JOINT MODELLING OF RECURRENT CLINICAL MALARIA EPISODES AND LONGITUDINAL PARASITEMIA DATA

By

CHRISTOPHER CHIKHOSI STANLEY

A Thesis Submitted to the Faculty of Health Sciences,

University of the Witwatersrand in fulfilment of

the requirements for the Degree of

DOCTOR OF PHILOSOPHY with publications in the

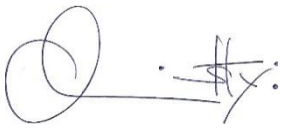
Major Area-Subject:

BIOSTATISTICS AND EPIDEMIOLOGY

December, 2020

Declaration

This thesis is a presentation of my own original and independent research work which has not been submitted, in whole or part, to this or any other institution of learning for the award of any other academic degree or diploma. Contributions of others are clearly indicated with acknowledgements for collaborative research or by explicit references where appropriate.



December 2020

Candidate signature **Date**

University of the Witwatersrand, Faculty of Health Sciences,

School of Public Health

7 York road, Johannesburg

December 2020

Dedication

This work is dedicated to my wife, children, parents, sisters and brothers for their moral and emotional support throughout this study. You have been my source of inspiration to work hard.

Above all, glory be to Almighty God for giving me a heathy life and innumerable blessings!

Thesis Summary

Background: Plasmodium falciparum infection is one of the most common parasitic infections in humans, with a disproportionately higher share of the problem in Sub-Saharan Africa (SSA). In Malawi, malaria is endemic and the prevalence of *Plasmodium falciparum* parasite is also high. Individuals in high-transmission settings are exposed to bites of infected Anopheles mosquitoes and develop frequent infections. With each infection, acquired immunity develops making subsequent disease episodes less likely. Therefore, longitudinal cohorts are ideal to study risk of malaria over time especially in high transmission areas because repeated infection is common. When modelling the risk of malaria in a cohort study, methods ought to account for this naturally acquired immunity which develops in individuals over time. Methods that focus on first incident alone (single-event joint model) excluding subsequent episodes may be inefficient since much information is lost and may underestimate the disease burden. A gap exists in literature on the use of joint modeling to analyze recurrent clinical malaria and parasitemia data collected during cohort studies.

Aim: To determine factors for recurrent malaria episodes in Malawi's Mfera cohort study through joint modelling of the recurrent episodes and longitudinal parasitemia data.

Methods: In order to investigate the trends of use of appropriate statistical methods among articles analyzing prospective longitudinal cohort data for clinical malaria episodes or Plasmodium infection published from 1996 until December 2017, a review of statistical

methods was conducted. Original articles in English, published over the 22-year period were reviewed in chronological order. Particular interest was on standard longitudinal analysis techniques including generalized estimating equations (GEE) and mixed-effects models which are considered appropriate for analyses of repeatedly measured data. For repeated time-to-episodes, recurrent Cox proportional hazards (PH)-based models such as frailty models, Andersen-Gill (AG), Prentice and Williams and Peterson (PWP), and the marginal means/rates models were considered appropriate. Other survival methods such as the Kaplan–Meier estimator, log-rank test and Cox PH model were considered only appropriate for analyses of time-to-single-episode.

Using a cohort data of participants enrolled with uncomplicated malaria in Malawi, a conventional Cox Proportional Hazards (PH) model of time-to-first clinical malaria episode with time-dependent parasitemia was compared with three competing joint models. The cohort enrolled 120 participants who presented with uncomplicated malaria at the Mfera health center in the Chikwawa district between June 2014 and March 2015. Participants underwent passive and active surveillance on a monthly basis and whenever sick for up to two years to assess the incidence of clinical malaria and *P. falciparum* infection. Four models of time-to-new clinical malaria disease were fitted. These included the reference model (model 1), a conventional Cox PH model fitted with time-dependent parasitemia measured at each visit, and three competing joint models. The three joint models had a common formulation only differing in the assumed association structure linking the hazard of clinical malaria with parasitemia. The hazard of clinical malaria disease at any time t was linked with: 1) the current underlying value of parasitemia at the same time point (model 2), 2) the rate of change in parasitemia trajectory at time t (model 3), and 3) the cumulative trajectory or area under

profile of parasitemia from baseline up to time t (model 4). Each joint model was formed of a mixed-effects sub-model of parasitemia longitudinal and Cox PH sub-model for time-to-first clinical malaria episode. Parameters were estimated using Markov chain Monte Carlo (MCMC) through the random walk Metropolis–Hastings (M-H) algorithm. Diffused normal priors were assumed for the covariates. Diagnostic assessments were conducted to assess the convergence of the MCMC samples using trace and kernel density estimator plots.

To assess whether jointly modelling recurrent clinical malaria episodes and parasitemia is more accurate than single-event joint models where the subsequent episodes are ignored, the joint model was applied to a real dataset from the Mfera malaria cohort. The models fitted were shared Gamma frailty model for the recurrent events and the mixed-effects model taking competing distributions for the longitudinal process: Poisson, quasi-Poisson and negative binomial (NB). Model fit was compared based on the Akaike information criterion (AIC). A model with lowest value of AIC is considered to be the best fitting model. In order to assess how the optimal recurrent event joint model can perform under different conditions, a simulation study was also conducted. In the simulations, data were generated with parameters that resemble the Mfera cohort. Age in years was assumed to be normally distributed on the log-scale, and values were transformed back to original scales by taking an exponential function to maintain the skewness that would reflect real data. The baseline hazard function was assumed to follow a Weibull distribution. Participants were assumed to be censored according to a uniform distribution on time of follow-up. Parasitemia data measurements were simulated from a mixed effects model with function of follow-up time. The model performance in terms of bias was assessed under different scenarios of: study sample sizes;

level of censoring; length of follow-up period; Gamma distributed frailty term variances; and correlation level of association between longitudinal and recurrent processes.

Results: In the statistical review, 197 (10.0%) articles met the inclusion criteria and were included in the analyses. The number of articles increased over the 22-year study period, with the highest number found between 2012 and 2016. Out of 197 articles reviewed, the most commonly reported methods included contingency tables which comprised Pearson Chi-square, Fisher exact and McNemar's tests (n = 102, 51.8%), Student's t-tests (n = 82, 41.6%), followed by Cox PH models (n = 62, 31.5%) and Kaplan–Meier estimators (n = 59, 30.0%). The longitudinal analysis methods GEE and mixed-effects models were reported in 41 (20.8%) and 24 (12.2%) articles, respectively, and increased in use over time. Six articles analyzed repeated time-to-episodes using appropriate recurrent models: Andersen-Gill (AG) (4); and frailty models (2).

There were 120 participants in the Mfera cohort, of which 69 (57.5%) were females. The overall median age was 7.5 years [inter-quartile range (IQR): 4.7 – 18.1], 6.3 years (IQR: 3.2 – 13.1) for males and 9.2 years (IQR: 5.3 – 18.5) females. Among the 115 participants who had at least one follow-up visit post enrolment included in the survival analyses, 372 recurrent clinical malaria episodes were experienced over the two-year follow-up period, of which 219 (58.9%) occurred among females. Out of the 372 episodes, 125 (33.6%) were among children under 5 years, 193 (51.9%) in school aged children of 5–15 years and 54 (14.5%) were experienced by adults above 15 years old. Overall, there was a decreasing trend in number of recurrent clinical malaria episodes during follow-up from 106 episodes in first month to 75

episodes in 24th month, for a drop of 3 participants in the period. In terms of precision, the recurrent events joint model outperformed single event joint model which underestimated the hazard ratios of clinical malaria in most cases with larger standard errors. Based on the recurrent event joint model, the hazard of recurrent clinical malaria decreased with increasing participants' age (hazard ratio [HR]=0.960 [95% confidence interval [CI]: 0.944, 0.976]), for 1-year increase in age. The hazard was also higher among participants who reported not to use LLINs every night compared to those who used the nets nightly (HR=1.424, [95% CI: 1.218, 2.032]). Compared to dry season, the hazard was higher during the rainy season (HR=1.355, [95% CI: 1.048, 1.753]).

In the simulation study, the performance of the recurrent event joint model improved as measured from decreasing bias with increased study sample size overall from 100, 200 to 400 participants, and by increasing length of study follow-up from 2 years to 3 and 4 years. The performance of the joint model worsened by increasing level of censoring from 10% to 20% and 50% in that order worsened the performance of the joint model.

Conclusions: Despite similar study designs across the reports, the statistical methods varied substantially and often represented overly simplistic models of risk. The results underscore the need for more effort to be channeled towards adopting standardized longitudinal methods to analyze prospective cohort studies of malaria infection and disease.

Modelling the risk of clinical malaria using the recurrent event joint models with improved precision may enhance study efficiency and allow for design of clinical trials with relatively lower sample sizes with increased power. Underestimation of risk by the single event joint

model may lead to incorrect inferences and wrong conclusions. The simulation study show that the recurrent events joint model can give parameter estimates with low bias. These results underline the need for models to utilize all collected follow-up data when estimating the risk of recurrent clinical malaria.

Keywords: Plasmodium falciparum, Longitudinal studies, Longitudinal analysis, Cohort studies, Prospective studies, longitudinal data, malaria parasitemia, time-to-event analysis, clinical malaria, Cox proportional hazards model, joint modelling, recurrent events, shared frailty model

Acknowledgements

This doctoral training was funded by the Fogarty International Center of the National Institutes of Health [Grant number D43TW010075]; National Institute of Allergy and Infectious Diseases [Grant number K24AI114996] to Profs Miriam Laufer and Don Mathanga; and partially supported through the DELTAS Sub-Saharan Africa Consortium for Advanced Biostatistics (SSACAB) program with funding from the Wellcome Trust [Grant number 107754/Z/15/Z] to Prof Tobias Chirwa.

I would like to express my sincere thanks to my main supervisor Prof Tobias Chirwa, and co-supervisors Profs Miriam Laufer, Mavuto Mukaka and Lawrence Kazembe who generously gave their time to provide guidance, critiques and insightful comments on the research. Your immense knowledge and enthusiastic encouragement pushed me to finish this work.

All co-authors on the papers in this thesis, thank you for accepting to be on the publications. I really appreciate for your contribution. Your expertise and insights propelled me through the course of this journey learning that the most effective way to do it, is to do it.

Special thanks to my fellow trainees whom we shared words of advice and encouragements during both low and high moments of this study.

My contributions to papers in the thesis

The following were my contributions as the first author in the thesis papers in chapters 3 to 5:

- Chapter 3: Conceptualization of research question, structuring of the paper, data collection, statistical data analysis, writing of all drafts of the paper, addressing co-authors' and peer reviewers' comments, and checking proofs for the final accepted manuscript.
- Chapter 4: Conceptualization of research question, structuring of the paper, statistical data analysis, writing of all drafts of the paper, addressing co-authors' and reviewers' comments, corresponding with journal and checking proofs for the final accepted manuscript.
- Chapter 5: Conceptualization of research question, structuring of the paper, statistical data analysis and writing of all drafts of the paper, addressing co-authors' and reviewers' comments, corresponding with journal and checking proofs for the final accepted and later published manuscript.

Table of contents

List of Figures

List of Tables

Acronyms / Abbreviations

AIDS	Acquired Immune Deficiency Syndrome
ANOVA	Analysis of Variance
ART	Antiretroviral therapy
GEE	Generalized Estimating Equations
HIV	Human Immunodeficiency Virus
HR	Hazard ratio
HREC	Wits Human Research Ethics Committee
IPTp	Intermittent preventive treatment in pregnant women
IPW	Inverse Probability Weighting
LLIN	Long-Lasting Insecticide-Treated Net
MALVAC	Malaria Vaccine Advisory Committee
MAR	Missing at Random
MCAR	Missing Completely at Random
MCMC	Markov Chain Monte Carlo
MEM	Mixed-effects models
MI	Multiple Imputation

ML	Maximum Likelihood
MNAR	Missing not at Random
RDT	Rapid diagnostic test
SSA	sub-Saharan Africa
WEE	Weighted Estimating Equations
WHO	World Health Organization

Chapter 1

Introduction

1.1 Background

Plasmodium falciparum infection is one of the most common parasitic infections in humans. Individuals in high-transmission settings are exposed to bites of infected Anopheles mosquitoes and develop frequent infections. In early childhood, infections are almost always associated with symptomatic disease (1). Globally, there were an estimated 219 million cases and 435 000 related deaths in 2017, with a disproportionately higher share of the problem in Sub-Saharan Africa (SSA), home to 92% of cases and 93% of malaria deaths as in appendix 1.1(2). In Malawi, malaria is endemic with approximately 1150 hospital admissions per 100 000 population, 215 cases per 1000 population and estimated 7 200 deaths in 2015 (3). Prevalence of *Plasmodium falciparum* parasite is also high in Malawi as shown in Appendix 1.2. (4). Data used in this study came from Chikwawa, a district in the rural area in southern Malawi. Although malaria is endemic throughout Malawi, the highest risk of the infection occurs along low-lying regions that are hotter, wetter and more humid, and the risk is lowest in highlands (5). Chikwawa district is located along a low-lying region which is usually hot, wet and humid with high rates of malaria parasite transmission by anopheline vectors (6,7) with an estimated entomological inoculation rate (EIR) of 172 infectious bites per year (8).

In Malawi, the National Malaria Control Program plan to control malaria has been focusing on prevention strategies including use of long-lasting insecticide treated bed nets (LLINs) (9–

11), prompt access to ACTs, intermittent preventive treatment in pregnancy (IPTp) and indoor residual spraying (IRS) in focal areas (12,13). Recently, in 2019, Malawi together with Kenya and Ghana rolled-out two pilot interventions namely: the use of piperonyl butoxide (PBO) nets; and RTS,S/AS01 (RTS,S) malaria vaccine in the context of routine use as one of World Health Organization (WHO) Malaria Vaccine Implementation Program (MVIP) (14). Despite the scale up of different combinations of the above prevention strategies, this has not resulted in a sustainable reduction of the risk of malaria and its morbidity in Malawi as well as the region (15–19).

In order to measure progress in the reduction of malaria burden, different studies are conducted on the disease prevalence and incidence which also serve to inform program implementation and better target interventions (4,6,20). These studies have often been cross-sectional (21) or repeated panel surveys (22,23) while cohort studies are scarce. Unlike other designs, cohort studies, with regular follow-ups, provide an opportunity to study the trend of outcomes such as clinical malaria or parasitemia for the same participants over time, thereby generating longitudinal and event-time data. Therefore, longitudinal cohorts are ideal to study risk of malaria over time especially in high transmission areas because repeated infection is common and with each infection, acquired immunity develops making subsequent disease episodes less likely (24,25). Over time, individuals acquire immunity to clinical disease as shown in appendix 1.3.

When modelling the risk of malaria in a cohort study, methods ought to account for this naturally acquired immunity which develops in individuals over time. In order to estimate the

risk of malaria, modelling clinical disease episodes (as recurrent events) jointly with the repeated measurements of *P. falciparum* parasitemia (as longitudinal data) may be crucial to capture the effect of the developed protective immunity to malaria. Applications of joint modelling of longitudinal and event-time data include HIV programs where evolutions of viral load or CD4 count have been modelled jointly with time to development of AIDS (26,27). In patients with cancer, performance scores have frequently been modelled jointly with time to death or tumor recurrence (28,29). Unlike in HIV and cancer (examples above) where once one develops the disease then mostly likely they will have it for life, malaria can have more than one disease episodes. In malaria, the joint modelling of longitudinal and event-time data have been on time to first episode (30). When following up individuals, methods that focus on first incident alone may underestimate the disease burden and therefore repeated episodes provide a better understanding of malaria characteristics over time. For malaria, there is an opportunity to study not only the first clinical malaria occurrence but more than one episode. There is clearly a gap in literature on the use joint modeling to analyze recurrent clinical malaria and parasitemia data collected during cohort studies. Meanwhile, malaria experts are urged to analyze recurrent event times to further methodological research and to evaluate malaria vaccines as per WHO Malaria Vaccine Advisory Committee (MALVAC) recommendation (31). This study aims to use a joint modelling approach to study risk factors for recurrent malaria episodes in presence of longitudinal measurements of parasitic count.

1.2 Problem statement and rationale

The burden of malaria is likely underestimated when methods used only consider first malaria episode for an individual. Therefore, there is a need to understand the risk factors for recurrent

episodes of malaria through models that can appropriately handle within-participant dependence of recurrent events and the naturally acquired disease immunity over time. Traditionally, time-to-event and longitudinal (non-survival) outcomes have been analyzed as two isolated fields, with much priority on the time-to-events. The longitudinal measurements data are usually downgraded to a secondary part as covariate. Since covariates are typically measured periodically, it is common to fill in the missing values by carrying forward the last observed values before each event time and this typically leads to bias especially when the last observation is distant in time (32). In order to address this and understand the risk of recurrent clinical malaria episodes among individuals in a cohort, we need methods that can jointly model the recurrent events and the longitudinal parasitemia measurements data. This is the research gap that the study aims to fill and is summarized in appendix 1.5.

1.3 Aims and specific objectives

1.3.1 Overall aim

To determine factors associated with recurrent malaria episodes in Malawi's Mfera cohort study through joint modelling of the recurrent episodes and longitudinal parasitemia data.

1.3.2 Specific objectives

i. Do a review of statistical methods for modelling of malaria cohort data.
ii. Determine whether jointly modelling time-to-new clinical malaria episode and parasitemia data improves precision compared to time-dependent Cox model.
iii. Assess whether jointly modelling recurrent clinical malaria episodes and parasitemia is more accurate than single-event joint models where the subsequent episodes are ignored.

iv. Predict probability of a new clinical malaria episode given participant's history of recurrent events and parasitemia trajectory using the optimal joint model.

v. Using simulations, assess the performance of the joint model under varying conditions of study sample size, length of follow-up and level of event censoring.

The next sections of the thesis are structured as follows: Chapter 2 presents the methodology with an overview of existing background literature of statistical models for the report. Chapter 3 is a peer reviewed article titled "*Systematic review of analytical methods applied to longitudinal studies of malaria"* which was published in Malaria Journal. Chapter 4 is a second paper on *"Joint modelling of time-to-clinical malaria and parasitemia in a cohort in an endemic area"* and is published in Journal of Medical Statistics and Informatics. Chapter 5 is a third paper from this thesis on *"Analysis of recurrent times-to-clinical malaria episodes and Plasmodium falciparum parasite from a cohort study: a joint modelling approach*" and is submitted for consideration to the BMC Research Methodology Journal. Lastly, Chapter 6 gives the overall thesis discussion and conclusions.

Chapter 2

Methodology

2.1 Data, variables and measurements

2.1.1 Data

This research used data from a prospective malaria cohort study conducted in a rural area of southern Malawi under Malawi International Center of Excellence for Malaria Research (ICEMR). The cohort enrolled 120 participants who presented with uncomplicated malaria at the Mfera health center in the Chikwawa district between June 2014 and March 2015. Initial diagnosis was made by rapid diagnostic test (RDT) and confirmed by microscopy using thick blood smears. Participants underwent passive and active surveillance monthly and whenever sick for up to two years to assess the incidence of clinical malaria and *P. falciparum* infection. In total, the cohort contributed a total of 2682 observations and accrued follow-up time of 73,514 days, median, 720 (interquartile range: 715–721) days.

2.1.2 Variables and measurements

Outcomes: The primary outcomes of the study were: 1) recurrent clinical disease defined as participants' self-reported fever and a positive RDT result, and 2) *P. falciparum* infection as parasitemia measured from thick smear during monthly or sick visits over the follow-up period.

Covariates: The models included participants' age, gender, season of visit categorized as dry (May-November) or rainy (December-April), and frequency of LLIN use in the previous month of the visit defined as whether one used the net every night. There were no missing data in all covariates considered in this study, except in 2 (0.1%) instances of LLIN net use.

2.2 Ethical consideration

This study was approved by the University of the Witwatersrand Human Research Ethics Committee (HREC) on 18th October 2017, with clearance certificate no. M170952 as attached in appendix 2.1. The data used in this thesis (as in section 2.1.1) came from a Mfera malaria cohort study which was approved by Institutional Review Boards (IRB)s and ethics committees of the College of Medicine of the University of Malawi and University of Maryland in USA. The data was anonymized during analysis.

2.3 Modelling longitudinal data

Longitudinal data which is generated by repeated measurements of same participant over time is becoming common in malaria and biomedical research overall. The main goal of conducting longitudinal studies is to characterize changes in the response of interest over time, in relation to some covariates. In malaria, examples include characterization of incidence of disease episodes, evolutions of parasite diversity, human host immune response and abundance of vectors of transmission. Covariates may include factors such as participants characteristics, environmental, social-economical, preventive and vaccine malaria interventions.

By design, the underlying key features in longitudinal data are: i) correlation – where response on one instance determines the likely value of the future response, and ii) heterogeneous variability – where the variance of the response varies over time. Together, the within-participant correlations of pairs of repeated measurements and the variability of the responses on different occasions is a covariance. The primary focus in models is normally on the mean response. However, inferences about change in the response in relation to covariates are also dependent on the covariance and if not properly accounted for, may lead to invalid inference (33).

Longitudinal studies are usually designed with pre-scheduled times of repeated measurements which consequently set common points for all participants. If this is achieved over the whole follow-up time, the data is considered to be balanced. However, in most settings of biomedical research it may not be practical for all study participants to have common sequence of measurement times rendering the data to be unbalanced over time. Participants can miss their scheduled visits for various reasons. Sometimes, by design studies such as those on drug safety measuring adverse events or characterizing disease incidence may require participants to come for unscheduled visits whenever sick.

Furthermore, missing data is also typical in longitudinal data, and can be due to several reasons including the participants missing visits, drop-out because of transfer from study area, withdrawal from study or death. Missing data threatens the validity of the results because if not handled properly may lead to bias and loss of efficiency, and results in decreased

statistical power and increased standard errors (34,35). The three patterns of missing data are univariate, monotone and arbitrary (36,37). Given m variables denoted as $x_1, x_2, x_3, \ldots, x_m$; data is said to be of univariate pattern if missing occurs in the same participant on at least one of the m variables. Monotone missing data pattern arises if a variable has missing due to missingness in variable preceding it and itself results in missing in the subsequent variable, i.e. when x_i is missing, $x_i + 1, x_i + 2, \ldots, x_m$ are also missing. Monotone pattern is frequent in longitudinal studies due to dropouts. When missing occurs randomly in any variable, it is called an arbitrary missing data pattern. When missing occurs randomly in any variable, it is called an arbitrary missing data pattern. The well documented three mechanisms of missing data according to Little and Rubin include, a) missing completely at random (MCAR), b) missing at random (MAR); and c) missing not at random (MNAR) (36,37). The MCAR occurs when the observed data are just random sample of all data and the failure to observe does not depend on any other unobserved or observed values. The MAR arises if failure to observe a value is independent of unobserved data but may depend on other observed values. Data are said to be MNAR if failure to observe a value depends on the value that would have been observed regardless of observed data. The ad-hoc methods to handle missing data include: complete case analysis - which uses only individuals who have all variables observed; list wise deletion (where an observation is dropped from analyses due to missing value in one or more of the specified variables) and the pairwise deletion (in which analysis procedure drops a particular variable due to a missing value, but it can still use the observation when analyzing other variables with non-missing values). However these methods are prone to yielding biased and inefficient estimates in most situations and are discouraged (38–40). Different contemporary and principled methods have been proposed for handling missing data. These methods include multiple imputation (MI), inverse probability weighting (IPW), doubly robust inverse probability weighting (DR-IPW), maximum

likelihood (ML), weighted estimating equations (WEE) and Bayesian methods. Excellent overviews on contributions in methodology for handling missing data include Ibrahim et al 2005 (41), Graham 2009 (42), Ibrahim and Molenberghs 2009 (35). Recently, Tseng et al 2016 have proposed latent random effects method for analyzing longitudinal outcomes in the context of MNAR arising from individuals intermittently missing scheduled visits or drop-outs (43). In practice, choice of the method is dependent on assumptions about missing mechanism present in the data at hand (35,44). Under frequentist inference, MCAR is commonly ignorable while in the likelihood and Bayesian paradigms, both the MCAR as well as MAR are ignorable. In this study MAR was assumed and so likelihood and Bayesian methods which were used are both valid.

Longitudinal data analysis methods should appropriately account for the above discussed attributes namely: covariance structure, balanced vs unbalanced design and missing data. The covariance structure contradicts the independence of within-participant measurements and homogeneity of variance over time which are fundamental assumptions of many ad hoc methods such as t-test, analysis of variance (ANOVA), and multiple linear regression. Additionally, the stated methods require that the data be balanced but this is usually not possible because of frequent unscheduled visits or data missingness.

The recent approaches common in practice for longitudinal data analysis that accommodate covariance structure and unbalanced designs are Mixed-effects models (MEMs) (45) and generalized estimating equations (GEEs) (46). The MEM is an individual-level (conditional) approach adopting random effects to capture the correlation between the observations of the same participant. For normally distributed outcomes, a linear mixed effect model can be

marginal or conditional and the choice between the two formulations is not particularly critical as the interpretation of the conditional is equivalent to marginal models. On the other hand, GEE is a population-level approach based on a quasi-likelihood function (47–50) and provides the population-averaged estimates of the parameters. Gibbons et al 2010 provide a detailed comparison between MEM vs. GEE in a review of recent advances in longitudinal data analysis (51). They advocated for MEM over GEE when the study has dropouts because unlike GEE, MEM uses all available data from individuals to model temporal responses that would be observed if not for dropouts. In order to accommodate missing data specifically under MAR, the weighted-GEE has been widely used (52). Under GEE, Wang 2014 gives detailed review and recent developments including choice of working correlation structures as summarized in appendix 2.2 (48). Recently, Shen and Chen 2012, 2013 also proposed a missing longitudinal information criterion (MLIC) for model specification, and the MLIC for correlation (MLICC) for selection of the correlation structure when there is dropout (53,54).

In the current study, the data are unbalanced because participant were encouraged to come to clinic whenever sick. Further than that, some participants did not show up on the exact scheduled dates with often a delay of few days, but dropout was infrequent. Therefore, in this study, mixed-effects model is utilized to model the longitudinal parasitemia process.

The mixed-effects sub-model: For the mixed-effects models, studies have frequently focused on continuous outcomes and often applied linear mixed-effects models (55–57). The linear mixed-effects model with Gaussian response can be expressed as follows; supposing data is available from N participants with n_i observations recorded for participant $i = 1, \dots, n$. The

response $y_{ij}, (j = 1, \dots, n_i)$, fixed-effect covariate vector $X_{ij} = (x_{1ij}, \dots, x_{pij})'$, and random-effect covariate vector $Z_{ij} = (z_{1ij}, \dots, z_{qij})'$ are recorded at times t_{ij}. The linear mixed-effects model becomes;

$$y_{ij} = \beta X'_{ij} + b_i Z'_{ij} + \varepsilon_{ij}, \qquad (2.1)$$

where β is the $p \times 1$ fixed-effect parameter vector, b_i is the $q \times 1$ vector of random effects for participant i which is usually assumed to be multivariate normal with mean zero, i.e., $b_i \sim N_q(0, \sum_b)$, and $\sum_b$ is the variance of the subject specific effects (58). The error vector $\varepsilon_i = (\varepsilon_{i1}, .., \varepsilon_{nij})'$ is assumed to be distributed - $\varepsilon_i \sim N_{ni}(0, \delta^2 I_{ni})$ where δ^2 is variance and I_{ni} is the $n_i \times n_i$ identity matrix; and b_i and ε: independent.

However, when the response follows a non-Gaussian distribution, the linear mixed-effects models may not be appropriate. In this study, the parasitemia is the longitudinal response following a negative binomial (NB) distribution as it fitted parasitemia better than Poisson or quasi-Poisson based on AIC. Taking that response y_{ij}follows NB distribution with mean μ and shape parameter ω, then it's distribution can be;

$$y_{ij} \sim NB\left(y_{ij} \middle| \mu_i, \omega\right) = \frac{\Gamma(y_i+\omega)}{\Gamma(\omega) y_{ij}!} . \left(\frac{\omega}{\mu_i+\omega}\right)^{\omega} . \left(\frac{\mu_i}{\mu_i+\omega}\right)^{y_{ij}} \qquad (2.2)$$

Where $\Gamma(.)$ is the Gamma function. The equation (2.2) can be expressed as in form of Gamma mixture of Poisson distribution (59) with $y_{ij} \sim Poisson(y_{ij}|\mu_i \varepsilon_i)$and $\varepsilon_{ij} \sim \Gamma(\omega, \omega)$. It is shown in appendix 2.2 that $E(y_{ij}) = \mu_i$, and $Var(y_{ij}) = \mu_i + \frac{\mu_i^2}{\omega}$, where $Var(y_{ij}) \geq E(y_i)$. The shape parameter ω determines the degree of over-dispersion, when $\omega \to +\infty$, then $Var(y_{ij}) = \mu_i$ consequently making the NB to converge to Poisson model which does not handle over-dispersion. In the current study, parasitemia data were over-dispersed. Taking notation by (60), The NB mixed-effects model links the mean of response to the fixed and random effects through a log-link function as;

$$\log(\mu_i) = \beta X_i' + b_i Z_i' + \log(M_i), \qquad (2.3)$$

where $\log(M_i)$ is offset representing total number for a set of measurements.

2.4 Modelling survival data

Time to event data popularly called survival data in biomedical research is a special case of longitudinal data in which the primary scientific interest is on time until occurrence of an event such as disease episode, hospital discharge or death. A main characteristic of survival data is that the event times may be censored, i.e. event of interest not observed during the study period due to drop-out or study closure. Two types of censoring mechanisms are; i) informative in which the chance of being censored is related to the expected failure time, and ii) non-informative whereby censoring can depend on covariates unrelated to the event process but not on the expected failure time. Typically, in survival analysis, the interest can be

to describe the survival function, $S(t)$ defined as the probability that a participant is event-free at t, $S(t) = \mathrm{P}\ (T > t)$, where T is the true survival time. However, the common scientific interest is to characterize the hazard function (hazards rate) $h(t)$ of an event. The hazard function can be defined as an instantaneous probability of experiencing an event at time T having survived up to time t and can be presented as;

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\mathrm{P}\ (t \leq T < t + \Delta t | T \geq t)}{\Delta t}, \quad t \geq 0\ . \qquad (2.4)$$

Several survival models exist particularly for modelling time-to-single event. These include fully parametric models namely exponential, Weibull, Gompertz, log-normal, log-logistic and Accelerated Failure Time (AFT) models, all which assume a particular distribution for the hazard function. Overviews about parametric survival models are many and some of them can be found here (61,62). However, in most settings in biomedical research, imposing a specific parametric assumption on survival times may not be practical. This study focuses on a semi-parametric Cox Proportional Hazards model (PH) model, the mostly used survival model in biomedical research introduced by Cox et al 1972 (63) and it is discussed below. The Cox PH model is semi-parametric because it does not make assumption about the baseline hazard function $\lambda_0(t)$ but assumes parametric form for the effect of the predictors on the hazard. This implies that the main advantage with the Cox PH model is that one can obtain covariate risk parameter estimates without need to specify the $\lambda_0(t)$, which may otherwise be mis-specified sometimes. The Cox PH assumes that the hazard is proportional over time and that its function is linearly related to the set of covariates. The Cox PH model with the hazard function $\lambda_i(t)$ for participant i at given time t can be expressed as;

$$\lambda_i(t) = \lambda_0(t)exp[\beta_s X'_{si}(t)], \qquad (2.5)$$

where and $\lambda_0(t)$ is the unspecified baseline hazard function and $X_{si}(t)$ represent the covariate vector for participant i, and $\beta_s$denote a vector of coefficients for $X_{si}(t)$.

2.5 Modelling recurrent events data

Recurrent events data arises when a participant experience repeated occurrence of same event over time such as repeated episodes of clinical malaria (in this study), visual acuity, heart attacks and hospitalizations. Statistical methods for analyzing recurrent events include, recurrent Cox-extended models such as frailty models (64–68), Andersen-Gill (AG) (69), Prentice and Williams and Peterson (PWP) (70), and the marginal means/rates model (71). While overviews on these models can be found in (30,67,72,73), this section discusses the recurrent events models are in relation to malaria data.

For the counting process of the recurrent clinical episodes, let $R_i^*(t)$ be the number of recurrent events for participant i occurring before or at time t over an interval $[0, \tau]$, where recurrent episodes could potentially be observed beyond prespecified maximum time point τ. Then the counting process may be stopped by the time to loss to follow-up or end of the study denoted by C_i, with the failure indicator Δ_i taking value 1 if $T_i \leq C_i$ and 0 otherwise. The observed counting process, $R_i = R_i^* min(t, C_i)$ has a known zero-one process $\{G_i(t): 0 \leq t \leq \tau\}$ indicating whether the participant i is at risk of experiencing an episode during period u.

Thus, the counting process R_i^* has a jump of size one i.e. $(R, R+1, \ldots)$ when an event occurs i.e. episode of clinical malaria is experienced.

The AG model uses a counting process time-scale for all episodes assuming that the correlation between episodes-times for each participant is explained by previous episodes where time-scale does not reset to zero after an episode. In the AG model, the hazard function of the k^{th} event for participant i at time t can be expressed as $\lambda_{ik}(t_i|X,\beta) = \lambda_0(t)exp[\beta X_{ik}{}']G_{ik}(t)$, where $\lambda_0(t)$ denote the baseline hazard for all events over follow-up time; $\lambda_0(t)$ is a vector of coefficients, and $G_{ik}(t)$ is a Bernoulli process denoting whether the participant i is in the risk period or not. This model is typically applicable when the interest is on the overall effect of covariates on the intensity of the recurrent episodes.

The PWP model is similar to the AG model but analyses ordered events by stratification and can be expressed as gap-time model where the time-scale is reset to zero after each episode occurs. Gap-times between malaria parasitemia detection to clearance is a good example where the PWP model can be valid. The formula for hazard function in the PWP model can take the form as; $\lambda_{ik}(t_i|X,\beta) = \lambda_{0k}(t)exp[\beta X_{ik}{}']G_{ik}(t)$, where λ_{0k}represents the event-specific baseline hazard for the event k over time.

The marginal means/rates model considers all recurrent episodes for each participant as a single counting process and interpreted in terms of mean number of episodes. However, this

model does not incorporate time-dependent covariates to reflect the historical process of the episodes (71) and thus may not be appropriate for recurrent malaria data.

Shared frailty models are the most common used method for recurrent events data. The shared frailty model was introduced by Clayton 1978 (68) and extensively studied in Hougaard 2000 (74), Therneau and Grambsch 2000 (66), Duchateau and Janssen 2004 (75), and Duchateau et al. 2008 (76), as extensions to the Cox PH model by introducing a random covariate that induces dependence among the recurrent episode times. By conditioning on covariates and the random effect, the recurrent episodes are assumed to become independent. There are different flavors of distributions of shared frailty models which include; Gamma, inverse Gaussian and positive stable (77,78), and their respective advantages and disadvantages have been discussed in Hougaard 2000 (74). The shared Gamma frailty model is the mostly used distribution for recurrent events data because of its mathematical convenience (79). Computationally, although the Gamma models have no closed form expressions for hazard functions, they fit well into the proportional hazard model. Consequently, it is easy to derive the closed form expressions for unconditional hazard functions. This makes estimation of the model parameters easier by integrating out the frailties appearing in the conditional likelihood to get marginal likelihood. One can obtain parameter estimates by maximizing the marginal likelihood. Further, Abbring and van den Berg 2007 showed that under some regularity conditions, the frailty distribution among the survivors at a specific time t converges to a Gamma distribution (79).

In this study, a shared Gamma frailty model was used to model the recurrent clinical malaria episodes. The model can be defined as follows; let β be a vector for covariate vector X_i for participant subject $i = 1, \dots, N$ and denote the recurrent event times as $t_{ij} = 1, \dots, n_i$. The shared frailty model with a Gamma frailty term can be expressed as follows;

$$\lambda_i(t_i|\gamma_i) = \lambda_0(t_i)\gamma_i exp[\beta X_i{}']G_i(t), \qquad (2.6)$$

the baseline hazard function λ_0 is assumed to follow a Weibull distribution. The term γ_i represents the unobservable random effects (frailty) term to account for dependence in within-participant episodes and is assumed to follow a Gamma distribution with unit mean and variance θ, i.e. $\gamma_i \sim \Gamma(1, \theta)$. The frailty variance θ, reflects the amount of the within-participant dependence of clinical episode times i.e. the correlation of the recurrent events is quantified by θ, with higher values corresponding to greater within-participant correlation. A Bernoulli process, $G_i(t)$, denotes whether the participant i is in the risk period of experiencing the malaria episode. The likelihood of the shared Gamma frailty model is derived in appendix 2.2.

2.5 Joint modelling of survival and longitudinal data

Joint modelling of the survival (time-to-single event) and longitudinal data become important when potentially the two data processes are associated. Typically, the joint model is composed of longitudinal and survival sub-models. Literature in this area has been reviewed by Hogan and Laird 1997 (80), Tsiatis and Davidian 2004 (27), Yu et al. 2004 (81) and Rizopoulos 2012 (82). Some of the contributions in the field include a joint model using a

parametric approach to model the time-to-event with likelihood inference studied by Schluchter 1992 (83) and DeGruttola and Tu 1994 (84). Self and Pawitan 1992 (85) have considered two-step approach to estimate parameters conditioned on the survival information when computing covariates expected values. In this method, estimates of the risk parameters were obtained through partial likelihood, but then to account for the uncertainty in the expected values of the covariates, corresponding variances were also derived by assuming fixed and known variance of the covariate random effects. Wulfsohn and Tsiatis 1997 (86) studied a fully likelihood based joint model using linear mixed-effects model and a Cox PH model with infinite-dimensional baseline hazard. Tsiatis et al. 1995 (87) proposed also a two-step process in which they assumed a growth curve random components model with normal errors for the longitudinal process and then use the modified expectation maximization (EM) algorithm for estimation. Then, these estimates are substituted into the proportional hazards model and the Cox PH model is used to obtain survival parameter estimates. Finally, most of the approaches reviewed focus on continuous longitudinal outcomes with linear distribution. Recently, Rizopoulos 2012 (82) has looked at a joint model which accommodates a case of non-linear and a non-Gaussian longitudinal outcome such as binary or count in the Bayesian approach in which joint model parameters are obtained using Markov chain Monte Carlo (MCMC) algorithms. This approach was considered in the current study since the longitudinal process, i.e. the parasitemia were neither continuous nor linear. Different association structures to link the longitudinal and survival outcomes have been studied (29,82,88). For example, Rizopoulos 2012 (82) assumed that the hazard for an event at any specific time of the follow-up is associated with: i) the current underlying value of the longitudinal outcome at the same time point, ii) the rate of increase/decrease of the biomarker's levels or iii) summary of the whole longitudinal trajectory. The joint model consists of a longitudinal sub-model taking the form of a mixed-effects model as in equation (2.1) while the survival sub-model

has a Cox PH model form. The hazard function $\lambda_i(t)$ for participant i at time t as given is modelled as;

$$\lambda_i(t) = \lambda_0(t)exp[\theta h(\beta, b_i, t) + \beta_s X'_{si}(t)], \qquad (2.6)$$

where $h(\beta, b_i, t)$ is a function of the fixed and random effects in the longitudinal sub-model, and θ is an association parameter linking the two sub-models: survival and longitudinal models (27). The β_s is a parameter vector for covariates unique to the survival sub-model. The survival covariate vector $X_{si}(t) = (x_{si1}, .., x_{sir})'$ includes baseline covariates for participant i with β_s representing $r \times 1$ parameter vector.

2.6 Joint modelling of recurrent events and longitudinal data

Joint modelling of longitudinal data with recurrent events (repeated time-to-events) becomes appropriate when study participants experience repeated episodes of the same outcome. In this scenario, considering time-to-single event while ignoring subsequent events may threaten the validity of the inferences. For example, in this study, the event history of clinical malaria episodes is particularly critical since each *P. falciparum* infection alters the host immune response against risk of subsequent episodes as detailed is section 1.1. Joint modelling of recurrent events and longitudinal data is receiving attention recently but literature is still in infancy stage. Among the proposed approaches, Han et al. 2007 (89) considered parametric latent class of joint models for a longitudinal outcome and recurrent events where the effects of accumulating events and possible interventions are captured through a latent class

structure. However, this method uses linear mixed-effects model for the longitudinal outcome. Mbognin et al. 2015 considered a joint modelling of longitudinal and recurrent events data using nonlinear mixed-effects models for continuous outcomes and the Stochastic Approximation Expectation-Maximization (SAEM) algorithm (90). However, this approach was suited for continuous outcomes only. Li 2016 studied a joint model of recurrent events process and an intermittently observed longitudinal binary covariate that is time-varying using Markov model for the binary covariate process and a random-effect proportional intensity model for the recurrent events process. In this approach, parameters were estimated using MCMC algorithm (91). In this study, the recurrent events model extends the single event semi-parametric proportional hazards model by introducing an unobservable (frailty) random term which the hazard function depends upon to account for within-participant dependence of events (92), i.e. clinical malaria episodes in this study. For recurrent clinical malaria episodes process, an intensity event model function is adopted as opposed to rate function because while the rate function only defines the occurrence of recurrent events unconditional on the event history, the intensity function conditions the occurrence of events on the event history (93). Thus, the intensity recurrent events model would account for the participants' strengthening immunity to episodes due to accumulating event occurrences over time, which is critical in recurrent events analysis (89). The shared Gamma frailty model representing an intensity recurrent events model as in Henderson et al. 2000 is expressed in equation (2.6) (55).

Likelihood of the joint model for recurrent and longitudinal data: Using generic terms Y to denote combined observed measurements of parasitemia data, R for combined recurrent episodes data, X for covariate information, and $\Phi = \{\varepsilon, \gamma\}$ for the random (or frailty)

processes, the joint distribution for the longitudinal measurements Y and recurrent events processes R are conditionally independent given X, ε and γ. The dependence between Y and R may arise as a result of the direct link between ε and γ, called latent association by Henderson et al (55), without which nothing can be gained through a joint analysis. Our interest is to model the subjects' recurrent processes of episodes together with their longitudinal measurements of parasitemia, through the association of ε and γ. Following framework proposed by Henderson et al (55), one can proceed to compute the joint likelihood as a product of the marginal distribution of observed measurements Y and the conditional distribution of the events (episodes) R and including a third term $\{t_{ij}\}$ for visit times process. However, if we assume that the timing of visits occurs independently of the measurements Y and events processes R, we can proceed to ignore this third term as non-informative. For each participant i, the observed data are $\{(t_i, X_i, u_i, \Delta_i, \tau_i), i = 1,2, \dots, n\}$. Computing the full joint likelihood $L = L(\pi, Y, K)$ where $\pi = (\beta, \sigma, \theta, \lambda_0, \alpha, \Phi)$ is a vector denoting a collection of all unknown parameters with $\lambda_0 = \{\lambda_0(t_i), i = 1,2, \dots, n\}$, one can proceed as follows:

$$L = L_Y \times L_{R|Y} = L_Y(\pi, Y) \times E_{Z_2|Y}\{L_{R|\gamma}(\pi, R|\gamma)\}, \quad (5)$$

Here $L_Y(\pi, Y)$ is the standard form of marginal distribution of Y for the parasitemia measurements process. The conditional likelihood of the recurrent episodes data, $L_{R|\gamma}(\pi, R|\gamma)$ captures the likelihood contribution of the longitudinal measurements up to any time of event. Suppose we denote $R_0 = \int_0^t \lambda_0(u)du$ as cumulative baseline intensity for the recurrent events process, then $L_{R|\gamma}(\pi, R|\gamma)$ can be expressed as;

$$L_{R|\gamma}(\pi, R|\gamma) = \prod_i \{(\prod_t [exp\{\beta X_i{}' + \gamma_i\}\lambda_0]^{\Delta R_i}) \times exp\big(-\int_0^\tau G_i exp\{\beta X_i{}' + \gamma_i\}\big) dR_0\}. \quad (6)$$

The Gamma distribution of the frailty γ with mean restricted to 1 and variance θ, i.e. $\gamma \sim \Gamma(1, \theta)$, can be expressed as $g(\gamma) = \frac{\theta^1 \gamma^{1-1} e^{-\theta y}}{\Gamma(1)}$, $\gamma_i \epsilon \{0, \infty\}$ which reduces to $g(\gamma_i) = \frac{\theta e^{-\theta y}}{\Gamma(1)}$.

Chapter 3

Systematic review of analytical methods applied to longitudinal studies of malaria. Malaria Journal. 2019; 18:254.

Christopher C. Stanley,[1,2] Lawrence N. Kazembe,[3] Mavuto Mukaka,[4,5] Kennedy N. Otwombe,[1,6] Andrea G. Buchwald,[7] Michael G. Hudgens,[8] Don P. Mathanga,[2] Miriam K. Laufer,[7] Tobias F. Chirwa,[1]

[1]School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa.

[2]Malaria Alert Centre, University of Malawi College of Medicine, Blantyre, Malawi.

[3]Department of Statistics, University of Namibia, Windhoek, Namibia.

[4]Oxford Centre for Tropical Medicine and Global Health, Oxford, United Kingdom.

[5]Mahidol-Oxford Tropical Medicine Research Unit, Bangkok, Thailand.

[6]Perinatal HIV Research Unit, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa.

[7]Center for Vaccine Development and Global Health, University of Maryland School of Medicine, Baltimore, USA.

[8]Department of Biostatistics, Center for AIDS Research, University of North Carolina Chapel Hill, North Carolina, USA.

This paper is published in the Malaria Journal (2019) 18:254, and it is attached in the appendices.

3.0 Abstract

Background: Modelling risk of malaria in longitudinal studies is common, because individuals are at risk for repeated infections over time. Malaria infections result in acquired immunity to clinical malaria disease. Prospective cohorts are an ideal design to relate the historical exposure to infection and development of clinical malaria over time, and analysis methods should consider the longitudinal nature of the data. Models must take into account the acquisition of immunity to disease that increases with each infection and the heterogeneous exposure to bites from infected *Anopheles* mosquitoes. Methods that fail to capture these important factors in malaria risk will not accurately model risk of malaria infection or disease.

Methods: Statistical methods applied to prospective cohort studies of clinical malaria or *Plasmodium falciparum* infection and disease were reviewed to assess trends in usage of the appropriate statistical methods. The study was designed to test the hypothesis that studies often fail to use appropriate statistical methods but that this would improve with the recent increase in accessibility to and expertise in longitudinal data analysis.

Results: Of 197 articles reviewed, the most commonly reported methods included contingency tables which comprised Pearson Chi-square, Fisher exact and McNemar's tests (n = 102, 51.8%), Student's t-tests (n = 82, 41.6%), followed by Cox PH models (n = 62, 31.5%) and Kaplan–Meier estimators (n = 59, 30.0%). The longitudinal analysis methods

generalized estimating equations and mixed-effects models were reported in 41 (20.8%) and 24 (12.2%) articles, respectively, and increased in use over time. A positive trend in choice of more appropriate analytical methods was identified over time.

Conclusions: Despite similar study designs across the reports, the statistical methods varied substantially and often represented overly simplistic models of risk. The results underscore the need for more effort to be channeled towards adopting standardized longitudinal methods to analyze prospective cohort studies of malaria infection and disease.

Keywords: Plasmodium falciparum, Longitudinal studies, Longitudinal analysis, Cohort studies

3.1 Introduction

Plasmodium falciparum infection is one of the most common parasitic infection in humans. Individuals in high-transmission settings are exposed to bites of infected Anopheles mosquitoes and develop frequent infections. In early childhood, infections are almost always associated with symptomatic disease (1). Over time, individuals acquire immunity to clinical disease and, to some degree, infection (24,25) as can be seen in appendix 1.3. Thus, each infection alters the host immune response that protects against the next episode. In addition, risk of exposure to infected bites is not uniformly distributed (94,95). Thus, risk of infection is not equally distributed across populations. While one infection may decrease the risk of disease or infection, an infection in an individual is also an indicator of more frequent exposure to infected mosquitoes. This provides an interesting challenge to modelling the risk of malaria disease and infection over time. Longitudinal cohorts, unlike other study designs such as cross-sectional surveys, can differentiate between infections that appear asymptomatic at time of detection but may become symptomatic over time. With efforts intensified to

prevent Plasmodium infection and reduce clinical malaria disease, prospective longitudinal cohorts are increasingly being conducted to assess risk of infection and disease over time.

Well-designed prospective longitudinal cohort studies can be key to understanding risk of malaria disease for participants over time, but appropriate analysis of the collected data is also critical for accurate results. Like any longitudinal designs, malaria cohorts are characterized by repeated measurements, resulting in possible correlated infection and disease episode data from same participant. The underlying assumption of no correlation among observations made by standard statistical methods such as t-tests, analyses of variance (ANOVA), and linear regression are violated (96–98). The use of inappropriate methods, particularly those that do not account for possible correlation of events when dealing with repeated episodes of malaria or infection, can lead to biased results. Ignoring correlation of events could result in the confidence intervals for the estimated rates being artificially narrow and the null hypothesis may be rejected more often than warranted (72,99). Appropriate longitudinal data analysis techniques should account for such possible within-participant correlation and different covariance structures of episodes of malaria or infection measurements over time. These include generalized estimating equations (GEE) and mixed-effects models. The superiority in consistency and efficiency of these methods over the traditional methods, such as ANOVA, t-tests and simple linear regression were demonstrated previously (100–103). One of the attractive features of GEE is that the method is robust to mis-specification of the correlation structure, such that the estimator is consistent even if the working correlation structure is mis-specified (50), although it can be inefficient under the mis-specified working correlation structure (104). When the research interest is on repeated time-to-episodes of disease or infection, it is common to analyze time-to first episode ignoring the subsequent events, but this approach fails to utilize all information available in the data (30,53,102,105).

Analyses of time-to-first event only may be preferred in some situations. For example, when the intention is to evaluate interventions or prognostic factors (e.g., some vaccines) as 'all-or-nothing' determinants of malaria risk, so occurrence of any events is considered a failure. Another scenario would be when the first event is considered to have 'reset' the host such that subsequent events may then be consequently considered in a different category. Recurrent time-to-event models malaria episode data because they take into account the within-participant non-independence of episodes. These techniques include frailty models (64,65) or recurrent Cox-extended models, such as the Andersen–Gill and Prentice–Williams–Peterson models (30).

The statistical models that incorporate all the repeated observations over follow-up are particularly important in malaria since over time, malaria infections result in individuals acquiring immunity to malaria disease. Each infection may introduce protective effect to future disease episodes. Therefore, appropriate analytical methods for repeated episodes of disease and infection should capture the impact of the developed protective immunity.

Recent advancements in statistical methods and computer software have improved the ease of access and application of statistical methods specifically designed for longitudinal studies, allowing one to handle complexity with cohort data (48,51,106–108). Limited information is available about whether the recent developments in computer software and statistical methods for longitudinal data has translated into more wide adoption and application of appropriate methods when analyzing prospective longitudinal malaria cohort data. This study tested the hypotheses that choice of statistical methods would change over time and that these changes would reflect more appropriate analytical methods. In order to address this hypothesis, a systematic review was conducted to investigate the trends of use of appropriate statistical

methods among articles analyzing prospective longitudinal cohort data for clinical malaria episodes or Plasmodium infection published over a 22-year period.

3.2 Methods

3.2.1 Search strategy

Titles and abstracts containing the terms “malaria prospective study” or “malaria cohort study” or “malaria longitudinal study” or “malaria years follow-up” or “malaria repeated measurements” indexed in PubMed, Medline or ScienceDirect were included in the search. Original articles in English, which was a familiar language to authors, published from 1996 until December 2017 were reviewed in chronological order.

3.2.2 Inclusion and exclusion criteria

Original research articles were included in the analysis if they were prospective cohort studies where participants were actively or passively followed up for repeated malaria episodes over time and described the following outcomes: clinical malaria based on clinical symptoms and confirmed by rapid diagnostic test (RDT) or microscopy; infection defined as positive results for polymerase chain reaction (PCR), microscopy, or RDT for Plasmodium parasites. Several publications reporting on the same study or data set were all included. Letters to the editor, editorials, reviews or systematic reviews, meta-analyses, repeated cross-sectional surveys, case reports and articles written in languages other than English were excluded. Detailed selection process of the studies under review is shown in Figure 3.1.

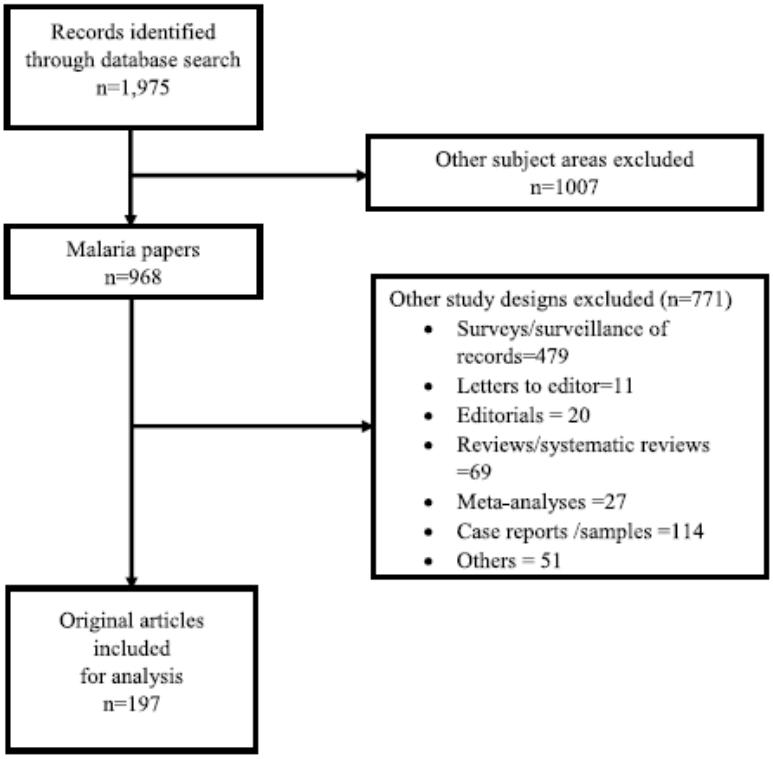


Figure 3.1. Flow chart for screening process of articles

3.2.3 Data extraction

For each article included in the analysis, the following information was extracted: year of publication, duration of follow-up, outcomes, sample size, population, location and type of statistical methods used. Statistical methods were grouped according to categorization by Colditz and Emerson (109,110) as presented in Appendix 2.1. In the cases where researchers studied multiple outcomes, only methods applied to the outcomes of interest-clinical malaria or infection—were considered. Particular interest was on standard longitudinal analysis techniques including GEE, mixed-effects models (random effects models) and repeated measures ANOVA which are considered appropriate for analyses of repeatedly measured data such as repeated episodes malaria disease or infection. For repeated time-to-episodes, recurrent Cox-based models and frailty models are considered appropriate, while other

survival methods such as the Kaplan–Meier estimator, log-rank test and Cox PH model are only appropriate for analyses of time-to-single-episode.

3.2.4 Statistical analysis

Logistic GEE was used to estimate the odds of using the types of statistical methods in the 2007–2017 period compared to 1996–2006 period. Each model adjusted for study sample size and year of publication was included as a panel variable. The outcome variable was binary, indicating whether in each article a particular method was used or not. Models were constructed for each of the following methods: contingency tables, descriptive statistics only, Kaplan–Meier estimator, Cox PH model, Poisson model, GEE and mixed-effects model. All analyses were done using Stata SE version 15.1 (Stata Corp, College Station, TX, USA).

3.3 Results

Out of the 1975 abstracts reviewed, 1778 were excluded because they were of other subject areas (n = 1007, 51.0%) and study designs (n = 771, 39.0%). One-hundred and ninety-seven (10.0%) articles met the inclusion criteria and were included in the analyses (Figure 3.1). Overall, the median follow-up time was 12 months [interquartile range (IQR): 9–24]. The median sample size per article was 351 (IQR: 206–700). The number of articles increased over the 22-year study period, with the highest number found between 2012 and 2016 (Figure 3.2).

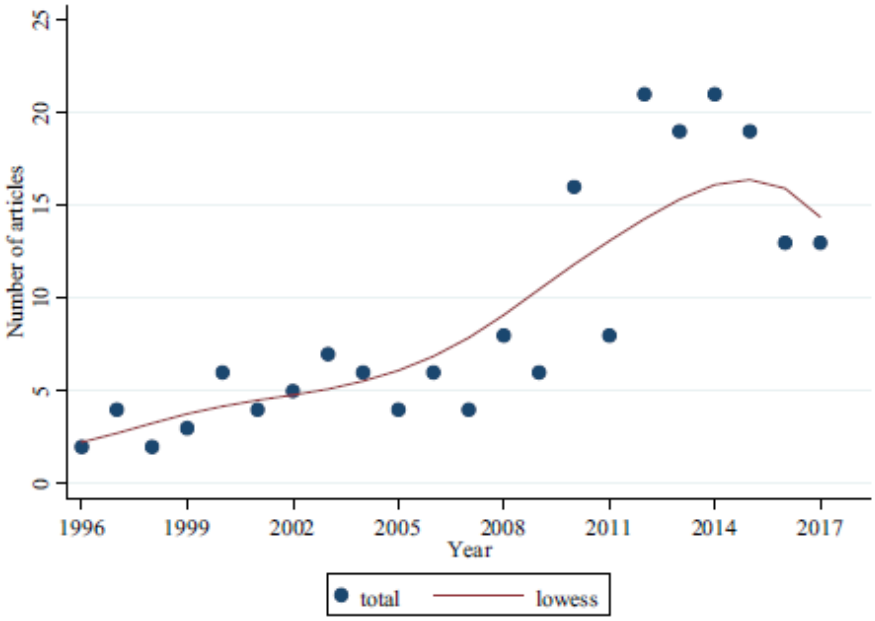


Figure 3.2. Total number of articles from 1996 to 2017 with lowess curve.

Overall, there were 128 (65.0%) articles that analyzed both clinical malaria episodes and infection as outcomes; 53 (26.9%) assessed clinical malaria episodes only, and 16 (8.1%) infections only. A table listing articles reviewed in this study with details of year of publication, duration of follow-up, outcomes, sample size, population, location, and analysis methods are provided as Additional file 3.1: Table S1. Among 197 articles reviewed, the most commonly reported methods included contingency tables comprising Pearson Chi-square, Fisher exact and McNemar's tests (n = 102, 51.8%), followed by Student's t-tests (n = 82, 41.6%), multiple linear regression (n = 71, 36.0%), Cox PH models (n = 62, 31.5%), and Kaplan–Meier estimators (n = 59, 30.0%) (Table 3.1).

Table 3.1. Frequency of articles using a particular statistical method

Method	All (n=197) n (%)	1996 - 2006 (n=49) n (%)	2007 - 2017 (n=148) n (%)
Descriptive statistics only	17 (8.6)	5 (10.2)	12 (8.1)
Contingency tables	102 (51.8)	32 (65.3)	70 (47.3)
Multiway tables	50 (25.4)	16 (32.7)	34 (23.0)
Epidemiologic statistics	37 (18.8)	12 (24.5)	25 (16.9)
t-tests	82 (41.6)	23 (46.9)	59 (39.9)
Analysis of variance	18 (9.1)	1(2.0)	17 (11.5)
Simple linear regression	6 (3.1)	5 (10.2)	1(0.7)
Non-parametric tests	48 (24.4)	11 (22.5)	37 (25.0)
Non-parametric correlation	21 (10.7)	3 (6.1)	18 (12.2)
Multiple regression	71 (36.0)	15 (30.6)	56 (37.8)
Poisson regression	49 (24.9)	8 (16.3)	41 (27.7)
Negative binomial regression	22 (11.2)	3 (6.1)	19 (12.8)
Survival analysis			

Kaplan-Meier estimator	59 (30.0)	11 (22.5)	48 (32.4)
Log-rank test	36 (18.3)	7 (14.3)	29 (19.6)
Cox model	62 (31.5)	9 (18.4)	53 (35.8)
Other survival methods	4 (2.0)	3 (6.1)	1 (0.7)
Longitudinal analysis			
Generalized estimating equations	41 (20.8)	9 (18.4)	32 (21.6)
Mixed-effects regression	24 (12.2)	2 (4.1)	22 (14.9)
Recurrent analysis			
Shared frailty model	2 (1.0)	-	2 (1.4)
Andersen-Gill (AG) model	4 (2.0)	-	4 (2.7)

The frequency of usage of different statistical methods changed between the two study periods: 1996–2006 versus 2007–2017. The proportion of articles using contingency tables was higher in the first period than in the second, 65.3 versus 47.3% as was simple linear regression, 10.2 versus 0.7%. Use of the survival analysis methods, Cox PH models and Kaplan–Meier estimators increased in the second period from 18.4 to 35.8% and 22.5 to 32.4%, respectively. Similarly, usage of mixed-effects models among longitudinal data analysis techniques was higher in the later period compared to the former, 4.1 versus 14.9%.

From the 197 articles included in the review, 131 (66.5%) analyzed repeated malaria episodes or infection, of which 60 (45.8%) used appropriate standard longitudinal analysis methods. Appropriate longitudinal analyses included 36 (60.0%) instances of GEE, 19 (31.7%) articles employing mixed-effects models, and 5 (8.3%) articles utilizing both the GEE and mixed-effects models. Seventy- one (54.2%) of the 131 articles used other methods, namely the Poisson (n = 49) and negative binomial (NB) regression models (n = 22). These models were based on the generalized linear modelling. Of the 41 studies that modelled the correlation structure, only 5 (12.2%) assumed exchangeable correlation, while 36 (87.8%) did not specify the correlation structure. Out of all 197 articles reviewed, there were 99 (50.3%) studies that assumed Poisson distribution, 4 of these assessed overdispersion, of which 2 adjusted for it. In 9 (4.7%) of all the 197 reviewed studies, authors assumed a normal distribution of count data without indicating if transformations were done.

There were 74 of 197 articles that analyzed time-to-first malaria episode or infection ignoring subsequent events and 61 (82.4%) of these employed survival analysis techniques that included Cox PH models, Kaplan–Meier estimators, and log-rank tests. There were 24 (12.2%) of 197 articles that analyzed repeated time-to-episodes and only 6 (25.0%) of these used appropriate recurrent event models with extensions of Cox PH model, namely Andersen–Gill (n = 4) and frailty models (n = 2) to evaluate risk factors for recurrent clinical malaria. Eighteen (9.1%) articles analyzed repeated time-to-malaria or infection episodes using Cox PH models and Kaplan–Meier estimators which are only valid for single-event analyses because they assume independence of events. Compared to the first period 1996–2006, the odds of an article using contingency tables adjusting for sample size were lower in

the second period 2007–2017 [odds ratio (OR) = 0.49, 95% confidence interval (CI) 0.27–0.91] (Table 2.2).

Table 3.2. Odds ratio of using a statistical method during 2007-2017 compared to 1996-2006 adjusted for article sample size

Method	Odds Ratio	95% Confidence interval
Descriptive statistics only	0.81	0.26 – 2.42
Contingency table	0.49	0.27 - 0.91
Kaplan-Meier estimator	1.80	0.88 – 3.67
Cox PH model	2.57	1.21 – 5.42
Log-rank	1.60	0.66 – 3.84
Poisson model	1.78	0.85 – 3.72
GEE	1.16	0.56 – 2.39
Mixed-effects model	3.85	1.13 - 6.39

The proportion of articles using Cox PH models, Kaplan–Meier estimators and mixed-effects models increased over time while GEE and mixed-effects models increased before remaining stable (Figure 3.3).

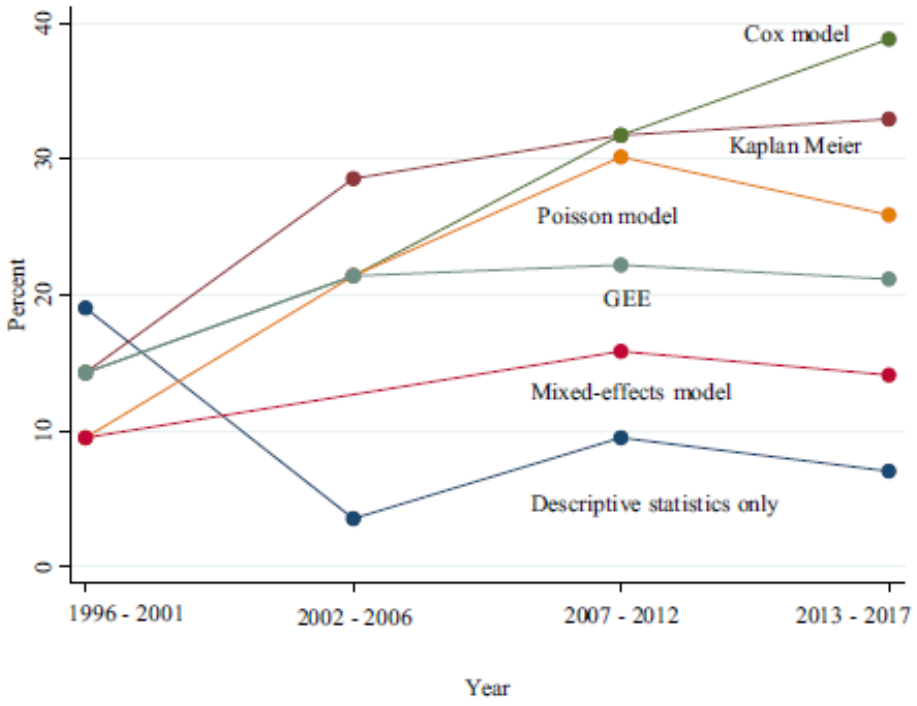


Figure 3.3. Time series plot with percentage of articles using a particular statistical method, from 1996 to 2017.

However, only the Cox PH models, and mixed-effects models had increased odds of being used in the second period compared to the first period, (OR = 2.57, 95% CI 1.21–5.42) and (OR = 3.85, 95% CI 1.13–6.39), respectively (Table 3.2).

3.4 Discussion

In this review of statistical methods among articles analyzing prospective longitudinal cohort data for clinical malaria episodes or Plasmodium infection published over a 22-year period,

the statistical methods substantially varied across articles despite that they reported analysis of the same general study design and outcome measurements. The most commonly used models for analyzing count number of malaria episodes included the Poisson and NB models, but these fall short with repeated measurements data as they cannot account for within-participant correlation of the episodes (96,111,112). Ignoring this possible correlation of events tends to bias the model results as demonstrated previously (45,49,50,113), yielding regression parameters with underestimated standards errors for time-dependent covariates or overestimated errors in time-dependent covariates (114). As hypothesized, trends in choice of appropriate methods improved over time, which included the GEE and mixed-effects models as well as recurrent event models, although their usage during the entire study period remained low. Furthermore, the majority of the studies that modelled the correlation structure did not specify the kind of assumed correlation structure, with only exchangeable correlation structure stated in few studies. Among studies that assumed Poisson distribution, the majority of them did not state whether or not overdispersion was checked.

Some articles analyzed the repeated time-to-disease episodes or infections but applied the Cox PH model (63) or the Kaplan–Meier estimator, which assume independent events and are only valid for modelling the time-to-single disease episode or infection. Such analyses may yield biased results. This is particularly true in malaria because over time, individuals acquire immunity to clinical disease and, to some degree, infection. Therefore, each infection may be dependent on the past, and in turn alter the host immune response that protects against the next episode. Additionally, there may be temporal variations in exposure which can alter the risk of clinical malaria disease over time. In some instances, data from the later episodes were discarded leading to loss of information which may limit the resulting conclusions. Modelling that focuses on time-to-first event only may not be generalizable to analyses examining all

repeated episodes of disease or infection over entire study period because the risk to subsequent episodes diminishes as a result of the gained immunity over time.

When analyzing repeated times-to-episodes malaria or infection, analytical methods should account for possible within-participant correlation of events as well as censoring over follow-up. Such appropriate methods include recurrent event survival models(67,73). In the current review, only six articles analyzed repeated time-to-episodes using appropriate recurrent models. Well-developed descriptions of the recurrent event models have been presented by several authors (30,67,72,73) while this section briefly discusses these models in relation to malaria data. The recurrent event models include extensions of the Cox PH model such as Andersen–Gill (AG) (69), Prentice–Williams–Peterson (PWP) (70) and the frailty model (66). The AG model uses a counting process time-scale for all episodes assuming that the correlation between episodes-times for each participant is explained by previous episodes where timescale does not reset to zero after an episode. This model is typically applicable when the interest is on the overall effect of covariates on the intensity of the recurrent episodes. The PWP model analyses ordered events by stratification and can be expressed as gap-time model where the time-scale is reset to zero after each episode occurs. Gap times between malaria parasitemia detection to clearance is a good example where the PWP model can be valid. Moreover, the PWP model can give both overall and episode-specific effects and so can be more applicable in malaria where covariate effects may be different in subsequent episodes due to the naturally acquired immunity. The shared frailty model extends the Cox PH model by introducing a random covariate that induces dependence among the recurrent episode times. By conditioning on covariates and the random effect, the recurrent episodes are assumed to become independent. Another challenge when analyzing recurrent events in malaria is the issue of exclusion of periods-at-risk after an event, and this may constrain the

choice of statistical methods. Depending on kind of treatment if any, recurrent events within a few days of each other are likely to be the result of the same infection or the second event may be considered to be due to treatment failure. Ignoring periods-at-risk after an event is also a potential source of bias. Thus, authors ought to consider exclusion of periods-at-risk after an event during the datasets structuring before analyses are conducted.

The review has demonstrated that publications on prospective longitudinal malaria cohorts are increasing over time, a trend reflecting the growing understanding of the importance of longitudinal cohort data coupled with advancements in computer software developments for analyzing such data (48,51). Optimal methods for longitudinal data analysis were in their infancy 20 years ago and both access to appropriate software and general acceptable of these methods by the research community have increased dramatically over this period. The increased trend in malaria cohort studies and the improved but slow adoption of appropriate methods underscores the need for standardized analytical methods for such data to avoid drawing erroneous conclusions based on inaccurate results.

This review may have included more than one report using the same data, which may have introduced bias due to duplication of datasets. For example, the same data may have been analyzed using inappropriate methods in first instance and then using appropriate techniques. Furthermore, some authors published more than one article and such articles may share similarities in the choice of analytical methods, but accounting for this was beyond the scope of this review as author background information was not collected. Future studies should consider testing whether more appropriate statistical methods improve or change understanding of malaria epidemiology by using among other factors, authors' background information. Lastly, further studies may explore and compare how results from this study would compare with trends specifically from field trials.

3.5 Conclusion

Cohort studies assessing risk of malaria infection and disease over time did not employ consistent analytic methods. The statistical methods varied substantially and often represented overly simplistic models despite similar designs, which may have led to biased outcomes and results that cannot be compared across studies. This review suggests that recent studies are adopting more appropriate statistical methods, though the statistical analyses are not uniform. These results underscore the need for more effort to be channeled towards adopting standardized longitudinal methods to model recurrent events for malaria infection and disease.

3.6 Authors' contributions

CCS and TFC conceptualized the study and collected, analyzed and interpreted the data; CCS, MKL and TFC led the writing of the manuscript; LKN, MM, KNO, AGB, MGH and DPM contributed to the study design, data interpretation and the writing of the manuscript; All authors read and approved the final manuscript.

3.7 Acknowledgements

This study was supported by the Fogarty International Center of the National Institutes of Health [Grant number D43TW010075]; and The Wellcome Trust [Grant number 107754/Z/15/Z]; National Institute of Allergy and Infectious Diseases [Grant number K24AI114996]. Funders have no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

3.8 Competing interests

The authors declare that they have no competing interests.

Chapter 4

Joint modelling of time-to-clinical malaria and parasitemia in a cohort in an endemic area. J Med Stat Inform. 2019; 7:1.

Christopher C. Stanley[1,2*], Lawrence N. Kazembe[3], Andrea G. Buchwald[4], Mavuto Mukaka[5,6], Don P. Mathanga[2], Michael G. Hudgens[7], Miriam K. Laufer[4] and Tobias F. Chirwa[1]

[1]School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa.

[2]Malaria Alert Center, University of Malawi College of Medicine, Blantyre, Malawi.

[3]Department of Statistics, University of Namibia, Windhoek, Namibia.

[4]Center for Vaccine Development and Global Health, University of Maryland School of Medicine, Baltimore, USA.

[5]Oxford Centre for Tropical Medicine and Global Health, Oxford, United Kingdom.

[6]Mahidol-Oxford Tropical Medicine Research Unit, Bangkok, Thailand.

[7]Department of Biostatistics, Center for AIDS Research (CFAR), University of North Carolina Chapel Hill, North Carolina, USA.

This paper is published in the Journal of Medical Statistics and Informatics (2019) 7:1, and it is attached in the appendices.

4.0 Abstract

Background: In malaria endemic areas such as sub-Saharan Africa, repeated exposure to malaria results in acquired immunity to clinical disease but not infection. In prospective studies, time-to-clinical malaria and longitudinal parasitemia trajectory are often analyzed separately which may result in inefficient estimates since these two processes can be associated. Including parasitemia as time-dependent covariate on time-to-clinical malaria episode may also be inaccurate because while clinical malaria disease frequently leads to treatment which may instantly affect the level of parasitemia, standard time-to-event models require that time-dependent covariates be external to the event process. We investigated whether jointly modelling time-to-clinical malaria disease and longitudinal parasitemia improves precision in risk factor estimates and assessed the strength of association between the hazard of clinical malaria and parasitemia.

Methods: Using a cohort data of participants enrolled with uncomplicated malaria in Malawi, a conventional Cox Proportional Hazards (PH) model of time-to-first clinical malaria episode with time-dependent parasitemia was compared with three competing joint models. The joint models had different association structures linking a quasi-Poisson mixed-effects of parasitemia and event-time Cox PH sub-models.

Results: There were 120 participants of whom 115 (95.8%) had $\geq$1 follow-up visit and 100 (87.5%) experienced the episode. Adults >15 years being reference, log hazard ratio for children <5 years was 0.74 (95% CI: 0.17, 1.26) in the joint model with best fit vs. 0.62 (95% CI: 0.04, 1.18) from the conventional Cox PH model. The log hazard ratio for the 5-15 years was 0.72 (95% CI: 0.22, 1.22) in the joint model vs. 0.63 (95% CI: 0.11, 1.17) in the Cox PH

model. The area under parasitemia trajectory was strongly associated with the risk of clinical malaria, with a unit increase corresponding to -0.0012 (95% CI: -0.0021, -0.0004) decrease in log hazard ratio.

Conclusions: Jointly modelling longitudinal parasitemia and time-to-clinical malaria disease improves precision in log hazard ratio estimates compared to conventional time-dependent Cox PH model. The improved precision of joint modelling may improve study efficiency and allow for design of clinical trials with relatively lower sample sizes with increased power.

Keywords: Prospective studies, longitudinal data, malaria parasite, time-to-event, clinical malaria, Cox proportional hazards model, joint modelling

4.1 Introduction

Malaria remains one of the most common parasitic infections globally with a disproportionately high burden in sub-Saharan Africa (115). In malaria-endemic areas, repeated exposure results in acquired immunity to clinical malaria disease but not infection. Of interest in many malaria studies is to estimate time-to-clinical malaria but repeated exposure may give rise to a relationship between the disease and time-dependent covariates such as parasitemia. However, time-to-clinical malaria and parasitemia data are often analyzed separately, mostly using Cox proportional hazards (PH) models and mixed-effects models or generalized estimating equations (GEE) respectively (18,116,117). Separate analysis of time-to-clinical malaria disease and parasitemia data may result in inefficient estimates when these two processes are strongly associated (118).

For accurate estimation of the risk of clinical malaria, analytical methods are required that account for historical exposure and the relationship between clinical malaria and infection parasitemia. In order to investigate the strength of this association, one approach would be including the parasitemia as a time-dependent covariate in the model of time-to-clinical malaria episode. However, this approach may be inaccurate because it does not account for the fact that parasitemia in this case is an endogenous covariate whose existence and future path can be directly related to the occurrence of the episodes (119). Standard time-to-event models require that time-dependent covariates be external to the event process (120) but clinical malaria disease frequently leads to treatment which may instantly affect the level of parasitemia.

A second approach is to fit a joint model of time-to-clinical malaria and longitudinal parasitemia profile. Previous applications of the joint modelling framework are many. These include, for example, analysis of CD4 count jointly with time-to-development of AIDS (26,27,121–123) and modelling quality of life performance scores jointly with time-to-death or disease progression among patients with cancer (28,29,122). In order to fit joint models to data from malaria studies, there are certain aspects of these types of data that require consideration. In particular, following treatment it is possible for the parasitemia to equal zero, so the joint model should allow for that. We investigated whether modelling time-to-new clinical malaria disease jointly with parasitemia trajectory data may improve precision in log hazard ratio estimates when compared to the conventional Cox PH model with time-dependent parasitemia. We also assessed the strength of the association between the hazard of clinical malaria and a time-varying parasitemia.

4.2 Methods

4.2.1 Data source

The study was motivated by data from the Mfera prospective cohort study conducted in Chikwawa district, southern Malawi, described previously by Buchwald et al. (124). Malaria disease is endemic in Malawi (3) and transmission of the Plasmodium falciparum parasite is high in Chikwawa (4,125). The cohort enrolled 120 participants who presented with uncomplicated malaria at the Mfera health center in the Chikwawa district between June 2014 and March 2015. Initial diagnosis was made by rapid diagnostic test (RDT) and confirmed by microscopy using thick blood smears. Exclusion criteria from the Mfera cohort included: acute illness requiring hospitalization, signs or symptoms of severe malaria or moderate to severe anemia, and chronic medication with any drug that has antimalarial activity e.g. HIV treatment. Participants underwent passive and active surveillance on a monthly basis and whenever sick for up to two years to assess the incidence of clinical malaria and *P. falciparum* infection.

4.2.2 Primary outcome

The outcome of interest was time-to-first new clinical malaria disease which was defined by participants' self-reported fever and a positive RDT result.

4.2.3 Notation and specification of models

The conventional Cox PH model with time-dependent parasitemia is defined as in (27) with the hazard function $\lambda_i(t)$ for participant i at given time t expressed as;

$$\lambda_i(t) = \lambda_0(t)exp[\beta_s X'_{si}(t)], \qquad (4.1)$$

where and $\lambda_0(t)$ is the unspecified baseline hazard function and the covariate vector $X_{si}(t)$ for participant i includes participant's age and frequency of LLIN use in the previous month of the visit. Covariates were included in multivariable models if they were significant at alpha level of 0.1 in univariate Cox PH models.

For the joint models, we utilized the Bayesian joint modelling approach for longitudinal and survival data proposed by Chen et al (123) which fits a model with Markov Chain Monte Carlo (MCMC) methods as presented by Ibrahim et al (126). The joint model is composed of longitudinal and survival sub-models. The longitudinal sub-model takes the form of a mixed-effects model as follows; supposing data is available from N participants with n_i observations recorded for participant $i, (i = 1, \dots, N)$. The response $y_{ij}, (j = 1, \dots, n_i)$, fixed-effect covariate vector $X_{ij} = (x_{1ij}, \dots, x_{pij})'$, and random-effect covariate vector $Z_{ij} = (z_{1ij}, \dots, z_{qij})'$ are recorded at times t_{ij}. The longitudinal sub-model is

$$y_{ij} = \beta X'_{ij} + b_i Z'_{ij} + \varepsilon_{ij}, \qquad (4.2)$$

where β is the $p \times 1$ fixed-effect parameter vector, b_i is the $q \times 1$ vector of random effects for participant i which is assumed to be multivariate normal with mean zero, i.e., $b_i \sim N_q(0, \sum_b)$, and $\sum_b$ is the variance of the subject specific effects. The error vector $\varepsilon_i = (\varepsilon_{i1}, .., \varepsilon_{nij})'$ is assumed to be distributed- $\varepsilon_i \sim N_{ni}(0, \delta^2 I_{ni})$ where δ^2 is variance and I_{ni} is the $n_i \times n_i$ identity matrix.

The survival sub-model takes a Cox PH model form (27) where the hazard function $\lambda_i(t)$ for participant i at time t as given in equation (1) is modelled as

$$\lambda_i(t) = \lambda_0(t)exp[\theta h(\beta, b_i, t) + \beta_s X'_{si}(t)], \qquad (4.3)$$

where $h(\beta, b_i, t)$ is a function of the fixed and random effects in the longitudinal sub-model, and θ is an association parameter linking the two sub-models: survival and longitudinal models. The β_s is a parameter vector for covariates unique to the survival sub-model. The survival covariate vector $X_{si}(t) = (x_{si1}, .., x_{sir})'$ may include baseline covariates for participant i with β_s representing $r \times 1$ parameter vector. The functional form of $h(\beta, b_i, t)$ determines the type of the association structure between longitudinal parasitemia and the time-to new clinical malaria episode. Taking the first derivative of $h(\beta, b_i, t)$ is interpreted in terms of an association of the rate of change in parasitemia at time t and the hazard of new clinical malaria episode at the same time, while the integral of $h(\beta, b_i, t)$ would relate the hazard and the cumulative parasitemia trajectory defined as area under parasitemia profile from baseline up to the time of the new episode.

Once the two sub-models are specified, the likelihood for the joint model can be constructed as follows. Let T_i and C_i represent potential failure and censoring times respectively for participant i. Let $S_i = \min\{T_i, C_i\}$ be the minimum of the observed failure and censoring times for participant i, and let τ_i be the failure indicator taking value 1 if $T_i \leq C_i$ and 0 otherwise. Then the value of the longitudinal trajectory for participant i at time t can be defines as $\varphi_i(\beta, b_i, t)$ and at visit j as $\varphi_{ij}(\beta, b_i)$. Using the full longitudinal trajectory, then the

likelihood of the joint distribution of the observed data and random effects for participant i can be decomposed as

$$L_i \propto f_i(Survival|longitudinal) \times f_i(longitudinal)$$

$$= f(S_i|\theta, \tau_i, \beta_s, \varphi_i(\beta, b_i, t), X_{si}) \times f(y_i|x_i, z_i, \beta, b_i) f(b_i) \qquad (4.4)$$

and the joint likelihood for all participants can be written as $L = \prod_{i=1}^{N} L_i$.

Denoting parasitemia as PC, the likelihood of the joint distribution of observed data, random effects and PC can be expressed as

$$L = f(S|\theta, \tau, \beta_s, \varphi(\beta, b, t, PC)) \times f(b, PC|\beta) \qquad (4.5)$$

and integrating out the random effects of the conditional likelihood yields the marginal likelihood. Under the Bayesian framework, the random effects b are sampled in MCMC algorithm as extra parameters. The survival and longitudinal sub-models are linked by sharing common random effects structure. The MCMC computations of the model parameters proceed assuming that given the random effects, the longitudinal parasitemia and time-to-clinical malaria process are independent as are the longitudinal responses of each participant.

4.2.4 Statistical analysis

Baseline data was summarized using frequencies and percentages if categorical and medians with inter-quartile ranges (IQR) were presented if continuous and skewed. Failure functions for time to new clinical malaria disease were summarized using Kaplan Meier plots, and compared across age and LLIN usage using log-rank test. Four models of time-to-new clinical malaria disease were fitted. These included the reference model (model 1), a conventional Cox PH model fitted with time-dependent parasitemia measured at each visit, and three competing joint models. The three joint models take a common formulation only differing in the assumed association structure linking the hazard of clinical malaria with parasitemia. The hazard of clinical malaria disease at any time t was linked with: 1) the current underlying value of parasitemia at the same time point (model 2), 2) the rate of change in parasitemia trajectory at time t (model 3), and 3) the cumulative trajectory or area under profile of parasitemia from baseline up to time t (model 4). Each joint model involved fitting a quasi-Poisson mixed-effects sub-model of parasitemia longitudinal trajectory with natural spline functions of time and including the resulting estimates as covariates in the Cox PH sub-model. Other covariates in the conventional Cox PH and survival sub-models included age of the participant and frequency of LLIN use in previous month. Parameters were estimated using MCMC through the random walk Metropolis–Hastings (M-H) algorithm. Diffused normal priors were assumed for the covariates including the association parameter θ. Diagnostic assessments were conducted to assess the convergence of the MCMC samples using trace and kernel density estimator plots, for the final optimal model. Analyses were done in Stata SE version 15.1 (Stata Corp., College Station, TX) (108) using program *bayesmh* and R version 3.4.3 using packages *JMbayes*, *survival* and *glmmPQL* (127).

4.3 Results

4.3.1 Baseline

There were 120 participants in the cohort, of which 69 (57.5%) were females. The overall median age was 7.5 years [inter-quartile range (IQR): 4.7 – 18.1], 6.3 years (IQR: 3.2 – 13.1) for males and 9.2 years (IQR: 5.3 – 18.5) females (Table 3.1).

Table 4.1. Baseline demographic, clinical conditions and vital signs for Mfera malaria cohort in Malawi

Variable	Total (n=120)
Gender, female, n (%)	69 (57.4)
Age, n (%)	
< 5 years	34 (28.3)
5-15 years	51 (42.5)
>15 years	35 (29.2)
Weight (kg), median (IQR)	21.5 (15.0 - 46.0)
Height (cm), median (IQR)	119.5 (103.0 - 151.8)
Temperature (℃), median (IQR)	36.7 (36.2 - 38.6)
Respiratory rate (breaths/minute), median (IQR)	28 (22 - 36)
Heart rate (beats/minute), median (IQR)	112 (92 - 139)
Hemoglobin (g/dl), median (IQR)	11.5 (10.2 - 12.4)
Parasitemia (number of parasites/μL), median (IQR)	11060 (840 - 54000)
Cough, n (%)	15 (12.5)
Musculoskeletal pain, n (%)	40 (33.3)
Headache, n (%)	36 (30.0)
Vomiting, n (%)	32 (26.7)
Abdominal pain, n (%)	15 (12.5)

LLIN use previous month*, n (%)	
Every night	48 (44.9)
Most nights (> half)	13 (12.1)
Some nights (< half)	8 (7.5)
No nights	38 (35.5)
Season enrolled, n (%)	
Dry: May – November	91 (75.8)
Rainy: December – April	29 (24.2)

*not adding up to column total due to missing

The median number of malarial parasites per µL was 11,060 (IQR: 840 – 54,000) overall, 24,840 (IQR: 1,600 – 68,600) in males, and 5,640 (IQR: 520 – 540,000) for females. During enrolment, 48 (44.9%) out of 107 participants reported to have been using LLINs every night in previous month.

4.3.2 Follow up and time-to-first new clinical malaria disease

Analyses of time-to-new clinical malaria disease included 115 participants who had a least one follow-up visit post enrolment and together contributed a total 894 observations. The median follow-up time to new clinical malaria disease episode was 3.5 months (IQR: 1.1 – 7.9). Out of the 115 participants, 100 (87.5%) experienced the episode while 15 (12.5%) were administratively right censored (Table 4.2).

Table 4.2. Characteristics by clinical malaria status for Mfera malaria cohort in Malawi

Variable	Clinical malaria episode (n=100)	No clinical malaria episode (n=15)
Sex, n (%)		
Male	42 (42.0)	8 (53.3)
Female	58 (58.0)	7 (47.7)
Age, n (%)		
< 5 years	27 (27.0)	4 (26.7)
5-15 years	48 (48.0)	2 (13.3)
>15 years	25 (25.0)	9 (60.0)
Net use previous month, n (%)		
Every night	37 (37.0)	8 (53.3)
Not every night	63 (63.0)	7 (46.7)
Hemoglobin (g/dl)	11.4 (10.0 - 12.3)	12.3 (10.9 - 13.9)
Parasitemia, number of parasites per µL, median (IQR)	13640 (840 - 52040)	2800 (560 - 60040)

Among 100 participants who experienced the episode, 58 (58%) were females, 48 (48%) were aged 5-15 years, 37 (37%) reported using LLIN nightly in prior month to enrolment and their median parasitemia was 13640 (IQR: 840 - 52040).

4.3.3 Parameter estimation

Regression coefficient estimates are log hazard ratios, log (HRs). Overall, the joint models gave larger log (HR) estimates with consistently smaller standard errors and narrower credible intervals when compared to the conventional Cox PH model with time-dependent parasitemia (Table 4.3).

Table 4.3. Log of hazard ratio estimates for time-to- new clinical malaria episode for Mfera cohort in Malawi

		Age at baseline*		LLIN use in previous month+	Parasitemia
Model	**Parameter**	$\theta_{<5 \text{ years}}$	$\theta_{5-15 \text{ years}}$	$\theta_{\text{not every night}}$	$\theta_{\text{parasitemia}}$
Model 1	Estimate	0.624	0.632	0.520	-1.77e-5
	Std error	0.282	0.259	0.292	1.86e-6
	95% CI	0.044, 1.183	0.115, 1.174	0.076, 1.085	-1.41e-5, 2.14e-5
Model 2	Estimate	0.672	0.707	0.582	-0.029
	Std error	0.016	0.014	0.024	0.053
	95% CI	0.094, 1.251	0.213, 1.265	0.110, 1.017	-0.345, 0.411
Model 3	Estimate	0.749	0.775	0.594	-0.103

	Std error	0.023	0.021	0.023	0.071
	95% CI	0.077, 1.485	0.193, 1.253	0.159, 1.156	-6.168, 6.124
Model 4	Estimate	0.735	0.722	0.575	-0.001
	Std error	0.022	0.028	0.030	0.0001
	95% CI	0.170, 1.263	0.219, 1.215	0.125, 1.066	-0.0021, -0.0004

Log hazard ratio estimates are posterior means. All models are multivariable. CI= Credible Interval. *reference: >15 years, $^{+}$reference: LLIN use nightly. Model 1= Conventional Cox PH model fitted with time-dependent parasitemia measured at each visit; Model 2= Joint model with association structure taking current underlying value of parasitemia at any time; Model 3= Joint model with association structure taking rate of change in parasitemia trajectory at any time; Model 4= Joint model with association structure taking cumulative parasitemia trajectory from baseline up to any time.

Comparing the Deviance Information Criteria (DIC) from the three joint models showed that joint model 4 with cumulative parasitemia had the lowest value (Table 4.4), suggesting the best fit.

Table 4.4. Deviance Information Criteria (DIC) from the competing joint models

Joint Model	DIC
Model 2: Joint model with current underlying value of parasitemia at any time	74763360
Model 3: Joint model with rate of change in parasitemia trajectory at any time	74763436
Model 4: Joint model with cumulative parasitemia from baseline up to any time	74763318

For the rest of the manuscript, the joint model is referencing to the joint model 4 with cumulative parasitemia which is being compared to the conventional Cox PH model (model 1). Considering adults above 15 years as reference group, the log (HR) of clinical malaria disease for children under 5 years, was 0.74 [95% Credible Interval (CI): 0.17, 1.26] in the joint model compared to 0.62 (95% CI: 0.04, 1.18) from the conventional Cox PH model. The log (HR) for participants aged 5-15 years was 0.72 (95% CI: 0.22, 1.22) in the joint model compared to 0.63 (95% CI: 0.11, 1.17) in the conventional Cox PH model. Considering participants who used LLINs nightly in the previous month as reference, the log (HR) of clinical malaria disease for those who did not use LLIN every night was 0.58 (95% CI: 0.13, 1.07) from the joint model compared to 0.52 (95% CI: 0.07, 1.08) in the conventional Cox PH model. From the joint model, the area under the longitudinal trajectory of parasitemia was strongly associated with the risk of clinical malaria, with a unit increase corresponding to a -

0.0012 (95% CI: -0.0021, -0.0004) decrease in the log hazard ratio. Neither current underlying value of parasitemia nor rate of change in parasitemia trajectory at the same time point were significantly associated with the occurrence of a new clinical malaria episode.

4.3.4 MCMC convergence of estimated log (HR) parameters

The MCMC sampler trace plots from the final joint model 4 seemed to mix well for parameters of children under 5 years, 5-15 years and infrequent LLIN use and never moved beyond 1.5 achieving convergence (Figure 4.1A-C).

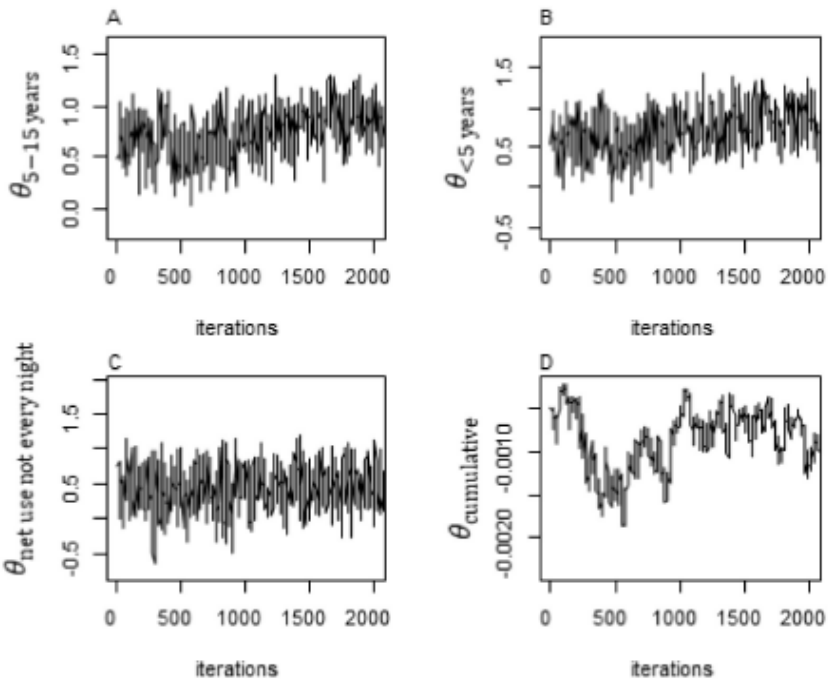


Figure 4.4. Trace plots for parameters from final joint model.

Trace plots show values that the parameter took during the runtime of the MCMC sampling until it reached convergence. Trace plots for parameters $\theta_{5-15\,\text{years}}$, $\theta_{<5\,\text{years}}$ and $\theta_{\text{net use not every night}}$ show that the M-H samplers explored the distribution by traversing to

areas where its density is very low with very small fluctuations, suggesting that the chains mixed well to the target distributions. In D, the sampling for the parameter of cumulative parasitemia $\theta_{\text{cumulative}}$ shows noisy pattern before converging at about 1000 iterations.

The sampler for cumulative parasitemia parameter, showed noisy pattern before converging at about 1000 iterations (Figure 4.1D).

From the Kernel density estimator plots, all the four parameters were roughly normal suggesting that the M-H algorithm sampled within the target normal distribution (Figure 4.2A-D).

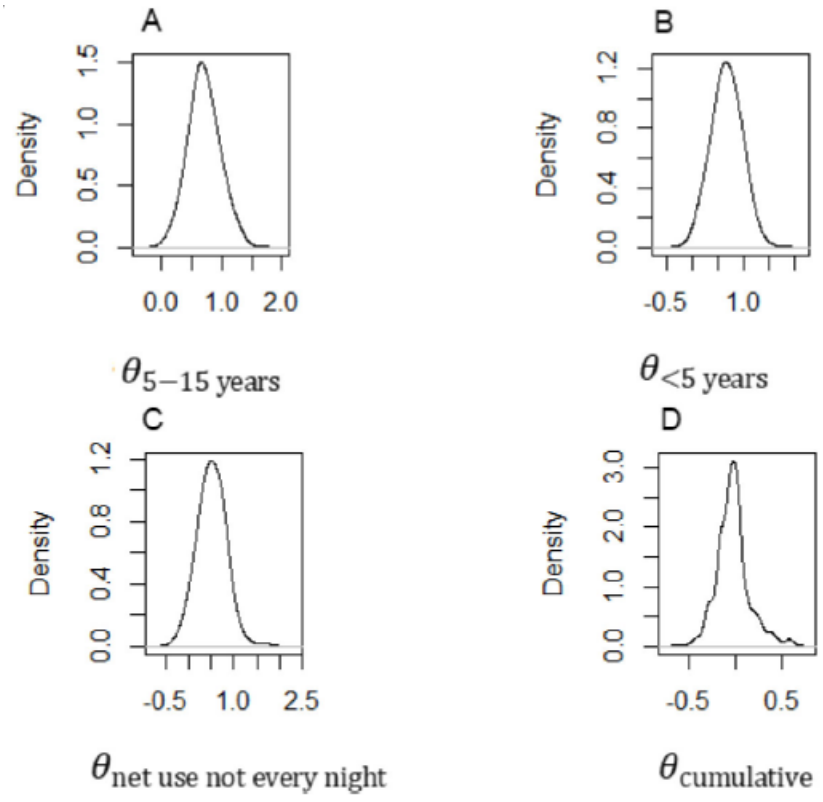


Figure 4.5. Kernel density estimator plots for the parameters of the final joint model.

The MCMC sampling process of the parameters portrays the target poster distribution. The plots suggest that algorithm sampled successfully within the assumed normal distribution for all parameters $\theta_{5-15\ years}$, $\theta_{<5\ years}$, $\theta_{net\ use\ not\ every\ night}$ and $\theta_{cumulative}$.

4.3.5 Factors associated with risk of new clinical malaria episode

In unadjusted analyses using Kaplan Meier failure estimator, the risk of experiencing new clinical malaria episode was higher in children under 5 years and the 5-15 years compared to adults above 15 years (log-rank p-value=0.016) (Figure 4.3 A).

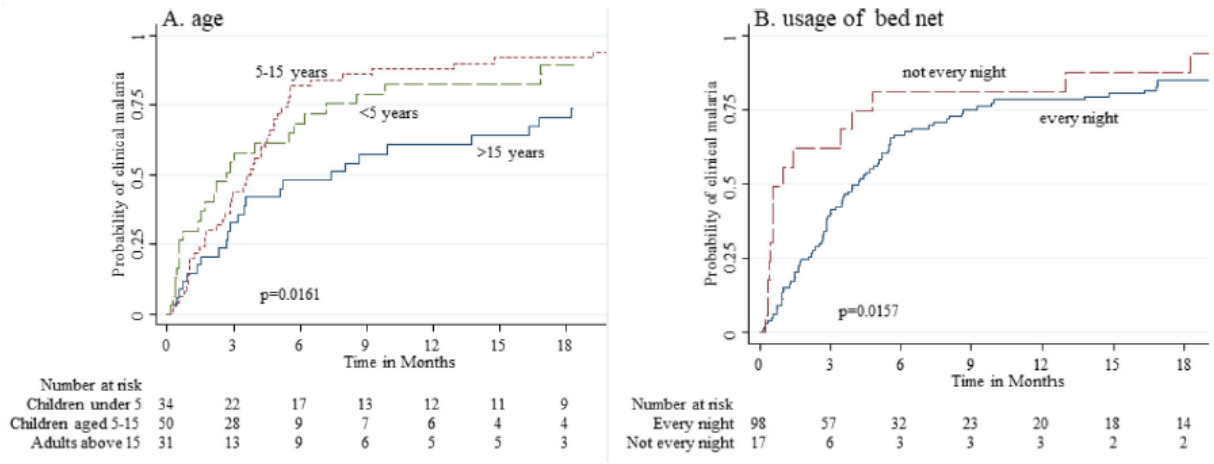


Figure 4.6. Kaplan-Meier estimator for time-to-new clinical malaria episode by A) baseline age, and B) use of LLIN in previous month.

The risk of getting a new clinical malaria episode was also higher in participants who did not use LLIN every night compared to those who used LLIN nightly in previous month (log-rank p-value=0.016) (Figure 4.3B). Based on the joint model, conditional on cumulative parasitemia profile, the hazard of getting new clinical malaria episode was higher in children under 5 years by 2.1-fold (95% CI: 1.2, 3.4) and those aged 5-15 years by 2.0-fold (95% CI:

1.2, 3.5) when compared to adults over 15 years old and controlling for LLIN usage. Among participants of the same age, the conditional hazard was also higher for participants who did not use LLIN every night by 1.8-fold (95% CI: 1.1, 3.3) compared to those who used the LLIN nightly in previous month.

4.3.6 Dynamic prediction of probability of a new clinical malaria episode

The dynamic predicted probabilities of (free-event) not experiencing a new clinical malaria episode conditional on the history of *P. falciparum* parasitemia trajectory for randomly selected participants from Mfera cohort are shown in Figure 4.4. The conditional predicted probabilities starting from the time of the first observed episode are obtained from the joint model. All graphs show that the free-event probability of developing a new episode soon after the first episode was low. The predictions seemed to be more certain closer to measured parasitemia data as the 95% pointwise confidence intervals are wider farther from the last measurements.

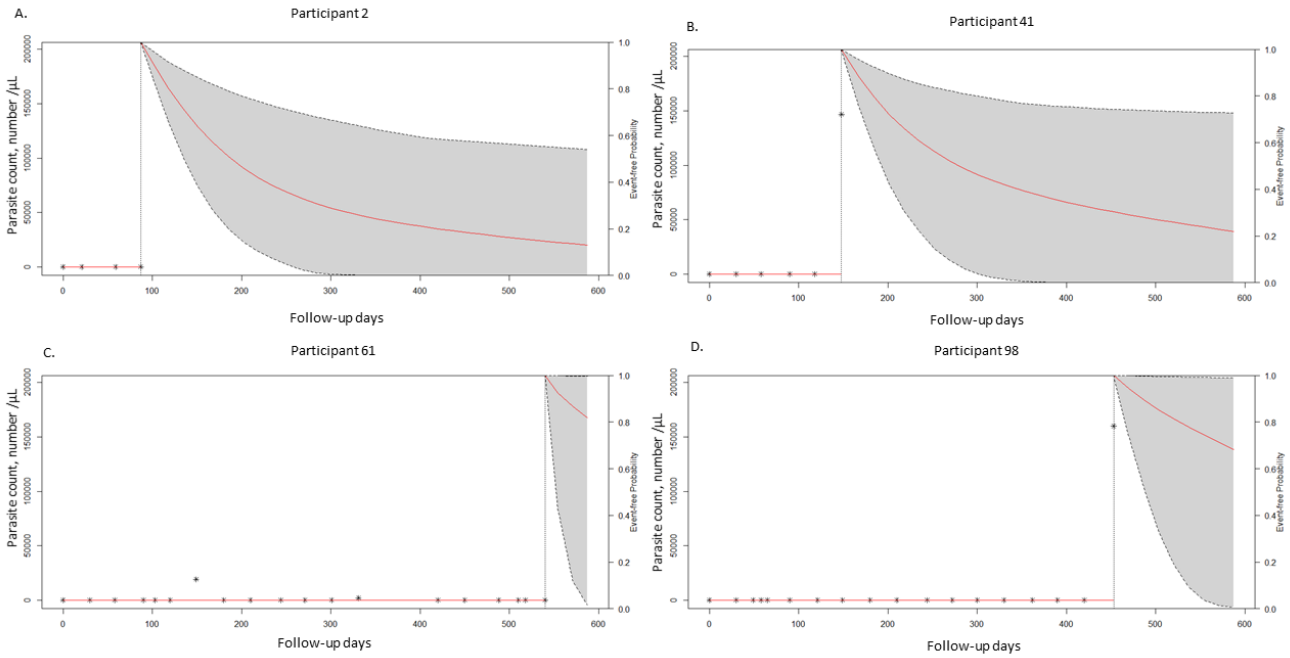


Figure 4.4. Predicted probability of experiencing a new clinical malaria episode of randomly selected participants from Mfera cohort with 95% pointwise confidence intervals.

4.4 Discussion

This study has demonstrated that jointly modelling longitudinal parasitemia and time-to clinical malaria episodes improves precision in risk factor estimates associated with clinical malaria disease. The joint model yielded larger parameter estimates with consistently smaller standard errors and narrower credible intervals when compared to a conventional Cox PH model with time-dependent parasitemia. These results are consistent with findings from other areas including HIV (26,123) and cancer (28,29,128) where joint modelling out-performs separate analyses in terms of optimal use of the available information giving both more precise and less biased estimates. The improved precision provided by the joint model may improve study efficiency, for example, clinical trials may be designed with relatively lower

sample sizes while still yielding high power. In general, the conventional time-dependent Cox PH model like any other standard time-to-event model assume that time-dependent covariates are external to the event process (120). However, parasitemia in this case is an endogenous covariate whose existence and future path can be directly related to the occurrence of the malaria episodes. By postulating a model for the joint distribution of the covariate parasitemia and the time-to-malaria processes, the joint model explicitly accounts for possible inter-dependence between the two processes through shared random effects (129). Moreover, the time-dependent covariate model assumes that the value of the parasitemia does not change until a new measurement is taken which may not be correct. When we modelled parasitemia using a quasi-Poisson mixed effects model, we were creating a model for the outcome at any time-point thereby indirectly addressing the measurement error. The joint model also offered flexibility to investigate the appropriate link structure between longitudinal parasitemia and time-to-new clinical malaria episode. Among the fitted joint models, only the model with cumulative parasitemia was strongly associated with the hazard of new clinical malaria episode suggesting that the risk can better be explained by conditioning on the cumulative effect of the longitudinal parasitemia trajectory from baseline up to the episode time. The underlying current value of the parasitemia or change in parasitemia at any time was not associated with the risk of experiencing new clinical malaria episode at the same.

Factors associated with a high risk of experiencing a new clinical malaria episode were young age and infrequent use of LLINs as have been reported in other studies (130–133). Children aged up to 15 years had higher risk for clinical malaria disease when compared to adults over 15 years. The higher risk of experiencing a clinical malaria episode among children is possibly be due to the naturally under-developed humoral immune responses to different

stage-specific antigens of *P. falciparum* otherwise acquired with age (25). In this study in an endemic area, high cumulative parasitemia was associated with lower risk of getting a new clinical malaria episode suggesting that increased exposure to malaria parasites with time may result into protective effect to future clinical malaria episodes.

This study may be limited by focusing on time-to-first malaria episode only, thus estimates obtained here may not be applicable to analyses examining all clinical malaria episodes over a follow up period. This paper has established the optimal way of incorporating parasitemia in estimating time-to-first new clinical malaria which may further be extended to study recurrent episodes over entire follow-up, but this may require different distributional assumptions. There were missing values for LLIN use which may have affected the findings. Future studies should include multiple imputation of missing covariate data and also investigate the role of measurement error due to detection limit of parasitemia in these joint models. Further work is required to consider joint modelling of parasitemia with recurrent episodes and to predict risk of future clinical malaria episodes.

4.5 Conclusion

In conclusion, jointly modelling longitudinal parasitemia and time-to-new clinical malaria improved precision in log hazard ratio estimates for clinical malaria when compared with the conventional Cox PH model with time-dependent parasitemia. The improved precision of joint modelling may improve study efficiency and allow for design of clinical trials with relatively lower sample sizes with increased power.

4.6 Authors' contributions

Authors' contributions	CCS	LNK	AGB	MM	DPM	MGH	MKL	TFC
Research concept and design	✓	✓	--	--	✓	--	✓	✓
Collection and/or assembly of data	--	--	✓	--	✓	--	✓	--
Data analysis and interpretation	✓	✓	✓	✓	✓	✓	✓	✓
Writing the article	✓	✓	✓	✓	✓	✓	✓	✓
Critical revision of the article	✓	✓	✓	✓	✓	✓	✓	✓
Final approval of article	✓	✓	✓	✓	✓	✓	✓	✓
Statistical analysis	✓	✓	--	✓	--	✓	--	✓

4.7 Acknowledgements

This Research reported in this publication was supported by the Fogarty International Center of the National Institutes of Health under Award Number D43TW010075. The work was

also partially supported through the DELTAS Sub-Saharan Africa Consortium for Advanced Biostatistics (SSACAB) program and the UK government with funding from the Wellcome Trust Award Number 107754/Z/15/Z.

4.8 Competing interests

The authors declare that they have no competing interests.

In the next chapter, the single-event joint model studied in chapter 4 is extended to include all the recurrent clinical malaria episodes jointly modelled with longitudinal parasitemia data observed over study follow-up time. For the purposes of this study, a single-event joint model refers to a joint model of longitudinal parasitemia and time-to-first new episode of clinical

malaria whereas a joint model of parasitemia and recurrent clinical malaria episodes refers to a recurrent events joint model.

Chapter 5

Analysis of recurrent times-to-clinical malaria episodes and *Plasmodium falciparum* parasite from a cohort study: a joint modelling approach.

5.0 Abstract

Recurrent clinical malaria episodes due to Plasmodium falciparum (Pf) parasite infection are common in endemic regions. With each infection, acquired immunity develops, making subsequent disease episodes less likely. To capture the effect of acquired immunity to malaria, it may be necessary to model recurrent clinical disease episodes jointly with Pf parasitemia data. A joint model of longitudinal parasitemia and time-to-first clinical malaria episode (single-event joint model) may be inaccurate since informativeness due to acquired immunity is lost when subsequent episodes are excluded. This study assessed whether joint modelling of recurrent clinical malaria episodes and parasitemia is more accurate than a single-event joint model where the subsequent episodes are ignored. The single event joint model is comprised of Cox Proportional Hazards (PH) sub-model for time-to-first clinical malaria episode and Negative Binomial (NB) mixed-effects sub-model for the longitudinal parasitemia. The recurrent events joint model extends the survival sub-model to a Gamma shared frailty model in order to include all recurrent clinical episodes. The models were applied to cohort data from Malawi. Simulations were also conducted to assess the performance of the model under different conditions. The recurrent events joint model was more precise, yielded higher hazard ratios, and mostly produced smaller standard errors than the single-event joint model with respective hazard ratio (HR) of clinical malaria of 1.42, [95% confidence interval [CI]: 1.22, 2.03] vs. HR=1.29, [95% CI:1.60, 2.45] among participants who reported not to use LLINs

every night compared to those who used the nets nightly; HR=0.96, [95% CI: 0.94, 0.98] vs. HR=0.81, [95% CI: 0.75, 0.88] for each 1-year increase in participants' age; and HR=1.36, [95% CI: 1.05, 1.75] vs. HR=1.10, [95% CI: 0.83, 4.11] for observations during the rainy season compared to dry season. The recurrent events joint model in this paper provides a way of estimating risk of malaria clinical malaria in a cohort where the effect of immunity on malaria disease acquired due to P falciparum parasitemia with aging is captured. The simulation study has shown that if correctly specified, the recurrent events joint model can give risk estimates with low bias.

5.1 Introduction

Joint modelling of longitudinal and time-to-event data has recently received increased attention in biomedical research. Typically, a joint model consists of two parts: a model for time-to-event process; and a model for the longitudinal measurements of a variable that share some parameters with the time-to-event model. The joint modelling approach is becoming popular because it offers an advantage of capturing important relationships between longitudinal outcomes and time-to-event processes that are otherwise lost. There is a gap in that methodologies to handle recurrent events are limited. Recent methodological and software developments in joint modelling have been extensively reviewed (27,82,129,134–136). Applications in the reviewed studies of joint modelling have mostly focused on time-to-single event only. For example, studies have frequently modelled longitudinal CD4 count jointly with time-to-HIV related outcomes in order to understand the relationships between the longitudinal history of CD4 and its effect on the risk of development of AIDS. (26,27,121,123). Among patients with cancer, repeated measurements of quality of life performance scores have been jointly modelled with time-to-death or disease progression

(28,29,122,137). However, for diseases that may have two or more episodes such as clinical malaria, chronic heart failure, epileptic seizures or asthma attacks, modelling that focus on time-to-first event only while ignoring the subsequent events may not be efficient since such approaches fail to utilize all information available in the data (30,57,134). In malaria, single-event models do not capture the role of protective immunity to subsequent episodes which develops with repeated *P. falciparum* infections over time (138). The WHO Malaria Vaccine Advisory Committee (MALVAC) has recently recommended analyzing recurrent event times to evaluate malaria vaccines and further methodological research (31).

In *P. falciparum* malaria studies, modelling of time-to-single malaria episode may not be adequate especially in endemic regions such as Sub-Saharan Africa because recurrent clinical disease is frequently observed. Instead, it might be more accurate to model the risk of the disease by including all the recurrent events during follow-up. Repeated infection is common and with each infection, acquired immunity develops making subsequent disease and infection episodes less likely (24,25). Therefore, modelling recurrent clinical disease episodes jointly with the longitudinal measurements of *P. falciparum* parasitemia data may be critical to capture the effect of the developed protective immunity to malaria.

The joint model of recurrent events and a longitudinal outcome typically consists of a recurrent events model and a mixed-effects model linked through either latent variables such as modelling prostate specific antigen as a biomarker for recurrent prostate cancer (89,139), or shared random effects (57,140). The most common approach used to model the recurrent events process is the shared frailty model introduced by Clayton 1978, 1994 (67,68), and

usually takes the Gamma distribution. For the longitudinal process, studies have frequently focused on continuous (Gaussian) outcomes and often applied linear mixed-effects models (55–57). However, it may not always be appropriate to use linear mixed-effects models for any longitudinal outcomes because some variables may follow a different distribution such as Poisson and negative binomial (NB). Joint models of recurrent events and a non-Gaussian longitudinal outcome such as the *P. falciparum* parasitemia remains a challenge.

In this study, a joint model is proposed which comprises of the shared frailty model for the recurrent malaria episodes and a NB mixed-effects model for the repeated measurements of *P. falciparum* parasitemia process. The proposed approach was motivated by data from a prospective malaria cohort in Malawi which has been described previously (124,141–143). Malaria is endemic in Malawi (3) and transmission of the *P. falciparum* parasite is high in the area of the study (4,125). We used data from a clinical study to investigate whether jointly modelling time-to-recurrent clinical malaria episodes with longitudinal parasitemia may yield more accurate risk factor estimates compared to single-event joint models where the subsequent episodes are ignored. Here, age groups of participants of under five children (<5 years), school-aged children (5-15 years) and adults (>15 years) was considered as proxy for immunity. The proposed joint model is also applied for prediction of a new clinical malaria episode given the history of recurrent events and *P. falciparum* parasitemia trajectory. Simulations were conducted in order to study the performance of the joint model under different conditions.

5.2 Methods

5.2.1 Data source

The proposed joint model is applied to data from the prospective cohort study conducted in a rural area of southern Malawi. The cohort enrolled 120 participants who presented with uncomplicated malaria at the Mfera Health Centre in Chikwawa district between June 2014 and March 2015. The study design was described previously (144). Briefly, initial diagnosis was made by rapid diagnostic test (RDT) and confirmed by microscopy using thick blood smears. Participants underwent passive and monthly active surveillance, and whenever sick, for up to two years to assess the incidence of clinical malaria and *P. falciparum* infection.

5.2.2 Outcomes

The primary outcomes of the study were: recurrent clinical disease defined as participants' self-reported fever and with a positive RDT result; and *P. falciparum* infection-the parasitemia measured as the number of parasites per microliter (μL) of blood. Measurements were taken from thick smears during monthly or sick visits over the follow-up period.

5.2.3 Covariates

The models included participants' age as continuous, gender, season of visit categorized as dry (May-November) or rainy (December-April), and frequency of long-lasting insecticide treated bed net (LLIN) use in the previous month of the visit defined as whether one used the LLINs every night.

5.2.4 Data structure

A sample of the data structure showing three hypothetical participants for the analysis of time-to-recurrent clinical malaria episodes is shown in supporting table 5.2.4. The time of

origin for the recurrent episodes analysis was set to be the day of enrolment. A clinical disease episode was considered new if it occurred >14 days after the previous episode based on the pharmacokinetics of the treatment in the study, artemisinin combined treatment (ACT) with Artemether–lumefantrine. In *P. falciparum* malaria, fever occurs after few days of prodromal symptoms usually at 14 days (145).

5.2.5 Notation and specification of models

The longitudinal sub-model: In the longitudinal setting, let Y_{ij} denote the longitudinal response of *P. falciparum* parasitemia for subject $i = 1, \dots, n$ at time j where $j = 1, \dots, J_i$. The measurements can be summarised as

$$Y_{ij} = \mu_{ij} + \varepsilon_i + \epsilon_{ij}, \qquad (5.1)$$

Where μ_{ij} is the mean response of parasitemia at time j, ε_i are subject-specific random effects accounting for within-subject correlation in each model part, and ϵ_{ij} represent error terms and are assumed to be normally distributed i.e. $\epsilon_{ij} \sim N_{ni}(0, \sigma^2 I_{ni})$ where σ^2 is variance and I_{n_i} is the $n_i \times n_i$ identity matrix. Postulating a model formulation proposed by Henderson et al (55), assuming that μ_i can be described by a linear mixed-effects (LMEM) sub-model with a Gaussian distribution can be;

$$\mu_i = \beta X_i' + b Z_i', \quad (5.2)$$

where β is the $p \times 1$ fixed-effect parameter vector for fixed-effect covariate vector X_i , b is the $q \times 1$ vector of random effects for random-effect covariate vector Z for participant i, assumed to be multivariate normal with mean zero, i.e., $b_i \sim N_q(0, \sum_b)$, and $\sum_b$ is the variance of the subject specific effects.

Assuming NB distribution for the parasitemia, then

$$(Y_i|\mu_i, \vartheta) = \frac{\Gamma(y_i+\vartheta)}{\Gamma(\vartheta)y_i!}.\left(\frac{\vartheta}{\mu_i+\theta}\right)^{\vartheta}.\left(\frac{\vartheta}{\mu_i+\vartheta}\right)^{y_i}, \qquad (5.3)$$

where μ_i is the mean and ϑ is the shape parameter that accounts for over-dispersion.

The NB mixed-effects model links the mean of response to the set of covariates through logarithm function can be expressed as;

$$\log(\mu_i) = \beta X_i' + b_i Z_i' + \log(M_i) \qquad (5.4)$$

where $\log(M_i)$ is offset representing total number for a set of measurements.

The NB distribution can be viewed as a Gamma mixture of Poisson distribution where in this case the parasitemia response y_i with mean μ_i follows Poisson and error term ε_i following

the Gamma distributions respectively (59); i.e. $y_i \sim Poisson(y_i|\mu_i\varepsilon_i)$ and $\varepsilon_i \sim Gamma(\vartheta,\vartheta)$. When the over-dispersion parameter is high, the NB model converges to a Poisson model and cannot deal with over-dispersion (60), and prone to nonconvergence problems. Further detail of the mixed-effects sub-model specifications for Gaussian, Poisson and NB distributions are presented in appendix 1.

The intensity recurrent event sub-model: The recurrent event model extends the single event semi-parametric proportional hazards model by introducing an unobservable (frailty) random term on which the hazard function depends upon to account for within-participant dependence of events (92), i.e. clinical malaria episodes. The single-event semi-parametric proportional hazards model can be expressed in terms of the hazard function λ_i for participant i as;

$$\lambda_i = \lambda_0 exp[\beta X_i'], \qquad (5.3)$$

where λ_0 is the unspecified baseline hazard function and X_i is the covariate vector for participant i. For ordinary Cox PH regression, the baseline hazard is usually left unspecified and still can offer valid statistical inference through the use of partial likelihood. However, in the context of joint modelling, a completely unspecified baseline hazard will generally lead to underestimation of the standard errors of the model parameters (82,146). Covariates here include participants' age as continuous, gender, season for each visit (dry or rainy) and frequency of LLIN use in the previous month of the visit. For recurrent clinical malaria episodes process, an intensity event model function is adopted as opposed to rate function

because while the rate function only defines the occurrence of recurrent events unconditional on the event history, the intensity function conditions the occurrence of events on the event history (93). In the current cohort, the event history is particularly critical because each *P. falciparum* infection alters the host immune response against threat of subsequent infections and disease episodes (24,25). Thus, the intensity recurrent event model would account for the participants' strengthening immunity to episodes due to accumulating event occurrences over time, which is critical in recurrent event analysis (89). The intensity recurrent event model at time t is given by the multiplicative intensity model following the structure proposed by Henderson et al (55) as follows;

$$\lambda_i = G_i \lambda_0 exp[\beta X_i' + \gamma_i], \qquad (5.4)$$

where G_i is assumed to follow Bernoulli distribution denoting whether the participant i is in the risk period of experiencing the malaria episode. As with the single event survival model in equation (4.3), the baseline hazard (intensity) function λ_0 is assumed to follow Weibull distribution. In the current cohort data, the vectors β and X_i contain different sets of elements from α and Z_i respectively in equation (2). The term γ_i represents the unobservable random effects (frailty) term to account for dependence in within-participant episodes and is assumed to follow a Gamma distribution with unit mean and variance θ, i.e. $\gamma_i \sim \Gamma(1, \theta)$. The frailty variance θ, reflects the amount of the within-participant dependence of clinical episode times i.e. the correlation of the recurrent events is quantified by θ, with higher values corresponding to greater within-participant correlation. When the variance is small, the values of the frailty are around one, implying no frailty effects and so recurrent events are independent.

For the counting process of the recurrent clinical episodes, let $R_i^*(t)$ be the number of recurrent events for subject i occurring before or at time t over an interval $[0, \tau]$, where recurrent episodes could potentially be observed beyond prespecified maximum time point τ. Then the counting process may be stopped by the time to loss to follow-up or end of the study denoted by C_i, with the failure indicator Δ_i taking value 1 if $T_i \leq C_i$ and 0 otherwise. The observed counting process, $R_i = R_i^* min(t, C_i)$ has a known zero-one process $\{G_i(u): 0 \leq u \leq \tau\}$ indicating whether the participant i is at risk of experiencing an episode during period u. Thus, the counting process R_i^* has a jump of size one i.e. $(R, R+1, \dots)$ when an event occurs i.e. episode of clinical malaria is experienced.

Likelihood of the joint model: Using generic terms Y to denote combined observed measurements of parasitemia data, R for combined recurrent episodes data, X for covariate information, and $\Phi = \{\varepsilon, \gamma\}$ for the random and frailty processes, the joint distribution for the longitudinal measurements Y and recurrent event processes R are conditionally independent given X, ε and γ. The dependence between Y and R may arise as a result of the direct link between ε and γ, called latent association, without which nothing can be gained through a joint analysis. Our interest is to model the subjects' recurrent processes of episodes together with their longitudinal measurements of parasitemia, through the association of ε and γ. Following framework proposed by Henderson et al 2000 (55), one can proceed to compute the likelihood of the joint model as a product of the marginal distribution of observed measurements Y and the conditional distribution of the events (episodes) R. For each participant i, the observed data are $\{(t_i, X_i, u_i, \Delta_i, \tau_i), i = 1,2, \dots, n\}$. Computing the full joint likelihood $L = L(\pi, Y, R)$ where $\pi = (\beta, \sigma, \theta, \lambda_0, \alpha, \Phi)$ is a vector denoting a collection of all unknown parameters with $\lambda_0 = \{\lambda_0(t_i), i = 1,2, \dots, n\}$, one can proceed as follows:

$$L = L_Y \times L_{R|Y} = L_Y(\pi, Y) \times E_{Z_2|Y}\{L_{R|\gamma}(\pi, R|\gamma)\}, \quad (5.5)$$

here $L_Y(\pi, Y)$ is the standard form of marginal distribution of Y for the parasitemia measurements process. The conditional likelihood of the recurrent episodes data, $L_{R|\gamma}(\pi, R|\gamma)$ captures the likelihood contribution of the longitudinal measurements up to any time of event. Suppose we denote $R_0 = \int_0^t \lambda_0(u)du$ as cumulative baseline intensity for the recurrent event process, then $L_{R|\gamma}(\pi, R|\gamma)$ can be expressed as

$$L_{R|\gamma}(\pi, R|\gamma) = \prod_i\{(\prod_t[exp\{\beta X_i{}' + \gamma_i\}\lambda_0]^{\Delta R_i}) \times exp\big(-\int_0^\tau G_i exp\{\beta X_i{}' + \gamma_i\}\big)dR_0\}. \quad (5.6)$$

The Gamma distribution of the frailty γ with mean restricted to 1 and variance θ, i.e. $\gamma \sim \Gamma(1, \theta)$, can be expressed as $g(\gamma) = \frac{\theta^1 \gamma^{1-1} e^{-\theta y}}{\Gamma(1)}$, $\gamma_i \epsilon \{0, \infty\}$ which reduces to $g(\gamma_i) = \frac{\theta e^{-\theta y}}{\Gamma(1)}$.

Parameter estimation: There are two ways in which one can estimate parameters of a specified full model: (i) the likelihood methods where maximum likelihood estimators are obtained using the expectation-maximization (E-M) algorithm (55,147), and (ii) the Bayesian methods (91,148). Here, estimates of the parameters are obtained by maximizing the joint likelihood for the parasitemia process and the recurrent episode times process using the EM

algorithm. Estimating the parameters by maximizing the likelihood of the observed data involves integrating over the random and frailty terms, γ. Since the joint likelihood contains an analytically intractable integral, numerical method of integration such as Bayesian approaches or quadrature approximation techniques are required for evaluation. The Gauss–Hermite quadrature method was used here. Further, a unit mean for the frailty term was assumed to make the parameters in the distribution and the baseline distribution λ_0 identifiable.

5.2.6 Statistical analysis

The proposed joint modelling approach was applied to a real dataset from the Mfera malaria cohort. In order to study how the joint model can perform under different conditions, simulations were also conducted. The models were fitted with shared Gamma frailty model for the recurrent events and the mixed-effects model taking competing distributions for the longitudinal process: Gaussian and NB. Model fit was compared based on the Akaike information criterion (AIC). A model with lowest value of AIC is considered to be the best fitting model. Under simulation, all data were simulated to resemble the Mfera cohort. Age in years was assumed to be normally distributed (mean: 2, Standard Deviation [SD]: 0.8) on the log-scale. To maintain the skewness of age that would reflect real data, the simulated log-normally distributed values were transformed back to original scales by taking an exponential function. The covariates sex, season and LLIN usage were assumed to be binomially distributed. Based on exploratory analyses of the Mfera cohort, the assumed true log of hazard values were; -0.04 for age, 0.02 for sex, 0.3 for season and 0.3 for LLIN usage. The baseline hazard function was assumed to follow a Weibull distribution with shape parameter lambda=1 and scale parameter =2. Participants were assumed to be censored according to a uniform

distribution on time of follow-up. After each clinical malaria episodes, a subject was assumed to be in risk-free period of 14 days, based on the pharmacokinetics of artemisinin combined treatment (ACT) with Artemether–lumefantrine therapy. Parasitemia data measurements were simulated from a mixed effects model with function of follow-up time. The model performance in bias was assessed under different scenarios that include: study sample sizes of 100, 200, 400 (representing small, medium and large sample sizes); level of censoring 10%, 20%, 50%; length of follow-up period of 1 year, 2 years, 4 years; Gamma distributed frailty term with variances 0.2 for low dependence of within-participant episodes, 1.5 for moderate and 2.5 for highly dependent episodes; and correlation level between longitudinal and recurrent processes 0.01 for weak association, 0.5 moderate and 0.8 strong association. For each scenario, 1000 runs were done. Simulations were conducted in R version 3.4.3 using package *simrec* (127). Data analysis was done in Stata SE version 15.1 (Stata Corp., College Station, TX) using *gsem* and user-written program *merlin* (108).

5.3 Results

5.3.1 Mfera cohort

Among 115 participants included in these analyses, 372 recurrent clinical malaria episodes were experienced over the two-year follow-up period, of which 219 (59%) occurred among females. Out of the 372 episodes, 125 (34%) were among children under 5 years, 193 (52%) in school aged children of 5–15 years and 54 (14%) were experienced by adults above 15 years old. Overall, there was a decreasing trend in number of recurrent clinical malaria episodes over follow-up from 106 episodes in first month to 75 episodes in 24th month (Figure 5.1).

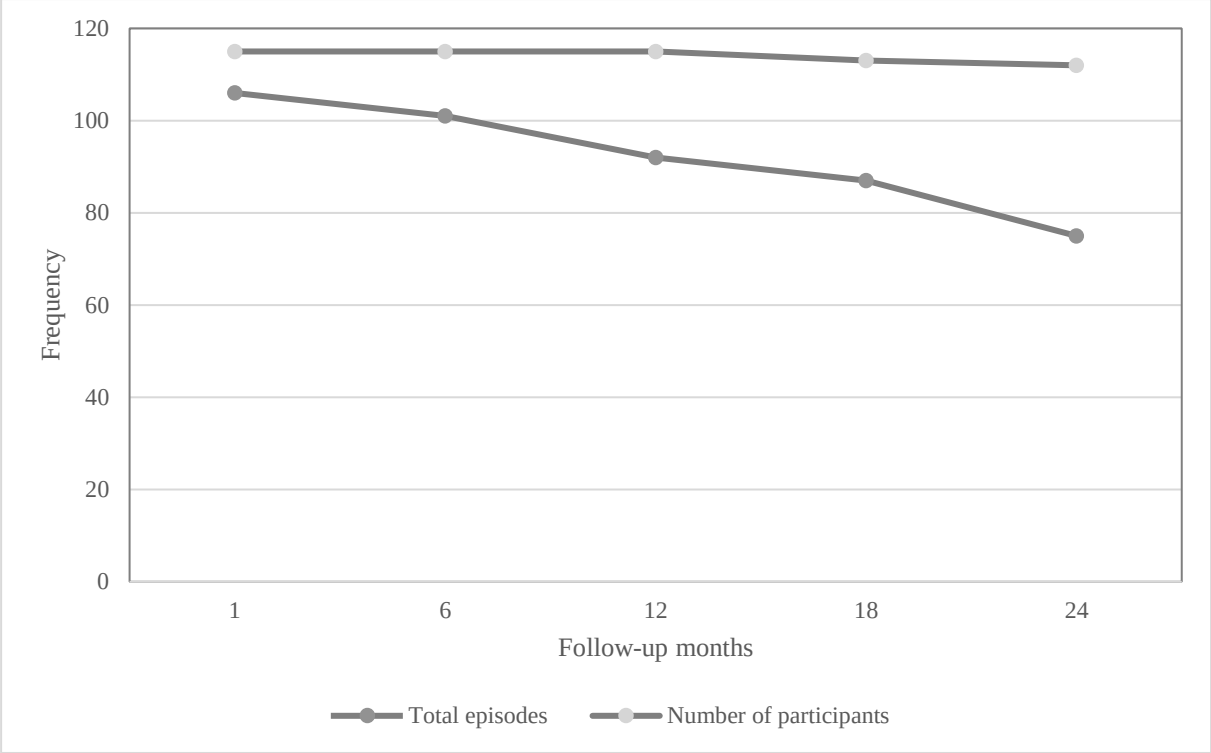


Figure 5.7. The observed number of recurrent clinical malaria episodes (dotted line) over follow-up time.

Hazard ratios for recurrent clinical malaria episodes obtained from the joint model of clinical malaria episodes and parasitemia are summarized in Table 5.1. The hazard of recurrent clinical malaria decreased with increasing participants' age (hazard ratio [HR]=0.960 [95% confidence interval [CI]: 0.944, 0.976]), for 1-year increase in age. The hazard was also higher among participants who reported not to use LLINs every night compared to those who used the nets nightly (HR=1.424, [95% CI: 1.218, 2.032]). Compared to dry season, the hazard was higher during the rainy season (HR=1.355, [95% CI: 1.048, 1.753]). Compared to results from the recurrent event joint model, the single-event joint model yielded lower hazard ratio estimates of clinical malaria and in most cases with larger standard errors.

Table 5.1. Hazard ratio estimates of clinical malaria among participants of Mfera cohort.

Variable	HR	SE	95 % CI
Joint model of shared Gamma frailty sub-model for the recurrent clinical malaria episodes and negative binomial mixed-effects sub-model for the parasitemia			
Age at enrolment, per year increase	0.960	0.010	0.944, 0.976
Gender, female	1.184	0.127	0.957, 1.466
Season, rainy*	1.355	0.209	1.048, 1.753
Less LLIN use in previous month^{+}	1.424	0.316	1.218, 2.032
Joint model of a Cox PH sub-model for time-to-first clinical malaria and negative binomial mixed-effects sub-model for the parasitemia			
Age at enrolment, per year increase	0.810	0.023	0.749, 0.876
Gender, female	1.070	0.235	0.675, 1.696
Season, rainy*	1.104	0.407	0.834, 4.108
Less LLIN use in previous month^{+}	1.290	0.327	0.680, 2.449

*reference category is dry season; +reference category is LLIN use nightly; HR=Hazard ratio; SE=Standard error; CI=Confidence Interval

The predicted conditional cumulative and marginal non-proportional hazards using the recurrent events joint model is shown in Figure 5.2. The expected number of clinical malaria episodes in the cohort increased sharply at the beginning of the follow-up period but later slowed beyond 1 year. This shows that there were less clinical malaria episodes in subsequent periods over time.

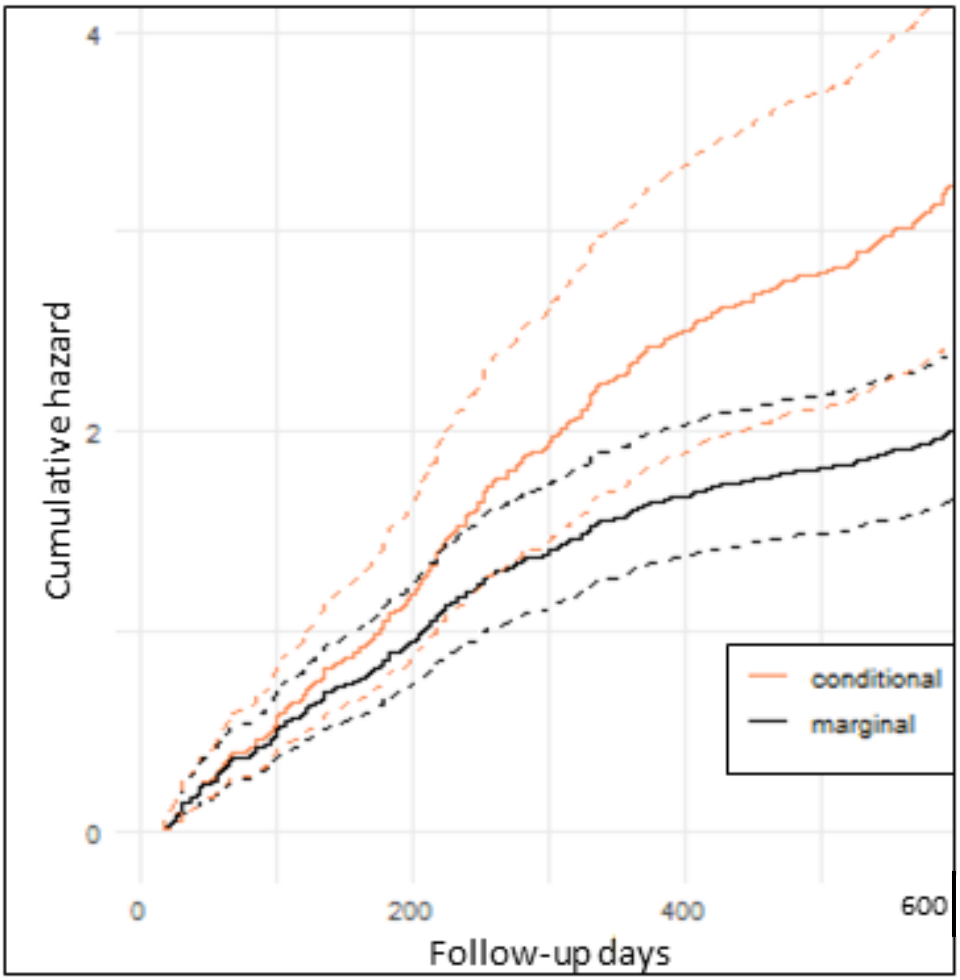


Figure 5.8. The predicted hazard of recurrent clinical malaria. The solid lines are the cumulative hazards which represent the expected number of clinical malaria episodes over follow-up time, and the dotted lines denote the 95% CIs.

So far, the results from the recurrent events joint model presented in this study are based on data from a cohort of 115 participants with 2 years of follow-up. In order to study the performance of the recurrent events joint model under varying sample sizes, length of follow-up time, and strength of association between recurrent events and longitudinal processes, simulations were also conducted.

5.3.2 Simulation study

Based on AIC, the joint models with mixed-effects sub-model of a NB distribution for the parasitemia fitted the data better than the linear mixed-effects model assuming a Gaussian response. For example, we considered a scenario assuming a cohort with sample size of 200 participants, followed up for 2 years with 10% censoring level, 0.05 correlation level between longitudinal and recurrent processes of 0.05, and frailty term variance for the dependence of within-participant clinical episodes being 0.2. A shared Gamma frailty sub-model with Weibull baseline hazard function is assumed for the recurrent process of clinical malaria episodes. Under this scenario, the joint model with the NB distribution for the parasitemia process yielded lower AIC value (48066) compared to that from the Gaussian distribution (549445). For this reason, subsequent analyses adopted a mixed-effects sub-model with a NB distribution for the parasitemia process.

Tables 5.3a and 5.3b below show log hazard ratio estimates of recurrent clinical malaria for simulated data obtained under different scenarios of study sample size, length of follow up time in years, level of censoring, level of correlation between longitudinal and recurrent processes, and frailty term variance for the dependence of within-participant clinical malaria episodes. The joint model consists of a shared Gamma frailty sub-model with Weibull baseline hazard function for the recurrent clinical malaria episodes and negative binomial mixed-effects sub-model for the parasitemia.

Table 5.2a. Log hazard ratio estimates of recurrent clinical malaria for simulated data under different scenarios for sample size of study participants; follow up time in years (τ); *c*ensoring level (C).

		N=100		N=200		N=400	
Variable/Parameter	True value	Bias	SE	Bias	SE	Bias	SE
Scenario1: C=10%, τ=2, Φ=0.01, var(γ)=0.2							
Age, per year increase	-0.04	-0.002	0.009	-0.002	0.007	-0.002	0.003
Gender, female	0.2	0.022	0.128	0.007	0.119	0.014	0.041
Season, rainy*	0.3	0.029	0.132	0.029	0.110	0.023	0.072
Less LLIN use in previous month+	0.3	0.071	0.184	0.043	0.151	0.025	0.069
Scenario 2: C=20%, τ=2, Φ=0.01, var(γ)=0.2							
Age, per year increase	-0.04	-0.003	0.006	-0.002	0.005	-0.002	0.003
Gender, female	0.2	0.028	0.106	0.012	0.085	0.012	0.054
Season, rainy*	0.3	0.035	0.107	0.035	0.095	0.034	0.051
Less LLIN use in previous month+	0.3	0.072	0.148	0.042	0.121	0.026	0.085
Scenario 3: C=50%, τ=2, Φ=0.01, var(γ)=0.2							
Age, per year increase	-0.04	-0.004	0.013	-0.003	0.005	-0.002	0.005
Gender, female	0.2	0.027	0.157	0.015	0.101	0.021	0.071
Season, rainy*	0.3	0.073	0.129	0.046	0.138	0.040	0.058
Less LLIN use in previous month+	0.3	0.086	0.204	0.051	0.116	0.042	0.095
Scenario 4: C=10%, τ=3, Φ=0.01, var(γ)=0.2							
Age, per year increase	-0.04	-0.002	0.006	-0.001	0.003	-0.001	0.002
Gender, female	0.2	-0.005	0.076	-0.001	0.064	-0.001	0.037
Season, rainy*	0.3	0.017	0.088	0.013	0.071	0.009	0.040
Less LLIN use in previous month+	0.3	0.036	0.115	0.020	0.087	0.017	0.041
Scenario 5: C=10%, τ=4, Φ=0.01, var(γ)=0.2							
Age, per year increase	-0.04	-0.001	0.005	0.000	0.003	-0.001	0.002

Gender, female	0.2	0.018	0.054	-0.001	0.041	-0.001	0.032
Season, rainy*	0.3	0.006	0.053	0.010	0.031	0.008	0.030
Less LLIN use in previous month+	0.3	0.022	0.072	0.020	0.060	0.017	0.044

*reference category is dry season; +reference category is LLIN use nightly; SE=standard error; CI= confidence interval

Table 5.2b. Log hazard ratio estimates of recurrent clinical malaria for simulated data under different scenarios for sample size of study participants; correlation level between longitudinal and recurrent processes ($\boldsymbol{\Phi}$); frailty term ($\boldsymbol{\gamma}$).

		N=100		N=200		N=400	
Variable/Parameter	True value	Bias	SE	Bias	SE	Bias	SE
Scenario1: C=10%, τ=2, Φ=0.01, var(γ)=0.2							
Age, per year increase	-0.04	-0.002	0.009	-0.002	0.007	-0.002	0.003
Gender, female	0.2	0.022	0.128	0.007	0.119	0.014	0.041
Season, rainy*	0.3	0.029	0.132	0.029	0.110	0.023	0.072
Less LLIN use in previous month+	0.3	0.071	0.184	0.043	0.151	0.025	0.069
Scenario 6: C=10%, τ=2, Φ=0.01, var(γ)=1.5							
Age, per year increase	-0.04	-0.003	0.016	-0.002	0.007	0.000	0.003
Gender, female	0.2	0.020	0.184	0.003	0.117	0.001	0.069
Season, rainy*	0.3	0.115	0.179	0.043	0.131	0.014	0.072
Less LLIN use in previous month+	0.3	0.033	0.258	0.006	0.126	0.000	0.103
Scenario 7: C=10%, τ=2, Φ=0.01, var(γ)=2.5							
Age, per year increase	-0.04	-0.003	0.016	-0.002	0.007	0.000	0.005
Gender, female	0.2	0.021	0.169	0.014	0.166	0.001	0.087
Season, rainy*	0.3	0.038	0.192	0.037	0.155	0.037	0.082
Less LLIN use in previous month+	0.3	0.036	0.256	0.016	0.192	0.000	0.116

Scenario 8: C=10%, τ=2, Φ=0.5, var(γ)=0.2							
Age, per year increase	-0.04	-0.002	0.010	-0.002	0.005	-0.001	0.003
Gender, female	0.2	0.018	0.121	0.014	0.112	0.014	0.038
Season, rainy*	0.3	0.027	0.130	0.032	0.108	0.022	0.055
Less LLIN use in previous month+	0.3	0.046	0.170	0.027	0.141	0.012	0.069
Scenario 9: C=10%, τ=2, Φ=0.8, var(γ)=0.2							
Age, per year increase	-0.04	-0.004	0.010	-0.003	0.005	-0.001	0.003
Gender, female	0.2	0.021	0.127	0.018	0.100	0.015	0.035
Season, rainy*	0.3	0.027	0.131	0.035	0.108	0.021	0.049
Less LLIN use in previous month+	0.3	0.063	0.170	0.031	0.141	0.013	0.130

*reference category is dry season; +reference category is LLIN use nightly; SE=standard error; CI= confidence interval

The performance of the recurrent events joint model improved with increased study sample size overall as evident from the decreased bias when changing number of participants from 100, 200 to 400 (Table 5.2a).

Level of censoring denotes the number of known outcomes during observation time. We hypothesized that as more outcomes become known, then the model bias would decrease. Considering scenario 1 as a reference, increasing level of censoring from 10%, 20% to 50% in that order worsened the performance of the joint model as seen from increased bias.

The length of study follow-up may determine the amount of measurements (information) that a model uses for estimation. Our hypothesis was that with longer follow-up, more data would better inform the model and improve the performance. The magnitude of bias decreased with increasing from 2 years to 3 and 4 years, as more measurements were available over time.

The level of association between recurrent events and longitudinal processes would also determine the performance of the joint model. We hypothesized that strong association between the two processes would lend the model perform better. As shown in Table 5.2b, the joint model performed best with moderate (Φ=0.5) association between recurrent and longitudinal processes when compared to weak (Φ=0.01) or strong (Φ=0.8). Referencing to a scenario with low dependence (var(γ)=0.2) of within-participant episodes, there was decrease in bias for moderately dependent episodes (var(γ)=1.5) but the performance did not improve further when episodes were assumed to be highly dependent (var(γ)=2.5).

5.4 Discussion and conclusions

This study has demonstrated that jointly modelling recurrent clinical malaria episodes and parasitemia estimates the hazard of clinical malaria with more precision (narrower confidence intervals and smaller standard errors) than a single-event joint model where the subsequent episodes are ignored. The single event joint model gave smaller estimates of hazard ratios in most cases with larger standard errors, when compared with the recurrent events joint model. The simulation study shows that if correctly specified, the recurrent events joint model can give parameter estimates with low bias. Exclusion of subsequent episodes by the single event joint model means loss of otherwise valuable information for estimation (149).

Underestimation of parameters may lead to incorrect inferences and wrong conclusions. In the Mfera cohort in this study for example, season was not associated with risk of time-to-first new clinical malaria episode when subsequent episodes were ignored yet rainy season was associated with increased risk of recurrent clinical malaria when the recurrent event joint model was used. These results underline the need for models to utilize all data collected during follow-up when estimating the risk of recurrent clinical malaria so as to capture the effect of protective immunity to subsequent episodes due to repeated *P. falciparum* infections. This study affirms the WHO MALVAC recommendation that urges the experts to evaluate the efficacy and safety of the vaccine using recurrent events models (31). The MVIP currently being implemented in Malawi, Kenya and Ghana (14) offers an opportunity to utilize these recurrent events joint models.

In this study, LLIN usage was protective of participants developing clinical malaria, consistent with other previous reports from Malawi and elsewhere (9,131,133,150–153). The protective effect of net usage was however more prominent at the beginning of the study, the time when the participant received LLINs but later started to diminish just about over a year follow-up. This could be explained by loss in physical integrity (due to worn out) and biological activity of insecticides of the nets which become apparent with ageing (154,155), insecticide resistance by the infectious mosquitoes (156–158), or changes in the biting behavior of malaria vectors following the use of LLINs (159,160). In order to understand the long-term protective effect of LLINs in current local conditions, longitudinal studies are required to assess the physical integrity and biological activity of insecticides over time. Well-designed longitudinal studies may inform experts in program implementation as to when nets need to be replaced or retreated.

Older age was associated with reduced risk of clinical malaria. Take aging as proxy for protective effect to clinical malaria, the trend may partially be attributed to acquired immunity over time (24,25). Being a cohort from a high transmission area, participants are continuously exposed to repeated bites of infected Anopheles mosquitoes (161,162). The partial immunity developed over time of exposure may not provide complete protection but it reduces the risk that malaria infection will cause the disease (163). The higher risk of developing a clinical malaria episode among children could possibly be due to the naturally under-developed humoral immune responses to different stage-specific antigens of *P. falciparum* otherwise acquired with age. Further, the school-aged children had the highest risk of developing clinical malaria, as also reported by other studies in SSA (125,164,165). The trend of high risk of malaria among school-aged children could partly be explained by low LLIN usage in this group (125,141,166,167). Results from this study underscore the need for longitudinal cohort studies with long follow-up periods in order to understand the changing patterns of risk of malaria.

Based on the joint model, the predicted conditional cumulative and marginal non-proportional hazards to clinical malaria shows that the expected number of clinical malaria episodes in the cohort increased sharply at the beginning of the follow-up period but later slowed beyond 1 year. Thus the trend shows less clinical malaria episodes in subsequent periods over time of the cohort, which is consistent with previous studies (1,25,168).

In the simulation study, the performance of the joint model as measured by amount of bias, improved with increasing study sample size and length of study follow-up, overall decreasing bias. These results are consistent with previous simulation-based studies in joint modelling (89,169–171). Thus, model performance improved as more data points were available over time. However, increasing level of censoring worsened the performance, a result in line with other joint modelling reports (171). The joint model performance improved by changing the strength of association between recurrent and longitudinal processes from weak to moderate but there was no further clear improvement when the two processes were assumed to be strongly associated. There was decrease in bias of the model by increasing the level of dependence of within-participant episodes from low to moderate but the performance did not improve further assuming highly dependence. Lack of the clear trends in model performance with change in strength of the association or level of dependence of within-participant episodes may be partly attributed to interaction among factors. In this study, factors were varied on a one-by-one basis and results compared to the reference scenario. This approach does not allow one to study the effect of interaction between factors. Morris et al. 2019 (172) recommend varying factors factorially as this approach may likely to be more informative since this allows for the exploration of interactions between factors. However, in this study, the extensive required computational time for the models renders the factorial approach infeasible.

Strengths of this study include a combined approach of using real data to fit the models and a simulation study to investigate how the model would perform under different conditions such as study sample size, follow-up period and level of censoring. Findings from the simulation study may guide researchers on setting these conditions appropriately when designing malaria

cohort studies for optimal results. Secondly, the real data used in this study had low level of missingness. Lastly, using the joint model, we were able to do predictions of risk of recurrent clinical episodes. The prediction ability can be crucial when designing malaria interventions.

The main limitation of this study was on the computational complexity of the likelihood for the joint models, resulting in nonconvergence problems of the EM algorithms. Non-convergence is a common problem in the field of joint modelling because of frequent high dimensional random effects and parameter space. Some of examples of joint model simulation studies with documented non-convergence problem include; Henderson et al. 2000 (55), Ferrer et al. 2016 (173) and Xu 2018 (174). The computational time further increased with increase in sample size and censoring. The computational time for some simulation models was long reaching up to 24 hours, using an Intel Core i7 2.5GHz CPU computer. The non-convergence problem prevented exploration of other simulation scenarios including larger sample sizes and longer study follow-up period which might be the practical conditions in most settings.

5.5 Authors' contributions

Study conceptualization and design: CCS, MM, LNK, MKL and TFC; Data collection: MKL, DPM and AGB; Statistical data analysis: CCS; Drafting of the manuscript: CCS; Led in interpretation of results: CCS, MM, MKL and TFC; Critical revision and approval of final version of the manuscript: CCS, MM, LNK, AGB, DPM, MGH, MKL and TFC.

5.6 Acknowledgements

This work was supported by the Fogarty International Center of the National Institutes of Health [Grant number D43TW010075]; and The Wellcome Trust [Grant number 107754/Z/15/Z]; National Institute of Allergy and Infectious Diseases [Grant number K24AI114996]. Funders have no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

5.7 Competing interests

The authors declare that they have no competing interests.

Chapter 6

Discussion and conclusions

6.1 Review of statistical analyses in malaria cohort studies

A review of malaria cohort studies that were published during 1996-2017 showed that most analyses have failed to capture the impact of the developed protective immunity in estimating the risk of repeated episodes of clinical malaria disease. Despite similar study designs across the reports, the statistical methods varied substantially and often represented overly simplistic models such as descriptive statistics, simple linear models, Poisson and negative binomial (NB) not accounting for within-participant correlation of the episodes (96,111,112). Ignoring this possible correlation of events tends to bias the model results (45,49,50,113) yielding regression parameters with underestimated standards errors for time-dependent covariates. Studies on prospective longitudinal malaria cohorts have increased over the time, a trend reflecting the growing understanding of the importance of longitudinal cohort data coupled with advancements in computer software developments for analyzing such data (48,51). The increased trend underscores the need for standardized analytical methods that can account for the impact of the developed protective immunity in estimating the risk of repeated episodes of clinical malaria disease.

6.2 Potential use of joint modelling

This study has demonstrated that a joint modelling longitudinal parasitemia and time-to first clinical malaria episodes (single-event joint model) outperformed the commonly used conventional time-dependent Cox PH model in precision when estimating the risk of clinical malaria. This is consistent with previous studies in other disease fields such as in HIV (26,123) and cancer (28,29,128). However, the single-event joint model where the subsequent episodes are ignored was outperformed yielding smaller estimates of hazard ratios in most cases with larger standard errors by a recurrent-events jointly model in which all clinical malaria episodes were included. Detailed results can be seen in papers 2 and 3 in appendences 4.1 and 5.1 respectively. These results underline the need for models to utilize all follow-up data borrowing strength of the entire history of the *P. falciparum* parasitemia process when estimating the risk of recurrent clinical malaria. These findings affirm the WHO MALVAC call for using recurrent events models to evaluate the efficacy and safety of the vaccine (31).

6.3 Simulation study

The simulation study of the recurrent events joint model shows that if correctly specified, it can give parameter estimates with low bias. The performance of the joint model improved, i.e. decreased bias by increasing study sample size and length of study follow-up, results which are consistent with previous simulation-based studies in joint modelling (89,169–171). However, increasing level of censoring worsened the performance of the joint model. This is in agreement with a previous joint modelling report by Efendi (2013) where they studied a marginalized joint model for continuous longitudinal and repeated time-to-event outcomes as well as a joint model for bivariate repeated time-to-event outcomes (171)

6.4 Prediction of risk of clinical malaria

The predicted risk-free period of a new clinical malaria episode conditional on the history of *P. falciparum* parasitemia trajectory were obtained using the single-event joint model. The prediction was more informative (with narrower confidence intervals) about the risk of clinical malaria at time closer to last parasitemia measurement than farther over follow-up. For longitudinal data, observations closer to the time of the event may be more informative about the event than distant ones (175). The ability to predict the risk of future clinical malaria episodes by the joint models may help malaria experts and policy makers in planning of malaria interventions.

6.5. Implication of results on existing malaria interventions

6.5.1 Long-lasting insecticide treated nets (LLIN) usage

In this study, LLIN usage was protective of participants developing clinical malaria, consistent with other previous reports from Malawi and elsewhere (9,131,133,150–153). The protective effect of net usage was however more prominent at the beginning of the study, the time when the participant received LLINs but later started to diminish just about over a year follow-up. This may be attributed to loss in physical integrity and biological activity of insecticides with ageing (154,155). Further, widespread insecticide resistance by the infectious mosquitoes has also been reported (156–158). In order to understand the long-term protective effect of LLINs in current local conditions, longitudinal studies are required to assess the physical integrity and biological activity of insecticides over time.

6.5.2 Age based malaria interventions

In this study, being of young age was associated with increased risk of experiencing clinical malaria episode. Considering aging as proxy for observation time (24,25), this study shows that there is protective effect to clinical malaria over time. Being a cohort from a high transmission area, participants are continuously exposed to repeated bites of infected Anopheles mosquitoes (161,162) and develop humoral immune responses to different stage-specific antigens of *P. falciparum* with aging (163). The school-aged children (5-15 years) had the high risk of developing clinical malaria, as also reported by other studies in SSA (125,164,165), and this may partly be explained by low LLIN usage (125,141,166,167). In Malawi, dedicated malaria preventive programs exist specially for under-five children and pregnant mothers. Examples include intermittent preventive treatment during pregnancy (IPTp), free LLINs for pregnant women at first antenatal care clinic (ANC) visit, and for infants attending first under-five clinic visit under the Expanded Program on Immunization (EPI). Results from this study underscore the need for expanding these programs and follow children longitudinally beyond school-aged children.

6.6 Study limitations

6.6.1 Inclusion criteria in review

Publications included in the critical review paper on statistical methods in this study were only those published in PubMed, Medline or ScienceDirect. Furthermore, the study only considered publications written in the English language. Hence, any other publications indexed in any other databases and written in any other language other than English were excluded.

6.6.2 Computational complexity of the likelihood

The other challenge in this study was on the computational complexity of the likelihood for the joint models which resulted in non-convergence problems of the EM algorithms. Non-convergence is a common problem in the field of joint modelling because of frequent high dimensional random effects and parameter space (55,173,174). This may be a barrier to adopting the joint modelling approach especially by malaria experts with limited background in statistics.

6.6.3 Variation of factors in simulation study

In the simulation study, factors were varied on a one-by-one basis and results compared to the reference scenario. This approach does not allow one to study the effect of interaction between factors. Morris et al. 2019 (172) recommend varying factors factorially as this approach may likely be more informative since this allows for the exploration of interactions between factors. However, in this study, the required extensive computational time for the models renders the factorial approach infeasible.

6.7 Strengths of the study

6.7.1 Design-specific review

Strengths of this study include a review of analytical methods dedicated to a specific study design - prospective cohort studies. This allowed for close assessment of the statistical methods content based on a uniform criterion and therefore comparable across all reviewed articles.

6.7.2 Low level of missing data

Mfera cohort the data which were used for model application had low level of missing with majority of the variables having more than 99% completeness. By using data with low level of missingness, threat on the validity of the results due to bias and loss of efficiency is minimized (34,35).

6.7.3 Optimal link structure between survival processes and longitudinal

Most joint modelling studies just settle for current value parameterization which is mostly the default association structure between the longitudinal and survival processes as elaborated in section 2.3. However, this study explored different structures which allowed for examining of the optimal way to link between parasitemia and clinical malaria processes.

6.7.4 Dynamic prediction ability

Thirdly, the ability of the joint models to be used for individual-level predictions of risk of new current clinical malaria episodes conditioned on longitudinal measurements of *P. falciparum* parasites offer additional value to planning of malaria interventions. Further details on dynamic prediction of future malaria episodes are in sections 4.3.6 and 6.4.

6.8 Areas of further research

The work in this thesis has a number of possible extensions for research. Firstly, this study has found that there is an increased trend in malaria cohort studies coupled with improved but slow adoption of appropriate methods. Future studies should consider conducting systematic reviews to test whether more appropriate longitudinal statistical methods improve or change

understanding of malaria epidemiology by using among other factors, authors' background information. Secondly, this study has utilized joint models in prediction of probability (risk) of future clinical malaria episodes at two levels using:

(i) Single-event joint model to predict the probability of a new episode conditional on history of parasitemia trajectory for specific study participants (individual-level dynamic prediction); and
(ii) Recurrent events joint models for the conditional prediction of the probability of future clinical malaria episodes for a group of participants (population-level prediction).

Future studies are required to investigate the individual-level prediction of risk of futures episodes using recurrent events joint models.

Further, the simulation study focused on bias but future studies should also consider assessing efficiency and consistency in recurrent events joint model particularly using relatively larger sample sizes. Lastly, future studies are also required to investigate how different data missing mechanisms can impact the estimation in the recurrent events joint models. In this study, likelihood and Bayesian methods were used, and both approaches were valid under MAR mechanism. However, further simulation studies should investigate the role of missingness mechanisms within the recurrent events joint modelling framework where the data are missing not at random (MNAR), which maybe a common applicable mechanism in some practical settings.

References

1. Aponte JJ, Menendez C, Schellenberg D, Kahigwa E, Mshinda H, Vountasou P, et al. Age interactions in the development of naturally acquired immunity to Plasmodium falciparum and its clinical presentation. PLoS Med. 2007;4(12):e242.

2. World Health Organization. World Malaria Report 2018. 2018.

3. World Health Organization (WHO). Epidemiological profile of malaria in Malawi in 2015 [Internet]. 2017 [cited 2017 Jul 11]. Available from: http://www.who.int/malaria/publications/country-profiles/profile_mwi_en.pdf?ua=1

4. Bennett A, Kazembe L, Mathanga DP, Kinyoki D, Ali D, Snow RW, et al. Mapping malaria transmission intensity in Malawi, 2000-2010. Am J Trop Med Hyg. 2013;89(5):840–9.

5. Malawi Malaria Indicator Survey (MIS) 2014. Lilongwe, Malawi, and Rockville, Maryland, USA; 2014.

6. Kazembe LN, Kleinschmidt I, Holtz TH, Sharp BL. Spatial analysis and mapping of malaria risk in Malawi using point-referenced prevalence of infection data. Int J Health Geogr. 2006;5(1):41.

7. Kabaghe AN, Chipeta MG, Gowelo S, Mburu M, Truwah Z, McCann RS, et al. Fine-scale spatial and temporal variation of clinical malaria incidence and associated factors in children in rural Malawi: A longitudinal study. Parasites and Vectors. 2018;11(129).

8. Mathanga DP, Walker ED, Wilson ML, Ali D, Taylor TE, Laufer MK. Malaria control in Malawi: Current status and directions for the future. Acta Trop. 2012;121(3):212–7.

9. Lindblade KA, Mwandama D, Mzilahowa T, Steinhardt L, Gimnig J, Shah M, et al. A cohort study of the effectiveness of insecticide-treated bed nets to prevent malaria in an area of moderate pyrethroid resistance, Malawi. Malar J. 2015;14(1):31.

10. Takken W, Knols BG. Malaria vector control: current and future strategies. Trends Parasitol. 2009;25(3):101–4.

11. Mathanga D, Molyneux ME. Bednets and malaria in Africa. Vol. 28, Malawi Medical Journal. 2016. p. 92–3.

12. Malawi Ministry of Health. Malawi National Malaria Indicator Survey 2010. Malariasurveys.org. 2010.

13. Boudová S, Divala T, Mawindo P, Cohee L, Kalilani-Phiri L, Thesing P, et al. The prevalence of malaria at first antenatal visit in Blantyre, Malawi declined following a universal bed net campaign. Malar J. 2015;14(1):422.

14. WHO. First malaria vaccine in Africa: A potential new tool for child health and improved malaria control. Who. 2018.

15. Roca-Feltrer A, Kwizombe CJ, Sanjoaquin MA, Sesay SSS, Faragher B, Harrison J, et al. Lack of decline in childhood Malaria, Malawi, 2001-2010. Emerg Infect Dis. 2012;18(2):272–8.

16. Okiro EA, Kazembe LN, Kabaria CW, Ligomeka J, Noor AM, Ali D, et al. Childhood Malaria Admission Rates to Four Hospitals in Malawi between 2000 and 2010. PLoS One. 2013;8(4).

17. Kepha S, Nikolay B, Nuwaha F, Mwandawiro CS, Nankabirwa J, Ndibazza J, et al. Plasmodium falciparum parasitaemia and clinical malaria among school children living in a high transmission setting in western Kenya. Malar J. 2016;15(157).

18. Loucoubar C, Grange L, Paul R, Huret A, Tall A, Telle O, et al. High number of previous Plasmodium falciparum clinical episodes increases risk of future episodes in a sub-group of individuals. PLoS One. 2013;8(2).

19. Guinovart C, Dobaño C, Bassat Q, Nhabomba A, Quintó L, Manaca MN, et al. The role of age and exposure to Plasmodium falciparum in the rate of acquisition of naturally acquired immunity: A randomized controlled trial. PLoS One. 2012;7(3).

20. Kazembe LN. Spatial modelling and risk factors of malaria incidence in northern Malawi. Acta Trop. 2007;102(2):126–37.

21. Chitunhu S, Musenge E. Direct and indirect determinants of childhood malaria morbidity in Malawi: a survey cross-sectional analysis based on malaria indicator survey data for 2012. Malar J. 2015;14(1):265.

22. Segula D, Frosch AP, SanJoaquin M, Taulo D, Skarbinski J, Mathanga DP, et al. Prevalence and spectrum of illness among hospitalized adults with malaria in Blantyre, Malawi. Malar J. 2014;13(1):391.

23. Lowe R, Chirombo J, Tompkins AM. Relative importance of climatic, geographic and socio-economic determinants of malaria in Malawi. Malar J. 2013;12(1):416.

24. Doolan DL, Dobaño C, Baird JK. Acquired immunity to malaria. Clin Microbiol Rev. 2009;22(1):13–36.

25. Biswas S, Seth RK, Tyagi PK, Sharma SK, Dash AP. Naturally acquired immunity and reduced susceptibility to falciparum malaria in two subpopulations of endemic Eastern India. Scand J Immunol. 2008;67(2):177–84.

26. Song X, Davidian M, Tsiatis AA. A Semiparametric Likelihood Approach to Joint Modeling of Longitudinal and Time-to-Event Data. Biometrics. 2002;58(2012):742–53.

27. Tsiatis AA, Davidian M. Joint modeling of Longitudinal and Time-to-Event Data: An Overview. Stat Sin. 2004;14:809–34.

28. Chi YY, Ibrahim JG. Joint models for multivariate longitudinal and multivariate survival data. Biometrics. 2006;62(2):432–45.

29. Brown ER, Ibrahim JG. A Bayesian semiparametric joint hierarchical model for longitudinal and survival data. Biometrics. 2003;59(2):221–8.

30. Sagara I, Giorgi R, Doumbo OK, Piarroux R, Gaudart J. Modelling recurrent events: comparison of statistical models with continuous and discontinuous risk intervals on recurrent malaria episodes data. Malar J. 2014;13(1):293.

31. Moorthy VS, Reed Z, Smith PG. Clinical trials to estimate the efficacy of preventive interventions against malaria in paediatric populations: A methodological review. Vol. 8, Malaria Journal. 2009. p. 23.

32. Albert PS. A transitional model for longitudinal binary data subject to nonignorable missing data. Biometrics. 2000;56(2):602–8.

33. Fitzmaurice GM, Ravichandran C. A Primer in Longitudinal Data Analysis Features of Longitudinal Studies Covariance Structure Statistical Primer for Cardiovascular Research. Circulation. 2008;118(19).

34. Peng C-YJ, Harwell M, Liou S-M, Ehman LH, others. Advances in missing data methods and implications for educational research. Real data Anal. 2006;10(1):31–78.

35. Ibrahim JG, Molenberghs G. Missing data methods in longitudinal studies: A review. Test. 2009;18(1):1–43.

36. Rubin DB. Inference and missing data. Biometrika. 1976;63(3):581–92.

37. Little RJA, Rubin DB. Statistical Analysis with Missing Data. John Wiley Sons Hoboken NJ. 2002. 381 p.

38. He Y. Missing data analysis using multiple imputation: Getting to the heart of the matter. Vol. 3, Circulation: Cardiovascular Quality and Outcomes. 2010. p. 98–105.

39. Dong Y, Peng C-YJ. Principled missing data methods for researchers. Springerplus. 2013;2(1):222.

40. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393.

41. Ibrahim JG, Chen M-H, Lipsitz SR, Herring AH. Missing-Data Methods for Generalized Linear Models: A Comparative Review. J Am Stat Assoc. 2005;100(469):332–46.

42. Graham JW. Missing data analysis: making it work in the real world. Annu Rev Psychol. 2009;60:549–76.

43. Tseng C, Elashoff R, Li N, Li G. Longitudinal data analysis with non-ignorable missing data. Stat Methods Med Res. 2016;25(1):205–20.

44. Mukaka M, White SA, Terlouw DJ, Mwapasa V, Kalilani-Phiri L, Faragher EB. Is using multiple imputation better than complete case analysis for estimating a prevalence (risk) difference in randomized controlled trials when binary outcome observations are missing? Trials. 2016;17(341).

45. Laird NM, Ware JH. Random-Effects Models for Longitudinal Data. Biometrics. 1982;38(4):963–74.

46. Liang K-LK&. ZS, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986;73(1):13–22.

47. Hedeker D, Gibbons RD. Missing Data in Longitudinal Studies. Longitud Data Anal. 2006;279–312.

48. Wang M. Generalized Estimating Equations in Longitudinal Data Analysis: A Review and Recent Developments. Adv Stat. 2014;2014:1–11.

49. Gardiner JC, Luo Z, Roman LA. Fixed effects, random effects and GEE: what are the differences? Stat Med. 2009;28:221–39.

50. Hubbard AE, Ahern J, Fleischer NL, Laan M Van Der, Lippman SA, Jewell N, et al. To GEE or not to GEE: Comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. Epidemiology. 2010;21(4):467–74.

51. Gibbons RD, Hedeker D, DuToit S. Advances in analysis of longitudinal data. Annu Rev Clin Psychol. 2010;6:79–107.

52. Troxel a B, Lipsitz SR, Brennan T a. Weighted estimating equations with nonignorably missing response data. Biometrics. 1997;53(3):857–69.

53. Shen CW, Chen YH. Model Selection for Generalized Estimating Equations Accommodating Dropout Missingness. Biometrics. 2012;68(4):1046–54.

54. Shen CW, Chen YH. Model selection of generalized estimating equations with multiply imputed longitudinal data. Biometrical J. 2013;55(6):899–911.

55. Henderson R, Diggle P, Dobson A. Joint modelling of longitudinal measurements and event time data. Biostatistics. 2000;1(4):465–80.

56. Shen Y, Huang H, Guan Y. A conditional estimating equation approach for recurrent event data with additional longitudinal information. Stat Med. 2016;35:4306–19.

57. Liu L, Huang X. Joint analysis of correlated repeated measures and recurrent events processes in the presence of death, with application to a study on acquired immune deficiency syndrome. J R Stat Soc Ser C Appl Stat. 2009;58(1):65–81.

58. McCulloch C, Searle S. Generalized, Linear, and Mixed Models Generalized, Linear, and Mixed Models. Wiley-Interscience. New York: John Wiley & Sons; 2001.

59. Gelman A, Carlin JBB, Stern HSS, Rubin DBB. Bayesian Data Analysis. third. Book. New York: Chapman & Hall/CRC Press; 2014.

60. Zhang X, Mallick H, Tang Z, Zhang L, Cui X, Benson AK, et al. Negative binomial mixed models for analyzing microbiome count data. BMC Bioinformatics. 2017;18(4).

61. Collett D. Modelling survival data in medical research. 3rd ed. Modelling Survival Data in Medical Research, Third Edition. London: Chapman and Hall/CRC Press; 2015.

62. Lee ET, Go OT. Survival analysis in public health research. Annu Rev Public Health. 1997;18:105–34.

63. Cox D. Regression models and life tables. J R Stat Soc B. 1972;34:182–220.

64. Hougaard P. Frailty models for survival data. Lifetime Data Anal. 1995;1(3):255–73.

65. Box-Steffensmeier JM, De Boef S. Repeated events survival models: The conditional frailty model. Stat Med. 2006;25(20):3518–33.

66. Therneau TM, Grambsch PM. Modeling Survival Data : Extending The Cox Model. 2nd ed.

Statistics for biology and health. New York: Springer; 2000.

67. Clayton D. Some approaches to the analysis of recurrent event data. Stat Methods Med Res. 1994;3(3):244–62.

68. Clayton DG. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. Biometrika. 1978;65(1):141–51.

69. Andersen PK, Gill RD. Cox's Regression Model for Counting Processes: A Large Sample Study. Ann Stat. 1982;10:1100–20.

70. Prentice RL, Williams BJ, Peterson A V. On the regression analysis of multivariate failure time data. Biometrika. 1981;68:373–9.

71. Lin DY, Wei LJ, Yang I, Ying Z. Semiparametric regression for the mean and rate functions of recurrent events. J R Stat Soc Ser B Stat Methodol. 2000;62:711–30.

72. Amorim LDAF, Cai J. Modelling recurrent events: A tutorial for analysis in epidemiology. Int J Epidemiol. 2015;44(1):324–33.

73. Kelly PJ, Lim LLY. Survival analysis for recurrent event data: an application to childhood infectious diseases. Stat Med. 2000;19(1):13–33.

74. Hougaard P. Analysis of multivariate survival data. New York: Springer Verlag; 2000.

75. Duchateau L, Janssen P. Penalized partial likelihood for frailties and smoothing splines in time to first insemination models for dairy cows. Biometrics. 2004;60(3):608–14.

76. Duchateau L, Janssen P, Gail M, Krickeberg K, Sarmet J, Tsiatis A, et al. The Frailty Model. New York: Springer; 2008.

77. Hougaard P. Survival models for heterogeneous populations derived from stable distributions. Biometrika. 1986;73(2):387–96.

78. Hougaard P. A class of multivanate failure time distributions. Biometrika. 1986;73(3):671–8.

79. Abbring JH, Van Den Berg GJ. The unobserved heterogeneity distribution in duration analysis. Biometrika. 2007;94(1):87–99.

80. Hogan JW, Laird NM. Model-based approaches to analysing incomplete longitudinal and failure time data. Stat Med. 1997;16:259–72.

81. Yu M, Law NJ, Taylor JMG, Sandler HM. Joint longitudinal-survival-cure models and their application to prostate cancer. Stat Sin. 2004;14:832–5.

82. Rizopoulos D. Joint models for longitudinal and time-to-event data: With applications in R. 1st ed. Joint Models for Longitudinal and Time-to-Event Data: With Applications in R. Boca Raton: Chapman and Hall/CRC Biostatistics Series; 2012.

83. Schluchter MD. Methods for the analysis of informatively censored longitudinal data. Stat Med. 1992;11(14):1861–70.

84. Gruttola V De, Tu XM. Modelling Progression of CD4-Lymphocyte Count and Its Relationship to Survival Time. Biometrics. 1994;50(4):1003–14.

85. Self S, Pawitan Y. Modeling a Marker of Disease Progression and Onset of Disease. In: AIDS Epidemiology. 1992. p. 231–55.

86. Wulfsohn MS, Tsiatis AA. A Joint Model for Survival and Longitudinal Data Measured with Error. Biometrics. 1997;53:330–9.

87. Tsiatis AA, Degruttola V, Wulfsohn MS. Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and cd4 counts in patients with aids. J Am Stat Assoc. 1995;90:27–37.

88. Rizopoulos D. JM: An R Package for the Joint Modelling of Longitudinal and Time-to-Event Data. J Stat Softw. 2010;35(9).

89. Han J, Slate EH, Peña EA. Parametric latent class joint model for a longitudinal biomarker and recurrent events. Stat Med. 2007;26(29):5285–302.

90. Mbogning C, Bleakley K, Lavielle M. Joint modelling of longitudinal and repeated time-to-event data using nonlinear mixed-effects models and the stochastic approximation expectation–maximization algorithm. J Stat Comput Simul. 2015;85(8):1512–28.

91. Li S. Joint modeling of recurrent event processes and intermittently observed time-varying binary covariate processes. Lifetime Data Anal. 2016;22(1):145–60.

92. Pickles A, Crouchley R. A comparison of frailty models for multivariate survival data. Stat Med. 1995;14(13):1447–61.

93. Wang MC, Qin J, Chiang CT. Analyzing recurrent event data with informative censoring. J Am Stat Assoc. 2001;96:455.

94. Milali MP, Sikulu-Lord MT, Govella NJ. Bites before and after bedtime can carry a high risk of human malaria infection. Malar J. 2017;16(1):91.

95. Rossati A, Bargiacchi O, Kroumova V, Zaramella M, Caputo A, Garavelli PL. Climate, environment and transmission of malaria. Infez Med. 2016;24(2):93–104.

96. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986;73(1):13–22.

97. Lin DY. Cox regression analysis of multivariate failure time data: the marginal approach. Stat Med. 1994;13(21):2233–47.

98. Crowder M. On the use of a working correlation matrix in using generalised linear models for repeated measures. Biometrika. 1995;82(2):407.

99. Gibbons RD, Hedeker D, DuToit S. Advances in Analysis of Longitudinal Data. SSRN. 2010.

100. Burton P, Gurrin L, Sly P, Gurrin L. Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling. Stat Med. 1998;17(11):1261–91.

101. Zeger S, Liang K-Y. Longitudinal data analysis for discrete and continuous outcomes. Biometrics. 1986;42(1):121–30.

102. Ballinger GA. Using Generalized Estimating Equations for Longitudinal Data Analysis. Organ Res Methods. 2004;7(2):127–50.

103. Hu XJ. Applied Longitudinal Data Analysis for Epidemiology: A Practical Guide. J Am Stat Assoc. 2004;99(468):1207–8.

104. Qu A. Improving generalised estimating equations using quadratic inference functions. Biometrika. 2000;87(4):823–36.

105. Heagerty PJ. Marginalized transition models and likelihood inference for longitudinal categorical data. Biometrics. 2002;58(2):342–51.

106. Halekoh U, Højsgaard S, Yan J. The R Package geepack for generalized estimating equations. J Stat Softw. 2006;15(2):1–11.

107. Bernal-Rusiel JL, Greve DN, Reuter M, Fischl B, Sabuncu MR. Statistical analysis of longitudinal neuroimage data with Linear Mixed Effects models. Neuroimage. 2013;66:249–60.

108. StataCorp. Stata Statistical Software: Release 15. College Station, TX: StataCorp LP. 2017.

109. Emerson JD, Colditz GA. Use of statistical analysis in the New England Journal of Medicine. N Engl J Med. 1983;309(12):709–13.

110. Colditz GA, Emerson JD. The statistical content of published medical research: some implications for biomedical education. Med Educ. 1985;19(3):248–55.

111. Bates DM. Models for longitudinal data. lme4 Mix Model with R. 2010;63–82.

112. Chan D. Data Analysis and Modeling Longitudinal Processes. Gr Organ Manag. 2003;28(3):341–65.

113. Pardo MC, Alonso R. GEEs for repeated categorical responses based on generalized residuals. J Stat Comput Simul. 2014;84(2):344–59.

114. Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, et al. Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol Evol. 2009;24(3):127–35.

115. World Health Organisation (WHO). World Malaria Report 2017 [Internet]. 2017 [cited 2017 Sep 8]. p. 186. Available from: http://www.who.int/malaria/publications/world-malaria-report-2017/en/

116. Barber BE, William T, Grigg MJ, Menon J, Auburn S, Marfurt J, et al. A prospective comparative study of Knowlesi, falciparum, and vivax malaria in Sabah, Malaysia: High proportion with severe disease from Plasmodium knowlesi and Plasmodium vivax but no mortality with early referral and artesunate therapy. Clin Infect Dis. 2013;56(3):383–97.

117. Kasirye R, Grosskurth H, Munderi P, Levin J, Anywaine Z, Nunn A, et al. Longitudinal effect of CD4 by cotrimoxazole use on malaria incidence among HIV-infected Ugandan adults on antiretroviral therapy: A randomized controlled study. Malar J. 2016;15(1).

118. Guo X, Carlin BP. Separate and Joint Modeling of Longitudinal and Event Time Data Using Standard Computer Packages. Am Stat. 2004;58(1):16–24.

119. White NJ. Malaria parasite clearance. Vol. 16, Malaria Journal. 2017. p. 194.

120. Kalbfleisch JD, Prentice RL. The Statistical Analysis of Failure Time Data. John Wiley and

Sons New York. New York: John Wiley and Sons; 2002. 462 p.

121. Taylor JM, Wang Y. Surrogate markers and joint models for longitudinal and survival data. Control Clin Trials. 2002;23(6):626–34.

122. Ibrahim JG, Chen M-H, Sinha D. Bayesian methods for joint modeling of longitudinal and survival data with applications to cancer vaccine trials. Stat Sin. 2004;14(3):863–884.

123. Chen Q, May RC, Ibrahim JG, Chu H, Cole SR. Joint modeling of longitudinal and survival data with missing and left-censored time-varying covariates. Stat Med. 2014;33(26):4560–76.

124. Buchwald AG, Sixpence A, Chimenya M, Damson M, Sorkin JD, Wilson ML, et al. Clinical implications of asymptomatic Plasmodium falciparum infections in Malawi. Clin Infect Dis. 2019;68(1):106–12.

125. Walldorf JA, Cohee LM, Coalson JE, Bauleni A, Nkanaunena K, Kapito-Tembo A, et al. School-age children are a reservoir of malaria infection in Malawi. PLoS One. 2015;10(7).

126. Ibrahim JG, Chen MH, Sinha D. Bayesian Survival Analysis. New York: Springer; 2001.

127. R Core Team. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2017.

128. Ibrahim JG, Chu H, Chen LM. Basic concepts and methods for joint models of longitudinal and survival data. J Clin Oncol. 2010;28(16):2796–801.

129. Lawrence Gould A, Boye ME, Crowther MJ, Ibrahim JG, Quartey G, Micallef S, et al. Joint modeling of survival and longitudinal non-survival data: Current methods and issues. Report of the DIA Bayesian joint modeling working group. Stat Med. 2015;34(14):2181–95.

130. Mathanga DP, Tembo AK, Mzilahowa T, Bauleni A, Mtimaukenena K, Taylor TE, et al. Patterns and determinants of malaria risk in urban and peri-urban areas of Blantyre, Malawi. Malar J. 2016;15(1):590.

131. Alemu A, Tsegaye W, Golassa L, Abebe G. Urban malaria and associated risk factors in Jimma town, south-west Ethiopia. Malar J. 2011;10:173.

132. Haque U, Glass GE, Bomblies A. Risk factors associated with multiple malaria infections in Bangladesh. Am J Trop Med Hyg. 2012;87(5 SUPPL. 1):403.

133. Haque U, Glass GE, Bomblies A, Hashizume M, Mitra D, Noman NN, et al. Risk Factors Associated with Clinical Malaria Episodes in Bangladesh: A Longitudinal Study. Am J Trop Med Hyg. 2013;88(4):7–9.

134. Hickey GL, Philipson P, Jorgensen A, Kolamunnage-Dona R. Joint models of longitudinal and time-to-event data with more than one event time outcome: A review. Int J Biostat. 2018;14(1).

135. Hickey GL, Philipson P, Jorgensen A, Kolamunnage-Dona R. Joint modelling of time-to-event and multivariate longitudinal outcomes: Recent developments and issues. BMC Med Res Methodol. 2016;16(1).

136. Proust-Lima C, Séne M, Taylor JMG, Jacqmin-Gadda H. Joint latent class models for longitudinal and time-to-event data: a review. Stat Methods Med Res. 2014;23(1):74–90.

137. Billingham L, Abrams K. Simultaneous analysis of quality of life and survival data. Stat

Methods Med Res. 2002;11(1):25–48.

138. Marsh K, Kinyanjui S. Immune effector mechanisms in malaria. Parasite Immunol. 2006;28(2):51–60.

139. Lin H, Turnbull BW, McCulloch CE, Slate EH. Latent class models for joint analysis of longitudinal biomarker and event process data: Application to longitudinal prostate-specific antigen readings and prostate cancer. J Am Stat Assoc. 2002;97(457):53–65.

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

141. Buchwald AG, Sorkin JD, Sixpence A, Chimenya M, Damson M, Wilson ML, et al. Association between Age and Plasmodium falciparum Infection Dynamics. Am J Epidemiol. 2019;188(1):169–76.

142. Earland D, Buchwald AG, Sixpence A, Chimenya M, Damson M, Seydel K, et al. Impact of Multiplicity of Plasmodium falciparum Infection on Clinical Disease in Malawi. Am J Trop Med Hyg. 2019;00(0):1–4.

143. Stanley CC, Kazembe LN, Buchwald AG, Mukaka M, Mathanga DP, Hudgens MG, et al. Joint modelling of time-to-clinical malaria and parasite count in a cohort in an endemic area. J Med Stat Informatics. 2019;7(1).

144. Buchwald AG, Sixpence A, Chimenya M, Damson M, Sorkin JD, Wilson ML, et al. Clinical implications of asymptomatic Plasmodium falciparum infections in Malawi. Clin Infect Dis. 2019;68(1):106–12.

145. Bartoloni A, Zammarchi L. Clinical aspects of uncomplicated and severe malaria. Mediterr J Hematol Infect Dis. 2012;4(1).

146. Hsieh F, Tseng YK, Wang JL. Joint modeling of survival and longitudinal data: Likelihood approach revisited. Biometrics. 2006;62(4):1037–43.

147. Elashoff RM, Li G, Li N. A joint model for longitudinal measurements and survival data in the presence of multiple failure types. Biometrics. 2008;64(3):762–71.

148. Wang Y, Taylor JMG. Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome. J Am Stat Assoc. 2001;96(455):895–905.

149. Huang X, Liu L. A joint frailty model for survival and gap times between recurrent events. Biometrics. 2007;63(2):389–97.

150. Holtz TH, Marum LH, Mkandala C, Chizani N, Roberts JM, Macheso A, et al. Insecticide-treated bednet use, anaemia, and malaria parasitaemia in Blantyre district, Malawi. Trop Med Int Heal. 2002;7:220–30.

151. Mathanga DP, Campbell CH, Taylor TE, Barlow R, Wilson ML. Socially marketed insecticide-treated nets effectively reduce Plasmodium infection and anaemia among children in urban Malawi. Trop Med Int Heal. 2006;11:1367–74.

152. Clark TD, Greenhouse B, Njama-Meya D, Nzarubara B, Maiteki-Sebuguzi C, Staedke SG, et al. Factors Determining the Heterogeneity of Malaria Incidence in Children in Kampala, Uganda. J Infect Dis. 2008;198(3):393–400.

153. De Beaudrap P, Turyakira E, White LJ, Nabasumba C, Tumwebaze B, Muehlenbachs A, et al. Impact of malaria during pregnancy on pregnancy outcomes in a Ugandan prospective cohort with intensive malaria screening and prompt treatment. Malar J. 2013;12(1):139.

154. Tusting LS, Willey B, Lucas H, Thompson J, Kafy HT, Smith R, et al. Socioeconomic development as an intervention against malaria: A systematic review and meta-analysis. Lancet. 2013;382(9896):963–72.

155. Mathanga DP, Mwandama DA, Bauleni A, Chisaka J, Shah MP, Landman KZ, et al. The effectiveness of long-lasting, insecticide-treated nets in a setting of pyrethroid resistance: a case-control study among febrile children 6 to 59 months of age in Machinga District, Malawi. Malar J. 2015;14(457).

156. Riveron JM, Chiumia M, Menze BD, Barnes KG, Irving H, Ibrahim SS, et al. Rise of multiple insecticide resistance in Anopheles funestus in Malawi: A major concern for malaria vector control. Malar J. 2015;14(344).

157. Mzilahowa T, Chiumia M, Mbewe RB, Uzalili VT, Luka-Banda M, Kutengule A, et al. Increasing insecticide resistance in Anopheles funestus and Anopheles arabiensis in Malawi, 2011-2015. Malar J. 2016;15(1):563.

158. Wiebe A, Longbottom J, Gleave K, Shearer FM, Sinka ME, Massey NC, et al. Geographical distributions of African malaria vector sibling species and evidence for insecticide resistance. Malar J. 2017;16(1):85.

159. Sougoufara S, Diédhiou SM, Doucouré S, Diagne N, Sembène PM, Harry M, et al. Biting by Anopheles funestus in broad daylight after use of long-lasting insecticidal nets: A new challenge to malaria elimination. Malar J. 2014;13(125).

160. Russell TL, Govella NJ, Azizi S, Drakeley CJ, Kachur SP, Killeen GF. Increased proportions of outdoor feeding among residual malaria vector populations following increased use of insecticide-treated nets in rural Tanzania. Malar J. 2011;10(80).

161. Guelbéogo WM, Gonçalves BP, Grignard L, Bradley J, Serme SS, Hellewell J, et al. Variation in natural exposure to anopheles mosquitoes and its effects on malaria transmission. Elife. 2018;7(e32625).

162. Bousema T, Sutherland CJ, Churcher TS, Mulder B, Gouagna LC, Riley EM, et al. Human immune responses that reduce the transmission of Plasmodium falciparum in African populations. Int J Parasitol. 2011;41(3):293–300.

163. Cohen S, McGregor IA, Carrington S. Gamma-globulin and acquired immunity to human malaria. Nature. 1961;192:733–7.

164. Mathanga DP, Halliday KE, Jawati M, Verney A, Bauleni A, Sande J, et al. The high burden of malaria in primary school children in Southern Malawi. Am J Trop Med Hyg. 2015;

165. Sissoko MS, Dicko A, Briët OJT, Sissoko M, Sagara I, Keita HD, et al. Malaria incidence in relation to rice cultivation in the irrigated Sahel of Mali. Acta Trop. 2004;89(2):161–70.

166. Noor AM, Kirui VC, Brooker SJ, Snow RW. The use of insecticide treated nets by age: implications for universal coverage in Africa. BMC Public Health. 2009;9(369).

167. Buchwald AG, Walldorf JA, Cohee LM, Coalson JE, Chimbiya N, Bauleni A, et al. Bed net

use among school-aged children after a universal bed net campaign in Malawi. Malar J. 2016;

168. Gatton ML, Cheng Q. Modeling the development of acquired clinical immunity to Plasmodium falciparum malaria. Infect Immun. 2004;72(11):6538–45.

169. Arisido MW, Antolini L, Bernasconi DP, Valsecchi MG, Rebora P. Joint model robustness compared with the time-varying covariate Cox model to evaluate the association between a longitudinal marker and a time-to-event endpoint. BMC Med Res Methodol. 2019;19(1):222.

170. Roustaei N, Ayatollahi SMT, Zare N. A Proposed Approach for Joint Modeling of the Longitudinal and Time-To-Event Data in Heterogeneous Populations: An Application to HIV/AIDS's Disease. Biomed Res Int. 2018;2018.

171. Efendi A, Molenberghs G, Njagi EN, Dendale P. A joint model for longitudinal continuous and time-to-event outcomes with direct marginal interpretation. Biometrical J. 2013;55(4):572–588.

172. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Stat Med. 2019;38(11).

173. Ferrer L, Rondeau V, Dignam J, Pickles T, Jacqmin-Gadda H, Proust-Lima C. Joint modelling of longitudinal and multi-state processes: application to clinical progressions in prostate cancer. Stat Med. 2016;35(22).

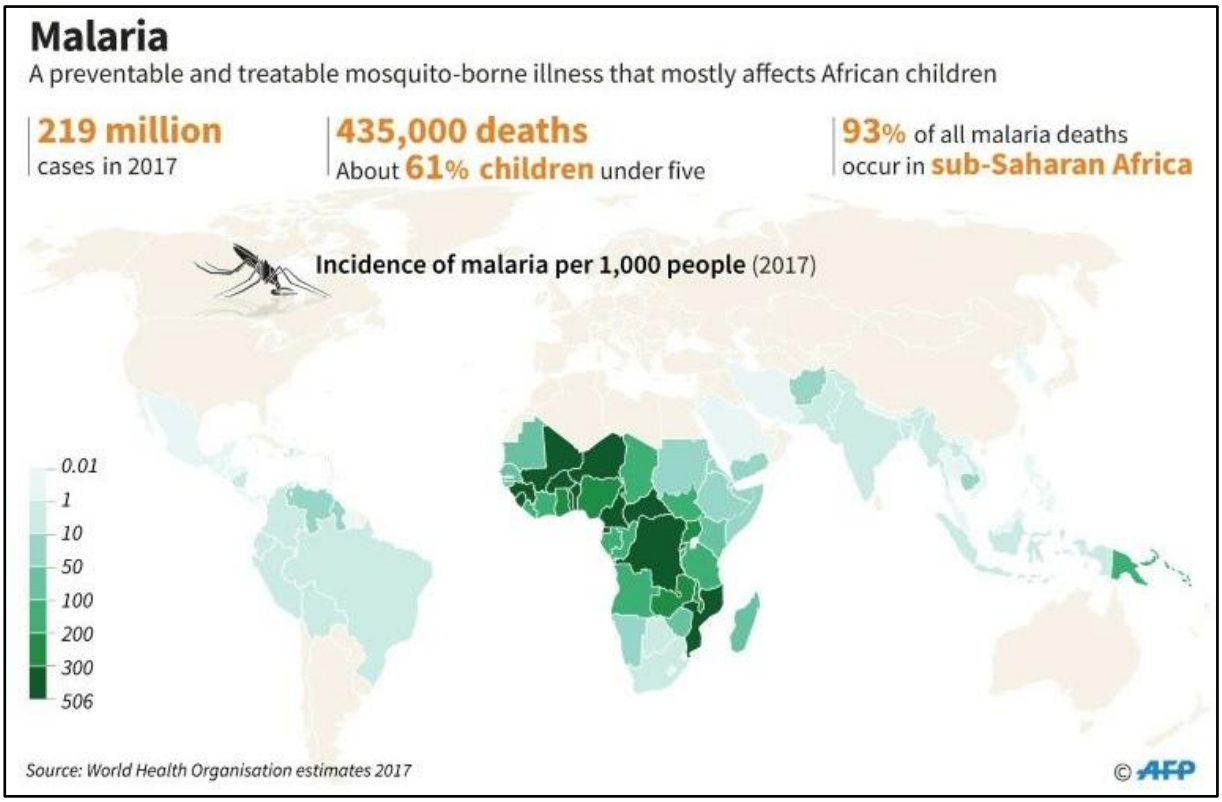
174. Xu C, Chinchilli VM, Wang M. Joint modeling of recurrent events and a terminal event adjusted for zero inflation and a matched design. Stat Med. 2018;37(18):2771–86.

175. Li S, Sun Y, Huang CY, Follmann DA, Krause R. Recurrent event data analysis with intermittently observed time-varying covariates. Stat Med. 2016;35(18):3049–65.

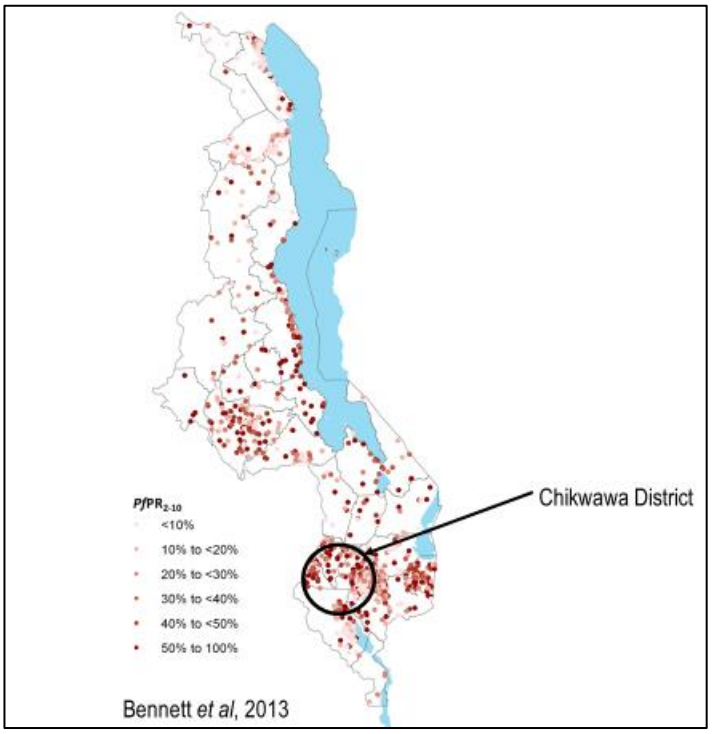
Appendices

The appendices for chapters 1-2; and the papers for chapters 3-5 as published or in-press online, as such page numbers and other detail are according to the specific journals.

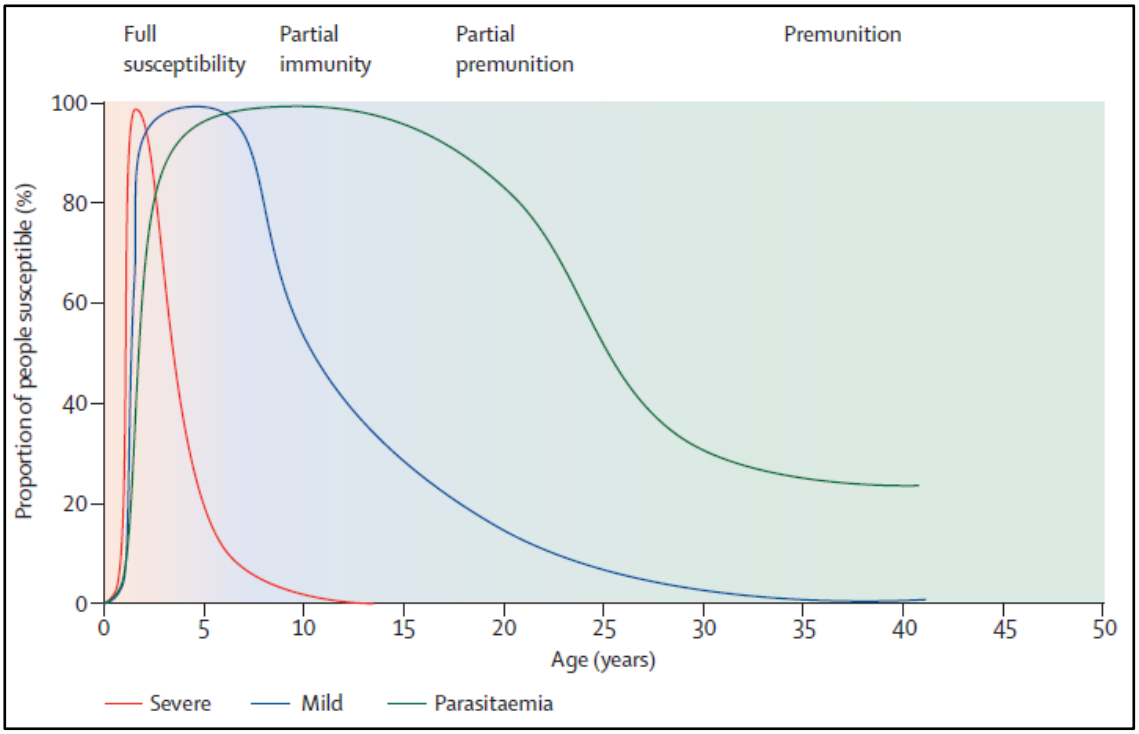
Appendix 1.1. World map showing the burden of Malaria



Appendix 1.2. Map of Malawi showing the distribution of age-standardized community Plasmodium falciparum parasite rate during 2000–2011.

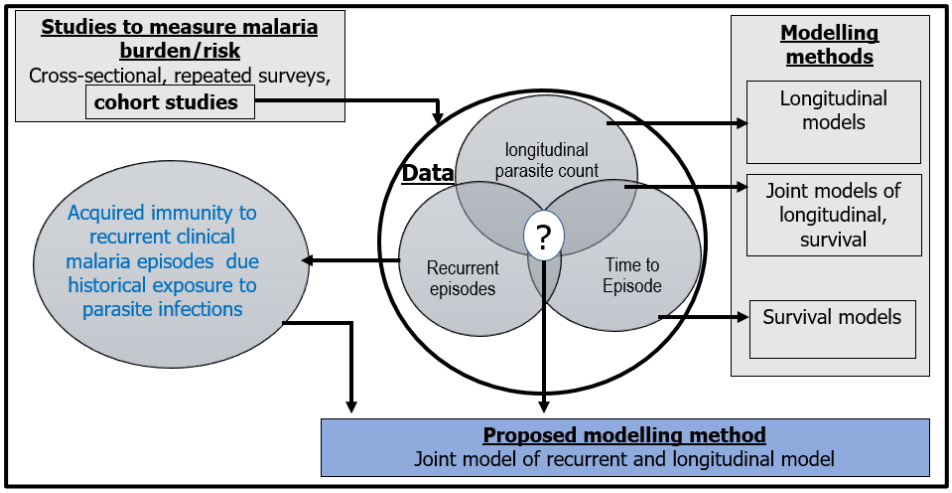


Appendix 1.3. Relationship between age and malaria severity in an area of moderate transmission intensity.



With repeated exposure protection is acquired, against severe malaria, then illness with malaria and much more slowly against microscopy-detectable parasitemia. Age-related acquired immunity protects against disease, not infection.

Appendix 1.4. Research gap



Appendix 2.1. Ethics Certificate

R14/49 Mr Christopher Stanley

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M170952

NAME: (Principal Investigator)	Mr Christopher Stanley
DEPARTMENT:	Public Health Malaria Alert Centre, Malawi
PROJECT TITLE:	Joint Modelling of Recurrent Malaria Episodes and Longitudinal Parasitic Density Incorporating Missing and Misclassification Error
DATE CONSIDERED:	29/09/2017
DECISION:	Approved unconditionally
CONDITIONS:	South African Human Research Ethics Committees (HRECs) have no standing outside South Africa. Ethics approval is also required from local HRECs in Malawi
SUPERVISOR:	Prof Tobias Chirwa
APPROVED BY:	Professor P. Cleaton-Jones, Chairperson, HREC (Medical)
DATE OF APPROVAL:	18/10/2017

This clearance certificate is valid for 5 years from date of approval. Extension may be applied for.

DECLARATION OF INVESTIGATORS

To be completed in duplicate and **ONE COPY** returned to the Research Office Secretary in Room 10004,10th floor, Senate House/3rd floor, Phillip Tobias Building, Parktown, University of the Witwatersrand. I/We fully understand the conditions under which I am/we are authorised to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated, from the research protocol as approved, I/we undertake to resubmit to the Committee. **I agree to submit a yearly progress report**. The date for annual re-certification will be one year after the date of convened meeting where the study was initially reviewed. in this case, the study was initially review September and will therefore be due in the month of September each year. Unreported changes to the application may invalidate the clearance given by the HREC (Medical).

Principal Investigator Signature

Date 20-10-2017

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

Appendix 2.2. Derivation of model likelihood

In order to obtain the marginal log-likelihood of the Gamma frailty model, one can proceed as follows: Let γ be a random variable distributed as Gamma, then its density function can be expressed as

$$g(\gamma) = \frac{\gamma^{\frac{1}{\theta}-1} e^{-\frac{\gamma}{\theta}}}{\Gamma\left(\frac{1}{\theta}\right)^{\frac{1}{\theta}}}, \gamma >, \theta > 0, \qquad (1)$$

Where the expected value, $E(\gamma)$ = 1 and variance, $Var(\gamma) = \theta$. Larger values of θ indicate that there is a higher degree of heterogeneity among participants and higher dependence of within-participant events.

The Gamma frailties can be integrated out in the conditional survival likelihood, and obtain explicit marginal likelihood function containing the parameters of interest only. The marginal likelihood function for the i^{th} participant can be given by;

$$L_i(\psi, \theta, \beta) = \prod_{j=1}^{n_i} \int_0^{\infty} \left(h_0(t)\gamma e^{\beta^t Z_{ij}}\right)^{C_{ij}} e^{-H_0(t)\gamma e^{\beta^t Z_{ij}}} \frac{\gamma^{\frac{1}{\theta}-1} e^{-\frac{\gamma}{\theta}}}{\Gamma\left(\frac{1}{\theta}\right)^{\frac{1}{\theta}}} d\gamma \quad (2)$$

where ψ contains the baseline hazard parameters and C represents censoring indicator. For the Weibull baseline hazard, $\psi = (\eta, \alpha)$,where $\eta > 0$ and $\alpha > 0$ denote shape and scale parameters respectively.

After re-arranging the terms in equation (1), the expression becomes

$$L_i(\psi,\theta,\beta)=\prod_{j=1}^{n_i} h_0(t)^{C_{ij}} e^{\beta^t Z_{ij} C_{ij}} \int_0^{\infty} \frac{\gamma^{\frac{1}{\theta+d_i}-1} e^{-\frac{\gamma}{\theta}} \exp\left(-\sum_{j=1}^{n_i} H_0(t)\gamma e^{\beta^t Z_{ij}}\right)}{\Gamma\left(\frac{1}{\theta}\right)^{\frac{1}{\theta}}} d\gamma$$

$$= \prod_{j=1}^{n_i} h_0(t)^{C_{ij}} e^{\beta^t Z_{ij} C_{ij}} \frac{\Gamma\left(\frac{1}{\theta+d_i}\right)\theta^{\frac{1}{\theta+d_i}}}{\Gamma\left(\frac{1}{\theta}\right)^{\frac{1}{\theta}}} \int_0^{\infty} \frac{\gamma^{\frac{1}{\theta+d_i}-1} \exp\left(-\gamma\left(\frac{1}{\theta}+\sum_{j=1}^{n_i} H_0(t)e^{\beta^t Z_{ij}}\right)\right)}{\Gamma\left(\frac{1}{\theta+d_i}\right)\theta^{\frac{1}{(\theta+d_i)}}} d\gamma,$$

where $d_i = \sum_{j=1}^{n_i} C_{ij}$.

By integrating out the frailty term γ, the expression becomes tractable, the term under the integral has a Gamma distribution with a probability density function $\Gamma\left(\frac{1}{\theta+d_i}, \frac{1}{\theta}\right)$. Thus, deriving the marginal likelihood function can proceed as;

$$L_i(\psi,\theta,\beta) = \prod_{j=1}^{n_i} h_0(t)^{C_{ij}} e^{\beta^t Z_{ij} C_{ij}} \frac{\Gamma\left(\frac{1}{\theta+d_i}\right)\theta^{\frac{1}{\theta+d_i}}}{\Gamma\left(\frac{1}{\theta}\right)\theta^{\frac{1}{\theta}}\theta^{\frac{1}{\theta+d_i}}\left(\frac{1}{\theta}+\sum_{j=1}^{n_i} H_0(t)e^{\beta^t Z_{ij}}\right)^{\left(\frac{1}{\theta+d_i}\right)}} \times$$

$$\int_0^{\infty} \frac{\gamma^{\frac{1}{\theta+d_i}-1} \mathrm{e}^{-\gamma\left(\frac{1}{\theta}+\sum_{j=1}^{n_i} H_0(t)e^{\beta^t Z_{ij}}\right)} \left(\frac{1}{\theta}+\sum_{j=1}^{n_i} H_0(t)e^{\beta^t Z_{ij}}\right)^{\left(\frac{1}{\theta+d_i}\right)}}{\Gamma\left(\frac{1}{\theta+d_i}\right)} d\gamma$$

$$= \prod_{j=1}^{n_i} h_0(t)^{C_{ij}} e^{\beta^t Z_{ij} C_{ij}} \frac{\Gamma\left(\frac{1}{\theta+d_i}\right)}{\Gamma\left(\frac{1}{\theta}\right)\theta^{\frac{1}{\theta}}\left(\frac{1}{\theta}+\sum_{j=1}^{n_i} H_0(t)e^{\beta^t Z_{ij}}\right)^{\left(\frac{1}{\theta+d_i}\right)}} \times$$

$$\int_0^{\infty} \frac{\gamma^{\frac{1}{\theta+d_i}-1} \mathrm{e}^{-\gamma\left(\frac{1}{\theta}+\sum_{j=1}^{n_i} H_0(t)e^{\beta^t Z_{ij}}\right)} \left(\frac{1}{\theta}+\sum_{j=1}^{n_i} H_0(t)e^{\beta^t Z_{ij}}\right)^{\left(\frac{1}{\theta+d_i}\right)}}{\Gamma\left(\frac{1}{\theta+d_i}\right)} d\gamma.$$

The term under the integral is considered to be the pdf of $\Gamma\left(\frac{1}{\theta+d_i}, \frac{1}{\theta}+\sum_{j=1}^{n_i} H_0(t)e^{\beta^t Z_{ij}}\right)$ whose integral reduces to 1. The marginal likelihood function becomes

$$L_i(\psi,\theta,\beta) = \frac{\Gamma\left(\frac{1}{\theta+d_i}\right)\prod_{j=1}^{n_i} h_0(t)^{C_{ij}} e^{\beta^t Z_{ij} C_{ij}}}{\left(\frac{1}{\theta}+\sum_{j=1}^{n_i} H_0(t)e^{\beta^t Z_{ij}}\right)^{\left(\frac{1}{\theta+d_i}\right)} \Gamma\left(\frac{1}{\theta}\right)\theta^{\frac{1}{\theta}}}$$

Applying a logarithm transformation to the above expression followed and summing over the N participants, we obtain the marginal loglikelihood function as,

$$L_i(\psi,\theta,\beta) = \sum_{i=1}^{N}\left(d_i \log(\theta) - log\left(\Gamma\left(\frac{1}{\theta}\right)\right) + log\left(\Gamma\left(\frac{1}{\theta}+d_i\right)\right) \right.$$
$$\left. - \left(\frac{1}{\theta}+d_i\right) log\left(1+\theta\sum_{j=1}^{n_i} H_0(t)e^{\beta^t Z_{ij}}\right) + \sum_{j=1}^{n_i} C_{ij}\left(\beta^t Z_{ij} + \log\left(h_0(t)\right)\right)\right).$$

The maximum likelihood estimators ψ, θ and β are obtained by maximizing the loglikelihood function. For the Weibull distribution, the hazard and cumulative hazard functions can be denoted by $\frac{\eta t^{\eta-1}}{\alpha^{\eta}}$ and $\left(\frac{t}{\alpha}\right)^{\eta}$ respectively, and therefore the marginal loglikelihood function for Gamma frailty with Weibull baseline hazard rate is

$$L_i(\psi,\theta,\beta) = \sum_{i=1}^{N}\left(d_i \log(\theta) - log\left(\Gamma\left(\frac{1}{\theta}\right)\right) + log\left(\Gamma\left(\frac{1}{\theta}+d_i\right)\right) \right.$$
$$\left. -\left(\frac{1}{\theta}+d_i\right) log\left(1+\theta\sum_{j=1}^{n_i}\left(\frac{t}{\alpha}\right)^{\eta} e^{\beta^t Z_{ij}}\right) + \sum_{j=1}^{n_i} C_{ij}\left(\beta^t Z_{ij} + \log\left(\frac{\eta t^{\eta-1}}{\alpha^{\eta}}\right)\right)\right)$$

Appendix 3.1. Original Paper I-Malaria Journal

Stanley et al. Malar J (2019) 18:254
https://doi.org/10.1186/s12936-019-2885-9

Malaria Journal

RESEARCH Open Access

Systematic review of analytical methods applied to longitudinal studies of malaria

Christopher C. Stanley[1,2], Lawrence N. Kazembe[3], Mavuto Mukaka[4,5], Kennedy N. Otwombe[1,6], Andrea G. Buchwald[7], Michael G. Hudgens[8], Don P. Mathanga[2], Miriam K. Laufer[7*] and Tobias F. Chirwa[1]

Abstract

Background: Modelling risk of malaria in longitudinal studies is common, because individuals are at risk for repeated infections over time. Malaria infections result in acquired immunity to clinical malaria disease. Prospective cohorts are an ideal design to relate the historical exposure to infection and development of clinical malaria over time, and analysis methods should consider the longitudinal nature of the data. Models must take into account the acquisition of immunity to disease that increases with each infection and the heterogeneous exposure to bites from infected *Anopheles* mosquitoes. Methods that fail to capture these important factors in malaria risk will not accurately model risk of malaria infection or disease.

Methods: Statistical methods applied to prospective cohort studies of clinical malaria or *Plasmodium falciparum* infection and disease were reviewed to assess trends in usage of the appropriate statistical methods. The study was designed to test the hypothesis that studies often fail to use appropriate statistical methods but that this would improve with the recent increase in accessibility to and expertise in longitudinal data analysis.

Results: Of 197 articles reviewed, the most commonly reported methods included contingency tables which comprised Pearson Chi-square, Fisher exact and McNemar's tests (n = 102, 51.8%), Student's t-tests (n = 82, 41.6%), followed by Cox models (n = 62, 31.5%) and Kaplan–Meier estimators (n = 59, 30.0%). The longitudinal analysis methods generalized estimating equations and mixed-effects models were reported in 41 (20.8%) and 24 (12.2%) articles, respectively, and increased in use over time. A positive trend in choice of more appropriate analytical methods was identified over time.

Conclusions: Despite similar study designs across the reports, the statistical methods varied substantially and often represented overly simplistic models of risk. The results underscore the need for more effort to be channelled towards adopting standardized longitudinal methods to analyse prospective cohort studies of malaria infection and disease.

Keywords: *Plasmodium falciparum*, Longitudinal studies, Longitudinal analysis, Cohort studies

Background

Plasmodium falciparum infection is one of the most common parasitic infection in humans. Individuals in high-transmission settings are exposed to bites of infected *Anopheles* mosquitoes and develop frequent infections. In early childhood, infections are almost always associated with symptomatic disease [1]. Over time, individuals acquire immunity to clinical disease and, to some degree, infection [2, 3]. Thus, each infection alters the host immune response that protects against the next episode. In addition, risk of exposure to infected bites is not uniformly distributed [4, 5]. Thus, risk of infection is not equally distributed across populations. While one infection may decrease the risk of disease or infection, an infection in an individual is also an indicator of more frequent exposure to infected mosquitoes. This provides an interesting challenge to modelling the risk of malaria disease and infection over time. Longitudinal cohorts, unlike other study designs such as

*Correspondence: mlaufer@som.umaryland.edu
[7] Center for Vaccine Development and Global Health, University of Maryland School of Medicine, 685 W. Baltimore St. HSF-1 Room 480, Baltimore, MD 21201, USA
Full list of author information is available at the end of the article

cross-sectional surveys, can differentiate between infections that appear asymptomatic at time of detection but may become symptomatic over time. With efforts intensified to prevent *Plasmodium* infection and reduce clinical malaria disease, prospective longitudinal cohorts are increasingly being conducted to assess risk of infection and disease over time.

Well-designed prospective longitudinal cohort studies can be key to understanding risk of malaria disease for participants over time, but appropriate analysis of the collected data is also critical for accurate results. Like any longitudinal designs, malaria cohorts are characterized by repeated measurements, resulting in possible correlated infection and disease episode data from same participant. The underlying assumption of no correlation among observations made by standard statistical methods such as t-tests, analyses of variance (ANOVA), and linear regression are violated [6–8]. The use of inappropriate methods, particularly those that do not account for possible correlation of events when dealing with repeated episodes of malaria or infection, can lead to biased results. Ignoring correlation of events could result in the confidence intervals for the estimated rates being artificially narrow and the null hypothesis may be rejected more often than warranted [9, 10]. Appropriate longitudinal data analysis techniques should account for such possible within-participant correlation and different covariance structures of episodes of malaria or infection measurements over time. These include generalized estimating equations (GEE) and mixed-effects models. The superiority in consistency and efficiency of these methods over the traditional methods, such as ANOVA, t-tests and simple linear regression were demonstrated previously [11–14]. One of the attractive features of GEE is that the method is robust to mis-specification of the correlation structure, such that the estimator is consistent even if the working correlation structure is mis-specified [15], although it can be inefficient under the mis-specified working correlation structure [16]. When the research interest is on repeated time-to-episodes of disease or infection, it is common to analyse time-to-first episode ignoring the subsequent events, but this approach fails to utilize all information available in the data [13, 17–19]. Analyses of time-to-first event only may be preferred in some situations. For example, when the intention is to evaluate interventions or prognostic factors (e.g., some vaccines) as 'all-or-nothing' determinants of malaria risk, so occurrence of any events is considered a failure. Another scenario would be when the first event is considered to have 'reset' the host such that subsequent events may then be consequently considered in a different category. Recurrent time-to-event models are the appropriate techniques for analysing recurrent malaria episode data because they take into account the within-participant non-independence of episodes. These techniques include frailty models [20, 21] or recurrent Cox-extended models, such as the Andersen–Gill and Prentice–Williams–Peterson models [17].

The statistical models that incorporate all the repeated observations over follow-up are particularly important in malaria since over time, malaria infections result in individuals acquiring immunity to malaria disease. Each infection may introduces protective effect to future disease episodes. Therefore, appropriate analytical methods for repeated episodes of disease and infection should capture the impact of the developed protective immunity.

Recent advancements in statistical methods and computer software have improved the ease of access and application of statistical methods specifically designed for longitudinal studies, allowing one to handle complexity with cohort data [9, 22–25]. Limited information is available about whether the recent developments in computer software and statistical methods for longitudinal data has translated into more wide adoption and application of appropriate methods when analysing prospective longitudinal malaria cohort data. This study tested the hypotheses that choice of statistical methods would change over time and that these changes would reflect more appropriate analytical methods. To address this hypothesis, a systematic review was conducted to investigate the trends of use of appropriate statistical methods among articles analysing prospective longitudinal cohort data for clinical malaria episodes or *Plasmodium* infection published over a 22-year period.

Methods

Search strategy

Titles and abstracts containing the terms "malaria prospective study" or "malaria cohort study" or "malaria longitudinal study" or "malaria years follow-up" or "malaria repeated measurements" indexed in PubMed, Medline or ScienceDirect were included in the search. Original articles in English, which was a familiar language to authors, published from 1996 until December 2017 were reviewed in chronological order.

Inclusion and exclusion criteria

Original research articles were included in the analysis if they were prospective cohort studies where participants were actively or passively followed up for repeated malaria episodes over time and described the following outcomes: clinical malaria based on clinical symptoms and confirmed by rapid diagnostic test (RDT) or microscopy, infection defined as polymerase chain reaction (PCR), microscopy, or RDT results positive for *Plasmodium* parasites. Several publications reporting on the same study or data set were all included. Letters to the editor, editorials, reviews or

systematic reviews, meta-analyses, repeated cross-sectional surveys, case reports and articles written in languages other than English were excluded. Detailed selection process of the studies under review is shown in Fig. 1.

Data extraction

For each article included in the analysis, the following information was extracted: year of publication, duration of follow-up, outcomes, sample size, population, location and type of statistical methods used. Statistical methods were grouped according to categorization by Colditz and Emerson [26, 27] as presented in Appendix. In the cases where researchers studied multiple outcomes, only methods applied to the outcomes of interest-clinical malaria or infection—were considered. Particular interest was on standard longitudinal analysis techniques including GEE, mixed-effects models (random effects models) and repeated measures ANOVA which are considered appropriate for analyses of repeatedly measured data such as repeated episodes malaria disease or infection. For repeated time-to-episodes, recurrent Cox-based models and frailty models are considered appropriate, while other survival methods such as the Kaplan–Meier estimator, log-rank test and Cox model are only appropriate for analyses of time-to-single-episode.

Statistical analysis

Logistic GEE was used to estimate the odds of using the types of statistical methods in the 2007–2017 period compared to 1996–2006 period. Each model adjusted for study sample size and year of publication was included as a panel variable. The outcome variable was binary, indicating whether in each article a particular method was used or not. Models were constructed for each of the following methods: contingency tables, descriptive statistics only, Kaplan–Meier estimator, Cox model, Poisson

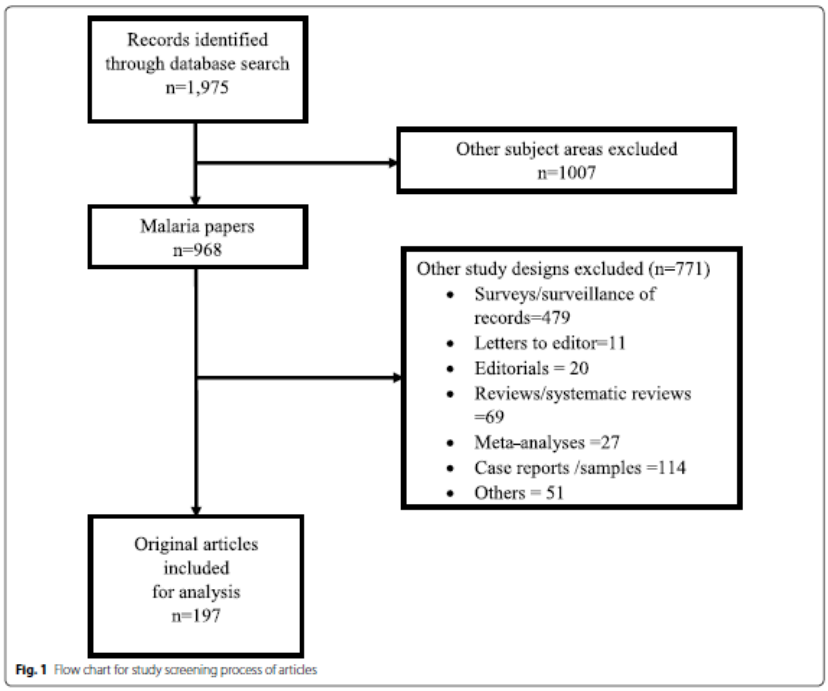


Fig. 1 Flow chart for study screening process of articles

model, GEE and mixed-effects model. All analyses were done using Stata SE version 15.1 (Stata Corp, College Station, TX, USA).

Results

Out of the 1975 abstracts reviewed, 1778 were excluded because they were of other subject areas (n = 1007, 51.0%) and study designs (n = 771, 39.0%). One-hundred and ninety-seven (10.0%) articles met the inclusion criteria and were included in the analyses (Fig. 1). Overall, the median follow-up time was 12 months [interquartile range (IQR): 9–24]. The median sample size per article was 351 (IQR: 206–700). The number of articles increased over the 22-year study period, with the highest number found between 2012 and 2016 (Fig. 2). Overall, there were 128 (65.0%) articles that analysed both clinical malaria episodes and infection as outcomes; 53 (26.9%) assessed clinical malaria episodes only, and 16 (8.1%) infections only. A table listing articles reviewed in this study with details of year of publication, duration of follow-up, outcomes, sample size, population, location, and analysis methods are provided as Additional file 1: Table S1.

Among 197 articles reviewed, the most commonly reported methods included contingency tables comprising Pearson Chi-square, Fisher exact and McNemar's tests (n = 102, 51.8%), followed by Student's t-tests (n = 82, 41.6%), multiple linear regression (n = 71, 36.0%), Cox models (n = 62, 31.5%), and Kaplan–Meier estimators (n = 59, 30.0%) (Table 1). The frequency of usage of different statistical methods changed between the two study periods: 1996–2006 versus 2007–2017. The proportion of articles using contingency tables was higher in the first period than in the second, 65.3 versus 47.3% as was simple linear regression, 10.2 versus 0.7%. Use of the survival analysis methods, Cox models and Kaplan–Meier estimators increased in the second period from 18.4 to 35.8% and 22.5 to 32.4%, respectively. Similarly, usage of mixed-effects models among longitudinal data analysis techniques was higher in the later period compared to the former, 4.1 versus 14.9%.

From the 197 articles included in the review, 131 (66.5%) analysed repeated malaria episodes or infection, of which 60 (45.8%) used appropriate standard longitudinal analysis methods. Appropriate longitudinal analyses included 36 (60.0%) instances of GEE, 19 (31.7%) articles employing mixed-effects models, and 5 (8.3%) articles utilizing both the GEE and mixed-effects models. Seventy-one (54.2%) of the 131 articles used other methods, namely the Poisson (n = 49) and negative binomial regression models (n = 22). These models were based on the

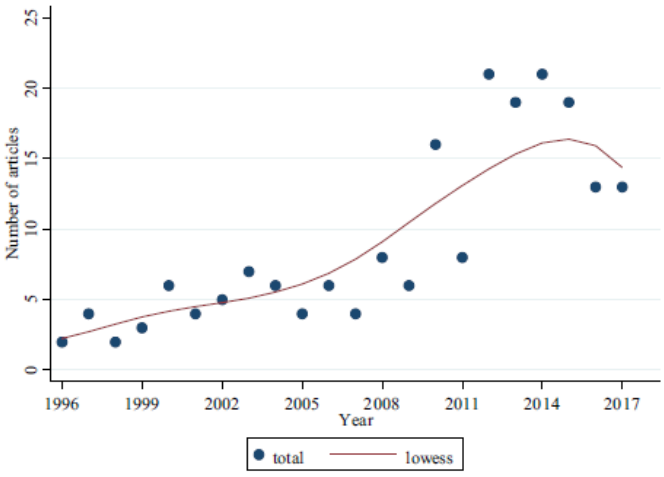


Fig. 2 Total number of articles from 1996 to 2017 with lowess curve. The number of articles published increased over the 22-year study period, with the highest number found between 2012 and 2016

Table 1 Frequency of articles using a particular statistical method

Method	All (n = 197)	1996–2006 (n = 49)	2007–2017 (n = 148)
	N (%)	N (%)	N (%)
Descriptive statistics only	17 (8.6)	5 (10.2)	12 (8.1)
Contingency tables	102 (51.8)	32 (65.3)	70 (47.3)
Multiway tables	50 (25.4)	16 (32.7)	34 (23.0)
Epidemiologic statistics	37 (18.8)	12 (24.5)	25 (16.9)
t-tests	82 (41.6)	23 (46.9)	59 (39.9)
Analysis of variance	18 (9.1)	1 (2.0)	17 (11.5)
Simple linear regression	6 (3.1)	5 (10.2)	1 (0.7)
Non-parametric tests	48 (24.4)	11 (22.5)	37 (25.0)
Non-parametric correlation	21 (10.7)	3 (6.1)	18 (12.2)
Multiple regression	71 (36.0)	15 (30.6)	56 (37.8)
Poisson regression	49 (24.9)	8 (16.3)	41 (27.7)
Negative binomial regression	22 (11.2)	3 (6.1)	19 (12.8)
Survival analysis			
Kaplan–Meier estimator	59 (30.0)	11 (22.5)	48 (32.4)
Log-rank test	36 (18.3)	7 (14.3)	29 (19.6)
Cox model	62 (31.5)	9 (18.4)	53 (35.8)
Other survival methods	4 (2.0)	3 (6.1)	1 (0.7)
Longitudinal analysis			
Generalized estimating equations	41 (20.8)	9 (18.4)	32 (21.6)
Mixed-effects regression	24 (12.2)	2 (4.1)	22 (14.9)
Recurrent analysis			
Shared frailty model	2 (1.0)	–	2 (1.4)
Andersen–Gill (AG) model	4 (2.0)	–	4 (2.7)

generalized linear modelling. Of the 41 studies that modelled the correlation structure, only 5 (12.2%) assumed exchangeable correlation, while 36 (87.8%) did not specify the correlation structure. Out of all 197 articles reviewed, there were 99 (50.3%) studies that assumed Poisson distribution, 4 of these assessed overdispersion, of which 2 adjusted for it. In 9 (4.7%) of all the 197 reviewed studies, authors assumed a normal distribution of count data without indicating if transformations were done.

There were 74 of 197 articles that analysed time-to-first malaria episode or infection ignoring subsequent events and 61 (82.4%) of these employed survival analysis techniques that included Cox models, Kaplan–Meier estimators, and log-rank tests. There were 24 (12.2%) of 197 articles that analysed repeated time-to-episodes and only 6 (25.0%) of these used appropriate recurrent event models with extensions of Cox model, namely Andersen–Gill (n = 4) and frailty models (n = 2) to evaluate risk factors for recurrent clinical malaria. Eighteen (9.1%) articles analysed repeated time-to-malaria or infection episodes using Cox models and Kaplan–Meier estimators which are only valid for single-event analyses because they assume independence of events.

Compared to the first period 1996–2006, the odds of an article using contingency tables adjusting for sample size were lower in the second period 2007–2017 [odds ratio (OR) = 0.49, 95% confidence interval (CI) 0.27–0.91] (Table 2). The proportion of articles using Cox models, Kaplan–Meier estimators and mixed-effects models increased over time (Fig. 3). However, only the Cox models, and mixed-effects models had increased odds of being used in the second period compared to the first period, (OR = 2.57, 95% CI 1.21–5.42) and (OR = 3.85, 95% CI 1.13–6.39), respectively (Table 2).

Discussion

In this review of statistical methods among articles analysing prospective longitudinal cohort data for clinical malaria episodes or *Plasmodium* infection published over a 22-year period, the statistical methods substantially varied across articles despite that they reported analysis of the same general study design and outcome measurements. The most commonly used models for analysing count number of malaria episodes included the Poisson and negative binomial models, but these fall short with repeated measurements data as they cannot account for

Table 2 Odds ratio of using a statistical method during 2007–2017 compared to 1996–2006 adjusted for article sample size

Method	Odds ratio (OR)	95% CI
Descriptive statistics only	0.81	0.26–2.42
Contingency table	0.49	0.27–0.91
Kaplan–Meier estimator	1.80	0.88–3.67
Cox model	2.57	1.21–5.42
Log-rank	1.60	0.66–3.84
Poisson model	1.78	0.85–3.72
GEE	1.16	0.56–2.39
Mixed-effects model	3.85	1.13–6.39

CI confidence interval

within-participant correlation of the episodes [6, 28, 29]. Ignoring this possible correlation of events tends to bias the model results as demonstrated previously [15, 30–32], yielding regression parameters with underestimated standards errors for time-dependent covariates or over-estimated errors in time-dependent covariates [33]. As hypothesized, trends in choice of appropriate methods improved over time, which included the GEE and mixed-effects models as well as recurrent event models, although their usage during the entire study period remained low. Furthermore, the majority of the studies that modelled the correlation structure did not specify the kind of assumed correlation structure, with only exchangeable correlation structure stated in few studies. Among studies that assumed Poisson distribution, the majority of them did not state whether or not overdispersion was checked.

Some articles analysed the repeated time-to-disease episodes or infections but applied the Cox model [34] or the Kaplan–Meier estimator, which assume independent events and are only valid for modelling the time-to-single disease episode or infection. Such analyses may yield biased results. This is particularly true in malaria because over time, individuals acquire immunity to clinical disease and, to some degree, infection. Therefore, each infection may be dependent on the past, and in turn alter the host immune response that protects against the next episode. Additionally, there may be temporal variations in exposure which can alter the risk of clinical malaria disease over time. In some instances, data from the later episodes were discarded leading to loss of information which may limit the resulting conclusions. Modelling that focuses on time-to-first event only may not be generalizable to analyses examining all repeated episodes of disease or infection over entire study period because the risk to subsequent episodes diminishes as a result of the gained immunity over time.

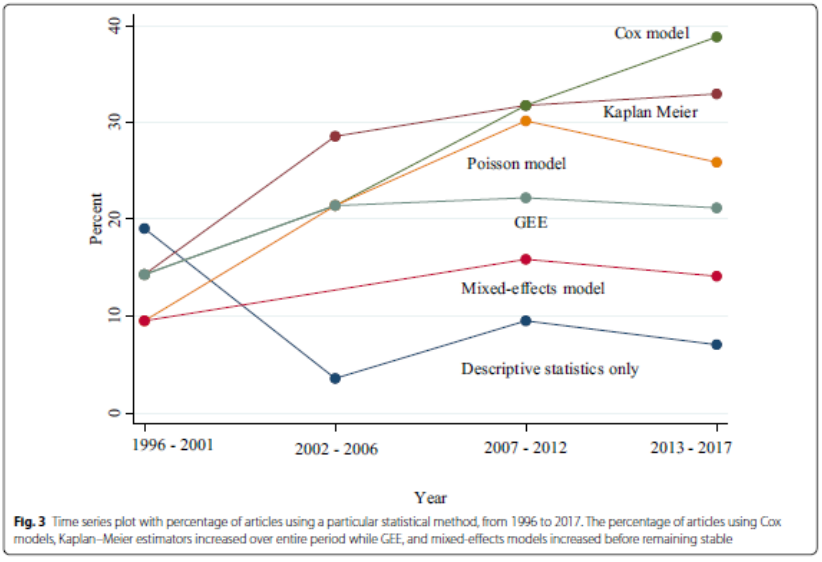


Fig. 3 Time series plot with percentage of articles using a particular statistical method, from 1996 to 2017. The percentage of articles using Cox models, Kaplan–Meier estimators increased over entire period while GEE, and mixed-effects models increased before remaining stable

When analysing repeated times-to-episodes malaria or infection, analytical methods should account for possible within-participant correlation of events as well as censoring over follow-up. Such appropriate methods include recurrent event survival models [35, 36]. In the current review, only six articles analysed repeated time-to-episodes using appropriate recurrent models. Well-developed descriptions of the recurrent event models have been presented by several authors [10, 17, 35, 36] while this section briefly discusses these models in relation to malaria data. The recurrent event models include extensions of the Cox model such as Andersen–Gill (AG) [37], Prentice–Williams–Peterson (PWP) [38] and the frailty model [39]. The AG model uses a counting process time-scale for all episodes assuming that the correlation between episodes-times for each participant is explained by previous episodes where time-scale does not reset to zero after an episode. This model is typically applicable when the interest is on the overall effect of covariates on the intensity of the recurrent episodes. The PWP model analyses ordered events by stratification and can be expressed as gap-time model where the time-scale is reset to zero after each episode occurs. Gap times between malaria parasitaemia detection to clearance is a good example where the PWP model can be valid. Moreover, the PWP model can give both overall and episode-specific effects and so can be more applicable in malaria where covariate effects may be different in subsequent episodes due to the naturally acquired immunity. The shared frailty model extends the Cox model by introducing a random covariate that induces dependence among the recurrent episode times. By conditioning on covariates and the random effect, the recurrent episodes are assumed to become independent. Another challenge when analysing recurrent events in malaria is the issue of exclusion of periods-at-risk after an event, and this may constrain the choice of statistical methods. Depending on kind of treatment if any, recurrent events within a few days of each other are likely to be the result of the same infection or the second event may be considered to be due to treatment failure. Ignoring periods-at-risk after an event is also a potential source of bias. Thus, authors ought to consider exclusion of periods-at-risk after an event during the datasets structuring before analyses are conducted.

This review has demonstrated that publications on prospective longitudinal malaria cohorts are increasing over time, a trend reflecting the growing understanding of the importance of longitudinal cohort data coupled with advancements in computer software developments for analysing such data [9, 22]. Optimal methods for longitudinal data analysis were in their infancy 20 years ago and both access to appropriate software and general acceptable of these methods by the research community have increased dramatically over this period. The increased trend in malaria cohort studies and the improved but slow adoption of appropriate methods underscores the need for standardized analytical methods for such data to avoid drawing erroneous conclusions based on inaccurate results.

This review may have included more than one report using the same data, which may have introduced bias due to duplication of datasets. For example, the same data may have been analysed using inappropriate methods in first instance and then using appropriate techniques. Furthermore, some authors published more than one article and such articles may share similarities in the choice of analytical methods, but accounting for this was beyond the scope of this review as author background information was not collected. Future studies should consider testing whether more appropriate statistical methods improve or change understanding of malaria epidemiology by using among other factors, authors' background information. Lastly, further studies may explore and compare how results from this study would compare with trends specifically from field trials.

Conclusions

Cohort studies assessing risk of malaria infection and disease over time did not employ consistent analytic methods. The statistical methods varied substantially and often represented overly simplistic models despite similar designs, which may have led to biased outcomes and results that cannot be compared across studies. This review suggests that recent studies are adopting more appropriate statistical methods, though the statistical analyses are not uniform. These results underscore the need for more effort to be channelled towards adopting standardized longitudinal methods to model recurrent events for malaria infection and disease.

Additional file

Additional file 1: Table S1. List of reviewed articles in the study with details of year of publication, duration of follow-up, outcomes, sample size, population, location and analysis methods.

Abbreviations

AIDS: acquired immune deficiency syndrome; ANOVA: analysis of variance; GEE: generalised estimating equations; HIV: human immunodeficiency virus; PCR: polymerase chain reaction; RDT: rapid diagnostic test.

Acknowledgements

Not applicable.

Authors' contributions

CCS and TFC conceptualized the study and collected, analysed and interpreted the data; CCS, MKL and TFC led the writing of the manuscript; LKN, MM, KNO, AGB, MGH and DPM contributed to the study design, data interpretation and the writing of the manuscript; All authors read and approved the final manuscript.

Funding
This work was supported by the Fogarty International Center of the National Institutes of Health [Grant number D43TW010075]; and The Wellcome Trust [Grant number 107754/Z/15/Z]; National Institute of Allergy and Infectious Diseases [Grant number K24AI114996]. Funders have no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials
All data generated and analysed during this study are included in additional files of this published article

Ethics approval and consent to participate
Not applicable.

Consent for publication
All authors reviewed and approved this manuscript for publication.

Competing interests
The authors declare that they have no competing interests.

Author details
[1] School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa. [2] Malaria Alert Centre, University of Malawi College of Medicine, Blantyre, Malawi. [3] Department of Statistics, University of Namibia, Windhoek, Namibia. [4] Oxford Centre for Tropical Medicine and Global Health, Oxford, UK. [5] Mahidol-Oxford Tropical Medicine Research Unit, Bangkok, Thailand. [6] Perinatal HIV Research Unit, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa. [7] Center for Vaccine Development and Global Health, University of Maryland School of Medicine, 685 W. Baltimore St. HSF-1 Room 480, Baltimore, MD 21201, USA. [8] Department of Biostatistics, Center for AIDS Research, University of North Carolina Chapel Hill, Chapel Hill, NC, USA.

Appendix

See Table 3.

Table 3 Categories of statistical methods used to assess the statistical content of articles

Category	Brief description	Include
No statistical methods or descriptive statistics only	Describe basic features of data to provide simple measures of summaries	No statistical content, or descriptive statistics only e.g., percentages, means, standard deviations, standard errors, histograms
Contingency tables	Cross tabulations used to summarize the relationship between categorical data	Chi-square test, Fisher's exact test, McNemar's test
Epidemiologic statistics	Measures of association for outcome of interest such as disease and some exposure(s)	Relative risk, risk ratios, rate ratios, risk difference, rate difference, odds ratio, log odds, risk difference, attributable risk fraction, sensitivity, specificity
Multiway tables	Extend two-way relationships to include three or more variables	Mantel–Haenszel procedure, log-linear models, logistic regression
t-test	Assess mean differences between groups	One-sample, matched-pair, two-sample t-tests
Pearson's correlation	Measures linear correlation between two variables	Classical product-moment correlation
Simple linear regression	Regression that summarizes relationships between two continuous variables, an explanatory and a response	Least-squares regression with one predictor and one response variable
Multiple regression	Extends the simple regression to include two or more explanatory variables for a response	Polynomial regression and stepwise regression
Analysis of variance	Assess within and between group differences in means	Analysis of variance, Analysis of covariance, simple linear contrasts, F-tests
Multiple comparisons	Methods for handling multiple inferences on same data sets	Bonferroni techniques, Scheffé's contrasts, Duncan multiple-range methods, Newman–Keuls procedure
Non-parametric test	Tests used when data is not assumed to follow a particular distribution, and are based on ranks of data	Sign test, Wilcoxon signed-rank test, Mann–Whitney test, Kruskal–Wallis test, Friedman test, Kolmogorov–Smirnov test
Non-parametric correlation	Measure strength and direction of association between two variables	Spearman's rho, Kendall's tau, monotone regression, test for trend
Survival analysis	Methods where outcome variable is the time until the occurrence of an event	Actuarial life table, Kaplan–Meier estimator for survival, survival function, Cox model, other parametric survival models, rate adjustment, log-rank test, Breslow's test
Sensitivity analysis	Examines sensitivity of outcome to small changes in parameters of model or in other assumptions	Sample size, multiple outcomes, model distribution assumptions
Transformation	Use of data transformation often in regression	Natural logarithm, square, cubic
Cluster analysis	Involves dividing a multivariate dataset into "natural" clusters (groups) for in-depth assessment	Hierarchical, K-means, two-step clustering
Repeated-measures analysis	Approaches that account for correlation for within-participant observations and non-constant variance in response over time	Generalized estimating equations (GEE), mixed-effects models, repeated measures ANOVA
Other	Other methods not specified in above	Receiver-operating characteristic, principal component analysis, power analysis, propensity score

Received: 18 April 2019 Accepted: 19 July 2019
Published online: 29 July 2019

References

1. Aponte JJ, Menendez C, Schellenberg D, Kahigwa E, Mshinda H, Vountasou P, et al. Age interactions in the development of naturally acquired immunity to *Plasmodium falciparum* and its clinical presentation. PLoS Med. 2007;4:e242.
2. Doolan DL, Dobaño C, Baird JK. Acquired immunity to malaria. Clin Microbiol Rev. 2009;22:13–36.
3. Biswas S, Seth RK, Tyagi PK, Sharma SK, Dash AP. Naturally acquired immunity and reduced susceptibility to falciparum malaria in two subpopulations of endemic Eastern India. Scand J Immunol. 2008;67:177–84.
4. Mlili MP, Sikulu-Lord MT, Govella NJ. Bites before and after bedtime can carry a high risk of human malaria infection. Malar J. 2017;16:91.
5. Rossati A, Bargiacchi O, Kroumova V, Zaramella M, Caputo A, Garavelli PL. Climate, environment and transmission of malaria. Infez Med. 2016;24:93–104.
6. Liang K, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986;73:13–22.
7. Lin DY. Cox regression analysis of multivariate failure time data: the marginal approach. Stat Med. 1994;13:2233–47.
8. Crowder M. On the use of a working correlation matrix in using generalised linear models for repeated measures. Biometrika. 1995;82:407.
9. Gibbons RD, Hedeker D, DuToit S. Advances in analysis of longitudinal data. Annu Rev Clin. 2010;6:79–107.
10. Amorim LDAF, Cai J. Modelling recurrent events: a tutorial for analysis in epidemiology. Int J Epidemiol. 2015;44:324–33.
11. Burton P, Gurrin L, Sly P, Gurrin L. Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling. Stat Med. 1998;17:1261–91.
12. Zeger S, Liang K-Y. Longitudinal data analysis for discrete and continuous outcomes. Biometrics. 1986;42:121–30.
13. Ballinger GA. Using generalized estimating equations for longitudinal data analysis. Organ Res Methods. 2004;7:127–50.
14. Hu XJ. Applied longitudinal data analysis for epidemiology: a practical guide. J Am Stat Assoc. 2004;99:1207–8.
15. Hubbard AE, Ahern J, Fleischer NL, Van Der Laan M, Lippman SA, Jewell N, et al. To GEE or not to GEE: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. Epidemiology. 2010;21:467–74.
16. Qu A. Improving generalised estimating equations using quadratic inference functions. Biometrika. 2000;87:823–36.
17. Sagara I, Giorgi R, Doumbo OK, Planroux R, Gaudart J. Modelling recurrent events: comparison of statistical models with continuous and discontinuous risk intervals on recurrent malaria episodes data. Malar J. 2014;13:293.
18. Shen CW, Chen YH. Model selection for generalized estimating equations accommodating dropout missingness. Biometrics. 2012;68:1046–54.
19. Heagerty PJ. Marginalized transition models and likelihood inference for longitudinal categorical data. Biometrics. 2002;58:342–51.
20. Hougaard P. Frailty models for survival data. Lifetime Data Anal. 1995;1:255–73.
21. Box-Steffensmeier JM, De Boef S. Repeated events survival models: the conditional frailty model. Stat Med. 2006;25:3518–33.
22. Wang M. Generalized estimating equations in longitudinal data analysis: a review and recent developments. Adv Stat. 2014. https://doi.org/10.1155/2014/303728.
23. Halekoh U, Højsgaard S, Yan J. The R Package geepack for generalized estimating equations. J Stat Softw. 2006;15:1–11.
24. Bernal-Rusiel JL, Greve DN, Reuter M, Fischl B, Sabuncu MR. Statistical analysis of longitudinal neuroimage data with linear mixed effects models. Neuroimage. 2013;66:249–60.
25. StataCorp. Stata Statistical Software: Release 15. College Station: StataCorp LP; 2017.
26. Emerson JD, Colditz GA. Use of statistical analysis in the New England Journal of Medicine. N Engl J Med. 1983;309:709–13.
27. Colditz GA, Emerson JD. The statistical content of published medical research: some implications for biomedical education. Med Educ. 1985;19:248–55.
28. Bates DM. Models for longitudinal data. lme4 mix model with R. 2010:63–82.
29. Chan D. Data analysis and modeling longitudinal processes. Gr Organ Manag. 2003;28:341–65.
30. Gardiner JC, Luo Z, Roman LA. Fixed effects, random effects and GEE: what are the differences? Stat Med. 2009;28:221–39.
31. Pardo MC, Alonso R. GEEs for repeated categorical responses based on generalized residuals. J Stat Comput Simul. 2014;84:344–59.
32. Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics. 1982;38:963–74.
33. Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, et al. Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol Evol. 2009;24:127–35.
34. Cox D. Regression models and life tables. J R Stat Soc B. 1972;34:182–220.
35. Clayton D. Some approaches to the analysis of recurrent event data. Stat Methods Med Res. 1994;3:244–62.
36. Kelly PJ, Lim LLY. Survival analysis for recurrent event data: an application to childhood infectious diseases. Stat Med. 2000;19:13–33.
37. Andersen PK, Gill RD. Cox's regression model for counting processes: a large sample study. Ann Stat. 1982;10:1100–20.
38. Prentice RL, Williams BJ, Peterson AV. On the regression analysis of multivariate failure time data. Biometrika. 1981;68:373–9.
39. Therneau TM, Grambsch PM. Modeling survival data: extending the Cox model. Statistics for biology and health. 2nd ed. New York: Springer; 2000.

Publisher's Note

Appendix 4.1. Original Paper II-Journal of Medical Statistics and Informatics

Journal of Medical Statistics and Informatics
ISSN 2053-7662 | Volume 7 | Article 1

HOAJ
Herbert Open Access Journals

Original Research Open Access

Joint modelling of time-to-clinical malaria and parasite count in a cohort in an endemic area

Christopher C. Stanley[1,2*], Lawrence N. Kazembe[3], Andrea G. Buchwald[4], Mavuto Mukaka[5,6], Don P. Mathanga[2], Michael G. Hudgens[7], Miriam K. Laufer[4] and Tobias F. Chirwa[1]

*Correspondence: stanleychristopher1@yahoo.com

[1]School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa.
[2]Malaria Alert Center, University of Malawi College of Medicine, Blantyre, Malawi.
[3]Department of Statistics, University of Namibia, Windhoek, Namibia.
[4]Center for Vaccine Development and Global Health, University of Maryland School of Medicine, Baltimore, USA.
[5]Oxford Centre for Tropical Medicine and Global Health, Oxford, United Kingdom.
[6]Mahidol-Oxford Tropical Medicine Research Unit, Bangkok, Thailand.
[7]Department of Biostatistics, Center for AIDS Research (CFAR), University of North Carolina Chapel Hill, North Carolina, USA.

Abstract

Background: In malaria endemic areas such as sub-Saharan Africa, repeated exposure to malaria results in acquired immunity to clinical disease but not infection. In prospective studies, time-to-clinical malaria and longitudinal parasite count trajectory are often analysed separately which may result in inefficient estimates since these two processes can be associated. Including parasite count as a time-dependent covariate in a model of time-to-clinical malaria episode may also be inaccurate because while clinical malaria disease frequently leads to treatment which may instantly affect the level of parasite count, standard time-to-event models require that time-dependent covariates be external to the event process. We investigated whether jointly modelling time-to-clinical malaria disease and longitudinal parasite count improves precision in risk factor estimates and assessed the strength of association between the hazard of clinical malaria and parasite count.

Methods: Using a cohort data of participants enrolled with uncomplicated malaria in Malawi, a conventional Cox Proportional Hazards (PH) model of time-to-first clinical malaria episode with time-dependent parasite count was compared with three competing joint models. The joint models had different association structures linking a quasi-Poisson mixed-effects of parasite count and event-time Cox PH sub-models.

Results: There were 120 participants of whom 115 (95.8%) had >1 follow-up visit and 100 (87.5%) experienced the episode. Adults >15 years being reference, log hazard ratio for children <5 years was 0.74 (95% CI: 0.17, 1.26) in the joint model with best fit vs. 0.62 (95% CI: 0.04, 1.18) from the conventional Cox PH model. The log hazard ratio for the 5-15 years was 0.72 (95% CI: 0.22, 1.22) in the joint model vs.0.63 (95% CI: 0.11, 1.17) in the Cox PH model. The area under parasite count trajectory was strongly associated with the risk of clinical malaria, with a unit increase corresponding to -0.0012 (95% CI: -0.0021, -0.0004) decrease in log hazard ratio.

Conclusion: Jointly modelling longitudinal parasite count and time-to-clinical malaria disease improves precision in log hazard ratio estimates compared to conventional time-dependent Cox PH model. The improved precision of joint modelling may improve study efficiency and allow for design of clinical trials with relatively lower sample sizes with increased power.

Keywords: Prospective studies, longitudinal data, malaria parasite, time-to-event, clinical malaria, Cox proportional hazards model, joint modelling

Introduction

Malaria remains one of the most common parasitic infections globally with a disproportionately high burden in sub-Saharan Africa [1]. In malaria-endemic areas, repeated exposure results in acquired immunity to clinical malaria disease but not infection. Of interest in many malaria studies is to estimate time-to-clinical malaria but repeated exposure may give rise to a relationship between the disease and time-dependent covariates such as parasite count. However, time-to-clinical malaria and parasite count data are often analysed separately, mostly using Cox proportional hazards (PH) models and mixed-effects models or generalised estimating equations (GEE) respectively [2-4]. Separate analysis of time-to-clinical malaria disease and parasite count data may result in inefficient estimates when these two processes are strongly associated [5].

For accurate estimation of the risk of clinical malaria, analytical methods are required that account for historical exposure and the relationship between clinical malaria and infection parasite count. To investigate the strength of this association, one approach would be including the parasite countas a time-dependent covariate in the model of time-to-clinical malaria episode. However, this approach may be inaccurate because it does not account for the fact that parasite count in this case is an endogenous covariate whose existence and future path can be directly related to the occurrence of the episodes [6]. Standard time-to-event models require that time-dependent covariates be external to the event process [7] but clinical malaria disease frequently leads to treatment which may instantly affect the level of parasite count.

A second approach is to fit a joint model of time-to-clinical malaria and longitudinal parasite count profile. Previous applications of the joint modelling framework are many. These include, for example, analysis of CD4 count jointly with time-to-development of AIDS [8-12] and modelling quality of life performance scores jointly with time-to-death or disease progression among patients with cancer [11,13,14]. In order to fit joint models to data from malaria studies, there are certain aspects of these types of data that require consideration. In particular, following treatment it is possible for the parasite count to equal zero, so the joint model should allow for that.

We investigated whether modelling time-to-new clinical malaria disease jointly with parasite count trajectory data may improve precision in log hazard ratio estimates when compared to the conventional Cox PH model with time-dependent parasite count. We also assessed the strength of the association between the hazard of clinical malaria and a time-varying parasite count.

Methodology

Data source

The study was motivated by data from the Mfera prospective cohort study conducted in Chikwawa district, southern Malawi, described previously by Buchwald and others [15]. Malaria disease is endemic in Malawi [16] and transmission of the *Plasmodium falciparum* parasite is high in Chikwawa [17,18]. The cohort enrolled 120 participants who presented with uncomplicated malaria at the Mfera health centre in the Chikwawa district between June 2014 and March 2015. Initial diagnosis was made by rapid diagnostic test (RDT) and confirmed by microscopy using thick blood smears. Exclusion criteria from the Mfera cohort included: acute illness requiring hospitalization, signs or symptoms of severe malaria or moderate to severe anemia, and chronic medication with any drug that has antimalarial activity e.g. HIV treatment. Participants underwent passive and active surveillance on a monthly basis and whenever sick for up to two years to assess re-infection, host response and parasite count.

Primary outcome

In the current analyses, the outcome of interest was time-to-first new clinical malaria disease which was defined by participants' self-reported fever and apositive RDT result.

Notation and specification of the models

The conventional Cox PH model with time-dependent parasite count is defined as in [9] with the hazard function $\lambda_i(t)$ for participant at given time expressed as.

$$\lambda_i(t) = \lambda_0(t)\exp\left[\beta_s X_{si}^{'}(t)\right], \quad (1)$$

where and $\lambda_0(t)$ is the unspecified baseline hazard function and the covariate vector $X_{si}(t)$ for participant includes participant's age and frequency of insecticide treated bed nets use in the previous month of the visit. Covariates were included in multivariable models if they were significant at alpha level of 0.1 in univariate Cox PH models.

For the joint models, we utilised the Bayesian joint modelling approach for longitudinal and survival data proposed by Chen et al [12] which fits a model with Markov Chain Monte Carlo (MCMC) methods as presented by Ibrahim et al [19]. The joint model is composed of longitudinal and survival sub-models. The longitudinal sub-model takes the form of a mixed-effects model as follows; supposing data is available from N participants with n_i observations recorded for participant i, $(i=1,.....,N)$. The response y_{ij} $(j=1,.....,n_i)$, fixed-effect covariate vector $X_{ij}=(X_{1ij},.....,X_{pij})'$, and random-effect covariate vector $Z_{ij}=(Z_{1ij},.....,Z_{qij})'$ are recorded at times t_{ij}. The longitudinal sub-model is

$$y_{ij} = \beta X_{ij}^{'} + b_i Z_{ij}^{'} + \varepsilon_{ij}, \quad (2)$$

where β is the $p\times1$ fixed-effect parameter vector, b_i is the $q\times1$ vector of random effects for participant which is assumed to be multivariate normal with mean zero, i.e., $b_i \sim N_q(0,\Sigma_b)$, and Σ_b is the variance-covariance matrix of the subject specific effects. The error vector $\varepsilon_i=(\varepsilon_{i1},..,\varepsilon_{ni})'$ is assumed to be distributed $\varepsilon_i \sim N_n^i(0,\delta^2 I_{ni})$ where δ^2 is variance and I_{ni} is the $n_i\times n_i$ identity matrix.

The survival sub-model takes a Cox PH model form [9] where

doi: 10.7243/2053-7662-7-1

the hazard function for $\lambda_i(t)$ participant *i* at time *t* as given in equation (1) is modelled as

$$\lambda_i(t) = \lambda_0(t)\exp\left[\theta h(\beta, b_i, t) + \beta_s X_{si}^{'}(t)\right], \quad (3)$$

where $h(\beta,b_i,t)$ is a function of the fixed and random effects in the longitudinal sub-model, and θ is an association parameter linking the two sub-models: survival and longitudinal models. The β_s is a parameter vector for covariates unique to the survival sub-model. The survival covariate vector $X_{si}(t)=(x_{si1},...,x_{sir})'$ may include baseline covariates for participant *i* with β_s representing $r\times 1$ parameter vector. The functional form of $h(\beta,b_i,t)$ determines the type of the association structure between longitudinal parasite count and the time-to new clinical malaria episode. Taking the first derivative of $h(\beta,b_i,t)$ is interpreted in terms of an association of the rate of change in parasite count at time and the hazard of new clinical malaria episode at the same time, while the integral of $h(\beta,b_i,t)$ would relate the hazard and the cumulative parasite count trajectory defined as area under parasite count profile from baseline up to the time of the new episode.

Once the two sub-models are specified, the likelihood for the joint model can be constructed as follows. Let T_i and C_i represent potential failure and censoring times respectively for participant *i*. Let $S_i=min\{T_i,C_i\}$ be the minimum of the observed failure and censoring times for participant *i*, and let τ_i be the failure indicator taking value 1 if $T_i \leq C_i$ and 0 otherwise. Then the value of the longitudinal trajectory for participant *i* at time *t* can be defined as $\varphi_i(\beta,b_i,t)$ and at visit *j* as $\varphi_{ij}(\beta,b_i)$. Using the full longitudinal trajectory, then the likelihood of the joint distribution of the observed data and random effects for participant can be decomposed as

$$\begin{aligned} L_i &\propto f_i(Survival \mid longitudinal)\times f_i(longitudinal) \\ &= f(S_i \mid \theta,\tau_i,\beta_s,\varphi_i(\beta,b_i,t),X_{si})\times f(y_i \mid x_i,z_i,\beta,b_i) f(b_i) \end{aligned} \quad (4)$$

and the joint likelihood for all participants can be written as $L=\Pi_{i=1}^{N} L_{i=1}$.

Denoting parasite count as PC, the likelihood of the joint distribution of observed data, random effects and PC can be expressed as

$$L = f\left(S \mid \theta,\tau,\beta_s,\varphi(\beta,b,t,PC)\right)\times f\left(b,PC \mid \beta\right) \quad (5)$$

and integrating out the random effects of the conditional likelihood yields the marginal likelihood. Under the Bayesian framework, the random effects are sampled in MCMC algorithm as extra parameters. The survival and longitudinal sub-models are linked by sharing common random effects structure. The MCMC computations of the model parameters proceed assuming that given the random effects, the longitudinal parasite countand time-to-clinical malaria process are independent as are the longitudinal responses of each participant.

Data analysis

Baseline data was summarised using frequencies and percentages if categorical and medians with ranges were presented if continuous and skewed. Failure functions for time to new clinical malaria disease were summarised using Kaplan Meier plots, and compared across age and bed net usage using log-rank test. Four models of time-to-new clinical malaria disease were fitted. These included the reference model (model 1), a conventional Cox PH model fitted with time-dependent parasite count measured at each visit, and three competing joint models. The three joint models take a common formulation only differing in the assumed association structure linking the hazard of clinical malaria with parasite count. The hazard of clinical malaria disease at any time was linked with: 1) the current underlying value of parasite count at the same time point (model 2), 2) the rate of change in parasite count trajectory at time (model 3), and 3) the cumulative trajectory or area under profile of parasite count from baseline up to time (model 4). Each joint model involved fitting a quasi-Poisson mixed-effects sub-model of parasite count longitudinal trajectory with natural spline functions of time and including the resulting estimates as covariates in the Cox PH sub-model. Other covariates in the conventional Cox PH and survival sub-models included age of the participant and frequency of insecticide treated bed net use in previous month. Parameters were estimated using MCMC through the random walk Metropolis–Hastings (M-H) algorithm. Diffused normal priors were assumed for the covariates including the association parameter. Diagnostic assessments were conducted to assess the convergence of the MCMC samples using trace and kernel density estimator plots, for the final optimal model. Analyses were done in Stata SE version 15.1 (Stata Corp., College Station, TX) [20] using programme *bayesmh* and R version 3.4.3 using packages *JMbayes, survival* and *glmmPQL* [21].

Results

There were 120 participants in the cohort, of which 69 (57.5%) were females. The overall median age was7.5 years [inter-quartile range (IQR): 4.7-18.1], 6.3 years (IQR: 3.2-13.1) for males and 9.2 years (IQR: 5.3-18.5) females (**Table 1**). The median number of malarial parasites per µL was 11,060 (IQR: 840-54,000) overall, 24,840 (IQR: 1,600-68,600) in males, and 5,640 (IQR: 520- 540, 000) for females. During enrolment, 48 (44.9%) out of 107 participants reported to have been using bed nets every night in previous month.

Follow up and time-to-first new clinical malaria disease

Analyses of time-to-new clinical malaria disease included 115 participants who had a least one follow-up visit post enrolment and together contributed a total 894 observations. The median follow-up time to new clinical malaria disease episode was 3.5 months (IQR: 1.1-7.9). Out of the 115 participants, 100 (87.5%) experienced the episode while 15 (12.5%) were administratively right censored (**Table 2**). Among 100 participants

Table 1. Baseline demographic, clinical conditions and vital signs for Mfera malaria cohort in Malawi.

Variable	Total (n=120)
Gender, female, n (%)	69 (57.4)
Age, n (%)	
< 5 years	34 (28.3)
5-15 years	51 (42.5)
>15 years	35 (29.2)
Weight (kg), median (IQR)	21.5 (15.0 - 46.0)
Height (cm), median (IQR)	119.5 (103.0 - 151.8)
Temperature (°C), median (IQR)	36.7 (36.2 - 38.6)
Respiratory rate (breaths/minute), median (IQR)	28 (22 - 36)
Heart rate (beats/minute), median (IQR)	112 (92 - 139)
Haemoglobin (g/dl), median (IQR)	11.5 (10.2 - 12.4)
Parasite count (number of parasites/μL), median (IQR)	11060 (840 - 54000)
Cough, n (%)	15 (12.5)
Musculoskeletal pain, n (%)	40 (33.3)
Headache, n (%)	36 (30.0)
Vomiting, n (%)	32 (26.7)
Abdominal pain, n (%)	15 (12.5)
Bed net use previous month*, n(%)	
Every night	48 (44.9)
Most nights (> half)	13 (12.1)
Some nights (< half)	8 (7.5)
No nights	38 (35.5)
Season enrolled, n(%)	
Dry: May - November	91 (75.8)
Rainy: December - April	29 (24.2)

*not adding up to column total due to missing

Table 2. Characteristics by clinical malaria statusfor Mfera malaria cohort in Malawi.

Variable	Clinical malaria episode (n=100)	No clinical malaria episode (n=15)
Sex, n (%)		
Male	42 (42.0)	8 (53.3)
Female	58 (58.0)	7 (47.7)
Age, n (%)		
<5 years	27 (27.0)	4 (26.7)
5-15 years	48 (48.0)	2 (13.3)
>15 years	25 (25.0)	9 (60.0)
Net use previous month, n(%)		
Every night	37 (37.0)	8 (53.3)
Not every night	63 (63.0)	7 (46.7)
Haemoglobin (g/dl)	11.4 (10.0 - 12.3)	12.3 (10.9 - 13.9)
Parasite count, number of parasites per μL, median (IQR)	13640 (840 - 52040)	2800 (560 - 60040)

Note: clinical malaria status data available for 115 participants who had a least one follow-up visit.

who experienced the episode, 58 (58%) were females, 48 (48%) were aged 5-15 years, 37 (37%) reported using bed net nightly in prior month to enrolment and their median parasite count was 13640 (IQR: 840 - 52040).

Parameter estimation

Regression coefficient estimates are log hazard ratios, log (HRs). Overall, the joint models gave larger log (HR) estimates with consistently smaller standard errors and narrower credible intervals when compared to the conventional Cox PH model with time-dependent parasite count (**Table 3**). Comparing the Deviance Information Criteria (DIC) from the three joint models showed that joint model 4 with cumulative parasite count had the lowest value (**Table 4**), suggesting the best fit [**22**]. For the rest of the manuscript, the joint model is referencing to the joint model 4 with cumulative parasite count which is being compared to the conventional Cox PH model (model 1). Considering adults above 15 years as reference group, the log(HR) of clinical malaria disease for children under 5 years,was 0.74 [95% Credible Interval (CI): 0.17, 1.26] in the joint model compared to 0.62(95% CI: 0.04, 1.18) from the conventional Cox PH model. The log (HR) for participants aged 5-15 years was 0.72 (95% CI: 0.22, 1.22) in the joint model compared to 0.63(95% CI: 0.11, 1.17) in the conventional Cox PH model. Considering participants who used a bed net nightly in the previous month as reference, the log (HR) of clinical malaria disease for those who did not use a bed net every night was 0.58 (95% CI: 0.13, 1.07) from the joint model compared to 0.52 (95% CI: 0.07, 1.08) in the conventional Cox PH model. From the joint model, the area under the longitudinal trajectory of parasite count was strongly associated with the risk of clinical malaria, with a unit increase corresponding to a -0.0012 (95% CI: -0.0021, -0.0004) decrease in the log hazard ratio. Neither current underlying value of parasite count nor rate of change in parasite count trajectory at the same time point were significantly associated with the occurrence of a new clinical malaria episode.

MCMC convergence of estimated log (HR) parameters from final optimal joint model

The MCMC sampler trace plots from the final joint model4 seemed to mix well for parameters of children under 5 years, 5-15 years and infrequent bed net use and never moved beyond 1.5 achieving convergence (**Figure 2A-2C**). The sampler for cumulative parasite count parameter, showed noisy pattern before converging at about 1000 iterations (**Figure 2D**). From the Kernel density estimator plots, all the four parameters were roughly normal suggesting that the M-H algorithm sampled within the target normal distribution (**Figure 3A-3D**).

Factors associated with risk of new clinical malaria episode

In unadjusted analyses using Kaplan Meier failure estimator, the risk of experiencing new clinical malaria episode was higher in children under 5 years and the 5-15 years compared

doi: 10.7243/2053-7662-7-1

Table 3. Log of hazard ratio estimates for time-to-new clinical malaria episode for Mfera cohort in Malawi.

		Age at baseline*		Bed net use in previous month+	Parasite count
Model	Parameter	$\theta_{<5\ years}$	$\theta_{5-15\ years}$	$\theta_{not\ every\ night}$	$\theta_{parasite\ count}$
Model 1	Estimate	0.624	0.632	0.520	-1.77e-5
	Std error	0.282	0.259	0.292	1.86e-6
	95% CI	0.044, 1.183	0.115, 1.174	0.076, 1.085	-1.41e-5, 2.14e-5
Model 2	Estimate	0.672	0.707	0.582	-0.029
	Std error	0.016	0.014	0.024	0.053
	95% CI	0.094, 1.251	0.213, 1.265	0.110, 1.017	-0.345, 0.411
Model 3	Estimate	0.749	0.775	0.594	-0.103
	Std error	0.023	0.021	0.023	0.071
	95% CI	0.077, 1.485	0.193, 1.253	0.159, 1.156	-6.168, 6.124
Model 4	Estimate	0.735	0.722	0.575	-0.001
	Std error	0.022	0.028	0.030	0.0001
	95% CI	0.170, 1.263	0.219, 1.215	0.125, 1.066	-0.0021, -0.0004

Log hazard ratio estimates are posterior means. All models are multivariable. CI= Credible Interval. *reference: >15 years, +reference: bed net use nightly.

Table 4. Deviance Information Criteria (DIC) from the competing joint models.

Joint Model	DIC
Model 2: Joint model with current underlying value of parasite count at any time	74763360
Model 3: Joint model with rate of change in parasite count trajectory at any time	74763436
Model 4: Joint model with cumulative parasite count from baseline up to any time	74763318

to adults above 15 years (log-rank p-value=0.016) (**Figure 1A**). The risk of getting a new clinical malaria episode was also higher in participants who did not use bed net every night compared to those who used a bed net nightly in previous month (log-rank p-value=0.016) (**Figure 1B**). Based on the joint model, conditional on cumulative parasite count profile, the hazard of getting new clinical malaria episode was higher in children under 5 years by 2.1-fold (95% CI: 1.2, 3.4) and those aged 5-15 years by 2.0-fold (95% CI: 1.2, 3.5) when compared to adults over 15 years old and controlling for bed net usage. Among participants of the same age, the conditional hazard was also higher for participants who did not use a bed net every night by 1.8-fold (95% CI: 1.1,3.3) compared to those who used the bed net nightly in previous month.

Discussion

This study has demonstrated that jointly modelling longitudinal parasite count and time-to clinical malaria episodes improves precision in risk factor estimates associated with clinical malaria disease. The joint model yielded larger parameter estimates with consistently smaller standard errors and narrower credible intervals when compared to a conventional Cox PH model with time-dependent parasite count. These results are consistent with findings from other areas including HIV [8,12] and cancer [13,14,23] where joint modelling out-performs separate analyses in terms of optimal use of the available information giving both more precise and less biased estimates. The improved precision provided by the joint model may improve study efficiency, for example, clinical trials may be designed with relatively lower sample sizes while still yielding high power. In general, the conventional time-dependent Cox PH model like any other standard time-to-event model assume that time-dependent covariates are external to the event process [7]. However, parasite count in this case is an endogenous covariate whose existence and future path can be directly related to the occurrence of the malaria episodes. By postulating a model for the joint distribution of the covariate parasite count and the time-to-malaria processes, the joint model explicitly accounts for possible inter-dependence between the two processes through shared random effects [24]. Moreover, the time-dependent covariate model assumes that the value of the parasite count does not change until a new measurement is taken which may not be correct. When we modelled parasite count using a quasi-Poisson mixed effects model, we were creating a model for the outcome at any time-point there by indirectly addressing the measurement error. The joint model also offered flexibility to investigate the appropriate link structure between longitudinal parasite count and time-to-new clinical malaria episode. Among the fitted joint models, only the model with cumulative parasite count was strongly associated with the hazard of new clinical malaria episode suggesting that the risk can

doi: 10.7243/2053-7662-7-1

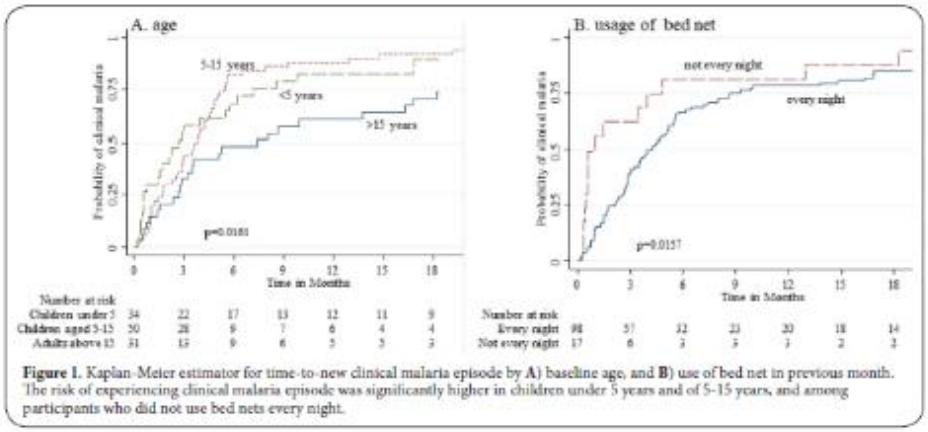


Figure 1. Kaplan-Meier estimator for time-to-new clinical malaria episode by **A)** baseline age, and **B)** use of bed net in previous month. The risk of experiencing clinical malaria episode was significantly higher in children under 5 years and of 5-15 years, and among participants who did not use bed nets every night.

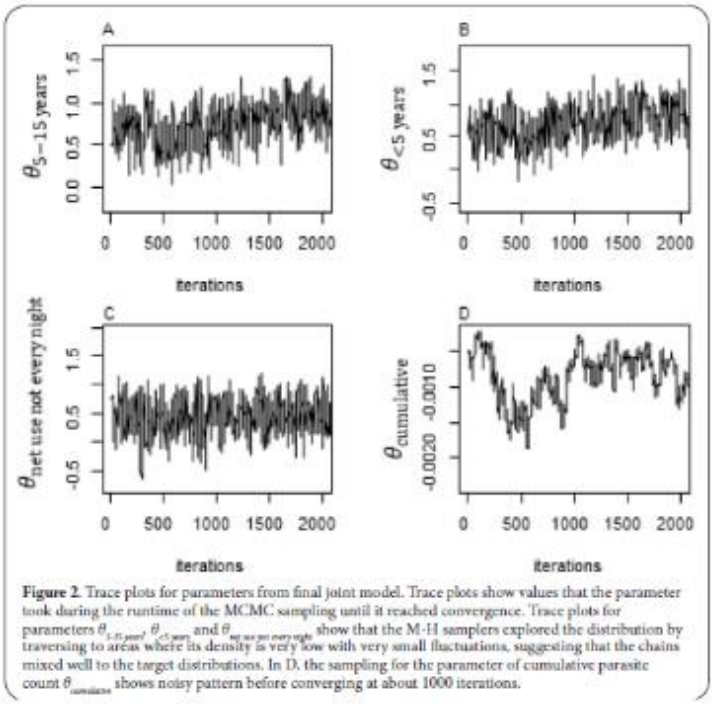


Figure 2. Trace plots for parameters from final joint model. Trace plots show values that the parameter took during the runtime of the MCMC sampling until it reached convergence. Trace plots for parameters $\theta_{5-15\ years}$, $\theta_{<5\ years}$ and $\theta_{net\ use\ not\ every\ night}$ show that the M-H samplers explored the distribution by traversing to areas where its density is very low with very small fluctuations, suggesting that the chains mixed well to the target distributions. In D, the sampling for the parameter of cumulative parasite count $\theta_{cumulative}$ shows noisy pattern before converging at about 1000 iterations.

doi: 10.7243/2053-7662-7-1

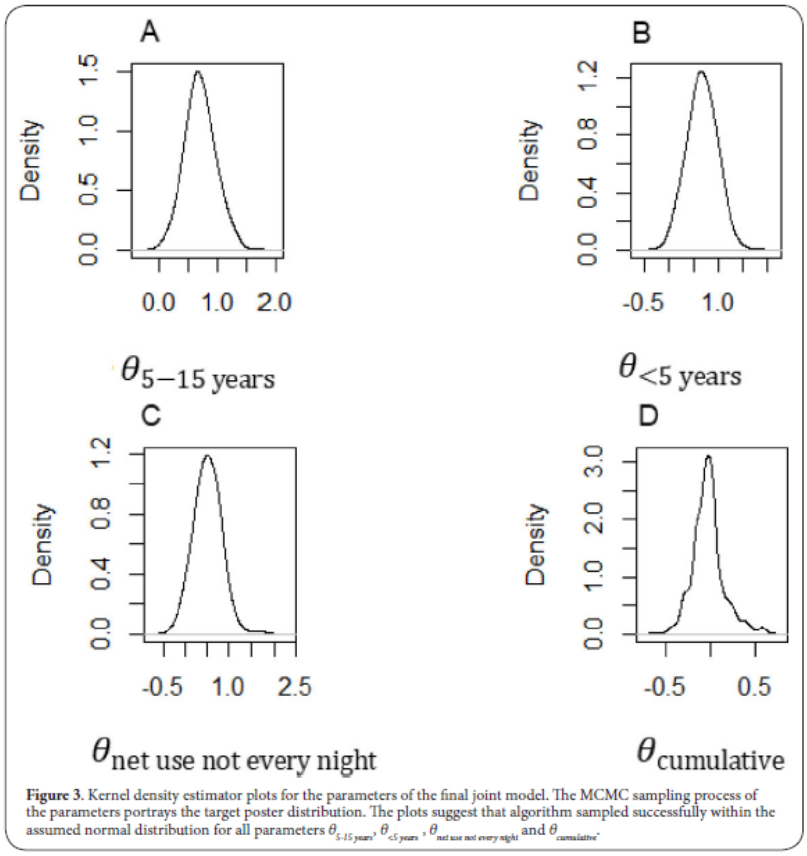


Figure 3. Kernel density estimator plots for the parameters of the final joint model. The MCMC sampling process of the parameters portrays the target poster distribution. The plots suggest that algorithm sampled successfully within the assumed normal distribution for all parameters $\theta_{5\text{-}15\ years}$, $\theta_{<5\ years}$, $\theta_{net\ use\ not\ every\ night}$ and $\theta_{cumulative}$.

better be explained by conditioning on the cumulative effect of the longitudinal parasite count trajectory from baseline up to the episode time. The underlying current value of the parasite count or change in parasite count at any time was not associated with the risk of experiencing new clinical malaria episode at the same.

Factors associated with a high risk of experiencing a new clinical malaria episode were young age and infrequent use of bed nets as have been reported in other studies [25-28]. Children aged up to 15 years had higher risk for clinical malaria disease when compared to adults over 15 years. The higher risk of experiencing a clinical malaria episode among children is possibly be due to the naturally under-developed humoral immune responses to different stage-specific antigens of *P. Falciparum* otherwise acquired with age [29]. In this study in an endemic area, high cumulative parasite count was associated with lower risk of getting a new clinical malaria episode suggesting that increased exposure to malaria parasites with time may result into protective effect to future clinical malaria episodes.

This study may be limited by focusing on time-to-first malaria episode only, thus estimates obtained here may not be applicable to analyses examining all clinical malaria episodes over a follow up period. This paper has established the optimal

way of incorporating parasite count in estimating time-to-first new clinical malaria which may further be extended to study recurrent episodes over entire follow-up, but this may require different distributional assumptions. There were missing values for bed net use which may have affected the findings. Future studies should include multiple imputation of missing covariate data and also investigate the role of measurement error due to detection limit of parasite count in these joint models. Further work is required to consider joint modelling of parasite count with recurrent episodes and to predict risk of future clinical malaria episodes.

Conclusions

In conclusion, jointly modelling longitudinal parasite count and time-to-new clinical malaria improved precision in log hazard ratio estimates for clinical malaria when compared with the conventional Cox PH model with time-dependent parasite count. The improved precision of joint modelling may improve study efficiency and allow for design of clinical trials with relatively lower sample sizes with increased power.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements and Funding

Research reported in this publication was supported by the Fogarty International Center of the National Institutes of Health under Award Number D43TW010075. The work was also partially supported through the DELTAS Sub-Saharan Africa Consortium for Advanced Biostatistics (SSACAB) programme and the UK government with funding from the Wellcome Trust Award Number 107754/Z/15/Z.

Authors' contributions

Authors' contributions	CCS	LNK	AGB	MM	DPM	MGH	MKL	TFC
Research concept and design	✓	✓	--	--	✓	--	✓	✓
Collection and/or assembly of data	--	--	✓	--	✓	--	✓	--
Data analysis and interpretation	✓	✓	✓	✓	✓	✓	✓	✓
Writing the article	✓	✓	✓	✓	✓	✓	✓	✓
Critical revision of the article	✓	✓	✓	✓	✓	✓	✓	✓
Final approval of article	✓	✓	✓	✓	✓	✓	✓	✓
Statistical analysis	✓	✓	--	✓	--	✓	--	✓

Publication history

EIC: Jimmy Efird, East Carolina University, USA.
Received: 04-Mar-2019 Final Revised: 17-Apr-2019
Accepted: 20-Apr-2019 Published: 05-Jun-2019

References

1. World Health Organisation (WHO). **World Malaria Report 2017**. 2017; 186. | Article
2. Barber BE, William T, Grigg MJ, Menon J, Auburn S, Marfurt J, Anstey NM and Yeo TW. **A prospective comparative study of knowlesi, falciparum, and vivax malaria in Sabah, Malaysia: high proportion with severe disease from Plasmodium knowlesi and Plasmodium vivax but no mortality with early referral and artesunate therapy**. *Clin Infect Dis*. 2013; 56:383-97. | Article | PubMed
3. Kasirye R, Grosskurth H, Munderi P, Levin J, Anywaine Z, Nunn A, Kamali A and Baisley K. **Longitudinal effect of CD4 by cotrimoxazole use on malaria incidence among HIV-infected Ugandan adults on antiretroviral therapy: a randomized controlled study**. *Malar J*. 2016; 15:361. | Article | PubMed Abstract | PubMed FullText
4. Loucoubar C, Grange L, Paul R, Huret A, Tall A, Telle O, Roussilhon C, Faye J, Diene-Sarr F, Trape JF, Mercereau-Puijalon O, Sakuntabhai A and Bureau JF. **High number of previous Plasmodium falciparum clinical episodes increases risk of future episodes in a sub-group of individuals**. *PLoS One*. 2013; 8:e55666. | Article | PubMed Abstract | PubMed FullText
5. Guo X and Carlin BP. **Separate and Joint Modeling of Longitudinal and Event Time Data Using Standard Computer Packages**. *Am Stat*. 2004; 58:16-24.
6. White NJ. **Malaria parasite clearance**. *Malar J*. 2017; 16:88. | Article | PubMed Abstract | PubMed FullText
7. Kalbfleisch JD and Prentice RL. **The Statistical Analysis of Failure Time Data**. *New York: John Wiley and Sons*. 2002.
8. Song X, Davidian M and Tsiatis AA. **A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data**. *Biometrics*. 2002; 58:742-53.. | Article | PubMed
9. Neuenschwander B and Branson M. **Modeling missingness for time-to-event data: a case study in osteoporosis**. *J Biopharm Stat*. 2004; 14:1005-19. | Article | PubMed
10. Taylor JM and Wang Y. **Surrogate markers and joint models for longitudinal and survival data**. *Control Clin Trials*. 2002; 23:626-34. | Article | PubMed
11. Ibrahim JG, Chen M-H and Sinha D. **Bayesian methods for joint modeling of longitudinal and survival data with applications to cancer vaccine trials**. *Stat Sin*. 2004; 14:863-884.
12. Chen Q, May RC, Ibrahim JG, Chu H and Cole SR. **Joint modeling of longitudinal and survival data with missing and left-censored time-varying covariates**. *Stat Med*. 2014; 33:4560-76. | Article | PubMed Abstract | PubMed FullText
13. Chi YY and Ibrahim JG. **Joint models for multivariate longitudinal and multivariate survival data**. *Biometrics*. 2006; 62:432-45. | Article | PubMed
14. Brown ER and Ibrahim JG. **A Bayesian semiparametric joint hierarchical model for longitudinal and survival data**. *Biometrics*. 2003; 59:221-8. | Article | PubMed
15. Buchwald AG, Sixpence A, Chimenya M, Damson M, Sorkin JD, Wilson ML, Seydel K, Hochman S, Mathanga DP, Taylor TE and Laufer MK. **Clinical Implications of Asymptomatic Plasmodium falciparum Infections in Malawi**. *Clin Infect Dis*. 2019; 68:106-112. | Article | PubMed

16. World Health Organization (WHO). **Epidemiological profile of malaria in Malawi in 2015**. 2017. | Pdf
17. Bennett A, Kazembe L, Mathanga DP, Kinyoki D, Ali D, Snow RW and Noor AM. **Mapping malaria transmission intensity in Malawi, 2000-2010**. *Am J Trop Med Hyg*. 2013; **89**:840-849. | Article | PubMed Abstract | PubMed FullText
18. Walldorf JA, Cohee LM, Coalson JE, Bauleni A, Nkanaunena K, Kapito-Tembo A, Seydel KB, Ali D, Mathanga D, Taylor TE, Valim C and Laufer MK. **School-Age Children Are a Reservoir of Malaria Infection in Malawi**. *PLoS One*. 2015; **10**:e0134061. | Article | PubMed Abstract | PubMed FullText
19. Ibrahim JG, Chen MH and Sinha D. **Bayesian Survival Analysis**. *New York: Springer*. 2001.
20. StataCorp. **Stata Statistical Software: Release**. 2017; **15**.
21. R Core Team. **A language and environment for statistical computing. R Found Stat Comput Vienna, Austria**. 2017; R Foundation for Statistical Computing. 2017.
22. Spiegelhalter DJ, Best NG, Carlin BP and Van Der Linde A. **Bayesian measures of model complexity and fit**. *J R Stat Soc Ser B Stat Methodol*. 2002; **64**:583-616. | Pdf
23. Ibrahim JG, Chu H and Chen LM. **Basic concepts and methods for joint models of longitudinal and survival data**. *J Clin Oncol*. 2010; **28**:2796-801. | Article | PubMed Abstract | PubMed FullText
24. Lawrence Gould A, Boye ME, Crowther MJ, Ibrahim JG, Quartey G, Micallef S and Bois FY. **Joint modeling of survival and longitudinal non-survival data: current methods and issues. Report of the DIA Bayesian joint modeling working group**. *Stat Med*. 2015; **34**:2181-95. | Article | PubMed Abstract | PubMed FullText
25. Mathanga DP, Tembo AK, Mzilahowa T, Bauleni A, Mtimaukenena K, Taylor TE, Valim C, Walker ED and Wilson ML. **Patterns and determinants of malaria risk in urban and peri-urban areas of Blantyre, Malawi**. *Malar J*. 2016; **15**:590. | Article | PubMed Abstract | PubMed FullText
26. Alemu A, Tsegaye W, Golassa L and Abebe G. **Urban malaria and associated risk factors in Jimma town, south-west Ethiopia**. *Malar J*. 2011; **10**:173. | Article | PubMed Abstract | PubMed FullText
27. Haque U, Glass GE and Bomblies A. **Risk factors associated with multiple malaria infections in Bangladesh**. *Am J Trop Med Hyg*. 2012; **87**:403. | Pdf
28. Haque U, Glass GE, Bomblies A, Hashizume M, Mitra D, Noman N, Haque W, Kabir MM, Yamamoto T and Overgaard HJ. **Risk factors associated with clinical malaria episodes in Bangladesh: a longitudinal study**. *Am J Trop Med Hyg*. 2013; **88**:727-732. | Article | PubMed Abstract | PubMed FullText
29. Biswas S, Seth RK, Tyagi PK, Sharma SK and Dash AP. **Naturally acquired immunity and reduced susceptibility to falciparum malaria in two subpopulations of endemic eastern India**. *Scand J Immunol*. 2008; **67**:177-84. | Article | PubMed

Citation:
Stanley CC, Kazembe LN, Buchwald AG, Mukaka M, Mathanga DP, Hudgens MG, Laufer MK and Chirwa TF. **Joint modelling of time-to-clinical malaria and parasite count in a cohort in an endemic area**. *J Med Stat Inform*. 2019; **7**:1.
http://dx.doi.org/10.7243/2053-7662-7-1

Appendix 6.1: Stata do-files and R-scripts

***#Chapter 3**

```
cd "C:\Users\Wits work\manuscripts\paper1" /*setting directory

use paper.dta, clear /*loading data

xtset year


xtgee descriptives_stats period samp_size, family(binomial)
link(logit) eform


xtgee contingency_table period samp_size, family(binomial)
link(logit) eform


xtgee kaplan_meier period samp_size, family(binomial) link(logit)
eform


xtgee coxmodel period samp_size, family(binomial) link(logit) eform


xtgee logrank period samp_size, family(binomial) link(logit) eform


xtgee poisson period samp_size, family(binomial) link(logit) eform


xtgee negativebinomial period samp_size, family(binomial)
link(logit) eform


xtgee gee period samp_size, family(binomial) link(logit) eform


xtgee mixedeffectmodel period samp_size, family(binomial)
link(logit) eform
```

```
*creating sub-dataset for trend analyis

bysort year: gen total_studies=_n /*counting number of articles
paper year

replace total_studies=. if year[_n]==year[_n+1]   /* retaining the
total number only per year



*collapsing obseravtions on year, for figure2-lowess plot

collapse (sum) total_studies - methods_other, by(period3)



*plot of all methods on the same graph

twoway (tsline descriptives_stats_percent kaplan_meier_percent
coxmodel_percent poisson_percent gee_percent
mixedeffectmodel_percent, recast(connected)), xlab(0(1)3)
ytitle(Percent)  xtitle(Year)title(Studies using a statistical
procedure)

graph save combined_ts, replace
```

***#Chapter 4**

```
#loading required packages

library(JM)

library(foreign)

library(JMbayes)

library("lattice")

library(lme4)

library(nlme)

library(pscl)


# loading data
```

```
Paper2 <- read.dta("C:/Users/wits/mfera1.dta")

Paper2.id <- read.dta("C:/Users/wits/mfera2.dta")


#mixed-effects model with Poisson distribution

poisson.malaria.glmmPQL <- glmmPQL(parasite ~ ns(obsdays,2), random
= ~ 1 | studyid, family = poisson, data = Paper2)

summary(poisson.malaria.glmmPQL)


#mixed model of quasipoisson

qpoisson.malaria.glmmPQL <- glmmPQL(parasite ~ ns(obsdays,2), random
= ~ 1 | studyid, family = quasipoisson(link='log'), data = Paper2)

summary(qpoisson.malaria.glmmPQL)


#mixed model with negative binomial

nbinomial.malaria.glmmPQL <- glmmPQL(parasite ~ obsdays, random = ~
1| studyid, family = negative.binomial(5), data = Paper2, control =
list(opt = "optim"))

summary(nbinomial.malaria.glmmPQL)


#rootogram to visualize the fit of a count regression model

library(aods3)     ## overdispersion diagnostics


#defining the densLong function

  #sample########################## for binary

  dLongBin <- function (y, eta.y, scale, log = FALSE, data) {

    dbinom(x = y, size = 1, prob = plogis(eta.y), log = log)

  }


  #for Poisson distribution

  dLongPois <- function (y, eta.y, scale, log = FALSE, data) {

  dpois(x=y, lambda = 1, log = log)
```

```
  }

  #work on the parameters

  #general form: dpois(x, lambda, log = FALSE)


  ############################### for negative binomal

    dLongnBinom <- function (y, eta.y, scale, log = FALSE, data) {

  dnbinom(x=y, size = 0.525, prob =0.5 , mu= , log = log)

  }

  #general form: dnbinom(x, size, prob, mu, log = FALSE)


#fitting cox model on time to first episode, with parasitemia as
time-dependent

td.cox.malaria <- coxph(Surv(start, stime, clinical_malaria) ~
parasite + age + pre_fu_netfreq + cluster(studyid), data = Paper2)

summary(td.cox.malaria)



#fitting cox model with parasitemia only as time-dependent

td.cox.malaria.PC <- coxph(Surv(start, stime, clinical_malaria) ~
parasite + cluster(studyid), data = Paper2)

summary(td.cox.malaria.PC)



#fitting cox model on time to first episode single record per id

survFit.malaria <- coxph(Surv(full_fu_days, clinical_malaria) ~
agecat+pre_netfreq, data = Paper2.id, x = TRUE)

summary(survFit.malaria)



#fitting joint model - with poisson

#set.seed(20)

jointFit.poisson.malaria <- jointModelBayes(poisson.malaria.glmmPQL,
survFit.malaria, timeVar = "obsdays", densLong = dLongPois)

summary(jointFit.poisson.malaria)
```

```
#fitting joint model - with negative binomial

#set.seed(30)

jointFit.nbinom.malaria <-
jointModelBayes(nbinomial.malaria.glmmPQL, survFit.malaria, timeVar
= "obsdays", densLong = dLongnBinom)

summary(jointFit.nbinom.malaria)



#fit final standard joint model (with quasi-Poisson) with only the
current value-m_i(t) term in the linear predictor of the survival
submodel

#set.seed(40)

jointFit.malaria1 <- jointModelBayes(qpoisson.malaria.glmmPQL,
survFit.malaria, timeVar = "obsdays", densLong = dLongPois)

summary(jointFit.malaria1)



#Including the time-dependent slopes term

dForm <- list(fixed = ~ 0 + dns(obsdays,2), random = ~ 0 ,indFixed =
2:3, indRandom = 0:0)

#set.seed(50)

jointFit.malaria2 <- update(jointFit.malaria1, param = "td-both",
extraForm = dForm, n.iter = 20000)

summary(jointFit.malaria2)



#Including the cumulative effect of the parasitemia

iForm <- list(fixed = ~ 0 + obsdays + ins(obsdays, 2), random = ~ 0
+ obsdays,indFixed = 1:3, indRandom = 1:1)

#set.seed(60)

jointFit.malaria3 <- update(jointFit.malaria1, param = "td-extra",
extraForm = iForm, n.iter =20000)

summary(jointFit.malaria3)



#diagnostics plots for mcmc samples
```

```
library("rvg")   #for editable plots

library("ggplot2")

library("officer")


plot(jointFit.malaria3)

plot(jointFit.malaria3, which="density")


#table of posterior means

xtable(jointFit.malaria3)

#Dynamic prediction of risk for new clinical malaria episode

#for participant 98

Data98 <- paper2[paper2$studyid == 98, ]

sfit.malaria <- survfitJM(jointFit.malaria3, newdata = Data98, idVar
= "studyid")

sfit.malaria

plot(sfit.malaria, estimator = "mean", include.y = TRUE,

     conf.int = TRUE, fill.area = TRUE, col.area = "lightgrey")
```

***#Chapter 5**

```
*cd "C:\Users\Wits work\paper3" /* setting directory

use paper3.dta, clear / loading data

*using merlin package to jointly model frailty with Weibull baseline
for recurrent events and longitudinal parasitemia

merlin (stime age_yrs sex season net_freq M1[studyid], ///
family(weibull, failure(clinical_malaria) ))(parasite age_yrs ///
sex season net_freq M1[studyid]@1, family(gaussian)), /// nolog
intmethod(mcarlo) iterate(9)

estat ic



*poisson distribution for parasitemia
```

```
merlin (stime age_yrs sex season net_freq M1[studyid], ///
family(weibull, failure(clinical_malaria) ))(parasite age_yrs ///
sex season net_freq M1[studyid]@1, family(poisson)), /// nolog
intmethod(mcarlo) iterate(9)

estat ic

*quasi-Poisson distribution for parasitemia

merlin (stime age_yrs sex season net_freq M1[studyid], ///
family(weibull, failure(clinical_malaria) ))(parasite age_yrs ///
sex season net_freq M1[studyid]@1, family(qpoisson)), /// nolog
intmethod(mcarlo) iterate(9)

estat ic



*Negative binomial distribution for parasitemia

merlin (stime age_yrs sex season net_freq M1[studyid], ///
family(weibull, failure(clinical_malaria) ))(parasite age_yrs ///
sex season net_freq M1[studyid]@1, family(nbinomial)), /// nolog
intmethod(mcarlo) iterate(9)

estat ic



*# Sample scenario for simulated data

import delimited simscenario1a.csv, clear

/*fitting joint model to simulated dataset

set seed 100

bootstrap: merlin     (syears age_yrs sex season net_freq
M1[studyid], family(weibull, failure(clinical_malaria))) ///

           (long_y age_yrs sex season net_freq M1[studyid]@1,
family(nbinomial)) ///

           , nolog intmethod(mcarlo) iterate(9)

estat ic

estat bootstrap
```

**Note: Other scenarios for the stated conditions in the report are skipped due to space.