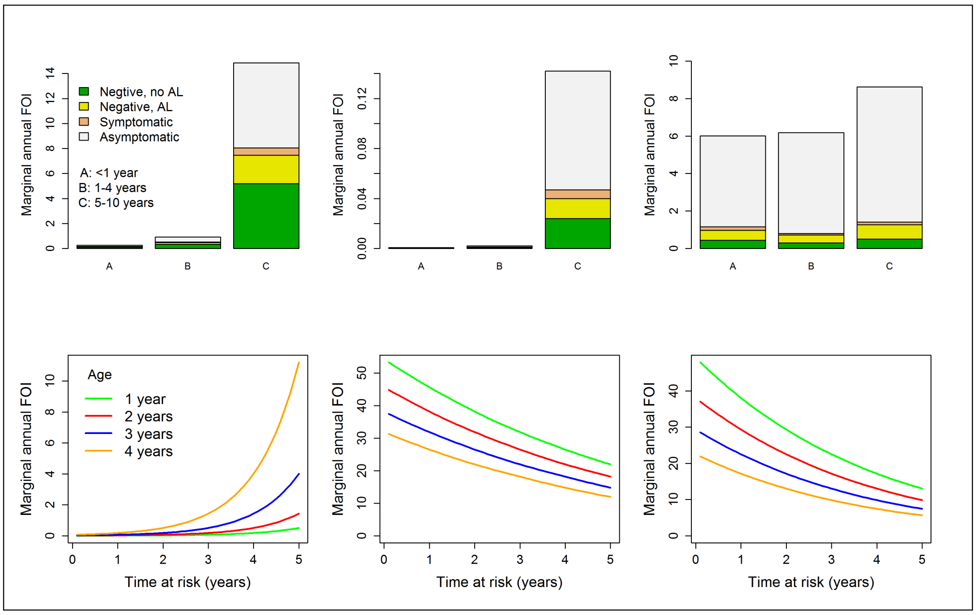

Figure 6.3: Estimated marginal FOI by time at risk and age when becoming at risk for the next malaria infection based on Scenario 4 and for children assumed to be born in the baseline year (2001). Top row: marginal FOI based on outcome process†. Bottom row: marginal FOI based on time process†. Left to right column: Nagongera, Kihihi and Walukuba.
†The terms outcome and time processes are used here to mean that the estimates are respectively obtained from the outcome and the time components of the joint model.

children who were tested for malaria parasites (parasitaemia) during such visits, was conducted in which different parametric functions were considered to model the age-specific malaria prevalence and FOI while accounting for both observed and unobserved heterogeneity. Though the simulation results indicate that scenario 1 and 3 perform as good as scenario 4, we preferred the latter because the former two use less data than are available, hence affecting precision, moreover scenario 1 was already applied by Mugenyi *et al.* [65] (see Chapter 5). The results clearly indicate that ignoring ODS leads to biased estimates for the marginal force of infection, hence, leads to an incorrect assessment and evaluation of malaria control strategies. We demonstrate that the bias can be reduced by using a joint model in which both outcome (routine) and observation-time (clinical) components are present. In order to reduce the bias, we propose to treat malaria events within a period of 35 days after a first malaria infection as being part of the same infection. This is supported by the results presented by Maiga *et al.* [57] and Ndiaye *et al.* [67].

The results show that both the malaria parasite prevalence and the FOI increase with increasing age in an area of high (Nagongera) transmision intensity. The FOI is highest in children aged 5–10 years and it becomes higher as children grow older or are at risk for a longer time. For an area with medium (Kihihi) transmision intensity, whereas the parasite prevalence and the FOI for the outcome process increase with increasing age, the FOI for the clinically observed infections (time process) is highest among children aged 1 year and it gradually decreases with increasing age and time at risk. In Walukuba which is an area of low transmission intensity, however, the prevalence and FOI at least for the time process peak at the age of about one year, after which the former remains constant while the latter decreases with increasing age and exposure time at least when based on the time process. Further, both the prevalence and FOI are highest among the children with asymptomatic infections, and lower among the symptomatic ones or the previously treated children. These results are in line with those reported previously by Mugenyi *et al.* [65]. The high prevalence and FOI estimated among the older children particularly in area with high transmission is in agreement with the work by Doolan *et al.* [21]. These authors show that children older than 5 years act as reservoirs for malaria parasites or asymptomatic infections and are rarely treated, hence leading to an increased infection risk. On the other hand, the decrease in the clinically observed infections (time process), that is FOI, as age increases in both the medium and low transmission intensities can be attributed to acquired immunity due to past infections or increase in age as discussed by Doolan *et al.* [21]. In our statistical analyses, we also estimated the probability of a malaria-like event π_0 which were quite in line with the empirical proportions in the three regions. However, π_0 also encompasses potential differences in reporting among the regions as individuals with symptoms will not always visit the clinic.

One way to avoid bias in estimating the epidemiological parameters of interest is the use of routine data only. This approach has been demonstrated in the past [65]. However, our methodology allows for a proper integration of all clinical data, including malarialike events, in the data analysis, thereby enabling the study of potential varying effects for symptomatic (detected at clinical visits) and asymptomatic (derived from routine data) infections. From our statistical analyses of the PRISM data, the hypothesis of differential age-effects for symptomatic and asymptomatic infections is highly supported as models forcing the effects to be the same are clearly outperformed by their unrestricted and more flexible counterparts. Though the estimated parasite prevalence is in line with the observed data, more flexible parametric or semi-parametric baseline hazard functions could be considered in both processes which is an interesting avenue for further research. Furthermore, Mugenyi *et al.* [65] used a generalized linear mixed model to model the observed parasite prevalence after which the force of infection is derived using equation (6.2). One of the shortcomings in this paper is the simplification of no parasite clearance when deriving the baseline hazard function for the time process. This could lead to an underestimation of the respective FOI. We consider this as an interesting avenue for future research.

The proposed joint model can be extended to have a shared parameter ψ to model the dependence between the outcome and observation time processes through individual- and process-specific random effects $\boldsymbol{b}_{i1}$ and $\boldsymbol{b}_{i2}$, respectively (see, e.g, Wulfsohn *et al.* [127]). In that way, one can allow the process-specific random effects to act at different levels. However, applying this approach to the PRISM data forced us to exclude the household-specific random effect for convergence reasons. The models presented in this manuscript ($\psi = 1$) outperformed the ones with different process-specific individual-level random effects in all regions, except for Nagongera, and the significance of covariates was not altered (not shown here).

In general, symptomatic and asymptomatic infections can be well mapped unto clinical and routine visits albeit with the limitation that whereas clinical visits are symptomatic, routine visits aren't necessarily all asymptomatic. Using a model in which we combine both routine and clinical visits, this can be explicitly taken into account. Admittedly, whereas combining data from both sources yields more efficient estimates as shown in our simulation study, our simulation study also shows that using routine and clinical data separately results in moderate to small bias and that the benefit of combining both seems to be minimal for settings in which the proportion asymptomatic is large compared to when the proportion asymptomatic is small. Ideally, with our approach (Scenario 4), we are working in between routine data (little information on time of exposure) and clinical data (more information on time of exposure) by nature of the data. Our approach is therefore generally applicable.

Chapter 7

Discussion and further research

The disease burden, largely due to malaria, in many African countries, including Uganda, is one of the major factors hindering development. Despite global efforts to eliminate malaria, less work has been done to estimate and understand how the various transmission parameters for the disease vary with various risk factors while accounting for unobserved heterogeneity. In this thesis, we therefore apply both statistical and mathematical models to estimate infectious disease parameters for the transmission of malaria among children in a sub-Saharan African country, Uganda. We focus on estimating the malaria force of infection (FOI), parasite prevalence and parasite clearance rate. Additionally, we estimated the malaria-related mortality hazard rate among Ugandan children dying from all causes between the age of 29 days and 14 years. We have described several methods for estimating some of these parameters while accounting for both observed and unobserved heterogeneity in the risk of acquiring the malaria infection. Whenever possible, relationships between mathematical and statistical models are derived, thereby enabling substitution of parameter estimates from the latter approach into the former, hence refining estimates in the former models where accounting for various factors and heterogeneity is rather difficult. The analyses in this thesis have been motivated by two datasets including data collected from different regions with varying malaria transmission intensities in Uganda.

According to Talisuna *et al.* [105], data on malaria death were still lacking in 2015 to enable future funding towards eradicating malaria in Uganda. The scant data on malaria death could be due to many deaths in Uganda and probably elsewhere occurring outside health centres and are thus missing in the national mortality registries [12]. To improve on the availability of these data, we applied survival methods using death data extracted from a verbal autopsy study to estimate the hazard of malaria-related death and the determinants in the presence of other causes of death among Ugandan children dying between the age of 29 days and 14 years. This analysis is presented in Chapter 4. The fact that some of these children died from causes other than malaria implied the use of competing risks analysis. Our results show that malaria was the leading cause of death, contributing to half of the total number of deaths in this population. Half of the deaths occurred before the age of 2 years, approximately 93.7% of the deaths happened in children under five, and more than half occurred outside hospitals or health facilities. Children who died at home were older than those who died in hospitals or health facilities, while those who died on the way to the hospital/HF were younger than those who died while admitted. This could be because children that are usually taken to the hospitals or health facilities are very sick, and their chances of dying are thus also higher. Children without a fever were older when they died compared to their counterparts who had a fever. For future research, we recommend that malaria surveillance at healthcare facilities and within communities be strengthened in Uganda to capture accurate data on malaria mortality. Additionally, educating caregivers about the symptoms of malaria and the importance of seeking care promptly to ensure appropriate diagnosis and effective treatment of malaria should continue to be a priority.

In Chapter 5, we estimate the age- and time-dependent hazards of acquiring malaria infection or the malaria force of infection while accounting for both observed and unobserved heterogeneity. Here, age and time, which represent the child's age and calendar time in years, are entered into the model as continuous variables. We used data from a cohort of children aged 0.5–10 years that was tested for the presence of malaria parasites at three sites in Uganda (see details of the data in Section 3.1). By assuming an SIS model, we show how the FOI relates to the point prevalence, allowing for the estimation of FOI by modelling the prevalence using a generalized linear mixed model (GLMM). We give two bounds for the FOI by using an Erlang distribution [20] to describe the clearance of parasites. The findings indicated that the malaria FOI significantly varied with both age and time, and it was highest among children aged 5–10 years in an area of high transmission and highest in those aged below 1 year in an area of low transmission. Additionally, heterogeneity was greater between households than within households, and it increased with decreasing risks of malaria infection. The finding about heterogeneity strengthens the work by Smith *et al.* [96], Smith [100] and White *et al.* [117], who stated that heterogeneity in malaria infection can arise due to several unobserved factors, including environmental, vector, and host-related factors. Therefore, estimating the malaria transmission parameters by ignoring the heterogeneity in acquisition of the infection may lead to the wrong results and conclusions. One limitation for the results in this chapter is that only routine data were used to avoid the potential bias introduced by outcome-dependent sampling (ODS), and less data were used than were actually available. ODS arises from unscheduled clinical visits which are triggered by the study outcome. This limitation was later dealt with in Chapter 6 by modelling both the routine and clinical data by using a joint model. For future research, we recommend that both observed and unobserved heterogeneity be accounted for when estimating the malaria FOI to refine existing mathematical models. Additionally, for future research, the methodology presented in this chapter can be extended to include an estimation of the reproduction number (R_0) when focusing on the underlying mechanistic modelling of the FOI. This could later be used for example, in guiding the calculation of the proportion of children that will need malaria vaccination to prevent the sustained spread of the infection. This proportion is calculated as $1 - 1/R_0$ [31].

In Chapter 6, we extend the work covered in Chapter 5 by accounting for outcome-dependent sampling (ODS) when estimating the age-specific malaria parasite prevalence and FOI. In this part of the analysis, we demonstrated that ignoring ODS in follow-up studies where the outcome triggers clinical observations can lead to biased estimates, with a consequence of making incorrect conclusions. We have proposed a methodology that allows for a proper integration of all clinical data, including malaria-like events, in the data analysis, thereby enabling the study of potential varying effects for symptomatic (detected at clinical visits) and asymptomatic (derived from routine data) infections. With the help of a simulation study motivated by malaria data from a cohort of Ugandan children (see data source in Section 3.1), our methodology gives the smallest bias, especially when positive malaria results observed within 35 days were considered to be of the same infection, a result that is supported by Maiga *et al.* [57] and Ndiaye *et al.* [67]. We explored different parametric functions of age to model the age-specific malaria parasite prevalence and the FOI while accounting for both observed and unobserved heterogeneity. The results indicate that parasite prevalence and FOI increase with age in the high intensity region with highest FOI for 5–10-year olds. For the medium intensity region, the prevalence increases with age, and the FOI for the routinely collected data is highest for 5–10-year olds; yet, for the clinical data, the FOI gradually decreases with increasing age. In the low intensity region, both prevalence and FOI peak at one year of age, after which the former remains constant, and the latter decreases with age for the clinical observations. At all study sites, the prevalence and FOI were highest among the previously asymptomatic children and lowest among their symptomatic counterparts. These results support the hypothesis of differential age-effects for symptomatic and asymptomatic infections as models forcing the effects to be the same were outperformed by their unrestricted and more flexible counterparts. These findings are in line with those presented by Mugenyi *et al.* [65] and the possible explanations for high prevalence and FOI among the olds or the asymptomatic cases therein, also discussed in Chapter 5, equally apply here. A shortcoming of the results presented in Chapter 6 is the simplification of no parasite clearance when deriving the baseline hazard function for the time process, which could lead to an underestimation of the respective FOI. For future research, a more realistic baseline hazard function for the time process should be considered, thereby allowing for the parasite clearance rate. We further recommend that for future research, ODS in clinical observations should be addressed in addition to heterogeneity when estimating the malaria transmission parameters. This consideration will lead to the correct assessment and evaluation of the impact of malaria control strategies, resulting in elimination of the disease, its burden and hinderance to both socio and economic development.

The works presented in Chapters 4, 5 and 6 are linked in the sense that they all use children's data on malaria in Uganda and that they all describe the burden of malaria as a function of age in years. Specifically, the age profiles for mortality presented in Chapter 4 (e.g., see Figure 4.2) are similar to those for the malaria FOI presented in Chapter 5 (e.g., see Figure 5.3) and in Chapter 6 for the clinical malaria FOI in areas of medium and low transmission (see Figure 6.3, bottom row). These figures show that the mortality and FOI both move together and are highest among young children, which is expected. However, mortality and the FOI from the routinely collected data or in areas with high transmission intensity tend to be inversely correlated (e.g., see Figure 4.2 vs Figure 6.3, top row). For example, on the top row of Figure 6.3, the FOI is highest in older children, yet mortality is lower among these ages (see Figure 4.2). Ideally, one would expect a higher mortality rate for a higher FOI. However, it has been documented elsewhere that older children act as reservoirs for malaria infection [114], and because of their increased immunity, they have a higher chance of surviving death due to malaria even if they frequently get infected.

In this thesis, we have used a binary outcome for the presence of malaria parasites and we have relied on only the human host when estimating malaria transmission parameters. Additionally, we have used compartmental models without demographic characteristics, such as birth, death and migration rates. However, for future research, the methodologies described in this thesis can be extended to handle other types of outcome data (e.g., count data) and to estimate transmission parameters in the mosquito-vector as well as to include demographic parameters.

Bibliography

[1] Aguas R, White LJ, Snow RW, Gomes MGM. Prospects for malaria eradication in sub-Saharan Africa. Plos One. 2008;3(3).

[2] Akaike H. New look at statistical-model identification. Ieee Transactions on Automatic Control. 1974;Ac19(6):716-23.

[3] Alberti C, Timsit JF, Chevret S. Survival analysis - the log rank test. Revue Des Maladies Respiratoires. 2005;22(5):829-32.

[4] Ambroisethomas P. Diagnosis and Seroepidemiologic Study of Malaria by Immunofluorescence, Indirect Hemagglutination and Immuno-Enzymology. Israel Journal of Medical Sciences. 1978;14(6):690-1.

[5] Andersen PK, Geskus RB, de Witte T, Putter H. Competing risks in epidemiology: possibilities and pitfalls. International Journal of Epidemiology. 2012;41(3):861-70.

[6] Bai XF, Tsiatis AA. A log rank type test in observational survival studies with stratified sampling. Lifetime Data Analysis. 2016;22(2):280-98.

[7] Baiden F, Bawah A, Biai S, Binka F, Boerma T, Byass P, Chandramohan D, Chatterji S, Engmann C, Greet D, Jakob R, Kahn K, Kunii O, Lopez AD, Murray CJ, Nahlen B, Rao C, Sankoh O, Setel PW, Shibuya K, Soleman N, Wright L, Yang G. Setting international standards for verbal autopsy. Bulletin of the World Health Organization. 2007;85(8):570-1.

[8] Bekessy A, Molineaux L, Storey J. Estimation of incidence and recovery rates of Plasmodium falciparum parasitaemia from longitudinal data. Bulletin of the World Health Organization. 1976;54(6):685-93.

[9] Boor D. A practical guide to splines. New York: Springer Verlag; 1978.

[10] Box GEP, Draper NR. Empirical Model-Building and Response Surfaces. Wiley, 1987.

[11] Bretscher MT, Maire N, Chitnis N, Felger I, Owusu-Agyei S, Smith T. The distribution of Plasmodium falciparum infection durations. Epidemics. 2011;3(2):109-18.

[12] Byass P, Herbst K, Fottrell E, Ali MM, Odhiambo F, Amek N, Hamel MJ, Laserson KF, Kahn K, Kabudula C, Mee P, Bird J, Jakob R, Sankoh O, Tollman SM. Comparing verbal autopsy cause of death findings as determined by physician coding and probabilistic modelling: a public health analysis of 54 000 deaths in Africa and Asia. Journal of Global Health. 2015;5(1):010402.

[13] Boyce MR, O'Meara WP. Use of malaria RDTs in various health contexts across sub-Saharan Africa: a systematic review. BMC Public Health. 2017;17(1):470.

[14] CDC Website https://www.cdc.gov/malaria/about/biology/. Access date: December 20, 2016

[15] Coleman RE, Sattabongkot J, Promstaporm S, Maneechai N, Tippayachai B, Kengluecha A, Rachapaew N, Zollner G, Miller RS, Vaughan JA, Thimasarn K, Khuntirat B. Comparison of PCR and microscopy for the detection of asymptomatic malaria in a Plasmodium falciparum/vivax endemic area in Thailand. Malaria Journal. 2006;5.

[16] Coleman RE, Sattabongkot J, Promstaporm S, Maneechai N, Tippayachai B, Kengluecha A, Rachapaew N, Zollner G, Miller RS, Vaughan JA, Thimasarn K, Khuntirat B. Comparison of PCR and microscopy for the detection of asymptomatic malaria in a Plasmodium falciparum/vivax endemic area in Thailand. Malaria Journal. 2006;5.

[17] Corran P, Coleman P, Riley E, Drakeley C. Serology: a robust indicator of malaria transmission intensity. Trends Parasitol. 2007;23(12):575-82.

[18] Coutinho FAB, Massad E, Lopez LF, Burattini MN, Struchiner CJ, Azevedo-Neto RS. Modelling heterogeneities in individual frailties in epidemic models. Mathematical and Computer Modelling. 1999;30(1-2):97-115.

[19] Cox DR. Regression Models and Life-Tables. Journal of the Royal Statistical Society Series B-Statistical Methodology. 1972;34(2):187.

[20] de Smith MJ. Statistical Analysis Handbook - a web-based statistics resource. Winchelsea , UK.: The Winchelsea Press; 2015.

[21] Doolan DL, Dobano C, Baird JK. Acquired immunity to malaria. Clin Microbiol Rev. 2009;22(1):13-36.

[22] Egger JR, Ooi EE, Kelly DW, Woolhouse ME, Davies CR, Coleman PG. Reconstructing historical changes in the force of infection of dengue fever in Singapore: implications for surveillance and control. Bulletin of the World Health Organization. 2008;86(3):187-96.

[23] Faes C, Hens N, Aerts M, Shkedy Z, Geys H, Mintiens K, Laevens H, Boelaert F. Estimating herd-specific force of infection by using random-effects models for clustered binary data and monotone fractional polynomials. Journal of the Royal Statistical Society Series C-Applied Statistics. 2006;55:595-613.

[24] Felger I, Maire M, Bretscher MT, Falk N, Tiaden A, Sama W, Beck HP, Owusu-Agyei S, Smith TA. The dynamics of natural Plasmodium falciparum infections. Plos One. 2012;7(9):e45542.

[25] Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. Journal of the American Statistical Association. 1999;94(446):496-509.

[26] Gallup JL, Sachs JD. The economic burden of malaria. American Journal of Tropical Medicine and Hygiene. 2001;64(1-2):85-96.

[27] Gentilini M, Richardlenoble D, Reviron J, Farragi M. Serological Diagnosis of Malaria by and Indirect Immunofluorescence Test Using Pl Bergher Antigen. Revue Francaise De Transfusion. 1976;19(2):363-7.

[28] Geskus RB. Data Analysis with Competing Risks and Intermediate States. Series CHCB, editor: Chapman and Hall/CRC; 2015.

[29] Ghai RR, Thurber MI, El Bakry A, Chapman CA, Goldberg TL. Multi-method assessment of patients with febrile illness reveals over-diagnosis of malaria in rural Uganda. Malaria Journal. 2016;15.

[30] Goel MK, Khanna P, Kishore J. Understanding survival analysis: Kaplan-Meier estimate. International Journal of Ayurveda Research. 2010;1(4):274-8.

[31] Goldstein E, Paur K, Fraser C, Kenah E, Wallinga J, Lipsitch M. Reproductive numbers, epidemic spread and control in a community of households. Mathematical Biosciences. 2009;221(1):11-25.

[32] Gray RJ. A Class of K-Sample Tests for Comparing the Cumulative Incidence of a Competing Risk. Annals of Statistics. 1988;16(3):1141-54.

[33] Halpern J, Brown BW, Jr. Designing clinical trials with arbitrary specification of survival functions and for the log rank or generalized Wilcoxon test. Contemporary Clinical Trials. 1987;8(3):177-89.

[34] Hastie T, Tibshirani R. Generalized aditive models. London: Chapman and Hall; 1990.

[35] Hens N, Aerts M, Faes C, Shkedy Z, Lejeune O, Van Damme P, Beutels P. Seventy-five years of estimating the force of infection from current status data. Epidemiology and Infection. 2010;138(6):802-12.

[36] Hens N., Shkedy Z., Aerts M., Faes C., Van Damme P., P. B. Modeling infectious disease parameters based on serological and social contact data: A modern statistical perspective: Springer; 2012.

[37] Hery, Alexandre. The Uganda poverty assessment report 2016. Uganda: Artfield Graphics Ltd; 2016.

[38] Howden BP, Vaddadi G, Manitta J, Grayson ML. Chronic falciparum malaria causing massive splenomegaly 9 years after leaving an endemic area. Medical Journal of Australia. 2005;182(4):186-8.

[39] Jeanne LD, Berry A, Dutoit E, Leclerc F, Beaudou J, Leteurtre S, Camus D, Benoit-Vical F. Molecular method for the diagnosis of imported pediatric malaria. Medecine Et Maladies Infectieuses. 2010;40(2):115-8.

[40] Jegede AS, Oshiname FO, Sanou AK, Nsungwa-Sabiiti J, Ajayi IO, Siribie M, Afonne C, Serme L, Falade CO. Assessing Acceptability of a Diagnostic and Malaria Treatment Package Delivered by Community Health Workers in Malaria-Endemic Settings of Burkina Faso, Nigeria, and Uganda. Clinical Infectious Diseases. 2016;63(suppl 5):S306-S11.

[41] Kain KC, Harrington MA, Tennyson S, Keystone JS. Imported malaria: Prospective analysis of problems in diagnosis and management. Clinical Infectious Diseases. 1998;27(1):142-9.

[42] Kamya MR, Arinaitwe E, Wanzira H, Katureebe A, Barusya C, Kigozi SP, Kilama M, Tatem AJ, Rosenthal PJ, Drakeley C, Lindsay SW, Staedke SG, Smith DL, Greenhouse B, Dorsey G. Malaria transmission, infection, and disease at three sites with varied transmission intensity in Uganda: implications for malaria control. American Journal of Tropical Medicine and Hygiene. 2015;92(5):903-12.

[43] Kaplan EL, Meier P. Nonparametric-Estimation from Incomplete Observations. Journal of the American Statistical Association. 1958;53(282):457-81.

[44] Kass RE, Raftery AE. Bayes Factors. Journal of the American Statistical Association. 1995;90(430):773-95.

[45] Katrak S, Murphy M, Nayebare P, Rek J, Smith M, Arinaitwe E, Nankabirwa JI, Kamya M, Dorsey G, Rosenthal PJ, Greenhouse B. Performance of Loop-Mediated Isothermal Amplification for the Identification of Submicroscopic Plasmodium falciparum Infection in Uganda. Am J Trop Med Hyg. 2017;97(6):1777-81.

[46] Katureebe A, Zinszer K, Arinaitwe E, Rek J, Kakande E, Charland K, Kigozi R, Kilama M, Nankabirwa J, Yeka A, Mawejje H, Mpimbaza A, Katamba H, Donnelly MJ, Rosenthal PJ, Drakeley C, Lindsay SW, Staedke SG, Smith DL, Greenhouse B, Kamya MR, Dorsey G. Measures of Malaria Burden after Long-Lasting Insecticidal Net Distribution and Indoor Residual Spraying at Three Sites in Uganda: A Prospective Observational Study. Plos Medicine. 2016;13(11):e1002167.

[47] Keeling MJ, Rohan P. Modeling infectious diseases in humans and animals. Princeton University Press, 41 William Street, Princeton, New Jersey 08540, 2008.

[48] Keiding N. Age-Specific Incidence and Prevalence - a Statistical Perspective. Journal of the Royal Statistical Society Series a-Statistics in Society. 1991;154:371-412.

[49] Kelly-Hope LA, McKenzie FE. The multiplicity of malaria transmission: a review of entomological inoculation rate measurements and methods across sub-Saharan Africa. Malaria Journal. 2009;8.

[50] Kilama M, Smith DL, Hutchinson R, Kigozi R, Yeka A, Lavoy G, Kamya MR, Staedke SG, Donnelly MJ, Drakeley C, Greenhouse B, Dorsey G, Lindsay SW. Estimating the annual entomological inoculation rate for Plasmodium falciparum transmitted by Anopheles gambiae s.l. using three sampling methods in three sites in Uganda. Malaria Journal. 2014;13(1):111.

[51] Klein JP, Andersen PK. Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. Biometrics. 2005;61(1):223-9.

[52] Laoboonchai A, Kawamoto F, Thanoosingha N, Kojima S, Scott Miller RR, Kain KC, Wongsrichanalai C. PCR-based ELISA technique for malaria diagnosis of specimens from Thailand. Tropical Medicine and International Health. 2001;6(6):458-62.

[53] Lee Y, Nelder JA, Pawitan Y. Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood. London: Chapman & Hall/CRC; 2006.

[54] Liang KY, Zeger SL. Longitudinal Data-Analysis Using Generalized Linear-Models. Biometrika. 1986; 73: 13-22.

[55] Lipsitz SR, Fitzmaurice GM, Ibrahim JG, Gelber R, Lipshultz S. Parameter estimation in longitudinal studies with outcome-dependent follow-up. Biometrics. 2002;58(3):621-30.

[56] Logan BR, Zhang MJ, Klein JP. Regression models for hazard rates versus cumulative incidence probabilities in hematopoietic cell transplantation data. Biology of Blood and Marrow Transplantation. 2006;12(1):107-12.

[57] Maiga AW, Fofana B, Sagara I, Dembele D, Dara A, Traore OB, Toure S, Sanogo K, Dama S, Sidibe B, Kone A, Thera MA, Plowe CV, Doumbo OK, Djimde AA. No evidence of delayed parasite clearance after oral artesunate treatment of uncomplicated falciparum malaria in Mali. American Journal of Tropical Medicine and Hygiene. 2012;87(1):23-8.

[58] Mathison BA, Pritt BS. Update on Malaria Diagnostics and Test Utilization. Journal of Clinical Microbiology. 2017;55(7):2009-17.

[59] McMorrow ML, Aidoo M, Kachur SP. Malaria rapid diagnostic tests in elimination settings–can they find the last parasite? Clin Microbiol Infect. 2011;17(11):1624-31.

[60] Miller LH, Ackerman HC, Su XZ, Wellems TE. Malaria biology and disease pathogenesis: insights for new treatments. Nature Medicine 2013;19(2):156-67.

[61] Molenberghs G, Verbeke G. Models for discrete longitudinal data. New York: Springer, Series in Statistics; 2005.

[62] Mpimbaza A, Filler S, Katureebe A, Kinara SO, Nzabandora E, Quick L, Ratcliffe A, Wabwire-Mangen F, Chandramohan D, Staedke SG. Validity of verbal autopsy

procedures for determining malaria deaths in different epidemiological settings in Uganda. Plos One. 2011;6(10):e26892.

[63] Mpimbaza A, Filler S, Katureebe A, Quick L, Chandramohan D, Staedke SG. Verbal Autopsy: Evaluation of Methods to Certify Causes of Death in Uganda. Plos One. 2015;10(6):e0128801.

[64] Mueller I, Schoepflin S, Smith TA, Benton KL, Bretscher MT, Lin E, Kiniboro B, Zimmerman PA, Speed TP, Siba P, Felger I. Force of infection is key to understanding the epidemiology of Plasmodium falciparum malaria in Papua New Guinean children. Proceedings of the National Academy of Sciences of the United States of America. 2012;109(25):10030-5.

[65] Mugenyi L, Abrams S, Hens N. Estimating age-time-dependent malaria force of infection accounting for unobserved heterogeneity. Epidemiology and Infection. 2017:1-18.

[66] Musca SC, Kamiejski R, Nugier A, Meot A, Er-Rafiy A, Brauer M. Data with hierarchical structure: impact of intraclass correlation and sample size on type-I error. Frontiers in Psychology. 2011;2:74.

[67] Ndiaye JL, Faye B, Gueye A, Tine R, Ndiaye D, Tchania C, Ndiaye I, Barry A, Cisse B, Lameyre V, Gaye O. Repeated treatment of recurrent uncomplicated Plasmodium falciparum malaria in Senegal with fixed-dose artesunate plus amodiaquine versus fixed-dose artemether plus lumefantrine: a randomized, open-label trial. Malaria Journal. 2011;10:237.

[68] Ndyomugyenyi R, Magnussen P, Lal S, Hansen K, Clarke SE. Appropriate targeting of artemisinin-based combination therapy by community health workers using malaria rapid diagnostic tests: findings from randomized trials in two contrasting areas of high and low malaria transmission in south-western Uganda. Tropical Medicine & International Health. 2016;21(9):1157-70.

[69] Nichols EK, Byass P, Chandramohan D, Clark SJ, Flaxman AD, Jakob R, Leitao J, Maire N, Rao C, Riley I, Setel PW, Group WHOVAW. The WHO 2016 verbal autopsy instrument: An international standard suitable for automated analysis by InterVA, InSilicoVA, and Tariff 2.0. Plos Medicine. 2018;15(1):e1002486.

[70] Noedl H, Yingyuen K, Laoboonchai A, Fukuda M, Sirichaisinthop J, Miller RS. Sensitivity and specificity of an antigen detection ELISA for malaria diagnosis. American Journal of Tropical Medicine and Hygiene. 2006;75(6):1205-8.

[71] Onori E, Grab B. Quantitative Estimates of the evolution of a malaria epidemic in Turkey If remedial measures had not been applied. Bulletin of the World Health Organization. 1980;58(2):321-6.

[72] Onori E, Grab B. Indicators for the forecasting of malaria epidemics. Bulletin of the World Health Organization. 1980;58(1):91-8.

[73] Orth H, Jensen BO, Holtfreter MC, Kocheril SJ, Mallach S, MacKenzie C, Muller-Stover I, Henrich B, Imwong M, White NJ, Haussinger D, Richter J. Plasmodium knowlesi infection imported to Germany, January 2013. Euro Surveill. 2013;18(40).

[74] Paola R, Agus S, Marie R. bshazard: A Flexible Tool for Nonparametric Smoothing of the Hazard Function. The R Journal. 2014;6/2.

[75] Phillips MA, Burrows JN, Manyando C, van Huijsduijnen RH, Van Voorhis WC, Wells TNC. Malaria. Nature Reviews Disease Primers 2017;3:17050.

[76] Pintilie M. An Introduction to Competing Risks Analysis. Revista Espanola De Cardiologia. 2011;64(7):599-605.

[77] Pintilie M. Analysing and interpreting competing risk data. Statistics in Medicine. 2007;26(6):1360-7.

[78] Poschl B, Waneesorn J, Thekisoe O, Chutipongvivate S, Karanis P. Comparative diagnosis of malaria infections by microscopy, nested PCR, and LAMP in northern Thailand. American Journal of Tropical Medicine and Hygiene. 2010;83(1):56-60.

[79] President's Malaria Initiative Report 2016, Uganda Malaria Operational Plan Financial Year 2016.

[80] Pull JH, Grab B. A simple epidemiological model for evaluating the malaria inoculation rate and the risk of infection in infants. Bulletin of the World Health Organization. 1974;51(5):507-16.

[81] Rich JT, Neely JG, Paniello RC, Voelker CCJ, Nussenbaum B, Wang EW. A practical guide to understanding Kaplan-Meier curves. Otolaryngology-Head and Neck Surgery. 2010;143(3):331-6.

[82] Riley EM, Wagner GE, Akanmori BD, Koram KA. Do maternally acquired antibodies protect infants from malaria infection? Parasite Immunology. 2001;23(2):51-9.

[83] Rizopoulos D, Verbeke G, Molenberghs G. Shared parameter models under random effects misspecification. Biometrika. 2008;95(1):63-74.

[84] Ross R. Application of the theory of probabilities to the study of priori pathometry In: proceedings of the Royal Society of Landon Series A, containing papers of a mathematical and physical character. 1916;92.

[85] Ross R. Report on the prevention of malaria in Mauritius. New York: E. P. Dutton & Company; 1908.

[86] Ryu D, Sinha D, Mallick B, Lipsitz SL, Lipshultz S. Longitudinal Studies With Outcome-Dependent Follow-up: Models and Bayesian Regression. Journal of the American Statistical Association. 2007;102:952-67.

[87] Sama W, Dietz K, Smith T. Distribution of survival times of deliberate Plasmodium falciparum infections in tertiary syphilis patients. Transactions of the Royal Society of Tropical Medicine and Hygiene. 2006;100(9):811-6.

[88] Schwarz GE. Estimating the dimension of a model. Annals of Statistics. 1978;6 (2):461-4.

[89] Scrucca L, Santucci A, Aversa F. Competing risk analysis using R: an easy guide for clinicians. Bone Marrow Transplant. 2007;40(4):381-7.

[90] Scrucca L, Santucci A, Aversa F. Regression modeling of competing risk using R: an in depth guide for clinicians. Bone Marrow Transplant. 2010;45(9):1388-95.

[91] Shah JA, Emina JB, Eckert E, Ye Y. Prompt access to effective malaria treatment among children under five in sub-Saharan Africa: a multi-country analysis of national household survey data. Malar J. 2015;14:329.

[92] Shkedy Z, Aerts M, Molenberghs G, Beutels P, Van Damme P. Modelling age-dependent force of infection from prevalence data using fractional polynomials. Statistics in Medicine. 2006;25(9):1577-91.

[93] Shkedy Z, Aerts M, Molenberghs G, Beutels P, Van Damme P. Modelling forces of infection by using monotone local polynomials. Journal of the Royal Statistical Society Series C-Applied Statistics. 2003;52:469-85.

[94] Singer B, Cohen JE. Estimating malaria incidence and recovery rates from panel surveys. Mathematical Biosciences. 1980;49:273-305.

[95] Smieja J, editor Advantages and pitfalls of mathematical modelling used for validation of biological hypotheses. 7th IFAC Symposium on Modelling and Control in Biomedical Systems; 2009; Aalborg, Denmark.

[96] Smith DL, Battle KE, Hay SI, Barker CM, Scott TW, McKenzie FE. Ross, Macdonald, and a theory for the dynamics and control of mosquito-transmitted pathogens. Plos Pathogens. 2012;8(4).

[97] Smith DL, Drakeley CJ, Chiyaka C, Hay SI. A quantitative analysis of transmission efficiency versus intensity for malaria. Nature Communications. 2010;1.

[98] Smith DL, Dushoff J, Snow RW, Hay SI. The entomological inoculation rate and Plasmodium falciparum infection in African children. Nature. 2005;438(7067):492-5.

[99] Smith DL, McKenzie FE. Statics and dynamics of malaria infection in Anopheles mosquitoes. Malaria Journal. 2004;3:13.

[100] Smith TA. Estimation of heterogeneity in malaria transmission by stochastic modelling of apparent deviations from mass action kinetics. Malaria Journal 2008; 7: 12.

[101] Snow RW, Armstrong JR, Forster D, Winstanley MT, Marsh VM, Newton CR, Waruiru C, Mwangi I, Winstanley PA, Marsh K. Childhood deaths in Africa: uses and limitations of verbal autopsies. Lancet. 1992;340(8815):351-5.

[102] Soleman N, Chandramohan D, Shibuya K. Verbal autopsy: current practices and challenges. Bulletin of the World Health Organization. 2006;84(3):239-45.

[103] Staedke SG, Maiteki-Sebuguzi C, DiLiberto DD, Webb EL, Mugenyi L, Mbabazi E, Gonahasa S, Kigozi SP, Willey BA, Dorsey G, Kamya MR, Chandler CI. The Impact of an Intervention to Improve Malaria Care in Public Health Centers on Health Indicators of Children in Tororo, Uganda (PRIME): A Cluster-Randomized Trial. Am J Trop Med Hyg. 2016;95(2):358-67.

[104] Streatfield PK, Khan WA, Bhuiya A, Hanifi SM, Alam N, Diboulo E, Sie A, Ye M, Compaore Y, Soura AB, Bonfoh B, Jaeger F, Ngoran EK, Utzinger J, Melaku YA, Mulugeta A, Weldearegawi B, Gomez P, Jasseh M, Hodgson A, Oduro A, Welaga P, Williams J, Awini E, Binka FN, Gyapong M, Kant S, Misra P, Srivastava R, Chaudhary B, Juvekar S, Wahab A, Wilopo S, Bauni E, Mochamah G, Ndila C, Williams TN, Desai M, Hamel MJ, Lindblade KA, Odhiambo FO, Slutsker L, Ezeh A, Kyobutungi C, Wamukoya M, Delaunay V, Diallo A, Douillot

L, Sokhna C, Gomez-Olive FX, Kabudula CW, Mee P, Herbst K, Mossong J, Chuc NT, Arthur SS, Sankoh OA, Tanner M, Byass P. Malaria mortality in Africa and Asia: evidence from INDEPTH health and demographic surveillance system sites. Global Health Action. 2014;7:25369.

[105] Talisuna AO, Noor AM, Okui AP, Snow RW. The past, present and future use of epidemiological intelligence to plan malaria vector control and parasite prevention in Uganda. Malaria Journal. 2015;14.

[106] Tan KS, French B, Troxel AB. Regression modeling of longitudinal data with outcome-dependent observation times: extensions and comparative evaluation. Statistics in Medicine. 2014;33(27):4770-89.

[107] Therneau TM, Grambsch, Patricia M. Modeling Survival Data: Extending the Cox Model. New York: Springer; 2000.

[108] Trampuz A, Jereb M, Muzlovic I, Prabhu RM. Clinical review: Severe malaria. Critical Care. 2003;7(4):315-23.

[109] Uganda Bureau of Statistics 2016, The National Population and Housing Census 2014 - Main Report, Kampala, Uganda

[110] Uganda Ministry of Health, Annual Health Sector Performance Report, Financial Year 2014/2015

[111] Vaupel JW, Manton KG, Stallard E. The impact of heterogeneity in individual frailty on the dynamics of mortality. Demography. 1979;16(3):439-54.

[112] Verbeke G, Molenberghs G. Linear mixed models for longitudinal data. Springer. 2000.

[113] Von Fricken ME, Weppelmann TA, Lam B, Eaton WT, Schick L, Masse R, Beau De Rochars MV, Existe A, Larkin J, 3rd, Okech BA. Age-specific malaria seroprevalence rates: a cross-sectional analysis of malaria transmission in the Ouest and Sud-Est departments of Haiti. Malaria Journal. 2014;13:361.

[114] Walldorf JA, Cohee LM, Coalson JE, Bauleni A, Nkanaunena K, Kapito-Tembo A, Seydel KB, Ali D, Mathanga D, Taylor TE, Valim C, Laufer MK. School-Age Children Are a Reservoir of Malaria Infection in Malawi. Plos One. 2015;10(7):e0134061.

[115] Wang J, Xie H, Fisher JH. Multilevel Models. Applications Using SAS. Library of Congress Cataloging-in-Publication Data: Mathematics Subject Classification. 2010.

[116] White IM, Thompson R, Brotherstone S. Genetic and environmental smoothing of lactation curves with cubic splines. Journal of Dairy Science. 1999;82(3):632-8.

[117] White MT, Griffin JT, Drakeley CJ, Ghani AC. Heterogeneity in malaria exposure and vaccine response: implications for the interpretation of vaccine efficacy trials. Malaria Journal. 2010;9.

[118] White NJ. Plasmodium knowlesi: the fifth human malaria parasite. Clin Infect Dis. 2008;46(2):172-3.

[119] White NJ, Pukrittayakamee S, Hien TT, Faiz MA, Mokuolu OA, Dondorp AM. Malaria. Lancet. 2014;383(9918):723-35.

[120] WHO. Global Technical Strategy for Malaria 2016 - 2030. United Kingdom; 2015.

[121] WHO. Scaling up Diagnostic Testing, Treatment and Surveillance for Malaria. Geneva; 2012.

[122] WHO. World Malaria Report 2016. Geneva: Licence: CC BY-NC-SA 3.0 IGO; 2016

[123] WHO. World Malaria Report 2017. Geneva: Licence: CC BY-NC-SA 3.0 IGO; 2017.

[124] Wilson ML. Laboratory Diagnosis of Malaria Conventional and Rapid Diagnostic Methods. Archives of Pathology & Laboratory Medicine. 2013;137(6):805-11.

[125] Worrall E, Basu S, Hanson K. Is malaria a disease of poverty? A review of the literature. Tropical Medicine and International Health. 2005;10(10):1047-59.

[126] Wu S, Crespi CM, Wong WK. Comparison of methods for estimating the intraclass correlation coefficient for binary responses in cancer prevention cluster randomized trials. Contemporary Clinical Trials. 2012;33(5):869-80.

[127] Wulfsohn MS, Tsiatis AA. A joint model for survival and longitudinal data measured with error. Biometrics. 1997;53(1):330-9.

[128] Zhang D., Lin X. Variance component testing in generalized linear mixed models for longitudinal/clustered data and other related topics. in: Dunson D.B. (eds) Random effect and latent variable model selection. lecture notes in statistics. New York: Springer. 2008;192.

Appendix A

Appendix – Chapter 4

Here, we provide fit statistics for selecting a subdistribution hazard model of Fine and Gray [25] under competing risks analysis and the R code used to generate the figures presented in Chapter 4 and to fit the subdistribution hazard model which is described in equation 4.1 of Chapter 4.

A.1 Fit statistics

Under this section, we present the fit statistics for candidate models and the corresponding AIC and ΔAIC_i estimates (see Table A.1 for details). A model containing hospitalization history in addition to the rest of the variables that had significant crude associations (except diarrhea), results in $\Delta AIC_i = 1.66$ which is within the recommended range of $0 < \Delta AIC_i < 2$. But since the adjusted effect of hospitalization history was not significant (p = 0.120), it was dropped from the model. On the other hand, a model that includes diarrhea in addition to the other candidate variables resulted in $\Delta AIC_i = 2.05$ (see Table A.1), which is outside the recommended range of ($0<\Delta AIC_i<2$), so it was also not considered for making inference. Finally, a model that included place of birth, convulsion, comorbidity, duration of illness, fever and headache offered a better fit to the data and was used to make inference.

A.2 R code

```
library(survival)
## Figure 4.1
proportion=(table(chlddata$causeofdeath)/sum(table(chlddata$causeofdeath)))*100
```

Table A.1: Overview of model building (number of observations in each case equal to 8645).

Model	Log-likelihood	AIC	ΔAIC_i
Null	-2525.68	5051.36	961.89
P + Cn + Cm	-2478.79	4967.59	878.11
P + Cn + Cm + D	-2088.38	4188.77	99.29
P + Cn + Cm + F	-2428.35	4868.70	779.23
P + Cn + Cm + H	-2476.24	4964.48	875.01
P + Cn + Cm + Hh	-2475.25	4964.50	875.04
P + Cn + Cm + D + F + H	-2036.74	4089.47	0.00
P + Cn + Cm + D + F + H + Hh	-2035.57	4091.13	1.66
P + Cn + Cm + D + F + H + Dr	-2034.76	4091.52	2.05

P = Place of death, Cn = Convulsion, Cm = Comorbidity, H = Headache
Hh = Hospitalization history, D = Duration of illness, F = Fever, Dr = Diarrhea
AIC = -2l + 2p, where l=loglikelihood, p=number of model parameters
$\Delta AIC_i = AIC_i - min(AIC)$

```
colors=c(”red”, ”yellow”, ”green”, ”violet”,”orange”, ”blue”, ”pink”, ”cyan”, ”black”, ”purple”, ”maroon”)
causeofdeath=c(”Malaria”, ”Malnutrition”, ”Anemia”, ”Diarrhea”, ”Pneumonia”, ”Road accident”, ”Measles”, ”Tetanus”, ”HIV/AIDS”, ”Undetermined”, ”Other causes”)
# Plot settings
angle1 = rep(c(45,45,135), length.out=11)
angle2 = rep(c(45,135,135), length.out=11)
density1 = seq(5,35,length.out=11)
density2 = seq(5,35,length.out=11)
col = 1 # rainbow(7)
par(mfrow=c(1,2))
hist(chlddata$age, col=’gray’,main=”(A)”, ylab=”Number of children dying”, xlab=”Age at death (years)”, cex.lab=1.5, cex.axis=1.3)
barplot(proportion, beside=TRUE, main=”(B)”,ylim=c(0,60), xlab=”Cause of death”, ylab=”Percent (%)”,cex.lab=1.5, cex.axis=1.3,xaxt=’n’, col=col, angle=angle1, density=density1) #xaxt=’n’ hides values on x-axis, and yaxt=’n’ for y-axis
barplot(proportion, add=TRUE, main=””,ylim=c(0,60), xlab=”Cause of death”, ylab=”Percent (%)”,cex.lab=1.5, cex.axis=1.3,xaxt=’n’, col=col, angle=angle2, density=density2) #xaxt=’n’ hides values on x-axis, and yaxt=’n’ for y-axis
legend(5,50,inset=c(-0.8,0),bty=’n’, fill=col, legend = causeofdeath, col=col, angle=angle1, density=density1, cex=1.3)
legend(5,50,inset=c(-0.8,0),bty=’n’, fill=col, legend = causeofdeath, col=col, angle=angle2, density=density2, cex=1.3)

## Prepare data for Figures 4.2, 4.3 & 4.4
source(”CumIncidence.R”)
fitCIF=CumIncidence(chlddata$age,chlddata$causeofdeath, cencode=0, xlab=”Age in years”, col=”red”, ylab = ”Probability of dying from malaria”, level=0.95, lwd=2)
anemia=fitCIF$est[1,] #Anemia is row 1
```

```
diarrhea=fitCIF$est[2,] # Malaria is row 2
hivaids=fitCIF$est[3,] #HIV/AIDS is row 3
malaria=fitCIF$est[4,] #Malaria is row 4
malnutrition=fitCIF$est[5,] #malnutrition is row 5
measles=fitCIF$est[6,] #measles is row 6
othercauses=fitCIF$est[7,] #othercauses is row 7
pneumonia=fitCIF$est[8,1:54] #Pneumonia is row 8
roadaccident=fitCIF$est[9,] #roadaccident is row 9
tetanus=fitCIF$est[10,] #tetanus is row 10
undetermined=fitCIF$est[11,] #Undetermined is row 11
#merge vectors into data frame
CIFdata0 = as.data.frame(cbind(anemia, diarrhea, hivaids, malaria, malnutrition, measles, othercauses, pneumonia, roadaccident, tetanus, undetermined))
#Extract time vector (stored as row labels in CIFdata0 data frame)
time=row.names(CIFdata0)
CIFdata=as.data.frame(cbind(time,CIFdata0))
#95% CI extracted from output for malaria (estimated at time point as in tci vector below)
tci=c(0,2,4,6,8,10,12,14)
lowermal=c(0.0005,0.3529,0.4380,0.4546,0.4611,0.4700,0.4713,0.4713)
uppermal=c(0.0088,0.4204,0.5061,0.5226,0.5290,0.5377,0.5391,0.5391)
## Figure 4.2
plot(CIFdata$time,CIFdata$malaria, type=”n”,xlab = ”Age in years”, ylab = ”Probability of dying”, cex.lab=1.5, cex.axis=1.3)
lines(CIFdata$time,CIFdata$malaria, lty=1, lwd=3, col=”red”)
lines(CIFdata$time,CIFdata$malnutrition, lty=2.5, lwd=3, col=”blue”)
lines(CIFdata$time,CIFdata$anemia, lty=3, lwd=3, col=”green”)
lines(CIFdata$time,CIFdata$diarrhea, lty=4, lwd=3, col=”violet”)
lines(CIFdata$time,CIFdata$pneumonia, lty=5, lwd=3, col=”orange”)
lines(CIFdata$time,CIFdata$roadaccident, lty=6, lwd=1, col=”black”)
lines(CIFdata$time,CIFdata$measles, lty=7, lwd=1, col=”skyblue”)
lines(CIFdata$time,CIFdata$tetanus, lty=8, lwd=1, col=”cyan”)
lines(CIFdata$time,CIFdata$hivaids, lty=9, lwd=1, col=”black”)
lines(CIFdata$time,CIFdata$undetermined, lty=10, lwd=1, col=”purple”)
lines(CIFdata$time,CIFdata$othercauses, lty=11, lwd=1, col=”maroon”)
legend(6,0.45,lty=c(1,2,3,4,5), col=c(”red”, ”blue”, ”green”, ”violet”,”orange”), c(”Malaria”, ”Malnutrition”, ”Anemia”, ”Diarrhea”, ”Pneumonia”), lwd=3, cex=1.2, title=””, bty=’n’)
legend(10, 0.45, lty=c(6,7,8,9,10,11), col=c(”black”, ”skyblue”, ”cyan”, ”black”, ”purple”, ”maroon”), c(”Road accident”, ”Measles”, ”Tetanus”, ”HIV/AIDS”, ”Undetermined”, ”Other cause”), lwd=1, cex=1, title=””, bty=’n’)
## Figure 4.3
library(rms)
deadsurv = npsurv(formula = Surv(age, malaria==1)   1, data = chlddata)
#Extract survival time, survival function, and corresponding 95%CI
t=summary(deadsurv)$time
s=summary(deadsurv)$surv
sL=summary(deadsurv)$lower
sU=summary(deadsurv)$upper
#Marginal CIF (1-KM)
plot(t, 1-s, type=”l”, col=”red”, lwd=2, ylim=c(0:1), xlab = ”Age in years”, ylab = ”Probability of dying from malaria”, cex.lab=1.5, cex.axis=1.3, lty=5)
lines(tt,1-sL, lty=3, col=”red”, lwd=2)
lines(tt,1-sU, lty=3, col=”red”, lwd=2)
legend(”topleft”, lty=c(5,1),lwd=2,col=c(”red”, ”blue”),cex=1.3, legend=c(”Marginal CIF (1-KM)”, ”Cause-specific CIF”), bty=’n’)
#Cause specific CIF
```

```
lines(CIFdata$time,CIFdata$malaria, type=”l”, col=”blue”, lwd=2, lty=1)
lines(tci,lowermal, lty=3, col=”blue”, lwd=2)
lines(tci,uppermal, lty=3, col=”blue”, lwd=2)
## Figure 4.4
#By Locdied, location of death
fitCIFLocdied=CumIncidence(chlddata$age,chlddata$causeofdeath,chlddata$Locdied,cencode=0,
xlab=”Age in years”, level=0.95, col=”red”, lwd=2)
#Extract estimates
malhome=fitCIFLocdied$est[1,]
malhosp=fitCIFLocdied$est[2,]
malonway=fitCIFLocdied$est[3,]
malother=fitCIFLocdied$est[4,]
#By convulsions/fits
fitCIFconvuls=CumIncidence(chlddata$age,chlddata$causeofdeath,chlddata$convuls,cencode=0,
xlab=”Age in years”, level=0.95, col=”red”, lwd=2)
malconvuls=fitCIFconvuls$est[1,]
malnoconvuls=fitCIFconvuls$est[2,]
#By Comorbidity
fitCIFcomorbid=CumIncidence(chlddata$age,chlddata$causeofdeath,chlddata$comorbid,cencode=0,
xlab=”Age in years”, level=0.95, col=”red”, lwd=2)
malcomorbid=fitCIFcomorbid$est[1,]
malnocomorbid=fitCIFcomorbid$est[2,]
#Ploting
par(mfrow=c(1,3))
#location plot(time,malhome,type=”l”,ylim=c(0:1),lty=1, lwd=2, col=”blue”, xlab = ”Age in years”,
ylab = ”Probability of dying from malaria”, cex.lab=1.5, cex.axis=1.3)
lines(time,malhosp,lty=2, lwd=2, col=”red”)
lines(time,malonway,lty=3, lwd=2, col=”orange”)
lines(time,malother,lty=4, lwd=2, col=”black”)
legend(”topleft”, lty=c(1,2,3,4),lwd=2,col=c(”blue”, ”red”, ”orange”, ”black”), cex=1.3,
legend=c(”Home”, ”Hospital/HF”, ”On way”, ”Other”), bty=’n’)
#Convulsions
plot(time,malconvuls,type=”l”,ylim=c(0:1),lty=1, lwd=2, col=”blue”,xlab = ”Age in years”, ylab =
”Probability of dying from malaria”, cex.lab=1.5, cex.axis=1.3)
lines(time,malnoconvuls,lty=2, lwd=2, col=”red”)
legend(”topleft”, lty=c(1,2),lwd=2,col=c(”blue”, ”red”), cex=1.3, legend=c(”Convulsion”, ”No
Convulsion”), bty=’n’)
#Comorbidity
plot(time,malcomorbid,type=”l”,ylim=c(0:1),lty=1, lwd=2, col=”blue”,xlab = ”Age in years”, ylab =
”Probability of dying from malaria”, cex.lab=1.5, cex.axis=1.3)
lines(time,malnocomorbid,lty=2, lwd=2, col=”red”)
legend(”topleft”, lty=c(1,2),lwd=2,col=c(”blue”, ”red”), cex=1.3, legend=c(”Comorbidity”, ”No
Comorbidity”), bty=’n’)
```

```
## Fine and Gray model

require(cmprsk)

require(lme4)

#load required function

source(”factor2ind.R”)

# First, we need to create a data matrix with dummy variables for all categorical variables as follows,

yet continuous retained.
```

```
# Start by creating a vector of variables to be used in modelling
causeofdeath=chlddata$causeofdeath
age=chlddata$age
place=chlddata$Locdied
fits=chlddata$fits
comorbid=chlddata$comorbid
logdaysick=chlddata$logdaysick
hospitaliz=chlddata$hospitaliz
fev=chlddata$fev
head=chlddata$head
vomi=chlddata$vomi
gender=chlddata$gender
rel=chlddata$rel
promptrt=chlddata$promptrt
whereseektrt=chlddata$whereseektrt
diarrhea=chlddata$diarr
# Create a matrix of fixed covariates, with reference level of categorical variables in double quotes(””)
x=cbind(factor2ind(place,”Hosp/health facility”), factor2ind(fits, ”0”), comorbid2, logdaysick, factor2ind(hospitaliz, ”No”), factor2ind(fev, ”0”), factor2ind(head, ”0”), factor2ind(vomi, ”0”), factor2ind(gender, ”Male”), factor2ind(rel, ”MOTHER”), factor2ind(promptrt, 1), factor2ind(whereseektrt, ”hospital”), diarrhea)
# To fit models we need function crr() which is inbuilt in package ”cmprsk”
# FINAL FIT considered for the adjusted analysis. Note: the values in c(1:3,4,5,6,10,11) below represent the selected columns in matrix x
model=crr(age,causeofdeath, x[,c(1:3,4,5,6,10,11)])
summary(model)
```

Appendix **B**

Appendix – Chapter 5

In this appendix, we provide additional methods and results plus both the R code and the SAS macro supporting the work presented in Chapter 5. Section B.1 of this appendix acknowledges the problem of getting a negative FOI and it provides possible solutions for avoiding this problem when using a flexible fractional polynomial, which we used in Chapter 5, to estimate age-time parasite prevalence and FOI. In Section B.2, we describe a methodology to estimate the subject-specific FOI and prevalence (hence accounting for unobserved heterogeneity), which we graphically present in Figure 5.3 (top row) of Chapter 5 and in Figure B.2 (top row) of this appendix. In Section B.3, we describe an approach to estimate the marginal FOI and prevalence using numerical averaging method for integrating out random effects. A SAS macro used to perform the numerical averaging is given in Section B.6 of this appendix. The results for the marginal FOI are presented in Figures 5.2, 5.3 and 5.4 in Chapter 5, and the results for the marginal prevalence are presented in Figure B.2 of this appendix. In Section B.4, we derive the lower and the upper bounds for the FOI and we present results for the marginal FOI using these two bounds in Figure 5.2. The rest of the results in this appendix are clearly referred to in Chapter 5.

B.1 Fractional polynomial and non-negative FOI

Though, fractional polynomials are very flexible, they can result into negative estimates for the FOI whenever the estimated probability to be infected before age a is a non-monotone function [36, 92]. A solution to this is to define a non-negative FOI, $\lambda_l(a_{ijk}|b_i) \geq 0$ for all a and to estimate $\pi_l(a_{ijk}|b_i)$ under these constraints [92]. From Table 5.1 in Chapter 5, for a logit link function, the condition $\eta^{'}(a_{ijk}|b_i) \geq -\gamma/(1-\pi_l(a_{ijk}|b_i))$ should be satisfied as to estimate a positive FOI. One option is to fit a constrained FP to ensure the above condition holds by applying a constraint on parameter estimates depending on the functional relationship with age. However, this approach becomes challenging especially if it involves constraining random effects. An alternative option is to find a probability of estimating a negative FOI using the model results. If this probability is considerably small, say less than 0.01, then one can consider the first option unnecessary. In this paper, the second option was applied. Indeed, all site-specific coefficients for age effect were negative (see Table 5.3), meaning that the site-specific derivatives for the linear predictors, $\eta^{'}(a_{ijk}|b_i) = [-(\hat{\beta}_6)a_{ijk}^{-2}, -(\hat{\beta}_7)a_{ijk}^{-2}, -(\hat{\beta}_8)a_{ijk}^{-2}] > 0$. This implies that the above condition always holds in our case since a_{ijk}^{-2}, γ and $(1-\pi_l(a_{ijk}|b_i))$ are always positive. Therefore, the probability to estimate a negative FOI was zero.

B.2 Conditional FOI and prevalence

For example, based on model results in Table 5.3, the conditional age-time dependent FOI for a subject from Walukuba, born in the baseline year (2001, that is, shifted year of birth = 0) and was symptomatic at the previous visit can be estimated as follows,

$$\hat{\lambda}_0(a_{ijk}|b_i) = \hat{\gamma}\text{exp}(\hat{\beta}_0 + \hat{\beta}_6 a_{ijk}^{-1} + \hat{\beta}_4 + b_{1ij} + b_{21j}) - \hat{\beta}_6 a_{ijk}^{-2}\hat{\pi}_0(a_{ijk}|b_i), \tag{B.1}$$

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

$$\hat{\pi}_0(a_{ijk}|b_i) = \frac{\text{exp}(\hat{\beta}_0 + \hat{\beta}_6 a_{ijk}^{-1} + \hat{\beta}_4 + b_{1ij} + b_{21j})}{1+\text{exp}(\hat{\beta}_0 + \hat{\beta}_6 a_{ijk}^{-1} + \hat{\beta}_4 + b_{1ij} + b_{21j})}, \tag{B.2}$$

and $\hat{\gamma}$ is an estimate for the clearance rate. The conditional FOI for other sites given the infection status at the previous visit and past use of AL can be estimated in a similar way.

B.3 Marginalisation

A sample of $M = 1000$ of the random affects vector $b_i = (b_{1i}, b_{2sj})^T$, $s = 1, 2, 3$ (sites), was generated from a multi-variate normal distribution,$N(0, \hat{L}\hat{L}^T)$, where for example, for Walukuba, $\hat{L} = (0.49, 1.67)^T$ whose elements are the square roots of $\hat{d}_{11}$

and $\hat{d}_{22}$, respectively as given in Table 5.3. A fine grid of age, $a = 0.5$ to 11 with interval 0.1 years (the age range in the data, though extrapolation is possible) was considered. For example, the marginalized FOI at each age value in the grid, again considering a subject from Walukuba, born in the baseline year and was symptomatic at the previous visit is calculated as in (6.8).

$$\hat{\lambda}_0(a) = \frac{1}{1000}\sum_{i=1}^{1000}\Big(\hat{\gamma}\exp(\hat{\beta}_0 + \hat{\beta}_6 a^{-1} + \hat{\beta}_4 + b_{1i} + b_{21i})\Big) - \hat{\beta}_6 a^{-2}\hat{\pi}_0(a), \tag{B.3}$$

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

$$\hat{\pi}_0(a) = \frac{1}{1000}\sum_{i=1}^{1000}\Big(\frac{\exp(\hat{\beta}_0 + \hat{\beta}_6 a^{-1} + \hat{\beta}_4 + b_{1i} + b_{21i})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_6 a^{-1} + \hat{\beta}_4 + b_{1i} + b_{21i})}\Big). \tag{B.4}$$

Extensions to estimate the marginal averages at different birth years, for different study sites and for different infection statuses at the previous visit, are straightforward. The SAS macro performing the numerical averaging for a case of $\hat{\gamma} = 1.643$ is given in Section B.6 of this appendix.

B.4 A general $S(I)_J(R)S$ system

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

$$\begin{aligned}
s'(t) &= \mu - \beta s i + \sigma r - \mu s, \\
i_1'(t) &= \beta s i - \gamma i_1 - \mu i_1, \\
i_2'(t) &= \gamma i_1 - \gamma i_2 - \mu i_2, \\
&\dots \\
i_J'(t) &= \gamma i_{J-1} - \gamma i_J - \mu i_J, \\
r'(t) &= \gamma i_J - \sigma r - \mu r,
\end{aligned} \tag{B.5}$$

where $i = \sum_{J=1}^{J} i_J$. Rewriting the system collapsing the infectious classes into i:

$$\begin{aligned} s'(t) &= \mu - \beta si + \sigma r - \mu s, \\ i'(t) &= \beta si - \gamma i_J - \mu i, \\ r'(t) &= \gamma i_J - \sigma r - \mu r. \end{aligned} \tag{B.6}$$

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

$$\begin{aligned} s'(t) &= \mu - \beta si + \gamma i_J - \mu s, \\ i'(t) &= \beta si - \gamma i_J - \mu i, \end{aligned} \tag{B.7}$$

yields (replacing $\lambda = \beta i$ and $s = 1 - i$)

$$i' = \lambda(1 - i) - \gamma i_J - \mu i \tag{B.8}$$

thus

$$\lambda = \frac{i' + \gamma i_J + \mu i}{1 - i} \approx \frac{i' + \gamma i_J}{1 - i}, \tag{B.9}$$

expressing time dependency,

$$\lambda(t) = \frac{i'(t) + \gamma i_J(t) + \mu i(t)}{1 - i(t)} \approx \frac{i'(t) + \gamma i_J(t)}{1 - i(t)}, \tag{B.10}$$

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

$$[\lambda_L(t), \lambda_U(t)] = \Big[\frac{i'(t)}{1 - i(t)}, \frac{i'(t) + \gamma i(t)}{1 - i(t)}\Big] \tag{B.11}$$

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

Table B.1: Overview of the fractional polynomial model selection.

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

Table B.2: Overview of model building (number of observations in each case equal to 8645).

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

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

Table B.3: Maximum values for the marginal annual FOI by study site, previous infection status and use of AL, and by age group.

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

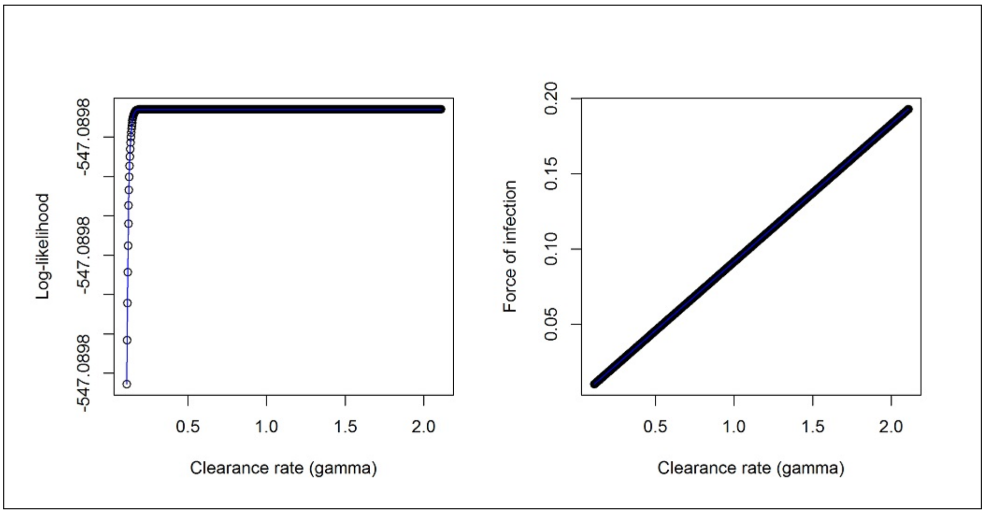

Figure B.1: Plots for log-likelihood verses the clearance rate (left panel) and force of infection verses the clearance rate (right panel) obtained after fitting 1000 models to the data according to $\pi = \frac{\lambda}{\lambda+\gamma} e^{-(\lambda+\gamma)a}$ as given by Pull and Grab (1974) [80] by choosing values for the annual clearance rate on a grid of 0.1 to 2.0 with a step size of 0.0019.

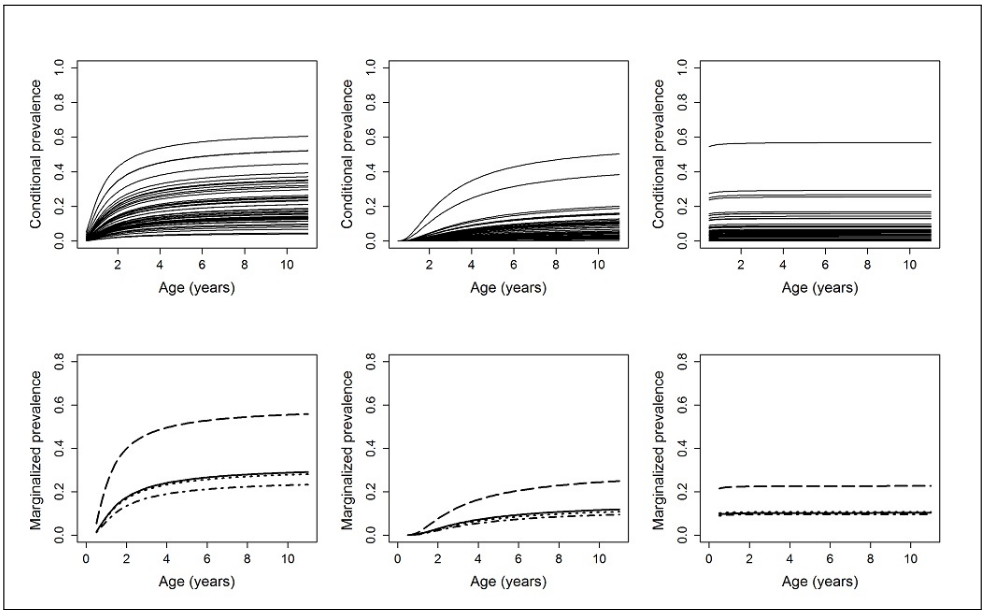

Figure B.2: Top row: Individual-specific evolutions for the conditional prevalence, by study site for children assumed to be symptomatic at the previous visit and were born in the baseline year (2001). Bottom row: Average evolutions for marginalized prevalence, by study site and the infection status at the previous visit and past use of AL (negative and no AL in the past (solid lines), negative and AL in the past (dotted lines), symptomatic (dash-dotted lines) and asymptomatic (long-dashed lines)). Left panel: Nagongera, middle panel: Kihihi, right panel: Walukuba.

Table B.4: Marginal FOI and the 95% confidence bounds for the age- and time-dependent marginal annual FOI by study site, previous infection status and use of AL, and by age group for children born in the baseline year (2001).

Infection status at		Nagongera	Kihihi	Walukuba
the previous visit	Age in	Marg. annual FOI	Marg. annual FOI	Marg. annual FOI
and past use of AL	years	(95% CI)x1000	(95% CI)x1000	(95% CI)x1000
Lower bound:				
Negative and no	< 1	143.8 (141.2 - 146.4)	9.3 (8.5 - 10.0)	10.2 (9.8 - 10.7)
AL in the past	1-4	53.7 (53.2 - 54.2)	22.7 (22.3 - 23.0)	1.0 (0.9 - 1.1)
	5-10	8.6 (8.5 - 8.6)	7.2 (7.2 - 7.3)	0.1 (0.1 - 0.1)
Negative and AL	< 1	137.4 (134.8 - 139.9)	7.6 (7.3 - 8.0)	10.7 (10.3 - 11.2)
the past	1-4	51.7 (51.2 - 52.1)	20.1 (19.8 - 20.4)	1.0 (1.0 - 1.0)
	5-10	8.3 (8.2 - 8.3)	6.6 (6.5 - 6.7)	0.1 (0.1 - 0.1)
Symptomatic	< 1	105.6 (103.7 - 107.5)	6.3 (6.0 - 6.5)	9.6 (9.2 - 10.0)
	1-4	41.4 (41.0 - 41.8)	16.9 (16.7 - 17.1)	0.9 (0.9 - 0.9)
	5-10	6.8 (6.8 - 6.9)	5.7 (5.7 - 5.8)	0.1 (0.1 - 0.1)
Asymptomatic	< 1	426.7 (420.3 - 433.1)	24.9 (23.7 - 26.1)	22.9 (22.1 - 23.6)
	1-4	123.3 (122.2 - 124.4)	55.2 (54.6 - 55.8)	2.1 (2.1 - 2.1)
	5-10	16.9 (16.8 - 16.9)	15.7 (15.6 - 55.8)	0.2 (0.2 - 0.2)
Upper bound:				
Negative and no	< 1	234.5 (229.7 - 239.3)	12.2 (11.2 - 13.3)	309.3 (285.2 - 333.5)
AL in the past	1-4	225.0 (223.3 - 226.7)	61.7 (60.1 - 63.2)	112.8 (109.4 - 116.3)
	5-10	445.7 (442.8 - 448.6)	161.2 (157.3 - 165.1)	191.4 (186.6 - 196.2)
Negative and AL	< 1	223.7 (219.2 - 216.4)	10.0 (9.5 - 10.5)	322.9 (298.1 - 347.7)
in the past	1-4	214.8 (213.2 - 216.4)	51.7 (50.9 - 52.4)	117.8 (114.2 - 121.3)
	5-10	424.5 (421.7 - 427.3)	131.3 (129.7 - 132.9)	199.7 (194.8 - 204.7)
Symptomatic	< 1	170.7 (167.3 - 174.0)	8.2 (7.8 - 8.6)	246.6 (232.0 - 261.1)
	1-4	164.2 (163.0 - 165.3)	42.7 (42.2 - 43.3)	89.5 (87.5 - 91.6)
	5-10	320.1 (318.2 - 322.1)	107.7 (106.5 - 108.9)	151.6 (148.8 - 154.5)
Asymptomatic	< 1	741.4 (728.1 - 754.6)	32.8 (31.2 - 34.5)	1134.5 (1034.6 - 1234.4)
	1-4	717.4 (712.3 - 722.3)	159.8 (157.6 - 162.1)	417.9 (403.5 - 432.4)
	5-10	1532.8 (1523.4 - 1542.1)	429.1 (423.7 - 434.6)	711.1 (691.0 - 731.3)

B.5 R code

```
## Figure 5.1
prevmidage=ddply(cohortdata, c(”siteid”, ”midage”), summarise, N = length(siteid), Totalpos=sum(parasitemia), observedPR = (Totalpos/N)*100)
prevmonthyear=ddply(cohortdata, c(”siteid”, ”year”, ”month”), summarise, N = length(siteid), Totalpos=sum(parasitemia), observedPR = (Totalpos/N)*100)
plotdata=subset(prevmidage, siteid==3)
par(mfrow=c(1,2))
# By age
plot(plotdata$midage, plotdata$observedPR, type=”n”, ylim=c(0,70), xlim=c(0,11), main=”(A)”, ylab=”Proportion infected (%)”, xlab=”Age in years”, cex.lab=1.3)
points(subset(prevmidage, siteid==3, select = c(midage, observedPR ) ), pch=c(16), col=”blue”,
cex=0.006*nagsize)
points(subset(prevmidage, siteid==2, select = c(midage, observedPR ) ), pch=c(16), col=”orange”,
cex=0.006*kihsize)
points(subset(prevmidage, siteid==1, select = c(midage, observedPR ) ), pch=c(16), col=”maroon”,
cex=0.006*walsize)
legend(-0.5,75, c(”Nagongera”, ”Kihihi”, ”Walukuba”), col=c(”blue”, ”orange”, ”maroon”),
pch=c(16,16,16), bty=’n’, cex=1.1)
# By calendar time
plotdata2=subset(prevmonthyear, siteid==3)
plot(plotdata2$monthyear, plotdata2$observedPR, type=”n”, ylim=c(0,100), main=”(B)”,
ylab=”Proportion infected (%)”, xlab=NA, cex.lab=1.3, xaxt=”n”)
axis(1, at=plotdata2$monthyear, labels=c(”Aug-11”, ””, ””, ”Nov-11” ,””, ””, ”Feb-12”, ””,””,”May-
12”,””,””,”Aug-12”,””,””,”Nov-12”,””,””,”Feb-13”,””,””,”May-13”,””,””,”Aug-13”,””,””,”Nov-
13”,””,””,”Feb-14”,””,””,”May-14”,””,””,”Aug-14”), cex.axis=0.8, las=2)
points( subset( prevmonthyear, siteid==3, select = c( monthyear, observedPR ) ), pch=c(16), col=”blue”,
cex=0.01*nagsize0)
points( subset( prevmonthyear, siteid==2, select = c( monthyear, observedPR ) ), pch=c(16),
col=”orange”, cex=0.01*kihsize0 )
points( subset( prevmonthyear, siteid==1, select = c( monthyear, observedPR ) ), pch=c(16),
col=”maroon”, cex=0.01*walsize0)
legend(0,105, c(”Nagongera”, ”Kihihi”, ”Walukuba”), col=c(”blue”, ”orange”, ”maroon”),
pch=c(16,16,16), bty=’n’, cex=1.1)
## Figure 5.2
datlowerdiffnag1= read.table(text = ”A B C
1 0.144 0.050 0.009
2 0.091 0.175 0.437”, header = T)
barplot(as.matrix(datlowerdiffnag1), col=terrain.colors(4), ylim=c(0,0.5), ylab=”Marginal annual FOI”,
cex.lab=1.3, cex.axis=1.2, xlab=”Negative, No AL”)
## Figure 5.3 (Nagongera site)
nsymp=subset(condprevfoi, L==0 & site==3 & pinfect==3)
# Conditional annual FOI
plot(nsymp$a, nsymp$foiL, type=”n”, xlab=”Age (years)”, ylim=c(0, 0.5), ylab=”Conditional annual
FOI”, main=””, cex.axis=1, cex.lab=1.5)
for (i in unique(nsymp$subject)) lines(nsymp[nsymp$subject==i, c(”a”)], nsymp$foiL[nsymp$subject==i])
# Marginalized FOI
plot(margprevfoi$a, margprevfoi$foiL, ylim=c(0,0.5), xlim=c(0,11), xlab=”Age (years)”, ylab=”Marginal
annual FOI”, main=””, type=”n”, cex.axis=1.3, cex.lab=1.5)
lines(subset(margprevfoi, L==0 & site==3 & pinfect==1, select = c(a, foiL)), lwd=2, lty=1)
lines(subset(margprevfoi, L==0 & site==3 & pinfect==2, select = c(a, foiL)), lwd=2, lty=3)
lines(subset(margprevfoi, L==0 & site==3 & pinfect==3, select = c(a, foiL)), lwd=2, lty=4)
lines(subset(margprevfoi, L==0 & site==3 & pinfect==4, select = c(a, foiL)), lwd=2, lty=5)
legend(”top”, c(”Negative, No AL”, ”Negative, AL”, ”Symptomatic”, ”Asymptomatic”), lty=c(1,3,4,5),
lwd=2, cex=1.5, title=”Previous status”, bty=’n’)


## Figure 5.4 (Nagongera site)

# Contour plot

# Start by forming matrix data

nagongera = subset(margfoigammasymp, site==3)

nagmatrix = xtabs(foi∼gamma+a, data=nagongera) # specify the variable to be put to vector using

[[’varname’]]

y1=subset(margfoigammasymp, site==1 & a==0.5, select=c(gamma))[[’gamma’]]
```

```
x1=subset(margfoigammasymp, site==1 & gamma==0, select=c(a))[['a']]
# Make sure length(x1) times length(y1)=length(nagmatrix)
length(x1)*length(y1); length(nagmatrix)
# Now define scale for x and y axis
x.at = seq(0, 11, by=1); x.at2 = seq(0, 5, by=1)
y.at = seq(0, 3, by=0.1)
# Now produce image plot
image(x1, y1, t(nagmatrix), ylim=c(0,3), xlim=c(0,11), col=terrain.colors(100),ylab=”Clearance rate”,xlab=”Age (years)”, cex.lab=1.4, axes=FALSE)
contour(x1, y1, t(nagmatrix), levels=seq(0.02, 1.09, by=0.1), add=TRUE, col=”brown”)
axis(1, at=x.at, cex.axis=1.3)
axis(2, at=y.at, cex.axis=1.3)
box()
title(main=””)
# Upper bound
datsympuppernag = read.table(text = ”A B C
1 0.171 0.164 0.320
2 0.192 0.184 0.361
3 0.215 0.206 0.406
4 0.241 0.231 0.458”, header = TRUE)
barplot(as.matrix(datsympuppernag), col=terrain.colors(4), ylim=c(0,1.6), ylab=”Marginal annual FOI”, cex.lab=1.4, cex.axis=1.2)
text(2, -0.3, cex=1.3, paste(”A:”, ’<1 year ’, ”B:”, ’1-4 years ’, ”C:”, ’5-10 years ’))
legend(”top”, inset=c(-0.2,0), legend = c(”2001”, ”2004”, ”2007”, ”2010”), fill=terrain.colors(4), cex=1.3, title=”Birth year”, bty=’n’)
## Figure 5.5 (Nagongera site)
library(stats4)
y=cohortdata$parasitemia[cohortdata$siteid==3]
datagamma=NULL
gamma=seq(0.11,2.11,0.002) #grid of step=0.002 from 0.11 to 2.11
for (i in 1:length(gamma))
{ cdat=NULL
levi1=function(w, I)
{pi=(1/(1+exp(-w)))+(I-(1/(1+exp(-w))))*exp(-gamma[i]*((1/exp(-w))+1)*t1)
ll=y*log(pi) + (1-y)*log(1-pi)
return(-sum(ll))
}
fit=mle(levi1,start=list(w=0.01,I=0.05))
# Store model estimates for each value of gamma in the sequence/grid loglikelihood=logLik(fit)
estw¡-coef(fit)[”w”]
estI¡-coef(fit)[”I”]
# Estimate FOI for each value of gamma in the sequence/grid
FOI=(gamma[i]/(exp(-estw)))
# Print estimates for w, I, FOI,loglikelihood for each value of gamma in the sequence/grid
cat(i,gamma[i],estw,estI,FOI,loglikelihood,”\n”)
cdat=cbind(i,gamma[i],estw, estI, FOI,loglikelihood)
datagamma=rbind(datagamma,cdat)
}
nonlineardata=as.data.frame(datagamma)
{# rename variables}
colnames(nonlineardata)[2] = ”gamma”
# Plot loglik vs gamma grid
par(mfrow=c(1,2))
plot(nonlineardata$gamma,nonlineardata$loglikelihood, ylab=”Log-likelihood”, xlab=”Clearance rate (gamma)”,cex.axis=1, cex.lab=1)
lines(nonlineardata$gamma, nonlineardata$loglikelihood, col=”blue”)
plot(nonlineardata$gamma,nonlineardata$FOI, ylab=”Force of infection”, xlab=”Clearance rate (gamma)”,cex.axis=1, cex.lab=1)
lines(nonlineardata$gamma, nonlineardata$FOI, col=”blue”)
```

## Figure 5.6

This can be produced in a similar way as Figure 6.3 but now for prevalence instead of FOI

B.6 SAS macro

***GLIMMIX code**

```
proc glimmix data=cohortdata method=laplace NOCLPRINT;
class hhid id siteid(ref=”1”) pinfectstatusandAL(ref=”0”);
model parasitemia = fpcohortage*siteid yearshift*siteid siteid pinfectstatusandAL/ dist=bin
oddsratio link=logit solution;
random intercept/ subject = hhid group=siteid solution;
random intercept / subject = id(hhid) solution;
COVTEST/ WALD;
run;
```

***Numerical averaging**

```
**Considering children born between 2001 to 2014 as they appear in the data;
```

```
data numaveragingprevfoinc;
do site =1 to 3 by 1; *study sites 1(walukuba),2(kihihi),3(nagongera);
do pinfect =1 to 4 by 1; *infection status 1(negative+no AL), 2(negative+AL), 3(symptomatic), 4(asymptomatic);
do subject=1 to 1000 by 1; *generate 1000 samples;
bi1=rannor(123); bi2=rannor(123); bi3=rannor(123); bi4=rannor(123); *used seed=123 to generate from standard normal;
d11=0.24;d22=2.80;d33=1.16;d44=0.21;*variances from the final fit, elements in D;
rd11=d11**0.5;rd22=d22**0.5;rd33=d33**0.5;rd44=d44**0.5; *sqrt(S2) to be used in Cholesky decomposition;
r1=rd11*bi1; r2=rd22*bi2; r3=rd33*bi3; r4=rd44*bi4; *using U+sqrt(S2)*rannor(seed): Note elements in here are sqrt of elements in D;
do a=0.5 to 11 by 0.1; *generate 1000 samples at each age point in the grid;
do L=0 to 13 by 1; *Repeat the above process for each value of birth year shift (L=year of birth - 2001);
*Parameter estimates;
B0=-3.04;B1=0.86;B2=2.19;B3=-0.01;B4=-0.24;B5=1.23;B6=-0.05;B7=-4.01;B8=-1.75;B9=-0.13;B10=0.11;B11=0.04;
ap=1/a; *Power of age, age-1;
*Linear Predictors;
lp11=B0+B6*ap+B9*L+r1+r2; lp12=B0+B6*ap+B9*L+B3+r1+r2;
lp13=B0+B6*ap+B9*L+B4+r1+r2;lp14=B0+B6*ap+B9*L+B5+r1+r2;
lp21=B0+B7*ap+B10*L+B1+r1+r3; lp22=B0+B7*ap+B10*L+B1+B3+r1+r3;
lp23=B0+B7*ap+B10*L+B1+B4+r1+r3;lp24=B0+B7*ap+B10*L+B1+B5+r1+r3;
lp31=B0+B8*ap+B11*L+B2+r1+r4; lp32=B0+B8*ap+B11*L+B2+B3+r1+r4;
lp33=B0+B8*ap+B11*L+B2+B4+r1+r4;lp34=B0+B8*ap+B11*L+B2+B5+r1+r4;
*Derivative of linear predictor;
lpder1=-(B6)*(ap*ap); lpder2=-(B7)*(ap*ap); lpder3=-(B8)*(ap*ap);
*Prevalence;
if site=1 and pinfect=1 then pi=exp(lp11)/(1+exp(lp11));
if site=1 and pinfect=2 then pi=exp(lp12)/(1+exp(lp12));
if site=1 and pinfect=3 then pi=exp(lp13)/(1+exp(lp13));
if site=1 and pinfect=4 then pi=exp(lp14)/(1+exp(lp14));
if site=2 and pinfect=1 then pi=exp(lp21)/(1+exp(lp21));
if site=2 and pinfect=2 then pi=exp(lp22)/(1+exp(lp22));
if site=2 and pinfect=3 then pi=exp(lp23)/(1+exp(lp23));
if site=2 and pinfect=4 then pi=exp(lp24)/(1+exp(lp24));
if site=3 and pinfect=1 then pi=exp(lp31)/(1+exp(lp31));
if site=3 and pinfect=2 then pi=exp(lp32)/(1+exp(lp32));
if site=3 and pinfect=3 then pi=exp(lp33)/(1+exp(lp33));
if site=3 and pinfect=4 then pi=exp(lp34)/(1+exp(lp34));
**FOI;
*Clearance rate of 1.643 for children <1 year as given by Bekessy et al.(1976) is demonstrated, a similar code can easily be adopted for ages 1-4 years and 5-10 years.;
if site=1 and pinfect=1 and a<1 then foi=1.643*exp(lp11)+ lpder1*exp(lp11)/(1+exp(lp11));
if site=1 and pinfect=2 and a<1 then foi=1.643*exp(lp12)+ lpder1*exp(lp12)/(1+exp(lp12));
if site=1 and pinfect=3 and a<1 then foi=1.643*exp(lp13)+ lpder1*exp(lp13)/(1+exp(lp13));
if site=1 and pinfect=4 and a<1 then foi=1.643*exp(lp14)+ lpder1*exp(lp14)/(1+exp(lp14));
if site=2 and pinfect=1 and a<1 then foi=1.643*exp(lp21)+ lpder2*exp(lp21)/(1+exp(lp21));
if site=2 and pinfect=2 and a<1 then foi=1.643*exp(lp22)+ lpder2*exp(lp22)/(1+exp(lp22));
if site=2 and pinfect=3 and a<1 then foi=1.643*exp(lp23)+ lpder2*exp(lp23)/(1+exp(lp23));
if site=2 and pinfect=4 and a<1 then foi=1.643*exp(lp24)+ lpder2*exp(lp24)/(1+exp(lp24));
if site=3 and pinfect=1 and a<1 then foi=1.643*exp(lp31)+ lpder3*exp(lp31)/(1+exp(lp31));
if site=3 and pinfect=2 and a<1 then foi=1.643*exp(lp32)+ lpder3*exp(lp32)/(1+exp(lp32));
if site=3 and pinfect=3 and a<1 then foi=1.643*exp(lp33)+ lpder3*exp(lp33)/(1+exp(lp33));
if site=3 and pinfect=4 and a<1 then foi=1.643*exp(lp34)+ lpder3*exp(lp34)/(1+exp(lp34));
output;
end; end; end; end; end; run;
*sort data;
proc sort data= numaveragingprevfoinc; by a site pinfect L;run;
*Get means;
proc means data= numaveragingprevfoinc; var pi foi; by a site pinfect L; output out=outpifoinc; run;
*Keep data for marginalized means;
data marginalizedprevandfoinc; set outpifoinc; where _stat_='MEAN'; run;
```

Appendix C

Appendix – Chapter 6

In this appendix, we present additional results and methods plus the R code and the SAS macro for the work covered in Chapter 6. In particular, Section C.1 presents extra results for the simulation study described in Section 6.5 of Chapter 6. Section C.2 presents a methodology accounting for interval-censored infection times with application to the PRISM study (see Section 3.1 in Chapter 3). Subsection C.2.2 presents the fit statistics (Table C.2) for the models fitted to the PRISM data by assuming different parametric distributions for the underlying age-specific malaria FOI. The rest of the content in this appendix is clearly referred to in Chapter 6.

C.1 Simulation study

Table C.1: Average number of malaria episodes, by varying percentage of assumed symptomatic infections (P). The labels C^+R^+ and R^+C^+ represent positive results at two near-by visits (C = clinical and R = routine) with the second observation deleted.

	All data		Data for scenario 4			
		Clinical	C^+R^+	R^+C^+		Clinical
P	N	%	%	%	N	%
20%	21832	8.4	0.2	0.03	21781	8.3
40%	22678	11.8	0.4	0.05	22576	11.8
60%	23520	15.0	0.6	0.07	23370	15.1
80%	24368	17.9	0.7	0.09	24169	18.2
100%	25218	20.7	0.9	0.10	24971	21.2

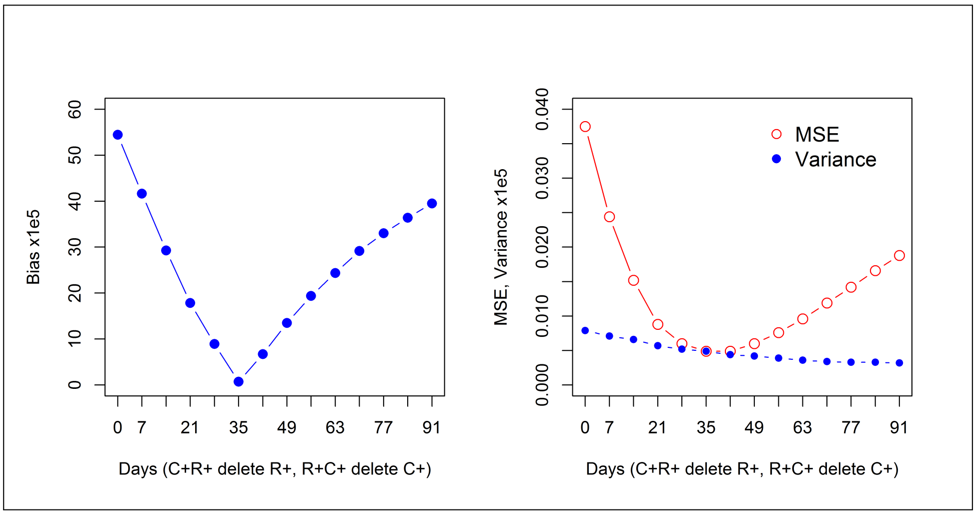

Figure C.1: Sensitivity analysis for bias, MSE and variance obtained using scenario 4 by considering different number of days (1 week interval) between two consecutive visits with positive results. Bias and MSE are minimal if positive results observed within 35 days are considered to be of the same infection. C+ and R+ represent positive result/infection at clinical and routine visits, respectively.

C.2 Data application

C.2.1 Interval-censored infection times

Interval censoring occurs if the time at risk T_{AR} is only known to lie between two time points. In the PRISM study, the time to the second, third or the n-th infection is only known to lie between the point the child is tested positive and the point he/she first tested negative after recovering from the previous infection. Generally, if the real time at risk t_{AR} for the n-th infection of an individual of age a when becoming susceptible again at calendar time $t_{(n-1)}$, lies between t_L and t_U, then the probability density function for the time at risk is given by

$$\begin{aligned} f_{IC}(t_{AR}|a) &= P(t_L \leq T_{AR} \leq t_U|a) = F(t_U|a) - F(t_L|a) \\ &= S(t_L|a) - S(t_U|a), \end{aligned} \tag{C.1}$$

where $f_{IC}(t_{AR}|a)$ is the modified density function for interval-censored data (t_{AR}, a); $S(t_L|a)$ and $S(t_U|a)$ are the conditional survival functions evaluated in t_L and t_U, respectively, i.e., for t_L,

$$S(t_L|a) = e^{-\int_a^{a+t_L} \lambda^*(u)du},$$

where $\lambda^*(u)$ is the infection hazard (for symptomatic infections). In case of **exponential** infection times, we have $\lambda^*(u) \equiv \lambda^*(u|\boldsymbol{x}) = \vartheta_1 e^{\boldsymbol{\zeta}'\boldsymbol{x}}$ and $S(t|a) = e^{-\vartheta_1 e^{\boldsymbol{\zeta}'\boldsymbol{x}} t}$, which implies

$$f_{IC}(t|a) = e^{-\vartheta_1 e^{\boldsymbol{\zeta}'\boldsymbol{x}} t_L} - e^{-\vartheta_1 e^{\boldsymbol{\zeta}'\boldsymbol{x}} t_U}.$$

Alternatively, for the **Weibull** and **Gompertz** distributions, it is straightforward to obtain similar expressions based on the expressions for the hazard functions in Table 6.1 in the main text. Finally, in case of the **fractional polynomial** model, we have $\lambda^*(u) \equiv \lambda^*(u|\boldsymbol{x}) = -\vartheta_2 u^{-2} e^{\vartheta_2 u^{-1}} e^{\boldsymbol{\zeta}'\boldsymbol{x}}$ and

$$S(t|a) = e^{-e^{\boldsymbol{\zeta}'\boldsymbol{x}}\left[e^{\vartheta_2 (a+t)^{-1}} - e^{\vartheta_2 a^{-1}}\right]}.$$

Note that in case the first event recorded for an individual of age a is a clinical malaria infection, the time at risk lies in the interval $[t_L, t_U] = [t^o_{AR}, (a-\nu) + t^o_{AR}]$ where t^o_{AR} is the observed time at risk, a is the age of the individual at the entry of the study, and $0 \leq \nu \leq a$ is the age of the individual when becoming susceptible after the last infection prior to the inclusion into the study, thereby giving rise to a contribution $S(t^o_{AR}|a, \nu = 0) - S((a-\nu) + t_{AR}|a, \nu)$ to the likelihood function. Since ν is unknown, we need to marginalize over the probability density function of the random variable ν. However, this leads to complicated expressions for the likelihood function, hence, in this manuscript, we take $S(t^o_{AR}|a) - S(a + t^o_{AR}|0)$ as likelihood contribution, implying that $[t_L, t_U] = [t^o_{AR}, a + t^o_{AR}]$, and we consider the aforementioned marginalization strategy as further research which is beyond the scope of this paper. Hereunder, we describe 4 possible situations for the treatment of interval censoring in the PRISM study. First, let $t_{(n)}$ be the calendar time at which one tests positive for the n-th infection ($n > 1$), $t_{(n-1)}$ the point at which one first tests negative from the $(n-1)$-th infection, and $t^*_{(n-1)}$ be the calendar time one was last observed positive for the $(n-1)$-th infection.

Situation 1: If $t_{(n-1)}$ and $t_{(n)}$ are exactly the points when one becomes susceptible and infected, respectively, then time at risk, $t_{AR} = t_{(n)} - t_{(n-1)}$. In this case there is no interval censoring and the contribution to the likelihood is simply $f(t_{AR}|a)$, where a is the age of the individual at time $t_{(n-1)}$.

Situation 2: If $t_{(n-1)}$ is exactly the point when one becomes susceptible, then the time at risk, $t_{AR} \in [0, t_{(n)} - t_{(n-1)}]$, meaning that $t_L = 0$ and $t_U = t_{(n)} - t_{(n-1)}$. Consequently, a represents the age of the individual at time $t_{(n-1)}$ in likelihood contribution (C.1).

Situation 3: If $t_{(n)}$ is exactly the point when one becomes infected, then the time at risk, $t_{AR} \in [t_{(n)} - t_{(n-1)}, t_{(n)} - t^*_{(n-1)}]$, meaning that $t_L = t_{(n)} - t_{(n-1)}$, $t_U = t_{(n)} - t^*_{(n-1)}$ and a represents the age of the individual at calendar time $t^*_{(n-1)}$.

Situation 4: If $t^*_{(n-1)}$ is exactly the point when one becomes susceptible, then the time at risk, $t_{AR} \in [0, t_{(n)} - t^*_{(n-1)}]$, meaning that $t_L = 0$, $t_U = t_{(n)} - t^*_{(n-1)}$ and a represents the age of the individual at calendar time $t^*_{(n-1)}$.

The statistical analysis presented in this paper is based on Situation 2, though the other situations are also plausible and worth considering, albeit that these scenarios are all approximations of the thruth. The impact of assumming Scenarios 3–4 on inference was found to be minor and the conclusions did not change.

C.2.2 Fit statistics

Table C.2: Fit statistics for models fitted to PRISM data based on scenario 2 and 4 by study site. Better fits for each site and scenario based on AIC are indicated in bold.

Site	Fit statistic	Exponential	Weibull	Gompertz	Fractional polynomial
SCENARIO 2:					
Walukuba:	AIC	1902.6	1867.7	**1867.0**	1867.8
	BIC	1921.9	1889.8	1889.1	1889.8
Kihihi:	AIC	6446.5	6440.1	**6411.1**	6492.6
	BIC	6465.2	6461.5	6432.5	6513.9
Nagongera:	AIC	9864.8	**9863.0**	9866.0	9889.1
	BIC	9883.6	9884.4	9887.4	9910.5
SCENARIO 4:					
Walukuba:	AIC	2012.0	2008.3	**1992.4**	2092.3
	BIC	2039.6	2049.0	2025.6	2122.7
Kihihi:	AIC	6683.8	6028.1	**5975.8**	7345.5
	BIC	6710.5	6060.2	6007.9	7384.1
Nagongera:	AIC	9554.4	9327.2	**9304.6**	10373
	BIC	9581.1	9359.3	9336.6	10403

C.3 R code

```
## Figure 6.2 (top left)
plot(subset(margprevfoiSc2Nag, L==0 & pinfect==3, select = c(a, pi)), ylim=c(0,1), type=’l’, lwd=2., xlab=”Age (years)”, ylab=”Marginal Prevalence”, col=”blue”, cex.axis=1.3, cex.lab=1.5)
lines(subset(margprevfoiSc4Nag, L==0 & pinfect==3, select = c(a, pi)), lwd=2, lty=2, col=”red”)
legend(”topleft”, c(”Scenario 2”, ”Scenario 4”), lty=c(1,2), col=c(”blue”, ”red”), lwd=2, cex=1.5, bty=’n’)
## Figure 6.3 (Nagongera)
# Top left figure
# Store data of mean FOI by age group
datmeansNag = read.table(text = ”A B C
1 0.086 0.330 5.184
2 0.053 0.137 2.285
3 0.024 0.045 0.577
4 0.100 0.407 6.817”, header = TRUE)
barplot(as.matrix(datmeansNag), col=terrain.colors(4), ylim=c(0,15), ylab=”Marginal annual FOI”, cex.lab=1.4, cex.axis=1.2)
legend(”topleft”, legend = c(”Negtive, no AL”, ”Negative, AL”, ”Symptomatic”, ”Asymptomatic”), fill=terrain.colors(4), cex=1.3, title=””, bty=’n’)
text(0.8, 6, cex=1.3, paste(”A:”, ’<1 year ’))
text(0.8, 4.8, cex=1.3, paste(”B:”, ’1-4 years ’))
text(0.8, 3.6, cex=1.3, paste(”C:”, ’5-10 years’))
# Bottom left figure
maxfoitNag=max(subset(margfoiSc4Nag, select = c(foit)))
plot(subset(margfoiSc4Nag, select = c(t, foit)), ylim=c(0,maxfoitNag), type=”n”, lwd=2, xlab=”Time at risk (years)”, ylab=”Marginal annual FOI”, cex.axis=1.3, cex.lab=1.5)
lines(subset(margfoiSc4Nag3, a==1, select = c(t, foit)),lwd=2, lty=1, col=”green”)
lines(subset(margfoiSc4Nag3, a==2, select = c(t, foit)),lwd=2, lty=1, col=”red”)
lines(subset(margfoiSc4Nag3, a==3, select = c(t, foit)),lwd=2, lty=1, col=”blue”)
lines(subset(margfoiSc4Nag3, a==4, select = c(t, foit)),lwd=2, lty=1, col=”orange”)
legend(0, 10, c(”1 year”, ”2 years”,”3 years”, ”4 years”), col=c(”green”,”red”,”blue”,”orange”),lwd=2, cex=1.5, bty=’n’)
legend(-0.2,11.5, ”Age”, bty = ”n”, cex=1.4)
```

C.4 SAS macro

****Scenario 2: Gompertz distribution, Walukuba site;**

```
proc nlmixed data=Cohortfulldata1 maxiter=100 NOAD;
where siteid=1;
parms B3=-0.3514 B4=-0.4549 B5=1.1101 B9=0.1079 logalpha = -4.3486 beta = -0.1 logsigma12=-2.1711
logsigma22=0.4327;
alpha = exp(logalpha);
sigma12 = exp(logsigma12);
sigma22 = exp(logsigma22);
b=bi1+bi2;
ha = log((alpha/beta)*(exp(beta*cohortage)-1));
fBX= B3*PT3 + B4*PT1 + B5*PT2 +B9*yearshift;
eta=ha + fBX;
p=1-exp(-exp(eta+b));
if p=0 then p=1e-10;
if p=1 then p=0.9999999;
ll = parasitemia*log(p) + (1-parasitemia)*log(1-p);
model parasitemia ~ general(ll);
random bi1 ~ normal(-sigma12/2,sigma12) subject = id(hhid);
random bi2 ~ normal(-sigma22/2,sigma22) subject = hhid;
estimate ’alpha’ alpha;
estimate ’B0’ logalpha;
run;
```

****Scenario 4: Gompertz distribution, Walukuba site;**

```
proc nlmixed data=PRISMdatamodel MAXITER=1000 lognote = 3 ITDETAILS;
where siteid=1;
parms logalpha1=-2.2623 logalpha2=1.2151 B3=0.08546 B4=-0.4951 B5=1.3604 B9=0.07841 E9=0.03941
logsigma12=-1.3128 logsigma22=-0.05639
beta1=-1 beta2=-0.1 logp0=-0.1;
sigma12 = exp(logsigma12);
sigma22 = exp(logsigma22);
alpha1=exp(logalpha1);
alpha2=exp(logalpha2);
p0=exp(logp0)/(1+exp(logp0));
*Random effects vector;
b=bi1+bi2;
*Outcome model;
ha=log(alpha1/beta1*(exp(beta1*cohortage)-1));
fBX= B3*PT3 +B4*PT1 +B5*PT2 +B9*yearshift;
eta=ha + fBX;
p=1-exp(-exp(eta+b));
if p=0 then p=1e-10; *Correction for termination in case of log(0) while maximizing the likelihood;
if p=1 then p=0.9999999; *Correction for termination in case of log(1-1) while maximizing the
likelihood;
*Time model;
fEX= E9*yearshift;
fLT=exp(-alpha2/beta2*exp(fEX+b)*(exp(beta2*(cohortage + timeyrsatrisk))-exp(beta2*cohortage))) -
exp(-alpha2/beta2*exp(fEX+b)*(exp(beta2*(cohortage + timeyrsatrisk))-1));
fIC=1-exp(-alpha2/beta2*exp(fEX+b)*(exp(beta2*(cohortage + timeyrsatrisk))-exp(beta2*cohortage)));
if delta=0 then
ll = parasitemia*log(p) + (1-parasitemia)*log(1-p);
else if delta=1 and parasitemia=0 then
ll = (1-parasitemia)*log(p0);
```

```
else if delta=1 and tau=1 then
ll = log(1-p0)+log(fLT);
else
ll = log(1-p0)+log(fIC);
model parasitemia   general(ll);
random bi1 ~ normal(-sigma12/2,sigma12) subject = id(hhid);
random bi2 ~ normal(-sigma22/2,sigma22) subject = hhid;
estimate 'alpha1' alpha1;
estimate 'alpha2' alpha2;
estimate 'B0' logalpha1;
estimate 'p0' p0;
estimate 'sigma12' sigma12;
estimate 'sigma22' sigma22;
run;

**Numerical averaging: Scenario 4: Gompertz distribution, Nagongera site;
**For the outcome process, a SAS macro similar to that in Section 6.8.6 was used
*Time process;
data numavgfoiSc4Nag;
*Add global parameters; zeta=0.92; vartheta1=0.0006; vartheta2=1.03; d11=0.94; d22=0.37;
do subject=1 to 1000 by 1;
bi1=rannor(123); bi2=rannor(123); *randomise from standard normal, for cholesky decomposition;
rd11=d11**0.5; rd22=d22**0.5; *For cholesky decomposition;
r1=rd11*bi1; r2=rd22*bi2; *cholesky decomposed variances;
do L=0 to 13 by 1;
do a=1 to 10 by 1;
do t=0.1 to 5.0 by 0.1;
*random effects;
b=r1+r2;
*Time component of the model;
lambda0at=vartheta1*exp(vartheta2*(a+t));
lp=zeta*L+b;
*FOI, lambdai(a+t—XB)=lambda0(a+t)Exp(XB);
foit=lambda0at*exp(lp);
output;end;end;end;end;
run;
*sort data;
proc sort data= numavgfoiSc4Nag; by t a L;run;
*Get means;
proc means data= numavgfoiSc4Nag; var foit; by t a L; output out=outfoitNag;
run;
*Keep data for marginalized means;
data margfoiSc4Nag; set outfoitNag; where _stat_='MEAN'; run;
```