University of Cape Town

Doctor of Philosophy Dissertation

Joint models for nonlinear longitudinal profiles in the presence of informative censoring

Author and supervisors

Author:
Tinashe Chatora

Supervisors:
A/Prof Francesca Little
Professor Karen Barnes

August 13, 2018

Contents

List of Tables

List of Figures

Acknowledgments

I would like to thank Assoc. Professor Francesca Little for her excellent supervision, statistical expertise and patient guidance throughout the course of this thesis. In addition I would like to thank Prof Barnes for her expert knowledge, clinical expertise and insightful feedback. Without the guidance and expertise of my supervisors this body of work would not have come to fruition.

I would also like to thank my loving wife Mildret Braundi, who undertook this long journey with me and who supported me every step of the way. She pushed me on when I thought I could go no further and for that I am eternally grateful.

PLAGIARISM DECLARATION

1. I know that plagiarism is wrong. Plagiarism is to use anothers work and pretend that it is my own.

2. I have used the APA referencing guide for citation and referencing. Each contribution to and quotation in this dissertation from the work(s) of other people has been contributed and has been cited and referenced.

3. I know the meaning of plagiarism and declare that all of the work in the dissertation, save for that which is properly acknowledged, is my own.

Name: Signed by candidate

Date: August 13, 2018

Abstract

Malaria is the parasitic disease that affects the most humans, with *Plasmodium falciparum* malaria being responsible for the majority of severe malaria and malaria related deaths. The asexual form of the parasite causes the signs and symptoms associated with malaria infection. The sexual form of the parasite, also known as a gametocyte, is the stage responsible for infectivity of the human host (patient) to the mosquito vector and thus ongoing transmission of malaria and the spread of antimalarial drug resistance. Historically malaria therapeutic efficacy studies have focused mainly on the clearance of asexual parasites. However, malaria in a community can only be truly combated if a treatment program is implemented that is able to clear both asexual and sexual parasites effectively.

In this thesis focus will be on the modeling of the key features of gametocytemia (the presence of gametocytes in a host's system). Particular emphasis will be on the modeling of the time to gametocyte emergence, the density of gametocytes and the duration of gametocytemia. It is also of interest to investigate the impact of the administered treatment on the aforementioned features.

Gametocyte data has several interesting features. Firstly, the distribution of gametocyte data is zero-inflated with a long-tail to the right. The observed longitudinal gametocyte profile also has a nonlinear relationship with time. In addition, since most malaria intervention studies are not designed to optimally measure the evolution of the longitudinal gametocyte profile, there are very few observation points in the time period where the gametocyte profile is expected to peak. Gametocyte data collected from malaria intervention studies are also affected by informative censoring, which leads to incomplete gametocyte profiles. An example of informative censoring is when a patient who experiences treatment failure is "rescued" and withdrawn, from the study in order to receive alternative treatment. This patient can be considered to be in worse health as compared to the patients who remain in this study. There are also competing risks of exit from the study, as a patient can either experience treatment failure or be lost-to-follow-up.

The aforementioned features of gametocyte data make it statistically appealing to analyze. In literature there are several modeling techniques that can be used to analyze individual features of the data. These techniques include standard survival models for modeling the time to gametocyte emergence and the duration of gametocytemia. The longitudinal nonlinear gametocyte profile would typically be modeled using nonlinear mixed effect models. These nonlinear models could then subsequently be extended to accommodate the zero-inflation in the data, by changing the

underlying assumption around the distribution of the response variable. However, it is important to note that these standard techniques do not account for informative censoring. Failure to account for informative censoring leads to bias in parameter estimates. Joint modeling techniques can be used to account for informative censoring. The joint models applied in this thesis combined the longitudinal nonlinear gametocyte densities and the time to censoring due to either being lost-to-follow-up or treatment failure.

The data analyzed in this thesis were collected from a series of clinical trials conducted between 2002 and 2004 in Mozambique and the Mpumulanga province of South Africa. These trials were a part of the South East African Combination Antimalarial Therapy (SEACAT) evaluation of the phased introduction of combination anti-malarial therapy, nested in the Lubombo Spatial Development Initiative. The aim of these studies was primarily to measure the efficacy of sulfadoxine-pyrimethamine (SP) and a combination of artesunate and sulfadoxine-pyrimethamine (ACT), in eliminating asexual parasites in patients. The patients enrolled in these studies had uncomplicated malaria[1], at a time of increasing resistance to sulfadoxine-pyrimethamine (SP) treatment. Blood samples were taken from patients during the course of 6 weeks on days 0, 1, 2, 3, 7, 14, 21, 28 and 42. Analysis of these blood samples provided longitudinal measurements for asexual parasite densities, gametocyte densities, sulfadoxine drug concentrations and pyrimethamine drug concentrations.

The gametocyte data collected in this study were initially analyzed using standard survival modeling techniques. Non-parametric Cox regression models and parametric survival models were applied to the data as part of this initial investigation. These models were used to investigate the factors that affected the time to gametocyte emergence. Subsequently, using the subset of the population that experienced gametocytemia, accelerated failure time models were applied to investigate the factors that affected the duration of gametocytemia. It is evident that the findings from the aforementioned duration investigation would only be able to provide valid duration estimates for patients who were detected to have gametocytemia. This work was extended to allow for population level duration estimates by incorporating the prevalence of gametocytemia into the estimation of duration, for generic patients with specific covariate patterns. The prevalence of gametocytemia was modeled using an underlying binomial distribution. The delta method was subsequently used to derive confidence intervals for the population level duration estimates that were associated with specific covariate patterns. An investigation into the fac-

[1]Malaria is defined as uncomplicated when the signs and symptoms of malaria are present but there are no clinical or laboratory signs to indicate severity or vital organ dysfunction.

tors affecting the early withdrawal of patients from the study was also conducted. Early exit from the study arose either through loss-to-follow-up (LTFU) or through treatment failure.

The longitudinal gametocyte profile was modeled using joint modeling techniques. The resulting joint model used shared random effects to combine a Weibull survival model, describing the cause-specific hazards of patient exit from the study, with a nonlinear zero-adjusted gamma mixed effect model for the longitudinal gametocyte profile. This model was used to impute the incomplete gametocyte profiles, after adjusting for informative censoring. These imputed profiles were then used to estimate the duration of gametocytemia.

It was found, in this thesis, that treatment had a very strong effect on the hazard of gametocyte emergence, density of gametocytes and the duration of gametocytemia. Patients who received a combination of sulfadoxine-pyrimethamine and artesunate were found to have significantly lower hazards of gametocyte emergence, lower predicted durations of gametocytemia and lower predicted longitudinal gametocyte densities as compared to patients who received sulfadoxine-pyrimethamine treatment only.

1 Introduction

Malaria is the parasitic disease that affects the most humans. It is estimated that 216 million episodes of malaria arose in 2016, with 194 million of these episodes originating from the African region (WHO, 2017b). These episodes are estimated to have resulted in 445 000 malaria related deaths, with Africa contributing to 91% of these deaths.

The malaria species, $Plasmodium falciparum$, is responsible for the majority of severe malaria cases and malaria related deaths. The asexual form of the parasite is responsible for the signs and symptoms attributed to malaria infection. Researchers who conduct malaria therapeutic efficacy studies are primarily interested in the time to asexual parasite clearance and the efficacy of treatment in curing patients, known as an adequate clinical and parasitological response.

The basic lifecycle of the $Plasmodium falciparum$ parasite, depicted graphically by Michalakis and Renaud (2009), is illustrated in this thesis in Figure 1. The basic lifecycle of the $Plasmodium falciparum$ parasite can be summarized as follows (Michalakis and Renaud, 2009, Barnes and White, 2005, Nacher et al., 2002)

- Infected female mosquitoes bite a human host leading to the sporozoite form of the parasite entering the host's blood stream.
- Once in the host's blood stream, sporozoites are transported to the liver where they multiply asexually over a period of 5-9 days, which is referred to as the "pre-erythrocytic" stage, before they become merozoites that invade the red blood cells of the host.
- A repetitive asexual cycle begins and this leads to the host experiencing chills and a fever, which are representative of clinical symptoms of malaria infection. Subsequently male and female gametocytes are produced, with mature gametocytes appearing in the host's blood stream 10-12 days after clinical symptoms appear.
- When a mosquito bites the host, gametocytes are transmitted into its system. Once in the gut of the mosquito, the male and female gametocytes fuse to form zygotes. These zygotes eventually give rise to sporozoites that allow the lifecycle of the parasite to continue.

Research into the development of treatments that quickly and efficiently eradicate gametocytes is thus of critical public health importance as decreasing

gametocyte carriage reduces ongoing malaria transmission. This can especially be advantageous in areas of increasing resistance to antimalarial treatment as a decrease in gametocyte carriage can slow the spread of the resistance that is threatening current malaria control and elimination efforts. It is thus essential that we improve our methods for the analysis of gametocyte data to better define risk factors associated with gametocyte carriage.

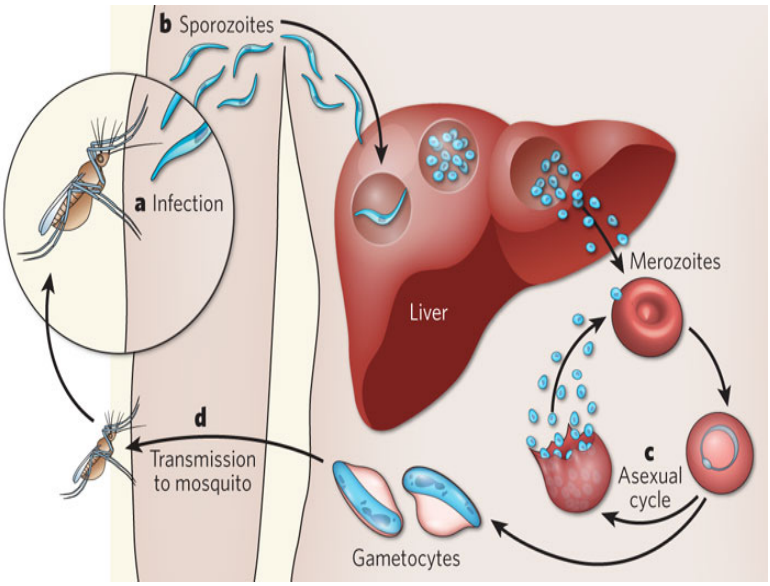

Figure 1: *Plasmodium falciparum* life cycle as depicted by Michalakis and Renaud (2009)

Host infectivity is defined as the probability of a mosquito becoming infected after feeding off an infected individual. Host infectivity is the key driver behind malaria transmission, though it is also important to note that there are additional factors that affect transmission like vector (organisms that carry pathogens from one host to another) characteristics, host susceptibility and climatic conditions (Draper, 1953, Diebner et al., 2000, Killeen et al., 2006).

Historically very little attention has focused on the modeling of gametocytes. Researchers commonly estimated the gametocyte density from the asexual parasite density. Examples of such methods include assuming a constant transition rate between asexual and sexual parasites (Pongtavornpinyo, 2006) and the use of models that rely on lagged values from asexual parasites (Ross et al., 2006). Distiller et al. (2010) directly modeled the observed gametocyte profiles over time, using nonlinear mixed effect models. In that paper a modified critical exponential model was applied as the underlying nonlinear function of the model.

Malaria intervention trials are generally not appropriately designed to measure the evolution of the gametocyte density in patients over time. This is because the patient visitation schedules are based on the clinical and asexual parasite response and do not coincide with the ideal times to observe the gametocyte life cycle. An additional hindrance to the observation of the gametocyte density, over time, is censoring. In this study censoring occurred due to two reasons. Firstly, a patient could be "rescued" and given an alternative treatment due to either failure to clear asexual parasites or late treatment failure (recurrence of asexual parasites due to recrudescence of the initial infection or reinfection). Secondly patients could exit the study due to being lost-to-follow-up. Rescuing patients from the study potentially leads to informative censoring. Patients who were rescued could be expected to have asexual parasites in their systems for prolonged periods of time. These asexual parasites would in turn be expected to develop into gametocytes. As a result it could be hypothesized that the prevalence of gametocytemia in patients who were rescued from the study was higher than that of patients who remained in the study. The loss-to-follow-up method of exit could potentially be either non-informative or informative censoring. If it is assumed that patients who quickly experience a successful clinical and parasitological outcome are more inclined to voluntarily exit the study early, it would imply that these patients would have a lower prevalence of gametocytemia as compared to the patients who remained in the study. It might also be hypothesized that patients who were lost-to-follow-up were in poor health and voluntarily exited the study to seek alternative treatment before they were rescued from the study. Alternatively, the loss-to-follow-up might be due to reasons completely unrelated to the study, which would imply non-informative censoring. This will be investigated further in this thesis.

Censoring leads to truncated longitudinal gametocyte profiles. It can reasonably be assumed that the gametocyte profiles of patients who were rescued from the study or who were lost-to-follow-up (assuming informative censoring) would differ from the observed profiles of patients who were successfully treated and observed to the end of the gametocyte cycle. Standard nonlinear mixed effects models are typically used to impute incomplete longitudinal profiles. However, these models do not account for informative censoring. Failure to account for informative censoring in such situations leads to biased parameter estimates in both survival and longitudinal models, which are fit to the data (Faucett and Thomas, 1996).

Joint modeling techniques can be applied to account for informative censoring. There are multiple approaches to joint models that have been documented.

Examples of these approaches are Wang and Taylor (2001) who used the longitudinal response as a time dependent covariate in the proportional hazards survival model. Lina et al. (2002) used a latent class model approach that involved using a logistic model to determine the patient's membership class and then subsequently modeling the longitudinal and the survival processes, for a given class, independently. Murawska et al. (2012) proposed a two-stage joint model, under the Bayesian framework, which firstly modeled the longitudinal process using a nonlinear mixed effect model and subsequently applied the Empirical Bayes estimates of the subject-specific parameters as predictors in the Cox proportional hazards model. Another approach, which will be expanded upon for the purposes of this research, was proposed by Henderson et al. (2000). This approach connected the longitudinal and survival processes together using bivariate random effects. This approach was designed for a longitudinal process that could be modeled using a linear mixed model and for a survival process modeled using an intensity model with frailty. This approach will be extended in this thesis to accommodate a nonlinear longitudinal process under the Bayesian framework. The advantages of using the Bayesian framework to develop the joint models used in this analysis are that asymptotic approximations of the complicated likelihood functions, that will be developed, are avoided and the framework is easily able to estimate complex functions of parameters arising from the nonlinear nature of the fitted models (Ghosh et al., 2006).

2 Data overview

The data used in this analysis was from a series of randomized clinical trials conducted between 2002 and 2004 in southern Mozambique and the Mpumulanga province of South Africa (Barnes et al., 2006, Allen et al., 2009). These trials were a part of the South East African Combination Antimalarial Therapy (SEACAT) Evaluation of the phased introduction of artemisinin-based combination anti-malarial therapy, nested within the Lubombo Spatial Development Initiative. The aim of these studies was primarily to measure the efficacy of two treatments, in eliminating asexual parasites in patients. The treatments used were sulfadoxine-pyrimethamine (SP) and a combination of sulfadoxine-pyrimethamine and artesunate (ACT). The patients enrolled in the study had uncomplicated malaria, at a time of increasing resistance to sulfadoxine-pyrimethamine (SP) treatment.

Over the period in that the study was conducted, malaria mortality rates had been seen to be increasing despite there being a decrease in all cause mortality rates (Snow et al., 2001, Korenromp et al., 2003). This increase was mainly attributed to a rise in the resistance of *Plasmodium falciparum* parasites to antimalarial treatment.

Historically, chloroquine was one of the most widely used treatments for malaria. However, due to an increase in resistance, it became largely ineffective in most malaria endemic countries (Trape, 2001). Sulfadoxine-pyrimethamine became the successor to chloroquine and for some time became one of the most widely used antimalarial drugs in the world. This was an ideal treatment as it could be administered through a single dose as a tablet. Unfortunately a resistance to this drug arose, thus limiting the effectiveness of the drug (Babiker et al., 2005, Snow et al., 2001, Nosten et al., 2000, Targett et al., 2001). Sulfadoxine-pyrimethamine is currently mostly used as a preventive treatment in high risk groups such as pregnant women (in malaria endemic areas) and infants (in areas of moderate to high malaria transmission) (WHO, 2015).

It is now recommended practice to administer combinations of two or more drugs when treating malaria, as this approach is believed to reduce the development of resistance to treatment. This is because it is unlikely that a parasite would be able to mutate successfully against multiple drugs that would have different mechanisms of action.

Artemisinin-based combination therapies are now the most recommended treatment for *Plasmodium falciparum* malaria. The combination of artesunate and sulfadoxine-pyrimethamine is currently one of the five options for the treatment of uncomplicated malaria, in areas where sulfadoxine-pyrimethamine re-

mains effective, that is recommended by the World Health Organization (WHO, 2015). The artesunate and sulfadoxine-pyrimethamine combination treatment is currently the first line treatment in India and some Eastern Mediterranean countries (WHO, 2017b). Artemisinin-based combination therapies are renowned for their ability to quickly reduce the number of parasites in a patient's blood. Artemisinin and its derivatives significantly reduce the number of asexual parasites in a patient's blood in the first three days of treatment, subsequently the drug administered in combination with artemisinin eliminates the remaining parasites. Due to an expanded access to artemisinin-based combination therapies, in malaria-endemic countries over the last 15 years, the global burden of malaria has reduced (WHO, 2017a).

Over the period when this study was conducted, *Plasmodium falciparum* had developed resistance to all classes of anti-malarial drugs with the possible exception of artemisinin derivatives (White, 2004, Pongtavornpinyo, 2006, Barnes et al., 2006). As a result, artemisinin was a preferred drug to add to a combination treatment. Unfortunately over the last few years artemisinin resistance has been confirmed in the Greater Mekong Subregion, in countries like Cambodia, the Lao Peoples Democratic Republic, Myanmar, Thailand and Vietnam. In most cases, patients with artemisinin-resistant parasites are still able to successfully clear their infection if the drug used in conjunction with the artemisinin derivate is still effective in the geographical area. However, it has been found that *Plasmodium falciparum* has become resistant to almost all available antimalarial treatments in areas along the Cambodia-Thailand border (WHO, 2013).

Patients who were considered for the analysis conducted in this thesis, did not show signs of gametocytemia (the presence of gametocytes in the host's system) at entry into the study. These trials had a 42 day follow up period with observations occurring on days 0, 1, 2, 3, 7, 14, 21, 28 and 42. On these observation days, blood samples were taken from patients. Analysis of these blood samples provided the following information on patients over the course of the study

- Sulfadoxine drug concentrations over time,
- Pyrimethamine drug concentrations over time,
- Asexual parasite density over time,
- Gametocyte density over time.

It is important to note that, due to study guidelines, gametocyte densities were

only recorded on days 0, 3, 7, 14, 21, 28 and 42. In total, 609 patients were considered for this analysis.

A range of baseline covariates were collected from patients on entry into the study, as shown in Table 1. These covariates were considered for their clinical significance to the prevalence and duration of gametocytemia.

Table 1: Covariates used in this analysis.

Covariate	Abbreviation
Baseline asexual parasite density per microliter, measured to the logarithm of base 10 ($\log_{10}$)	*pzero*
Indicator variable for the presence (1) or absence (0) of quintuple mutations	*mut*5
Indicator variable for the treatment given to a patient, with a value of 0 for sulfadoxine-pyrimethamine treatment only (SP) and 1 for artesunate and sulfadoxine-pyrimethamine combination treatment (ACT)	*trt*
Parasite reduction ratio at 24 hours	*ratio*
Gender of patient, with males given a value of 1 and females given a value of 0	*gender*
Indicator variable for the presence (1) or absence (0) of moderate anaemia, with patients who had a haemoglobin density of less than 11g/dL defined by the World Health Organization (WHO) as having moderate anaemia	*anaemia*
Patient age in years, measured to the logarithm of base 2 ($\log_2$)	*lage*

Treatment (*trt*) was used as a covariate because the implementation of artesunate and sulfadoxine-pyrimethamine combination treatment (ACT) policies was predicated on their more rapid asexual parasite clearance, higher cure rates and reduced gametocyte carriage (White, 1997, Adjalley et al., 2011, Price et al., 1996). Since gametocytes develop from asexual parasites, a treatment that rapidly clears asexual parasites from a patient would be expected to have a low gametocyte prevalence. The efficacy of asexual parasite clearance can be measured by the rate at that parasites are cleared. In this study, the percentage of baseline asexual parasites cleared within 24 hrs (*ratio*) was used to measure the rate of parasite clearance. In addition to the rate of clearance, the number of baseline asexual parasites also plays a part in the prevalence of gametocytemia as it would be expected that a high baseline asexual parasite density would result in higher gametocyte prevalence. This assumption was applied in the models by Pongtavornpinyo (2006) and Ross et al. (2006) that relied on asexual parasite densities to estimate gametocyte densities in patients.

This analysis was conducted in the presence of increasing resistance to SP treatment in southern and eastern Africa. Pyrimethamine and sulfadoxine target the dihydropteroate synthase (dhps) and dihydrofolate reductase (dhfr) enzymes in the folate synthesis pathway of *Plasmodium falciparum*. When SP is given to a patient, the combination of the pyrimethamine and sulfadoxine drugs act in unison to disrupt folate synthesis and kill the *Plasmodium falciparum* parasite. Resistance to SP arises due to point mutations that accumulate at several sites in the dhfr and dhps genes. According to Roper et al. (2003), there is a direct correlation between the number of pretreatment dhfr and dhps mutations and the level of parasite resistance; with a high number of mutations leading to an increase in treatment resistance. In this analysis, treatment resistance was measured using a dichotomous variable ($mut5$) that takes a value of 1 if a patient has 5 mutations (3 dhfr and 2 dhps) and 0 if the patient has fewer than 5 mutations. There were no infections in this study with greater than 5 mutations.

Researchers have found that the frequency of infection and density of asexual parasites declines as the age of a population increases. This occurs primarily in areas of moderate to high intensity transmission (e.g. in Mozambique) but not in areas of low intensity transmission (e.g. Mpumulanga). This decline has been attributed to partial immunity of the population due to repeated infections (Despommier et al., 1994). The number of repeat infections that a patient gets, can be considered as being proportional to the age of that patient specifically in areas of more intense malaria transmission. Age can thus be considered as a covariate that provides information about patient immunity. In this thesis the logarithm to the base two of age was used. As a result, a one unit increase in the logarithm to the base 2 of age can be interpreted as the doubling of age. This approach seemed appropriate as the age distribution was skew to the right with a long-tail. Taking the logarithm of distributions with such characteristics narrows the range of a covariate and potentially leads to an acceleration in the time to Bayesian model convergence. Patient gender was also included as a covariate in this analysis. Similarly to age, gender can harbor a latent relationship of partial immunity to gametocytemia if associated with more frequent malaria infections. In addition it allows for the population demographics to be included in the analysis.

Authors like von Seidlein et al. (2001), Price et al. (1999) and Nacher et al. (2002) found that the patients with low concentrations of haemoglobin, which can be considered as the presence of anaemia, had a higher prevalence of gametocytemia. In this analysis moderate anaemia was defined as having a haemoglobin

concentration of less than 11g/dL. This definition is consistent with the guidelines outlined by the World Health Organization (WHO, 2011). Applying the 11g/dL cut-off level allowed for a balanced distribution of patients with anaemia, across the various categories of covariates used in this analysis.

2.1 Treatment outcome

The simplified treatment outcomes from the study are shown in Table 2, along with their respective definitions. The treatment failure classification used in this analysis was a combination of the early treatment failure, recrudescence and reinfection classifications outlined by the World Health Organization. These classifications were combined for the purpose of this thesis in order to improve the statistical power of the analysis.

Table 2: Simplified definition of treatment outcomes.

Outcome	Definition
Success (Adequate Clinical and Parasitological Response)	• Asexual parasites are cleared from the patient's system before day 7 with no recurrence
Failure	• Asexual parasites fail to clear from the patient's system before day 7 (Early treatment failure) • Asexual parasites clear before day 7 but they recur during the course of the study (Recrudescence) • Asexual parasites are cleared from the patient's system before day 7 and the patient contracts a new infection during the course of the study (Reinfection)
LTFU (Loss-to-follow-up)	• The patient is lost-to-follow-up during the course of the study

The distribution of the patients included in the analysis, across the three treatment outcome categories, is shown in Table 3. This table also provides a comparison of the three treatment outcomes with respect to the baseline covariates. It can be seen that there was an overall 14.1 % treatment failure rate (86 out of 609 patients). However, the overall treatment failure rate for patients receiving a combination of artesunate and sulfadoxine-pyrimethamine (ACT) treatment was only 3.3 % as compared to a 18.7% failure rate for patients receiving sulfadoxine-pyrimethamine (SP) treatment. A Fisher's Exact test revealed that there was a strong association between the treatment administered

to a patient and treatment outcome. It can also be seen that there was a strong association between treatment outcome and the presence or absence of the quintuple mutations. There also appears to be an association between treatment outcome and patient gender, with males being seen to have a higher treatment failure rate as compared to females. The prevalence of moderate anaemia can be seen to not have an association with treatment outcome. Kruskal-Wallis tests revealed that all the distributions of the continuous covariates were significantly different across treatment outcomes.

Table 3: Descriptive statistics for the covariates considered in this analysis by treatment outcome.

Variable		Failure	LTFU	Success	P-value
Treatment (*trt*)	SP	80 (*19%*)	23 (*5%*)	325 (*76%*)	<0.001
	ACT	6 (*3%*)	22 (*12%*)	153 (*85%*)	
Quintuple mutation (*mut*5)	N	47 (*10%*)	36 (*7%*)	410 (*83%*)	<0.001
	Y	39 (*34%*)	9 (*8%*)	68 (*59%*)	
Moderate anaemia (*anaemia*)	N	47 (*13%*)	23 (*6%*)	289 (*81%*)	0.316
	Y	39 (*16%*)	22 (*9%*)	189 (*76%*)	
Gender (*gender*)	F	38 (*11%*)	24 (*7%*)	278 (*82%*)	0.053
	M	48 (*18%*)	21 (*8%*)	200 (*74%*)	
Baseline asexual parasitaemia (*pzero*)	Median	14.42	12.89	13.50	0.002
	IQ Range	13.17 ; 15.64	10.19 ; 15.21	10.63 ; 15.37	
Parasite reduction ratio at 24hours (*ratio*)	Median	7.4%	45.4%	31.8%	<0.001
	IQ Range	-5.0% ; 33.5%	-2.1% ; 100%	4.7% ; 100%	
$\log_2$ age in years (*lage*)	Median	3.17	3.59	3.64	0.006
	IQ Range	1.69 ; 4.25	2.32 ; 4.59	2.59 ; 4.64	

Row percentages are provided in brackets

Fisher's Exact test were applied to categorical variables

Kruskal Wallis test were applied to continuous variables

Figure 2 provides boxplots for the distribution of continuous covariates, analyzed in this thesis, by treatment outcome. From this plot it can be seen that, as compared to patients who achieved an adequate clinical and parasitological response, patients who experienced treatment failure were

- generally younger
- had higher baseline asexual parasite densities
- cleared the smallest percentage of baseline asexual parasites in the first 24 hours.

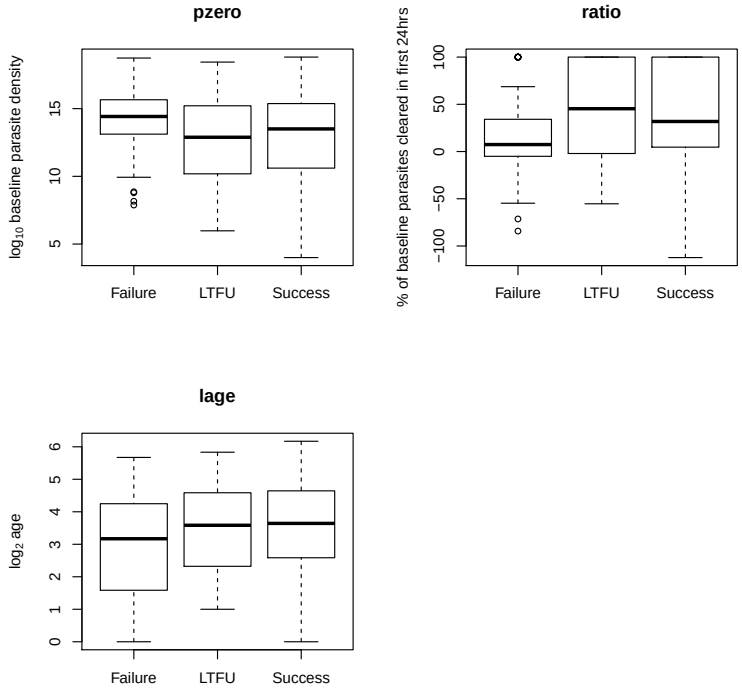

Figure 2: Boxplots of continuous covariates by treatment outcome

The probability that a patient survived without experiencing either treatment failure or loss-to-follow-up was assessed graphically using Kaplan-Meier survival plots shown in Figure 3. In the derivation of these plots patients who were lost-to-follow-up were conservatively combined with patients who experienced treatment failure. It is evident that the highest survival rates are associated with female patients, patients receiving a combination of artesunate and

sulfadoxine-pyrimethamine (ACT) treatment, patients with less than 5 mutations and patients with a haemoglobin density greater than 11g/dL. It is important to note that there was a marginal difference between the survival probabilities of patients receiving sulfadoxine-pyrimethamine treatment only (SP) and patients receiving artesunate and sulfadoxine-pyrimethamine combination treatment (ACT). The reason for this weak effect is shown in Figure 4 and Figure 5. These figures show the Kaplan-Meier survival plots when treatment failure and loss-to-follow-up are considered as separate events.

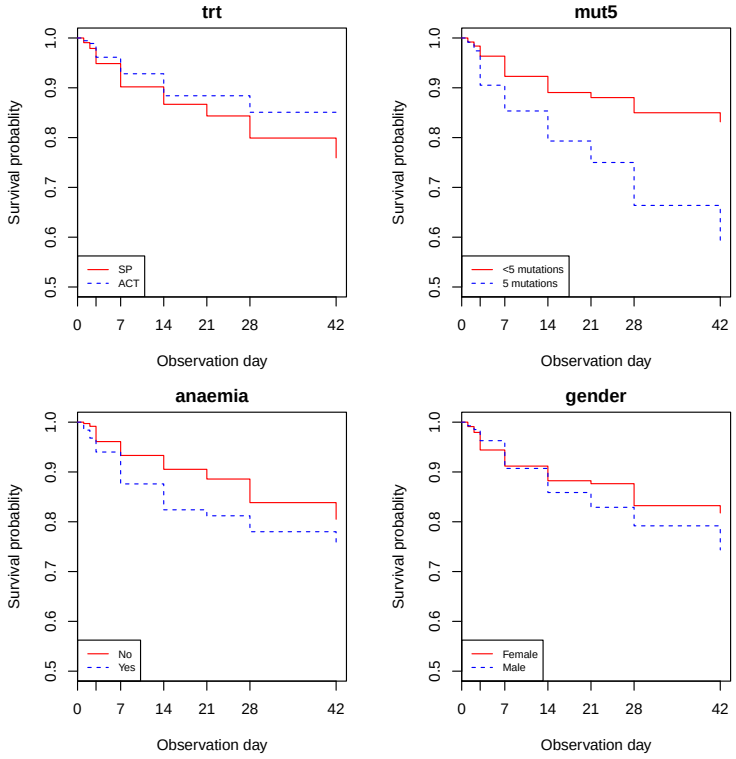

Figure 3: Kaplan-Meier survival plots, by categorical covariates, when treatment failure and loss-to-follow-up were combined and considered as the event of interest

In deriving Figure 4, loss-to-follow-up events were considered as censored observations while in Figure 5 treatment failures were considered as censored observations. It is evident that a patient receiving sulfadoxine-pyrimethamine (SP) treatment is more likely to experience treatment failure and less likely to be lost-to-follow-up, as compared to a patient receiving a combination of artesunate and sulfadoxine-pyrimethamine (ACT) treatment. It is interesting to note that treatment appears to be the only categorical covariate that has an

association with the probability of surviving loss-to-follow-up.

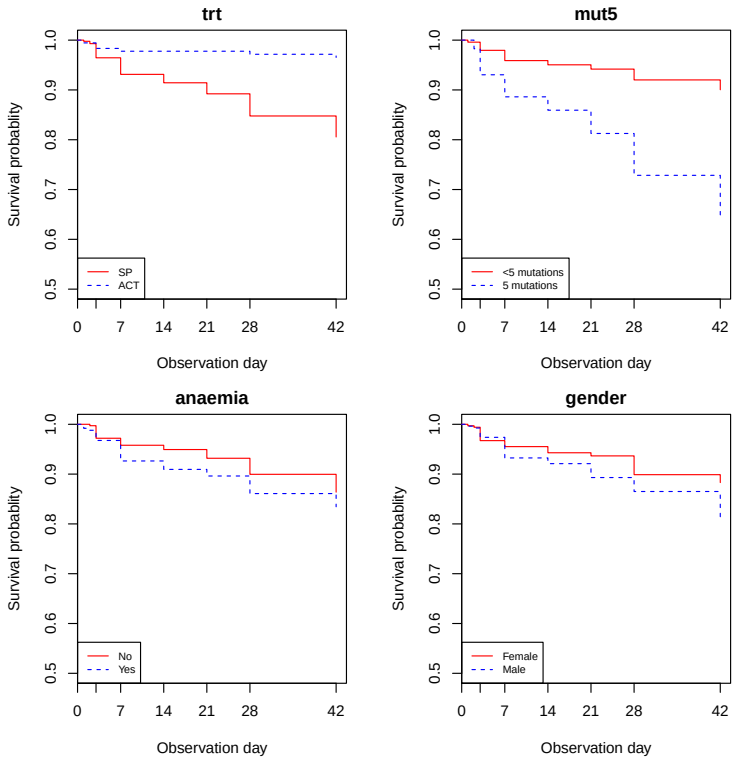

Figure 4: Kaplan-Meier survival plots, by categorical covariates, when treatment failure was considered as the event of interest

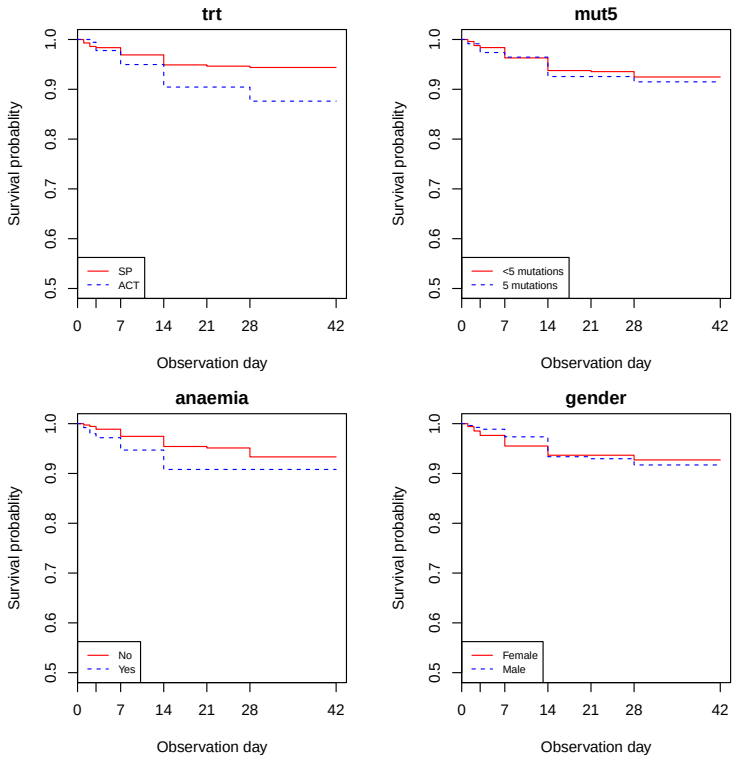

Figure 5: Kaplan-Meier survival plots, by categorical covariates, when loss-to-follow-up was considered as the event of interest

2.2 Gametocyte prevalence

The association between gametocyte prevalence and the covariates considered in this analysis, is summarized in Table 4. It can be seen that 83% of patients who received a combination of artesunate and sulfadoxine-pyrimethamine (ACT) treatment did not develop any gametocytes. In comparison only 51% of patients receiving sulfadoxine-pyrimethamine (SP) treatment did not develop any gametocytes. A Fisher's Exact test reveals that there is a strong association between treatment and the prevalence of gametocytemia. It is also shown that mutation prevalence (*mut*5) and age (*lage*) do not have any association with the prevalence of gametocytemia. It can be seen that patient gender (*gender*) and the prevalence of moderate anaemia (*anaemia*) appear to have some association with gametocyte prevalence. Male patients can be seen to have a higher prevalence of gametocytemia as compared to females. In addition, patients with moderate anaemia have a higher prevalence of gametocytemia as compared to patients with a haemoglobin density greater than 11g/dL. Table 4 also reveals

that the prevalence of gametocytemia is strongly associated with treatment outcome, with the majority of patients who are lost-to-follow-up being seen to not have experienced gametocytemia.

Table 4: Descriptive statistics for the prevalence of gametocytemia.

Variable		Absent	Present	P-value
Treatment (*trt*)	SP	219 (*51%*)	209 (*49%*)	<0.001
	ACT	151 (*83%*)	30 (*17%*)	
Quintuple mutation (*mut5*)	N	307 (*62%*)	186 (*38%*)	0.139
	Y	63 (*54%*)	53 (*46%*)	
Gender (*gender*)	F	220 (*65%*)	120 (*35%*)	0.030
	M	150 (*56%*)	119 (*44%*)	
Moderate anaemia (*anaemia*)	N	230 (*64%*)	129 (*36%*)	0.052
	Y	140 (*56%*)	110 (*44%*)	
Treatment outcome	Failure	39 (*45%*)	47 (*55%*)	<0.001
	LTFU	36 (*80%*)	9 (*20%*)	
	Success	295 (*62%*)	183 (*38%*)	
Baseline asexual parasitaemia (*pzero*)	Median	12.86	14.43	<0.001
	IQ Range	9.68 ; 14.95	13.01 ; 15.79	
Parasite reduction ratio at 24hours (*ratio*)	Median	47.7%	10.0%	<0.001
	IQ Range	8.7% ; 100.0%	-2.2% ; 41.2%	
$\log_2$ age in years (*lage*)	Median	3.700	3.459	0.087
	IQ Range	2.585 ; 4.644	2.161 ; 4.459	
Row percentages are provided in brackets				
Fisher's Exact test were applied to categorical variables				
Kruskal Wallis test were applied to continuous variables				

Figure 6 provides boxplots for the distribution of continuous covariates, by the prevalence of gametocytemia. This figure reveals that patients with a high baseline asexual parasite density generally exhibit gametocytemia more often than patients with a low baseline asexual parasite density. Figure 6 also reveals that patients with a low first 24 hour parasite reduction ratio are more likely to experience gametocytemia as compared to patients with a high first 24 hour parasite reduction ratio. It is evident from this figure that patient age does not appear to have an association with the prevalence of gametocytemia as the age distributions for the patients, who either experienced or did not experience gametocytemia, are similar.

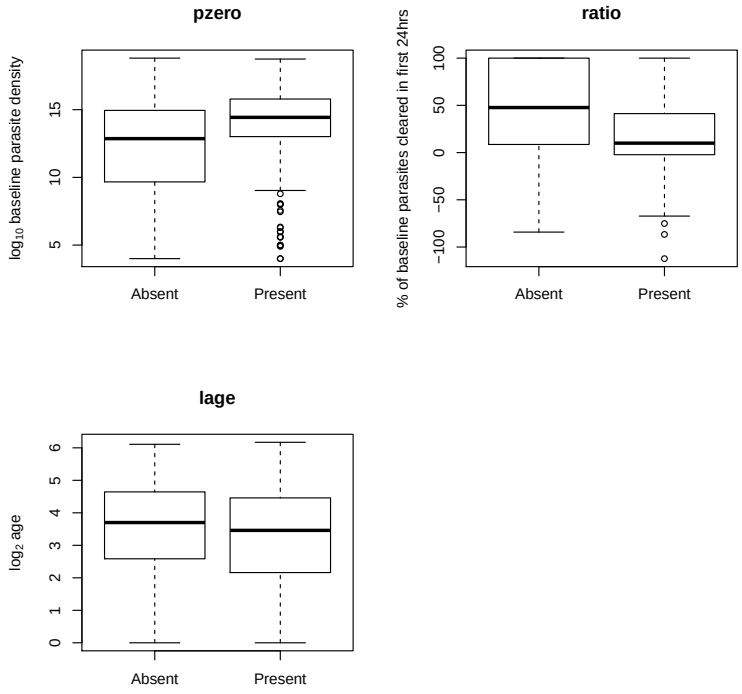

Figure 6: Boxplots for the continuous covariates, by gametocyte prevalence

Table 5 shows that gametocytes were present in only 538 of the 3771 observations recorded during the study, with these observations arising from only 239 patients (Table 6). Table 5 reveals that the incidence rate for gametocytemia peaks between days 7 and 21. This table also highlights that there is a nonlinear relationship between gametocyte prevalence and time.

Table 5: Gametocyte presence over time.

Day	Gametocytes absent	Gametocytes present	Total observations	Incidence rate
0	609	0	609	0.000
3	524	58	582	0.100
7	427	141	568	0.248
14	398	136	534	0.255
21	388	103	491	0.210
28	413	78	491	0.159
42	474	22	496	0.044
Total	3233	538	3771	0.143

Table 6: Distribution of the number of gametocytes observed per patient during the study.

Number of occasions gametocytes were observed	Number of patients	Percentage of patients
0	370	60.8%
1	97	15.9%
2	51	8.4%
3	43	7.1%
4	31	5.1%
5	16	2.6%
6	1	0.2%
Total	609	100.0%

2.3 Gametocyte density

Gametocyte density measurements were taken microscopically, by counting the number of observed gametocytes on thick blood smears contrasted against 1,000 leukocytes, assuming 8,000 leukocytes per microlitre (μL). The implication of the data collection method was that the lowest detectable density was 8 per μL. For the purposes of this analysis values below this limit were considered as being zero. In addition the response variable that will be considered for analysis in this investigation, will be the logarithm to the base two of the observed number of gametocytes. Taking the logarithm of a long-tailed measurement is a commonly applied method of analysis, as it improves the symmetry of the

underlying distribution that generally leads to an acceleration in the time to Bayesian model convergence. The implication is that a one unit increase in the logarithm to the base two values of the gametocyte density will be equivalent to the doubling of the gametocyte count on the original scale.

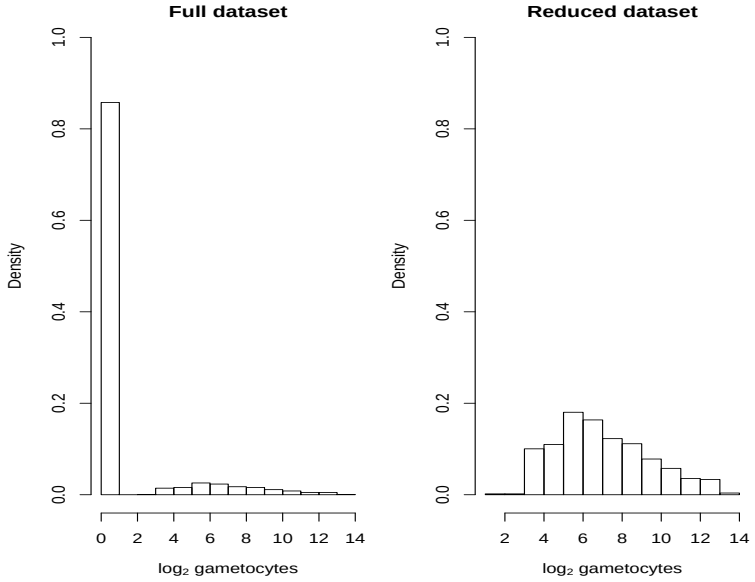

Figure 7: Distribution of $\log_2$ gametocyte density for the full dataset and the reduced dataset, where only observations with gametocytes recorded were considered

Figure 7 shows the distribution of $\log_2$ gametocytes for all patients considered in this analysis. It can be seen that the data is zero-inflated as over 80% of the observations had zero values. Such data can be analyzed using a mixture distribution that models the prevalence of gametocytemia using a logistic model and assigns a continuous distribution to the logged nonzero gametocyte measurements.

Figure 7 indicates that the distribution of the logged nonzero gametocyte measurements are still right-skewed. The *fitdistrplus* package (Delignette-Muller et al., 2014) in R Core Team (2015) was applied to determine the distribution of the nonzero component of the gametocyte data, using graphical assessment. Two distributions, the log-normal distribution and the gamma distribution, were considered as candidates. These distributions were considered because they are able to accommodate long-tailed continuous data. Figure 8 provides a graphical fit of these two distributions to the nonzero component of the data. It can be seen from this figure that there is very little difference, with

regards to fit, between the two distributions. Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) statistics were used to assess the goodness of fit of these distributions. The results of these tests are shown in Table 7. It can be seen that the gamma distribution is marginally better than the log-normal distribution. As a result the gamma distribution was applied to the logged nonzero gametocyte measurements, when a mixture distribution was used as a means to handle the zero-inflation in the data.

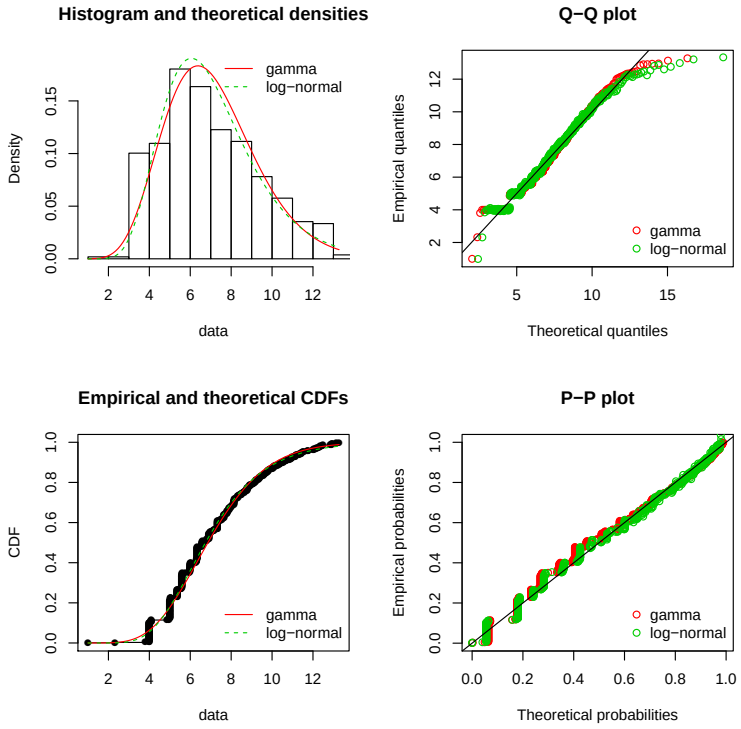

Figure 8: Graphical goodness of fit tests for the log-normal and gamma distributions, to the non-zero gametocyte densities

Table 7: Goodness of fit test for the log-normal and gamma distributions.

Distribution	Log-normal	Gamma
AIC	2384	2378
BIC	2392	2386

Figure 9 illustrates the mean profiles across the categorical covariates considered in this analysis. It can been seen that the mean profile for patients receiving

a combination of artesunate and sulfadoxine-pyrimethamine (ACT) treatment was considerably lower than for those receiving sulfadoxine-pyrimethamine (SP) treatment only. It can also be seen that the gametocyte density profile over time across all covariates follows a non-linear trajectory, which peaks between days 7 and 21. The implication of these findings is that a nonlinear mixed effect model would need to be applied during the course of this investigation.

In the next section the results of an analysis into the time-to-event processes that were generated during this study will be presented.

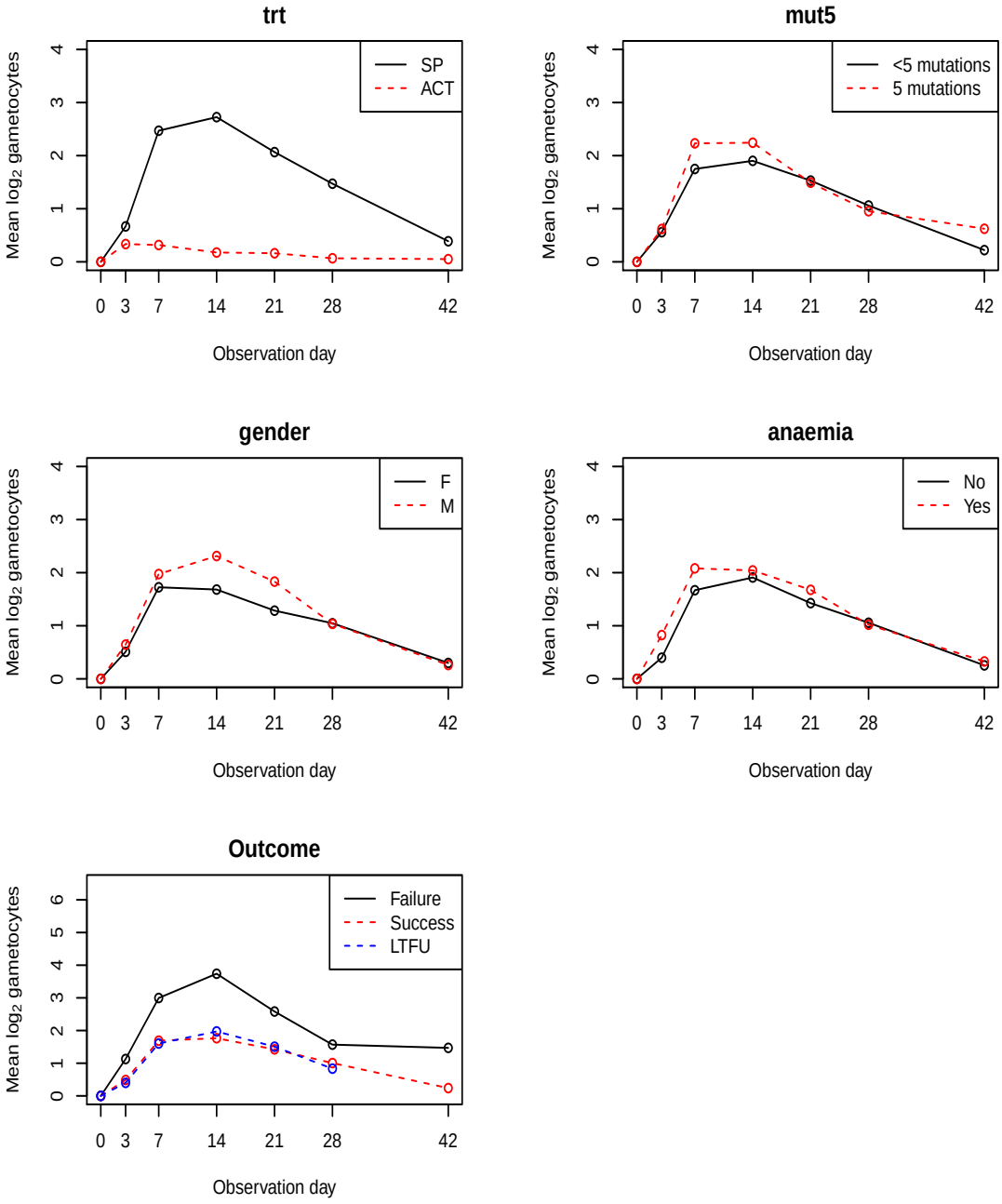

Figure 9: Gametocyte mean profiles by categorical covariates

3 Survival Models

The aim of many clinical trials is to investigate the time until a predefined event occurs and to determine significant risk factors that have an impact on this observed time. In the context of this analysis, the time-to-event processes can be considered as the time to early exit from the study, time to gametocyte emergence or time to gametocyte clearance. Throughout the remainder of this section the time to early exit from the study will be referred to as the event of interest.

Historically survival analysis was used to estimate the probability of survival. Popular non-parametric estimators of the probability of survival are the Kaplan-Meier (Kaplan and Meier, 1958) and the Nelson-Aalen (Aalen, 1976) estimators. Another area of interest is the statistical modeling of survival data. This approach allows for the impact of multiple risk factors on the survival process to be analyzed simultaneously.

In survival data analysis, the time-to-event (T) random variable is defined as strictly positive with a range of $(0, \infty)$. As a result this random variable usually has a positively skew distribution. This skewness makes the assumption of normality, required for certain statistical methods, invalid.

A key characteristic of survival data is censoring. Censoring occurs when the event of interest is not observed for an individual. An example of censoring is when the study comes to an end before the event of interest is observed. Censoring results in incomplete data, as not all times to the event of interest are recorded for individuals in the study. Failure to account for this censoring results in biased estimates. Inferences from incomplete data are also more sensitive to the misspecification of the distribution of survival times as compared to complete data (Rizopoulos, 2012). As a result, the use of standard statistical methods on survival data is unreasonable.

There are several types of censoring, namely

- Right censoring - The event of interest is only known to have occurred after a given timepoint
- Left censoring - The event of interest is only known to have occurred before a given timepoint
- Interval censoring - The event of interest is only known to have occurred between two specified timepoints. This type of censoring is thus a combination of left and right censoring

In the presence of censoring, survival data consists of two components. These

components are the duration that an individual was at risk in a study (T) and an indicator of whether the event of interest occurred (δ). The indicator component takes on a value of 1 if the event of interest occurred and 0 otherwise.

When accounting for censoring it is important to consider whether the probability of censoring depends on the time-to-event process. Informative censoring occurs when the reasons for a patient being censored are related to the time-to-event process. An example of informative censoring is when a patient withdraws from a treatment efficacy study because their health deteriorates and they seek alternative treatment. On the other hand non-informative censoring occurs when the reasons for censoring are not related to the time-to-event process. An implication of non-informative censoring is that the patients who remain in the study have the same risk of hazard as compared to the censored lives.

An area of interest in this analysis is the measurement of the duration of gametocytemia. This involves observing a patient from the time they develop gametocytemia, to the time of gametocyte clearance. Considering only patients who developed gametocytemia; right censoring occurs when either the study ends or a patient exits the study before gametocyte clearance has occurred. In this study, early exit from the study occurs due to either loss-to-follow-up or treatment failure. In this scenario, early exit from the study can be viewed as informative censoring. This is explained by discussing the impact of censoring due to treatment failure and loss-to-follow-up outcomes separately.

Patients who experience treatment failure are rescued from the study and given alternative treatment. These patients can be expected to have asexual parasites in their systems for prolonged periods of time. Assuming that there is a direct correlation between the asexual parasite and the sexual parasite densities, it can be hypothesized that the prevalence of gametocytemia in patients who are rescued from the study is higher than that of patients who remain in the study. This would imply that early exit from the study due to treatment failure is a form of informative censoring.

Patients who are lost-to-follow-up, can be assumed have quickly experienced a successful clinical and parasitological outcome. These patients would thus be more inclined to voluntarily exit the study early, which would imply that these patients would have a lower prevalence of gametocytemia as compared to the patients who remain in the study. Assuming that this hypothesis is correct, loss-to-follow-up would be a form of informative censoring. Additionally, assuming that the hypothesis regarding the health status of patients lost-to-follow-up holds, treatment failure and loss-to-follow-up would potentially be competing risks of early exit from the study. However, it is important to note that the rea-

sons for loss-to-follow-up could potentially be completely unrelated to the study (e.g. accidental death). In such a scenario loss-to-follow-up would be considered as random censoring. This will be investigated further in this chapter.

This section will proceed as follows: firstly an overview of the methodology used in survival analysis will be provided. The techniques that will be discussed, will then be applied to the analysis of the time to gametocyte emergence, time to gametocyte clearance and the time to early exit from the study.

3.1 Overview of survival analysis functions

Several survival functions are introduced in this section. These functions are used to develop the methodology behind the survival modeling techniques implemented in this analysis. The notation and theory used in this section is derived from Collett (1994) and Rizopoulos (2012). The event of interest will be considered as the time to early exit from the study.

The random variable for the time to early exit from the study is given as T, with t being the observed time. T is considered as being a non-negative continuous variable. This variable is assumed to have a probability density function denoted by $f(t)$. The resulting cumulative distribution function $F(t)$ is defined as the probability of exit occurring before time t and it is defined as

$$F(t) = Pr(T \leq t) = \int_0^t f(s)ds. \tag{1}$$

On the other hand the survival function $S(t)$ is the probability of exit not occurring by time t. The survival function is given as

$$S(t) = Pr(T > t) = \int_t^\infty f(s)ds. \tag{2}$$

The survival function is a non-increasing function, over time, which is bounded over the range $[0, 1]$ with the probability of survival at time 0 given as 1.

The hazard function gives the instantaneous risk of exit over the time interval $[t, t + \delta t)$, given that the individual survived (exit did not occur) to time t. The hazard function is defined as

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{Pr(t \leq T < t + \delta t | T \geq t)}{\delta t} \right\}, \tag{3}$$

with $h(t)\delta t$ being the approximate probability that the event of interest occurs in the small interval $[t, t + \delta t)$. The relationship between the hazard function and the survival function is given as

$$\begin{aligned} S(t) &= \exp\{-\int_0^t h(s)ds\} \\ &= \exp\{-H(t)\}, \end{aligned} \tag{4}$$

where $H(t)$ is the cumulative hazard function.

The duration until a patient exits the study can be estimated by taking the area under the survival curve. This implies that the duration (s) is given as

$$s = \int_0^\infty S(t)dt. \tag{5}$$

3.2 Modeling the hazard function

Statistical modeling of the hazard function allows for the impact of multiple risk factors on the survival process to be simultaneously assessed. The resulting models allow for the derivation of individual hazard functions for patients, based on their covariate patterns. Insights from these models can provide guidance to clinicians with regards to the appropriateness of treatment protocols and medical intervention strategies. In this section semi-parametric and parametric hazard models will be discussed.

3.2.1 Semi-parametric proportional hazards models

Given that n patients are observed in a treatment efficacy study, with the hazard of exit for the i^{th} patient depending on a set of p observed covariates $\mathbf{x}_i^T = (x_{i1}, \ldots, x_{ip})$, the hazard function under the Cox proportional hazards model framework (Cox, 1972) is described as

$$h_i(t) = h_0(t) \exp(\mathbf{x}_i^T \boldsymbol{\beta}), \tag{6}$$

where $\boldsymbol{\beta}$ is a p dimensional vector of fixed effect coefficients and $h_0(t)$ is the baseline hazard at time t. The baseline hazard is the underlying hazard of exit when all covariates are set to 0. Equation 6 can be rearranged as

$$\begin{aligned} \frac{h_i(t)}{h_0(t)} &= \exp(\mathbf{x}_i^T \boldsymbol{\beta}) \\ \log \frac{h_i(t)}{h_0(t)} &= \mathbf{x}_i^T \boldsymbol{\beta}, \end{aligned} \tag{7}$$

where $\frac{h_i(t)}{h_0(t)}$ is defined as the hazard ratio that represents the relative change in the hazard of exit due to the predictor variables in the fitted model. The logarithm of the hazard ratio is seen to have linear relationship with the predictor variables in the fitted model. The model is semi-parametric in nature as its coefficients ($\boldsymbol{\beta}$) can be estimated without needing a distributional assumption for $h_0(t)$.

Considering the case where a single continuous predictor variable is included in the fitted model, e^β represents a relative change in the hazard of exit due to

a one unit increase in the predictor variable. In the case of a binary indicator variable, specifying the presence (1) or absence (0) of the predictor variable, e^{β} represents the relative change in the hazard of exit due to the presence of the predictor variable. When e^{β} is greater than 1, it implies that there is a relative increase in the hazard of exit occurring.

A key assumption of the Cox proportional hazards model is that any change in the hazard ratio, due to the effect of a particular predictor variable, is constant over time. This is known as the proportional hazards assumption. To illustrate this, consider the SP and ACT treatments defined in Chapter 1. The proportional hazard assumption implies that the following relationship is constant over time

$$h_{ACT}(t) = h_{SP}(t)\psi, \tag{8}$$

where ψ is the ratio of the hazards of exit for a patient receiving ACT treatment relative to SP treatment. This result implies that

$$S_{ACT}(t) = [S_{SP}(t)]^{\psi}. \tag{9}$$

The coefficients of the proportional hazards model, given in equation 6, can be estimated using the method of maximum partial likelihood. In the absence of ties, multiple exits occurring at the same time, the partial likelihood function derived by Cox (1972), for equation 6 is given as

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \Big\{ \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{\sum_{l \in R(t_{(i)})} \exp(\mathbf{x}_l^T \boldsymbol{\beta})} \Big\}^{\delta_i}, \tag{10}$$

with δ_i being an indicator variable that has a value of 0 if the i^{th} survival time is right censored and 1 otherwise; $R(t_{(i)})$ is the risk set at time t_i. The resulting partial log-likelihood function, $l(\boldsymbol{\beta})$, is given as

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \Big\{ (\mathbf{x}_i^T \boldsymbol{\beta}) - \log \Big(\sum_{l \in R(t_{(i)})} \exp(\mathbf{x}_l^T \boldsymbol{\beta}) \Big) \Big\}. \tag{11}$$

Numerical methods like the Newton-Raphson procedure (Lange, 2004) can be used to obtain maximum likelihood estimates for the $\boldsymbol{\beta}$ coefficients given in equation 11.

The appropriateness of applying the Cox proportional hazards model, to a dataset, depends on the validity of the proportional hazards assumption. This assumption should thus be validated before fitting the Cox model. In this thesis a graphical assessment of the validity of the proportional hazards assumption will be employed. A plot of the cumulative hazard function against time can be used to test this assumption for each covariate used. This can be illustrated by

considering a model with only one covariate. The cumulative hazard function for this model can be derived as

$$\begin{aligned} h_i(t) &= h_0(t)e^{\beta x_i} \\ \int_0^t h_i(s) &= e^{\beta x_i} \int_0^t h_0(s)ds \\ H_i(t) &= e^{\beta x_i} H_0(t) \\ \log H_i(t) &= \beta x_i + \log H_0(t) \end{aligned} \tag{12}$$

This can be rearranged to give

$$\beta x_i = \log H_i(t) - \log H_0(t).$$

It is evident that the difference between the log-cumulative hazard plots of patients with different covariate patterns does not depend on time. Given two patients with covariates values of x and $x + 1$, the difference between their log cumulative plots will be constant over time with a value of β.

3.2.2 Parametric proportional hazards models

The Cox proportional hazards model does not assume a probability distribution for the survival times. As a result the baseline hazard is unspecified, with the shape of the function defined by the underlying data used in the analysis. This leads to a flexible hazard function that can be applied in a wide range of scenarios. Models that assign a distribution to the baseline hazard function are called parametric proportional hazards models. Inferences from these models are more precise as compared to those from the Cox model. In addition the standard errors from these models are smaller than those from the Cox model. Two popular parametric proportional hazards models are the Weibull and the exponential models. These models assume that the survival times (t), defined over the range $0 \leq t < \infty$, follow Weibull and exponential distributions respectively. The underlying survival functions for these parametric models are given in Table 8.

Table 8: Exponential and Weibull survival functions.

Function	Exponential	Weibull
$f(t)$	$\lambda \exp(-\lambda t)$	$\lambda\rho t^{\rho-1} \exp(-\lambda t^{\rho})$
$S(t)$	$\exp(-\lambda t)$	$\exp(-\lambda t^{\rho})$
$h(t)$	λ	$\lambda\rho t^{\rho-1}$

The exponential parametric model assumes that the baseline hazard function is constant over time. This implies that once the study has began, the hazard of exit remains constant over time.

The Weibull parametric model allows the baseline hazard function to depend on two strictly positive parameters λ and ρ, where λ is the scale parameter and ρ is the shape parameter. The baseline hazard function can be seen to simplify to that of the exponential model when $\rho = 1$. In the case when $\rho \neq 1$, the hazard function is monotonically increasing (when $\rho > 1$) or decreasing (when $\rho < 1$) over time.

Equation 6 can be expanded to assume that survival times of the n patients observed in the study, follow a Weibull distribution. As a result the baseline hazard function can be defined as

$$h_0(t) = \lambda\rho t^{\rho-1}.$$

The definition of the proportional hazards model can be extended as follows

$$\begin{aligned} h_i(t) &= h_0(t)\exp(\mathbf{x}_i^T\boldsymbol{\beta}) \\ &= \lambda\rho t^{\rho-1}\exp(\mathbf{x}_i^T\boldsymbol{\beta}). \end{aligned} \tag{13}$$

It is evident that $h_i(t)$ also follows a Weibull distribution with scale parameter $\lambda\exp(\mathbf{x}_i^T\boldsymbol{\beta})$ and shape parameter ρ. This result verifies that the proportional hazards property is upheld for the Weibull distribution. It follows that the resulting survival function is given as

$$S_i(t_i) = \exp\{-\lambda t_i^{\rho}\exp(\mathbf{x}_i^T\boldsymbol{\beta})\}.$$

The likelihood function for the Weibull proportional hazards model is given as

$$\begin{aligned} L(\boldsymbol{\beta},\rho,\lambda) &= \prod_{i=1}^{n}\{h_i(t_i)\}^{\delta_i}S_i(t_i) \\ &= \prod_{i=1}^{n}\{\lambda\rho t_i^{\rho-1}\exp(\mathbf{x}_i^T\boldsymbol{\beta})\}^{\delta_i}\exp\{-\lambda t_i^{\rho}\exp(\mathbf{x}_i^T\boldsymbol{\beta})\} \end{aligned} \tag{14}$$

where δ_i is an indicator variable having a value of 1 if exit occurs and 0 otherwise. The parameters $\boldsymbol{\beta}, \rho$ and λ are estimated by maximizing the logarithm of the likelihood function, $l(\boldsymbol{\beta},\rho,\lambda)$, which is given as

$$\begin{aligned} l(\boldsymbol{\beta},\rho,\lambda) &= \sum_{i=1}^{n}[\delta_i\log\{h_i(t_i)\} + \log S_i(t_i)] \\ &= \sum_{i=1}^{n}[\delta_i\{\mathbf{x}_i^T\boldsymbol{\beta} + \log(\lambda\rho) + (\rho-1)\log t_i\} - \lambda\exp(\mathbf{x}_i^T\boldsymbol{\beta})t_i^{\rho}]. \end{aligned} \tag{15}$$

The Newton-Raphson algorithm can be used to maximize the log-likelihood function given in Equation 15. Setting $\rho = 1$, gives the corresponding result for the exponential model.

3.2.3 Accelerated failure time models

Parametric and semi-parametric proportional hazards models depend on the validity of the proportional hazards assumption. In situations where this assumption is not valid, accelerated failure times (AFT) models can be used. AFT models are able to accommodate a larger pool of survival time distributions as compared to the proportional hazards models. In this section the general form of the AFT model will be provided along with the specific formulations for the exponential and Weibull models.

Collett (1994) provides the following log-linear representation of the AFT models

$$\begin{aligned} \log T_i &= \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \cdots + \alpha_p x_{ip} + \sigma\epsilon_i \\ &= \alpha_0 + \mathbf{x}_i^T \boldsymbol{\alpha} + \sigma\epsilon_i, \end{aligned} \tag{16}$$

where T_i is the time to early exit for the i^{th} patient, x_{ij} is the j^{th} covariate recorded for the i^{th} patient, $\alpha_1, \ldots, \alpha_p$ are unknown coefficients, α_0 is the value of $\log T_i$ when all covariates are equal to 0 (baseline value of $\log T_i$), σ is the scale parameter and ϵ_i is the residual error that is assumed to follow a particular probability distribution. The survival function for the i^{th} individual is given as

$$\begin{aligned} S_i(t) &= Pr[T_i \geq t] \\ &= Pr[e^{(\alpha_0 + \mathbf{x}_i^T\boldsymbol{\alpha} + \sigma\epsilon_i)} \geq t] \\ &= Pr[e^{(\alpha_0+\sigma\epsilon_i)} \times e^{\mathbf{x}_i^T\boldsymbol{\alpha}} \geq t] \\ &= Pr[e^{(\alpha_0+\sigma\epsilon_i)} \geq te^{-\mathbf{x}_i^T\boldsymbol{\alpha}}] \\ &= S_0(te^{-\mathbf{x}_i^T\boldsymbol{\alpha}}) \\ &= S_0(t\phi^{-1}), \end{aligned} \tag{17}$$

where $\phi = e^{\mathbf{x}_i^T\boldsymbol{\alpha}}$ and $S_0(.)$ is the baseline survival function. From Equation 17 it can be seen that the survival probability of the i^{th} individual at time t is equivalent to the baseline survival probability at time $t\phi^{-1}$. This relationship is referred to as the accelerated failure time property that defines this class of models. The factor ϕ is known as the acceleration factor. If $\phi < 1$ then the time to early exit from the study is accelerated. If $\phi > 1$ then the time to early exit from the study is decelerated. The general AFT hazard function can be derived from Equation 17 as follows

$$\begin{aligned} S_i(t) &= S_0(te^{-\mathbf{x}_i^T\boldsymbol{\alpha}}) \\ -\log[S_i(t)] &= -\log[S_0(te^{-\mathbf{x}_i^T\boldsymbol{\alpha}})] \\ -\frac{d}{dt}\log[S_i(t)] &= -\frac{d}{dt}\log[S_0(te^{-\mathbf{x}_i^T\boldsymbol{\alpha}})] \\ h_i(t) &= e^{-\mathbf{x}_i^T\boldsymbol{\alpha}}[h_0(te^{-\mathbf{x}_i^T\boldsymbol{\alpha}})]. \end{aligned} \tag{18}$$

The Weibull and exponential models possess both the accelerated failure time and proportional hazards properties. The Weibull model will be used to illustrate this. Consider n survival times that are assumed to follow a Weibull distribution, with a baseline hazard function of

$$h_0(t) = \lambda\rho t^{\rho-1}.$$

Using the result from Equation 18, the hazard function for the i^{th} patient is given as

$$\begin{aligned} h_i(t) &= e^{-\mathbf{x}_i^T\boldsymbol{\alpha}} h_0(te^{-\mathbf{x}_i^T\boldsymbol{\alpha}}) \\ &= e^{-\mathbf{x}_i^T\boldsymbol{\alpha}} \lambda\rho(te^{-\mathbf{x}_i^T\boldsymbol{\alpha}})^{\rho-1} \\ &= e^{-\mathbf{x}_i^T\boldsymbol{\alpha}\rho} \lambda\rho t^{\rho-1} \end{aligned} \tag{19}$$

Under the proportional hazards (PH) framework, the hazard function for the i^{th} patient is given as

$$\begin{aligned} h_i(t) &= h_0(t)e^{(\mathbf{x}_i^T\boldsymbol{\beta})} \\ &= \lambda\rho t^{\rho-1} e^{(\mathbf{x}_i^T\boldsymbol{\beta})}, \end{aligned} \tag{20}$$

where $\beta_1, \ldots, \beta_p$ are unknown coefficients estimated under the PH framework. The $\boldsymbol{\beta}$ coefficients (estimated under for the PH model) can be equated to the $\boldsymbol{\alpha}$ coefficients (estimated under the AFT framework), by using the relationship

$$\boldsymbol{\beta} = -\boldsymbol{\alpha}\rho.$$

The Weibull model can thus be seen to possess both the proportional hazard and the accelerated failure time properties. This also holds for the exponential model, as it is just a special case of the Weibull model with $\rho = 1$.

The Weibull and exponential models are used to model monotonically increasing or decreasing hazard functions. If the hazard function is not monotonically increasing or decreasing over time then the log-logistic, log-normal and generalized-gamma models can be used. The density, survival and hazard functions of the log-logistic model are

$$f(t) = \frac{\lambda^{\frac{1}{\gamma}} t^{\frac{1}{\gamma}-1}}{\gamma\left\{1 + (\lambda t)^{\frac{1}{\gamma}}\right\}^2},$$

$$S(t) = \{1 + (\lambda t)^{\frac{1}{\gamma}}\}^{-1}\}$$

and

$$h(t) = \frac{\frac{1}{\gamma}\lambda^{\frac{1}{\gamma}} t^{\frac{1}{\gamma}-1}}{\{1 + (\lambda t)^{\frac{1}{\gamma}}\}}. \tag{21}$$

The density, survival and hazard functions of the log-normal model are

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\Big[-\frac{1}{2\sigma^2}\{\log(t) - \mu\}^2\Big],$$

$$S(t) = 1 - \Phi\Big\{\frac{\log(t) - \mu}{\sigma}\Big\}$$

and

$$h(t) = \frac{\frac{1}{t\sigma\sqrt{2\pi}} \exp\Big[-\frac{1}{2\sigma^2}\{\log(t) - \mu\}^2\Big]}{1 - \Phi\Big\{\frac{\log(t)-\mu}{\sigma}\Big\}}, \tag{22}$$

where $\Phi(z)$ is the standard normal cumulative distribution function.

The exponential, log-normal and Weibull distributions are all special cases of the generalized-gamma distribution (Collett, 1994). As a result the generalized-gamma distribution can be used to test for the functional form of a standard parametric survival model. The probability density function of the generalized-gamma distribution is given as

$$f(t) = \frac{\theta\lambda^{\gamma\theta}t^{\gamma\theta-1}\exp\{-(\lambda t)^{\theta}\}}{\Gamma(\gamma)},$$

for $0 \leq t < \infty$, where $\lambda > 0$ is the scale parameter; $\gamma > 0$ and $\theta > 0$ are shape parameters, with $\Gamma(x)$ as the gamma function defined as

$$\Gamma(x) = \int_0^{\infty} s^{x-1}e^{-s}ds.$$

The corresponding survival function is given as

$$S(t) = 1 - I\{\gamma, (\lambda t)^{\theta}\},$$

where $I\{\gamma, (\lambda t)^{\theta}\}$ is the incomplete gamma function given as

$$I\{\gamma, (\lambda t)^{\theta}\} = \frac{1}{\Gamma(\gamma)} \int_0^{(\lambda t)^{\theta}} u^{\gamma-1}e^{-u}du.$$

The corresponding hazard function is given as

$$f(t) = \frac{[\theta\lambda^{\gamma\theta}t^{\gamma\theta-1}\exp\{-(\lambda t)^{\theta}\}/\Gamma(\gamma)]}{1 - I\{\gamma, (\lambda t)^{\theta}\}} \tag{23}$$

Prentice (1974) highlighted that the above parametrization of the generalized-gamma distribution led to log-likelihood equations that did not always have solutions. He thus proposed an alternative parametrization of this distribution. This alternative approach was applied in this thesis. The notation used to describe this approach is consistent with the notation used in the *flexsurv*

package (Jackson, 2016) of the R statistical program (R Core Team, 2015). The properties of this function are

$$f(t) = \begin{cases} \frac{\gamma^\gamma}{\sigma t \Gamma(\gamma)\sqrt{\gamma}} \exp(z\sqrt{\gamma} - u) & \text{if } q \neq 0 \\ \frac{1}{\sigma t \sqrt{2\pi}} \exp(-z^2/2) & \text{if } q = 0 \end{cases},$$

$$S(t) = \begin{cases} 1 - I(\gamma, u) & \text{if } q > 0 \\ 1 - \Phi(z) & \text{if } q = 0 \\ I(\gamma, u) & \text{if } q < 0 \end{cases},$$

$$h(t) = \begin{cases} \frac{\frac{\gamma^\gamma}{\sigma t \Gamma(\gamma)\sqrt{\gamma}} \exp(z\sqrt{\gamma} - u)}{1 - I(\gamma, u)} & \text{if } q > 0 \\ \frac{\frac{1}{\sigma t \sqrt{2\pi}} \exp(-z^2/2)}{1 - \Phi(z)} & \text{if } q = 0 \\ \frac{\frac{\gamma^\gamma}{\sigma t \Gamma(\gamma)\sqrt{\gamma}} \exp(z\sqrt{\gamma} - u)}{I(\gamma, u)} & \text{if } q < 0 \end{cases},$$

where

- $I(\gamma, u)$ is the incomplete gamma function
- Φ is the standard normal cumulative distribution
- $u = \gamma e^{(|q|z)}$
- $z = \frac{\text{sign}(q)\{\log t - \mu\}}{\sigma}$
- $\gamma = |q|^{-2}$

The special cases of the generalized-gamma distribution are outlined in Table 9.

Table 9: Special cases of the generalized-gamma distribution.

Distribution	Parameterization
Exponential	$q = 1$ and $\sigma = 1$
Gamma	$\sigma = \|q\|$
Log-Normal	$q = 0$
Weibull	$q = 1$

3.2.4 Model checking

Once a survival model has been fit to a dataset it is important to assess the adequacy of the model fit. An analysis of the Cox-Snell residuals, from the

fitted model, can be used for this purpose. The Cox-Snell residual for the i^{th} individual, r_{Ci}, is defined as the estimated value of the cumulative hazard function for the i^{th} observation at time t_i (Collett, 1994). The Cox-Snell residual is thus given as

$$\begin{aligned} r_{Ci} &= \hat{H}_i(t_i) \\ &= -\log \hat{S}(t_i). \end{aligned} \tag{24}$$

The Cox-Snell residuals follow a unit exponential distribution. A fitted model is deemed adequate if a plot of the cumulative hazard of the Cox-Snell residuals against the Cox-Snell residuals, produces a straight line through the origin with a gradient of 1. If these conditions are not met, the fitted model is deemed inappropriate and an alternative model is fit to the data.

Martingale (r_{Mi}) and deviance residuals (r_{Di}) are additional residuals that can used to determine model adequacy. Martingale residuals are defined as

$$\begin{aligned} r_{Mi} &= \delta_i - \hat{H}_i(t_i) \\ &= \delta_i - r_{Ci}, \end{aligned} \tag{25}$$

where δ_i is an indicator value taking a value of 1 if the event of interest occurs and 0 otherwise. These residuals are derived using martingale methods outlined in Fleming and Harrington (1991). Martingale residuals have a mean of 0 and a range of $(-\infty\ ,\ 1]$, with censored observations having negative residuals. These residuals can be used to assess the validity of the functional form of covariates, in addition they can be used to identify outliers. A key feature of martingale residuals is that they are asymmetric, which makes them difficult to interpret. Therneau et al. (1990) introduced deviance residuals that are a transformation of martingale residuals. These residuals closely resemble standardized residuals from linear regression in that they are symmetric, have a mean of 0 and a standard deviation of 1. Deviance residuals are defined as

$$r_{Di} = \text{sign}(r_{Mi})\sqrt{-2[r_{Mi} + \delta_i \log(\delta_i - r_{Mi})]}. \tag{26}$$

These residuals are negative when the observed times to event are smaller than expected. If a model is adequate, a plot of deviance residuals against fitted values would not have a systematic trend. Since deviance residuals are derived from martingale residuals, it was deemed sufficient to only consider Cox-Snell and deviance residuals when assessing model adequacy.

3.2.5 Model selection

Model selection involves the comparison of multiple alternative models using various tests. The deviance test is an example of a test that can be applied. It

is used when the models being compared are nested, that is when one model contains explanatory variables that are a subset of these in the alternative model. The deviance test statistic is defined as the difference between the fitted log-likelihood functions of the models being compared. The derived maximum likelihood estimates, from the individual model fits, are used as the parameters for the respective models in the deviance test. To illustrate this, consider two nested models $m1$ and $m2$. Model $m1$ consists of p explanatory variables, while $m2$ consists of $p+q$ explanatory variables. The maximized log-likelihood functions for the models are given as $\hat{l}_1$ and $\hat{l}_2$. The deviance statistic is defined as

$$D = -2\{\log \hat{l}_1 - \log \hat{l}_2\}, \tag{27}$$

with the distribution of D being approximately chi-squared with degrees of freedom equal to q under the null hypothesis that the additional q parameters are equal to 0.

In the event that two non-nested models are being compared, the AIC (Akaike, 1974) statistic can be computed per model. The model with the lower AIC statistic is considered as the best model. The AIC statistic for $m1$ is defined as

$$\text{AIC}_1 = -2\{\log \hat{l}_1\} + 2p. \tag{28}$$

The AIC statistic is designed to penalize complex models as each additional parameter incurs a penalty on the statistic.

3.3 Competing Risks Models

The discussion around survival analysis has focused on the time until a single predefined event, that is time to early exit from the study. As previously stated, early exit from the study can arise either through treatment failure or through loss-to-follow-up. In reality researchers who conduct treatment efficacy studies are mainly interested in the hazard of treatment failure, being lost-to-follow-up prevents researchers from observing treatment failure. As a result loss-to-follow-up can be considered as a competing risk to treatment failure, as it hinders the observation of treatment failure. When analyzing competing risk data a researcher has the choice of either accounting for the competing event or ignoring it. Pintilie (2007) highlighted that ignoring the competing risk can be beneficial as long as the interpretation of the results from the analysis is correct. The decision regarding the treatment of competing risks is driven by the research question on hand. In this study a researcher could seek to answer any one of the following questions:

1. How many more patients experienced treatment failure after taking a combination of artesunate and sulfadoxine-pyrimethamine treatment (ACT) as compared to sulfadoxine-pyrimethamine (SP) treatment only?

2. Is the rate of treatment failure different between the ACT and SP treatments?

Given that a researcher is attempting to solve the first question, consideration must be made for the number of patients who were lost-to-follow-up as they reduce the number of treatment failures that could have been observed. A Fine-Gray competing risk models that makes use of the sub-distribution hazard function (Fine and Gray, 1999, Lau et al., 2009), can be used in this situation. When attention is turned to the second question, the presence of a competing risk is not desirable, thus it is ignored in the analysis. The cause-specific hazard model (Kalbfleisch and Prentice, 2002), would be applied in this scenario. The key difference between the cause-specific and the Fine-Gray model is the definition of the hazard function for the two models.

Given that the random variable for the time to early exit from study is given as T, with t being the observed time. The overall hazard function is as previously defined,

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{Pr(t \leq T < t + \delta t | T \geq t)}{\delta t} \right\}. \tag{29}$$

It can be assumed that early exit from the study may occur due to a set of m distinct causes that are indexed by $j \in \{1, 2, \ldots, m\}$, with J being a random variable that indicates the specific cause of exit. Based on this, the cause-specific hazard rate is given as

$$h_j(t) = \lim_{\delta t \to 0} \left\{ \frac{Pr(t \leq T < t + \delta t, J = j | T \geq t)}{\delta t} \right\}. \tag{30}$$

The cause-specific hazard function can be interpreted as the approximate probability that an individual exits the study in a small interval $[t, t + \delta t)$ due to the j^{th} cause of exit. The overall hazard function, given in equation 29, can be defined as the sum of the m distinct cause-specific hazard functions through the application of the law of total probability,. The relationship is shown as

$$h(t) = \sum_{j=1}^{m} h_j(t). \tag{31}$$

The Fine-Gray model would look at the instantaneous risk of exit from the j^{th} event in patients who have either not exited or who have exited due to any of

the other causes of exit except for cause j. The sub-distribution hazard function is defined as

$$\bar{h}_j(t) = \lim_{\delta t \to 0} \left\{ \frac{Pr(t \leq T \leq t + \delta t, J = j | T \geq t \cup [T \leq t \cap J \neq j])}{\delta t} \right\}. \quad (32)$$

The definition of the sub-distribution hazard function incorporates the use of an "unnatural" risk set, which contains patients who are still in the study and patients who exited the study due to reasons other than j. In order to account for this mixture of lives, weights are applied in the sub-distribution likelihood function to distinguish between patients who are still in the study and those that exited due to reasons other than the j^{th} cause. Details around the application of these weights can be found in the paper by Fine and Gray (1999).

Based on the characteristics of these models, Lau et al. (2009) proposed that the cause-specific model should be used to answer epidemiological questions around the causes of diseases. These authors also proposed that the Fine-Gray model should be used for the development of predictive models for clinical use. This was because the Fine-Gray model allows for the direct modeling of the effect of covariates on the incidence of a particular event, while accounting for competing events.

One of the aims of this thesis is the imputation of incomplete gametocyte profiles in the presence of censoring, which arises due to exit from the study either from loss-to-follow-up or treatment failure. The hazard rates of these two causes of exit are expected to provide information that will be used in the imputation of the incomplete gametocyte profiles. Since interest is in the hazard rates of the two causes of exit, only cause-specific hazard models were considered in this analysis. These models will be expanded upon for the remainder of this section.

3.3.1 Cause-specific hazard model

In the previous section it was shown that the hazard function for the j^{th} cause of exit was

$$h_j(t) = \lim_{\delta t \to 0} \left\{ \frac{Pr(t \leq T < t + \delta t, J = j | T \geq t)}{\delta t} \right\},$$

with the overall hazard function given as

$$h(t) = \sum_{j=1}^{m} h_j(t).$$

It can be seen that the resulting survival function, interpreted as the probability of surviving all m distinct causes of exit up to time t, is defined as

$$\begin{aligned} S(t) &= \exp\{-\int_0^t h(s)ds\} \\ &= \exp\{-H(t)\}; \end{aligned}$$

where $H(t)$ is the cumulative hazard function.

The cause-specific probability density of exit, at time t, is the unconditional risk that a patient exits the study due to the j^{th} cause of exit at time t. This density is defined as

$$\begin{aligned} f_j(t) &= \lim_{\delta t \to 0} \frac{Pr(t \leq T < t + \delta t, J = j)}{\delta t}, \\ &= h_j(t)S(t). \end{aligned} \tag{33}$$

The overall probability density function of exit is given by applying the law of total probability. This overall probability density function is given as

$$f(t) = \sum_{i=1}^{m} f_j(t).$$

The likelihood function for the cause-specific hazard model can be derived from equations 31 and 33. Given that n patients are observed in the study and that the following data is recorded for the i^{th} patient

- t_i the observed exit time
- δ_i an indicator variable with a value of 1 if exit occurred
- j_i the cause of exit, with $j_i \in \{1, 2, \ldots, m\}$ when an exit occurred and undefined otherwise
- x_i covariates recorded for the i^{th} patient.

The likelihood function can be defined as

$$\begin{aligned} L &= \prod_{i=1}^{n} \{h_{ji}(t_i)\}^{\delta_i} S_i(t_i) \\ &= \prod_{i=1}^{n} \{h_{ji}(t_i)\}^{\delta_i} \prod_{j=1}^{m} \exp\{-\int_0^{t_i} h_j(s)ds\}. \end{aligned} \tag{34}$$

The defined likelihood is derived by considering the following probability contributions

- The probability contribution of a patient who is censored at time t_i, thus contributing a probability of $S_i(t_i)$

- The probability contribution of a patient who has exited at time t_i due to cause j, thus contributing a probability of $h_{ji}(t_i)S_i(t_i)$

An indicator variable, δ_{ij} that has a value of 1 when the i^{th} patient exits due to the j^{th} cause and 0 otherwise, can be introduced into equation 34. Given the restriction that a patient can exit from only one cause (if multiple causes occur simultaneously they can be combined into a new composite cause of exit) the following relationship unfolds

$$\delta_i = \sum_{j=1}^{m} \delta_{ij},$$

which simplifies equation 34 to

$$L = \prod_{i=1}^{n} \prod_{j=1}^{m} \{h_{ji}(t_i)\}^{\delta_{ij}} \exp\{-\int_0^{t_i} h_j(s)ds\}. \tag{35}$$

The overall likelihood function can be seen to be the product of m likelihood functions derived for each of the causes of exit. The implication is that $h_j(t)$ can be maximized using separate likelihood functions. Additionally, when considering a specific cause of exit, the likelihood for that cause is exactly the same as the likelihood obtained if the other $m-1$ causes of exit we treated as censored observations. As a result the m likelihood functions can be solved using any of the modeling methods previously outlined in this chapter.

3.4 Interval censoring

The gametocyte data used in this analysis was collected over a 42 day follow up period with observations occurring on days 0, 3, 7, 14, 21, 28 and 42. On these observation days, blood samples were taken from patients. Subsequently gametocyte density measurements were taken microscopically, by counting the number of observed gametocytes on thick blood smears contrasted against 1,000 leukocytes, assuming 8,000 leukocytes per microlitre (μL). Due to the design of the study, it is evident that interval censoring would be expected to arise. This would affect both the time to gametocyte emergence and time to gametocyte clearance survival analyses.

When considering a time to gametocyte emergence analysis, a patient can be observed to not have gametocytes at entry into the study then subsequently on day 3 it can be found that gametocytes have developed in the patient's blood. It is clear that the gametocytes would have developed at some point between day 0 and 3. This is an example of interval censoring.

Table 10 reveals the number of patients who experienced gametocytemia and the respective periods when gametocytemia first occurred. It is evident that gametocytes first emerged within the first 7 days of follow up, in 70% of the patients who experienced gametocytemia. This is the period when the intervals between observation days was the shortest. As a result it can be assumed that interval censoring would have a small impact on the time to gametocyte emergence analysis. As a result interval censoring techniques were not considered for that analysis.

Table 10: Number of patients who experienced gametocytemia, by period in which gametocytemia occurred.

Interval	(0 , 3]	(3 , 7]	(7 , 14]	(14 , 21]	(21 , 28]	(28 , 42]	Total
Number of patients	58	110	42	14	13	2	239

Table 11 reveals that 89% of the gametocyte clearances, in patients who were able to clear their gametocyte infection during the course of the study, were observed after day 7. This coincides with the period that had the longest intervals between observation days. It is clear that interval censoring would be expected to have an impact on the time to gametocyte clearance analysis.

Table 11: Number of patients who cleared gametocytemia, by period when clearance occurred.

Interval	(0 , 3]	(3 , 7]	(7 , 14]	(14 , 21]	(21 , 28]	(28 , 42]	Total
Number of patients	0	22	37	42	38	56	195

Authors like Huang (1996), Kooperberg and Clarkson (1997), Younes and Lachin (1997) and Lindsey and Ryan (1998) developed non-parametric approaches to apply when analyzing interval censored data. Work has also gone into the parametric modeling of interval censored data (Samuelson and Kongerud, 1994, Klein and Moeschberger, 2005). Lindsey (1998) highlighted that since parametric models naturally apply smoothing to the observed data, as these models use information from adjacent points as part of their estimation procedure, interval censoring would be expected to have less of an impact on parametric models as compared to non-parametric models.

In this chapter it will be assumed that gametocyte clearance occurs in the middle of an interval. Subsequently the midpoint of this interval would be used as a point estimate for the gametocyte clearance time. It was shown by Lindsey (1998) that this approach can provide good results for the estimation of model parameters. However, this approach is not always reliable.

It is acknowledged that the use of the midpoint is a simplistic approach to apply when analyzing gametocyte clearance data. However, it is important to note that the main aim of this thesis is the imputation of the incomplete gametocyte profile in the presence of informative censoring. These completed profiles would provide estimates for all missing data, including missing data arising in-between the predefined study observation days. Subsequently the complete profiles can be used to conduct an analysis into the time to gametocyte clearance. The use of these imputed profiles would mitigate the effect of interval censoring on the resulting analysis. The results of the time to gametocyte clearance analysis, conducted in this chapter, will be used for comparative purposes. These results will be compared with the results of a survival analysis conducted using imputed gametocyte data.

3.5 Model fitting results

The methodology around survival analysis, which has previously been discussed, was applied to the treatment efficacy study dataset outlined in Chapter 1. Multiple survival (time-to-event) processes were generated in this study. These processes were the time to gametocyte emergence, the time to gametocyte clearance and the time to early exit from the study due to either treatment failure or loss-to-follow-up. The selection of covariates discussed in Chapter 1 was used to fit various parametric and Cox regression models to the data.

A basic model building approach was applied in this study. This approach involved firstly fitting models with different underlying distribution assumptions and the same set of covariates, to the data. These fitted models were then compared using AIC statistics as well as Cox-snell residual plots. The selected model was then refined by excluding non-significant covariates. The resulting AIC and deviance test statistics were used to assess the impact of each covariate removal. In addition Cox-Snell residual plots were generated for each of the fitted models to assess model fit, with the results helping to guide the decision into the final model to apply to the data.

Focus is only on the modeling of the location parameter, when analyzing parametric survival models in this investigation, with no modeling applied to any of the ancillary parameters. This approach is applied for ease of interpretation

of results across models.

This section will proceed as follows, firstly the time until gametocyte emergence will be investigated. This investigation will lead to a better understanding of the risk factors that influence the emergence of gametocytes. Subsequently, an analysis into the duration of gametocytemia will be conducted. Finally an investigation will be conducted into the factors that influence the exit of patients from the study, with the aim of trying to understand the censoring mechanism affecting the observation of the gametocyte profile.

3.5.1 Time to gametocyte emergence

The factors that affect the emergence of gametocytes are of particular interest in combating the transmission of malaria. In this section, the main interest is in evaluating the factors that influence whether or not gametocytes emerge and not necessarily those for predicting the time until gametocytes emerge. As a result proportional hazard models were used in this analysis. These include the Cox, exponential and Weibull proportional hazard models.

The first step in this analysis involved testing the validity of the proportional hazards assumption. Figure 10 provides a graphical test of the proportional hazards assumption. This figure illustrates the relationship between the log cumulative hazard function and time, across the different strata of the categorical variables used in this analysis. This plot reveals that the difference between the log cumulative hazard plots of the two treatments, is smaller at observation day 3 as compared to observation day 7. However from observation day 7 onwards the log cumulative hazard plots of the two treatments are fairly parallel. It can also be seen that the difference between the log cumulative hazard plots for the presence and absence of moderate anaemia is greater at observation day 3 as compared to observation day 7, with the difference remaining fairly constant from observation day 7 onwards. The covariates for parasite resistance ($mut5$) and patient gender also exhibit minor deviations from the proportional hazards assumption. However, none of the deviations from the categorical covariates shown in Figure 10 appear to be significant enough to make the overall proportional hazards assumption invalid.

Proportional hazard models were fit to the data and the estimated β coefficients, derived under the proportional hazards framework, with their associated confidence intervals are shown in Table 12. It can be seen that the Weibull model had the lowest AIC statistic. However, an evaluation of the Cox-snell residuals, shown in Figure 11, revealed that the Cox model provided a better fit to the data. Based on the results of the model fits and the Cox-snell residual

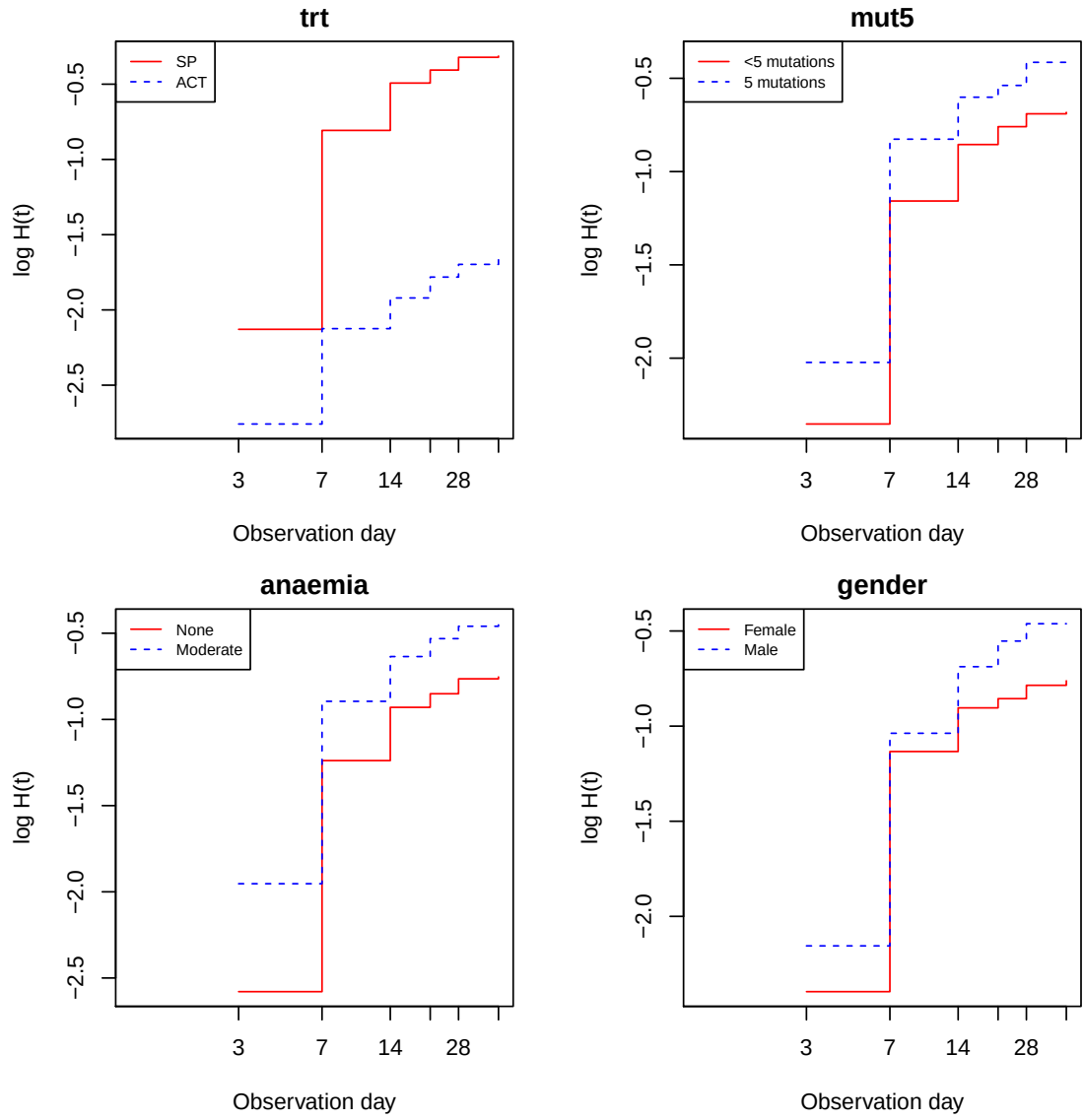

Figure 10: Plots of $\log H(t)$ against time for the time to gametocyte emergence process, stratified by categorical predictor variables

plots, it was decided that the Cox model was the most appropriate model to consider for further investigation.

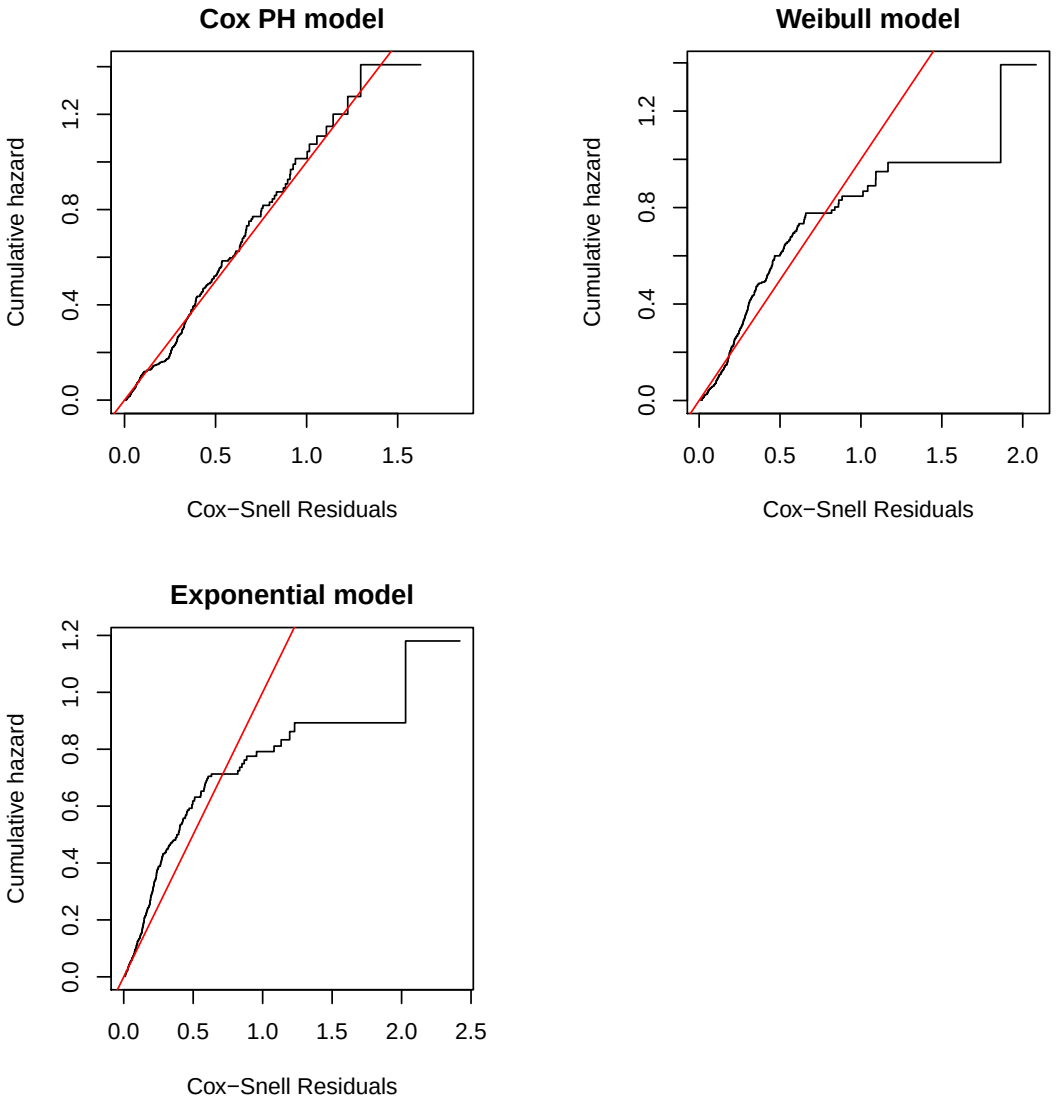

Figure 11: Cox-Snell residual plots for the time to gametocyte emergence process

Table 12: Parameter estimates (95% CI) for the time to gametocyte emergence PH survival models.

Definition	Parameter	Cox	Exponential	Weibull
scale	λ		0.007 (0.002 ; 0.012)	0.015 (0.003 ; 0.027)
trt	β_1	-0.908 (-1.334 ; -0.482)	-1.008 (-1.430 ; -0.586)	-0.968 (-1.391 ; -0.545)
ratio	β_2	-0.737 (-1.104 ; -0.370)	-0.898 (-1.270 ; -0.526)	-0.836 (-1.205 ; -0.467)
anaemia	β_3	0.322 (0.047 ; 0.597)	0.387 (0.111 ; 0.663)	0.358 (0.082 ; 0.634)
*mut*5	β_4	0.286 (-0.021 ; 0.593)	0.321 (0.015 ; 0.627)	0.299 (-0.008 ; 0.606)
lage	β_5	-0.001 (-0.097 ; 0.095)	-0.010 (-0.106 ; 0.086)	-0.007 (-0.103 ; 0.089)
pzero	β_6	0.068 (0.023 ; 0.113)	0.072 (0.027 ; 0.117)	0.071 (0.026 ; 0.116)
gender	β_7	0.243 (-0.017 ; 0.503)	0.310 (0.052 ; 0.568)	0.288 (0.029 ; 0.547)
shape	ρ		1	0.797 (0.710 ; 0.884)
AIC		2826	2342	2326

The fitted Cox model was refined by systematically removing non-significant covariates from the model and subsequently assessing the impact of this action on the AIC statistic. Since the models being compared are nested, the deviance test was also applied in the model building process. The results of this process are shown in Table 13. It can be seen that removing non-significant covariates had a small impact on the AIC statistics across all the models. In addition there is a small change in the parameter estimates and confidence intervals, for the covariates remaining after variable deletion. The deviance test revealed that there was a moderately significant difference between models $C2$ and $C3$. A comparison of the Cox-Snell residuals plots, for the fitted models, is provided in Figure 12. This figure reveals that model $C2$ provides a better fit to the data as compared to model $C3$. Based on the model fitting results, it was decided that Model $C2$ was the most appropriate model and warranted further investigation.

Table 13: Parameter estimates (95% CI) for the time to gametocyte emergence Cox PH survival models.

Definition	Parameter	C1	C2	C3
trt	β_1	-0.908 (-1.334 ; -0.482)	-0.908 (-1.334 ; -0.482)	-0.894 (-1.319 ; -0.469)
ratio	β_2	-0.737 (-1.104 ; -0.37)	-0.738 (-1.105 ; -0.371)	-0.750 (-1.117 ; -0.383)
anaemia	β_3	0.322 (0.047 ; 0.597)	0.323 (0.067 ; 0.579)	0.316 (0.060 ; 0.572)
*mut*5	β_4	0.286 (-0.021 ; 0.593)	0.286 (-0.021 ; 0.593)	0.305 (-0.001 ; 0.611)
lage	β_5	-0.001 (-0.097 ; 0.095)		
pzero	β_6	0.068 (0.023 ; 0.113)	0.068 (0.023 ; 0.113)	0.070 (0.025 ; 0.115)
gender	β_7	0.243 (-0.017 ; 0.503)	0.244 (-0.011 ; 0.499)	
AIC		2826	2824	2826
Deviance test statistic			(C2 vs C1) 0.001	(C3 vs C2) 3.498
Deviance test p-value			(C2 vs C1) 0.977	(C3 vs C2) 0.061

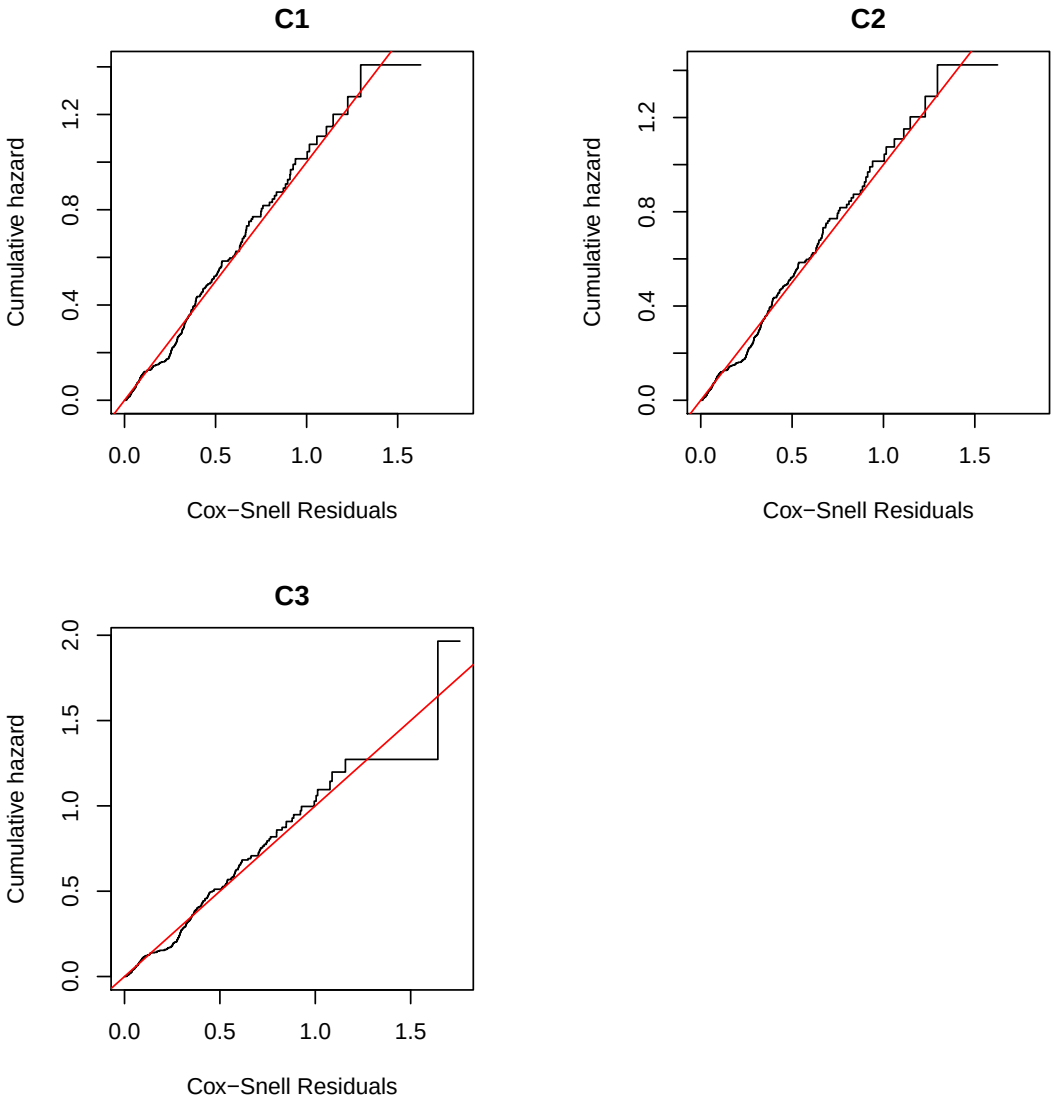

Figure 12: Cox-Snell residual plots for the Cox PH models

Deviance residual plots were created for model $C2$ in order to assess its overall model fit. These plots are shown in Figure 13. These residual plots indicate that the deviance residuals have a symmetric distribution. In addition there do not appear to be any significantly large outliers, with the range of the residuals being -1.80 to 2.81. The residuals have a mean -0.07 and a standard deviation of 1.08 that is in line with expectations as these residuals are assumed to follow a standard normal distribution with a mean of 0 and a standard deviation of 1. Looking at each of the categorical covariates separately, it can be

seen that the medians of the deviance residuals are generally less than 0 implying that the observed times to emergence are smaller than expected. However, this deviation from 0 is small, thus model $C2$ is an adequate model.

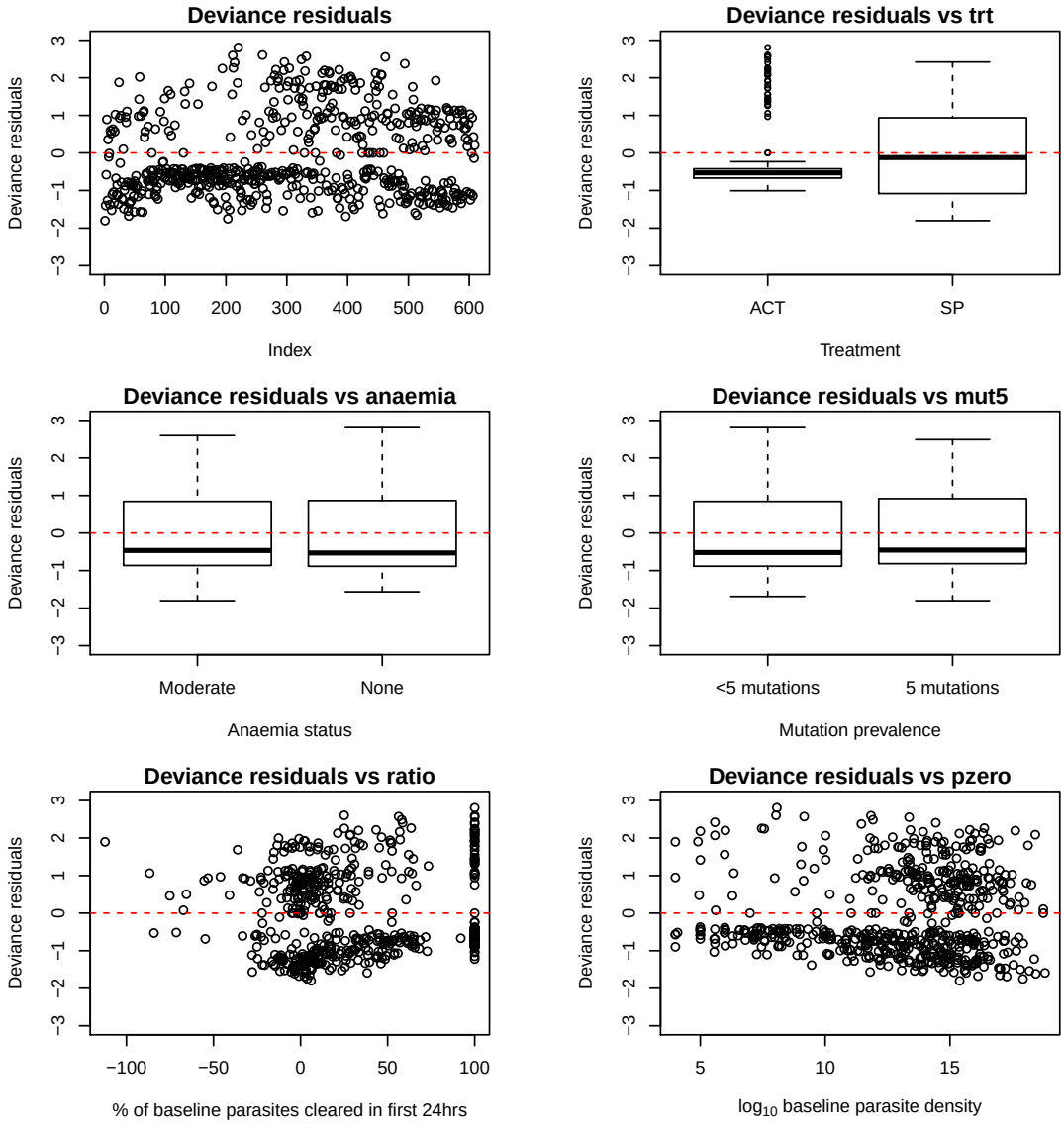

Figure 13: Deviance residual plots for the, Model $C2$, time to gametocyte emergence Cox PH survival model

The model fitting results from Model $C2$ revealed that treatment, parasite reduction ratio at 24 hours, the prevalence of moderate anaemia and baseline asexual parasite density all had strong associations with the hazard of gameto-

cyte emergence. Additionally the prevalence of quintuple mutations and gender were found to have a moderate association with the hazard of gametocyte emergence as their confidence intervals can be seen to be predominately positive. Model $C2$, for the i^{th} patient, is defined as

$$\log \frac{h_i(t)}{h_0(t)} = (trt_i \times \beta_1) + (ratio_i \times \beta_2) + (anaemia_i \times \beta_3) + (mut5_i \times \beta_4) + (pzero_i \times \beta_6) + (gender_i \times \beta_7).$$

This model can be interpreted as follows

- Receiving a combination of artesunate and sulfadoxine-pyrimethamine treatment as compared to sulfadoxine-pyrimethamine treatment only, reduces the hazard of gametocyte emergence by 59.6% ($= 1 - e^{-0.908}$).
- Clearing all baseline asexual parasites in the first 24 hours leads a 52.2% ($= 1 - e^{-0.738}$) reduction in the hazard of gametocyte emergence.
- Patients with moderate anaemia have a 38.2% ($= e^{0.323} - 1$) higher hazard of gametocyte emergence as compared to patients with a haemoglobin density > 11g/dL.
- Patients with 5 mutations have a 33.1% ($= e^{0.286} - 1$) higher hazard of gametocyte emergence as compared to patients who have less than 5 mutations.
- Every ten-fold increase in the baseline asexual parasite density results in a 7.0% ($= e^{0.068} - 1$) increase in the hazard of gametocyte emergence.
- Being a male patient increases the hazard of gametocyte emergence by 27.6% ($= e^{0.244} - 1$) as compared to being a female.

3.5.2 Time to gametocyte clearance

Time to gametocyte clearance and hence the duration of gametocytemia, was treated as the time-to-event process in this section. The aim of this section was to estimate the duration of gametocytemia, thus only AFT models were considered for this analysis since they can be explicitly used to estimate time. The exponential, generalized-gamma, log-logistic, log-normal and Weibull AFT models were used in this section. The *flexsurv* (Jackson, 2016) package, in R (R Core Team, 2015), was used to fit the aforementioned AFT models. This package was selected because it can accommodate the 3 parameter generalized-gamma AFT model.

In order to model the duration of gametocytemia, only patients who experienced gametocytemia were considered in the analysis. As a result the risk factors identified in this analysis would only be applicable to this cohort of patients. From the 609 patients included in the study, only 239 patients experienced gametocytemia. 195 of these patients cleared their gametocyte infection during the course of the study. As previously stated, it will be assumed that gametocyte clearance occurs at the midpoint of an observation interval. That is if a patient was observed to have gametocytes in their blood on day 7 but not on day 14, it would be assumed that gametocyte clearance occurred on day 10.5.

In this analysis an additional adjustment was required for patients who were observed only once in the study. It was assumed that these patients were observed for a period of 1 day before they were censored. This assumption affected 18 patients included in the analysis. This adjustment is not expected to have a significant impact on the model fitting results.

Table 14 shows the results of the AFT models fit to the data. It can be seen that the generalized-gamma, log-normal and Weibull models had the lowest AIC statistics, with there being a small difference between these models. It was previously stated that several of the AFT models used in this section were simplifications of the generalized-gamma model. Table 9 showed the constraints required to simplify the generalized-gamma distribution into the exponential, log-normal and Weibull distributions. These constraints were

- Exponential $\{q = 1 \text{ and } \sigma = 1\}$
- Log-Normal $\{q = 0\}$
- Weibull $\{q = 1\}$.

Based on the results of the generalized-gamma AFT model fit, it can be con-

cluded that the null hypothesis of $q = 0$ cannot be rejected; as the 95% confidence interval for the parameter overlaps with 0. This would imply that the log-normal model would be an adequate model to apply to the data. However, an evaluation of the Cox-Snell residuals for these fitted models (Figure 14) indicates that both the generalized-gamma and the log-normal models do not fit the data well. This is because the Cox-Snell residual plots for these models deviate from the straight line that runs through the origin with a gradient of 1. It can be seen, from Figure 14, that the Weibull model provides an adequate fit to the data. Based on the preceding information it was decided that the Weibull model would be considered for further investigation.

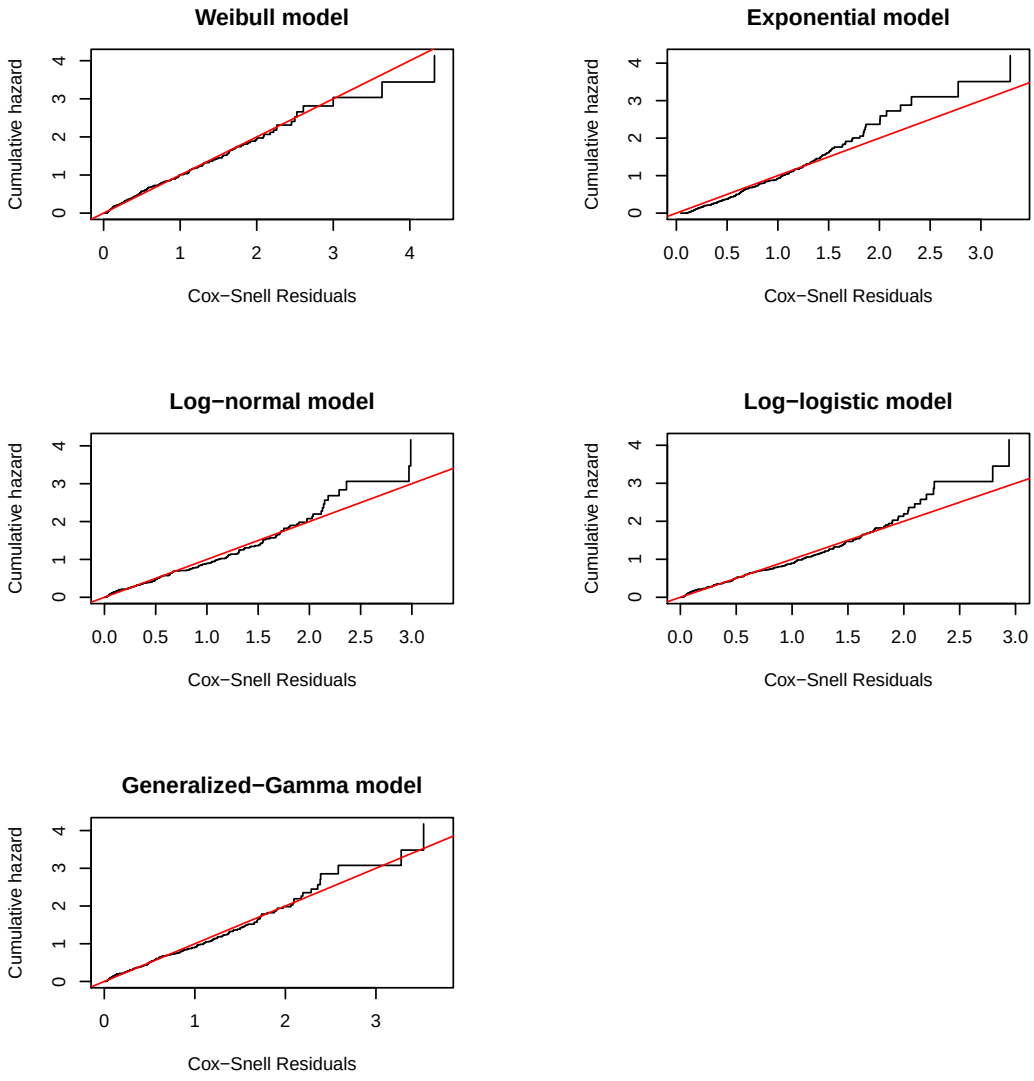

Figure 14: Cox-Snell residual plots for the time to gametocyte clearance process

Table 14: Parameter estimates (95% CI) for the time to gametocyte clearance AFT survival models.

Definition	Parameter	Generalized-gamma	Exponential	Weibull	Log-normal	Log-logistic
q	q	0.370 (-0.137 ; 0.876)				
intercept	α_0	0.933 (0.260 ; 1.607)	1.081 (0.244 ; 1.918)	1.281 (0.661 ; 1.902)	0.818 (0.163 ; 1.474)	0.720 (0.054 ; 1.386)
trt	α_1	-0.921 (-1.318 ; -0.523)	-0.940 (-1.454 ; -0.427)	-1.022 (-1.401 ; -0.643)	-0.862 (-1.253 ; -0.470)	-0.915 (-1.299 ; -0.530)
ratio	α_2	-0.228 (-0.519 ; 0.064)	-0.270 (-0.633 ; 0.093)	-0.224 (-0.484 ; 0.037)	-0.224 (-0.528 ; 0.080)	-0.221 (-0.540 ; 0.098)
anaemia	α_3	-0.057 (-0.302 ; 0.188)	-0.024 (-0.328 ; 0.280)	0.005 (-0.214 ; 0.223)	-0.100 (-0.347 ; 0.146)	-0.067 (-0.319 ; 0.184)
mut5	α_4	0.036 (-0.255 ; 0.328)	0.098 (-0.281 ; 0.476)	0.041 (-0.229 ; 0.311)	0.036 (-0.263 ; 0.335)	0.002 (-0.302 ; 0.306)
lage	α_5	0.046 (-0.039 ; 0.132)	0.043 (-0.066 ; 0.153)	0.046 (-0.034 ; 0.125)	0.045 (-0.042 ; 0.133)	0.051 (-0.036 ; 0.139)
pzero	α_6	0.106 (0.066 ; 0.145)	0.109 (0.058 ; 0.16)	0.096 (0.058 ; 0.134)	0.104 (0.065 ; 0.144)	0.112 (0.071 ; 0.152)
gender	α_7	0.117 (-0.115 ; 0.349)	0.107 (-0.187 ; 0.401)	0.120 (-0.091 ; 0.330)	0.119 (-0.122 ; 0.360)	0.116 (-0.131 ; 0.363)
scale	σ	0.819 (0.716 ; 0.938)	1	0.711 (0.631 ; 0.790)	0.862 (0.775 ; 0.949)	0.508 (0.449 ; 0.566)
AIC		1382	1415	1386	1382	1391

The Weibull model shown in Table 14 was refined by systematically excluding non-significant variables and assessing the impact on the AIC statistic of the resulting model. Deviance tests and Cox-snell residual plots were used in conjunction with the AIC statistic, to determine the final model. Table 15 shows the results of the model fits from the model building process along with the resulting AIC statistics. It can be seen that there is a small change in the AIC statistics across all the models. It is also evident that parameter estimates do not change dramatically as non-significant variables are systematically excluded from the models being fit. Deviance test reveals that there is no significant difference between any of the fitted models. These findings lend support to model $W6$ as it is a simplified model that is not significantly different from the other more complex models that were fit to the data. However, Cox-snell residual plots (Figure 15) reveal that $W4$ was the simplest model that had an adequate Cox-snell residual plot. Taking all the preceding information into account, it was decided that Model $W4$ was the most appropriate model to apply to the data.

Table 15: Parameter estimates (95% CI) for the time to gametocyte clearance Weibull AFT survival models.

Definition	Parameter	W1	W2	W3	W4	W5	W6
intercept	α_0	1.281 (0.661 ; 1.902)	1.285 (0.682 ; 1.887)	1.279 (0.678 ; 1.879)	1.422 (0.869 ; 1.975)	1.455 (0.906 ; 2.004)	1.452 (0.912 ; 1.992)
trt	α_1	-1.022 (-1.401 ; -0.643)	-1.022 (-1.401 ; -0.643)	-1.022 (-1.401 ; -0.643)	-1.036 (-1.416 ; -0.656)	-1.034 (-1.413 ; -0.656)	-1.153 (-1.494 ; -0.813)
ratio	α_2	-0.224 (-0.484 ; 0.037)	-0.222 (-0.477 ; 0.032)	-0.216 (-0.467 ; 0.035)	-0.197 (-0.447 ; 0.053)	-0.190 (-0.440 ; 0.060)	
anaemia	α_3	0.005 (-0.214 ; 0.223)					
*mut*5	α_4	0.041 (-0.229 ; 0.311)	0.040 (-0.228 ; 0.308)				
lage	α_5	0.046 (-0.034 ; 0.125)	0.045 (-0.030 ; 0.120)	0.046 (-0.029 ; 0.121)			
pzero	α_6	0.096 (0.058 ; 0.134)	0.096 (0.058 ; 0.134)	0.097 (0.059 ; 0.135)	0.099 (0.061 ; 0.137)	0.100 (0.062 ; 0.137)	0.098 (0.061 ; 0.135)
gender	α_7	0.120 (-0.091 ; 0.330)	0.120 (-0.090 ; 0.330)	0.124 (-0.085 ; 0.332)	0.088 (-0.113 ; 0.289)		
scale	σ	0.711 (0.631 ; 0.790)	0.711 (0.632 ; 0.790)	0.710 (0.631 ; 0.789)	0.713 (0.633 ; 0.792)	0.715 (0.635 ; 0.794)	0.717 (0.637 ; 0.796)
AIC		1386	1384	1382	1381	1380	1380
Deviance test statistic			(W2 vs W1) 0.002	(W3 vs W2) 0.088	(W4 vs W3) 1.403	(W5 vs W4) 0.801	(W6 vs W5) 2.168
Deviance test p-value			(W2 vs W1) 0.968	(W3 vs W2) 0.767	(W4 vs W3) 0.236	(W5 vs W4) 0.371	(W6 vs W5) 0.141

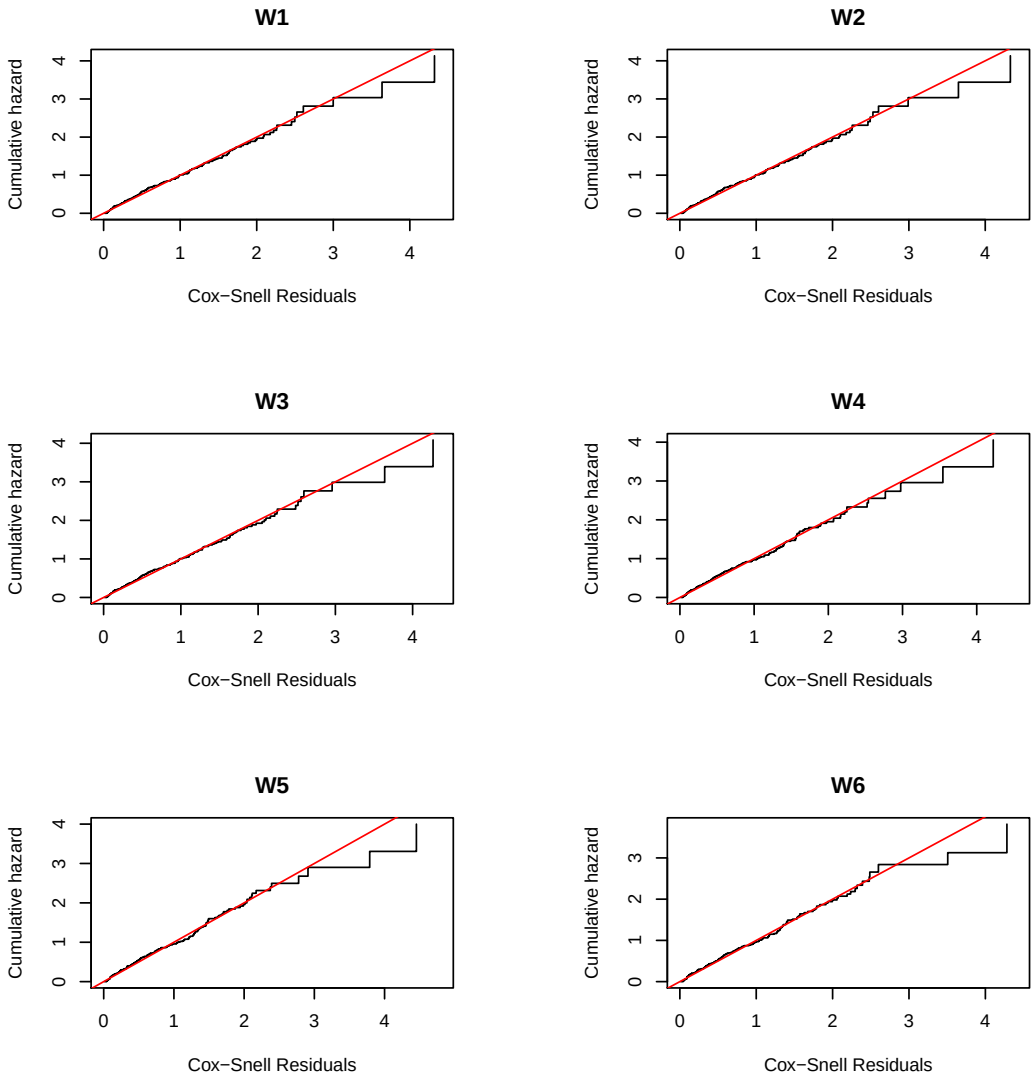

Figure 15: Cox-Snell residual plots for the Weibull time to gametocyte clearance models

Deviance residual plots were created for model $W4$ in order to assess its overall model fit. These plots are shown in Figure 16. The range of these residuals is -2.15 to 2.52, which implies that there are no significantly large outliers. The residuals have a mean of -0.23 and a standard deviation of 1.09, which is in line with the distributional assumptions of deviance residuals that assume normality with a mean of 0 and standard deviation of 1. It is acknowledged that the negative mean for the deviance residuals implies that the model is overestimating

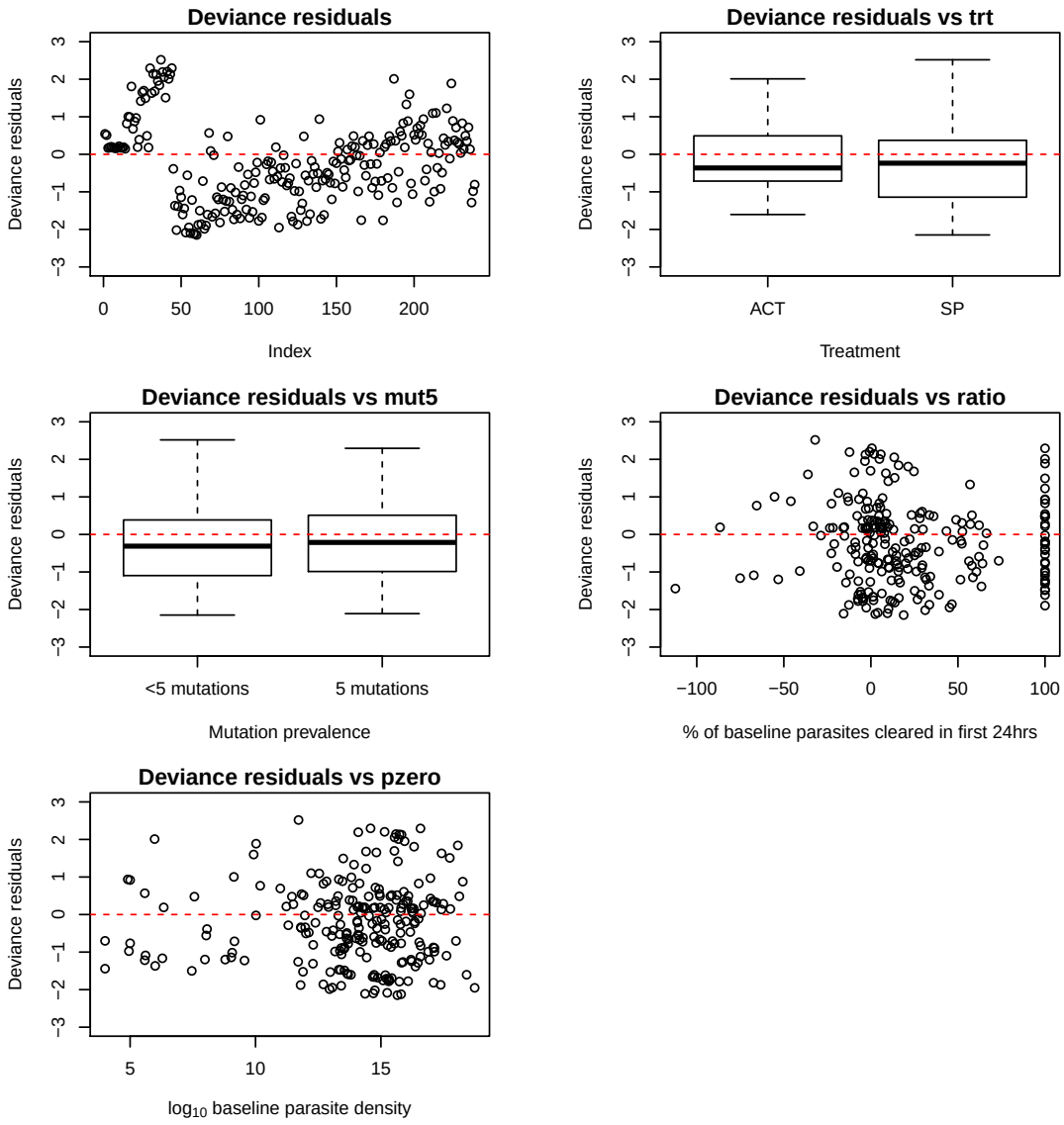

Figure 16: Deviance residual plots for the, Model $W4$, time to gametocyte clearance Weibull AFT survival model

the time to clearance. However, this overestimation is small thus, model $W4$ is an appropriate model to apply to this cohort of lives.

Model $W4$, for the i^{th} patient, is defined as

$$\log(T_i) = \alpha_0 + (trt_i \times \alpha_1) + (ratio_i \times \alpha_2) + (pzero_i \times \alpha_6) + (gender_i \times \alpha_7),$$

and it can be interpreted as follows

- Receiving a combination of artesunate and sulfadoxine-pyrimethamine

treatment as compared to sulfadoxine-pyrimethamine treatment only, decreases the time to gametocyte clearance by 64.5% ($= 1 - e^{-1.036}$).

- Clearing all baseline asexual parasites in the first 24 hours decreases the time to gametocyte clearance by 17.9% ($= 1 - e^{-0.197}$).
- The time to gametocyte clearance increases by 10.4% ($= e^{0.099}$) with every ten-fold increase in the baseline asexual parasite density.
- Being a male patient increases the time to gametocyte clearance by 9.2% ($= e^{0.088}$) as compared to being a female patient.

3.5.3 Estimation of the duration of gametocytemia

The time to gametocyte clearance AFT models, outlined above, were derived using a sample of the population that experienced gametocytemia. It can be argued that this group of patients is the least healthy segment of the population. As a result the estimated durations from this group cannot be used to make predictions on the overall population, as these predictions would be biased toward longer durations. A possible solution to this problem would involve incorporating the prevalence of gametocytemia into the estimation of the duration of gametocytemia. This methodology is outlined below.

Given that a Weibull AFT model has been fit to the dataset of individuals who exhibited gametocytemia, the duration of gametocytemia for the i^{th} patient (s_i) is determined as the area under the graph of the Weibull survival curve. This is defined as

$$\begin{aligned} s_i &= \int_0^\infty S_i(t)dt \\ &= \int_0^\infty \exp(-\lambda_i t^\gamma)dt. \end{aligned} \tag{36}$$

The estimated durations from Equation 36 would be expected to overestimate the duration of gametocytemia for the full population. In order to adjust the estimated durations to a population level, the prevalence of gametocytemia must be considered. The prevalence of gametocytemia is incorporated into the estimation of duration as follows

$$d_i = s_i \times \pi_i, \tag{37}$$

where d_i is the adjusted duration and π_i is the estimated probability of that individual developing gametocytes, which is based on the patient's covariate pattern. The prevalence of gametocytemia (π_i) is modeled using logistic regression. Since the Weibull AFT model is derived on the log-scale, it is more appropriate to model the prevalence of gametocyte on the same scale as it allows for a simpler derivation of the variance structure for the adjusted duration estimate. Taking this approach leads to the use of the log-link function, as opposed to the logit link function, in the modeling of gametocyte prevalence. This is shown in Equation 38 below.

$$\log(\pi_i) = \mathbf{x}_i^T \boldsymbol{\beta}_B, \tag{38}$$

where x_i is a set of p observed covariates $\mathbf{x}_i^T = (x_{i1}, \ldots, x_{ip})$, for the i^{th} patient and $\boldsymbol{\beta}_B$ is a p dimensional vector of fixed effect coefficients. Assuming that the

prevalence of gametocytemia is independent of the duration of gametocytemia, the following relationship can be defined on the log-scale

$$\begin{aligned} \log d_i &= \log(s_i \times \pi_i) \\ &= \log s_i + \log \pi_i. \end{aligned} \tag{39}$$

Applying a simplifying assumption that s_i and π_i are independent leads to a variance structure of

$$\text{Var}(\log d_i) = \text{Var}(\log s_i) + \text{Var}(\log \pi_i). \tag{40}$$

Equation 40 can thus be used to derive the confidence interval for the adjusted duration on the log-scale. Confidence intervals for d_i can subsequently be derived by using the delta method (Oehlert, 1992). The delta method allows for the approximation of the mean and variance of a function that is a transformation other random variables. The method applies a first-order Taylor series expansion about the mean of a function to derive an approximation for its variance. The Taylor series expansion for a differentiable function $G(X)$, where X is an asymptotically normally distributed random variable with a mean of μ, is given as

$$G(X) = G(\mu) + \{(X - \mu) \times G^{'}(\mu)\}.$$

It follows that the variance of $G(X)$, using the Taylor series approximation, is given by

$$\begin{aligned} \text{Var}[G(X)] &= \text{Var}[G(\mu) + \{(X - \mu) \times G^{'}(\mu)\}] \\ &= \text{Var}(X) \times [G^{'}(\mu)]^2. \end{aligned}$$

In this analysis X is $\log d_i$, $G(.)$ is the exponential function and μ is the linear predictor of the fitted Weibull AFT model on the log-scale. This implies that

$$\text{Var}[\hat{d}_i] = \text{Var}(\log \hat{d}_i) \times [\hat{d}_i]^2.$$

Since the adjusted duration is based on patient covariate patterns and not the actual observed patients, it is possible to derive the prevalence model based of the full dataset of 609 patients. The parameter estimates from this model would then be combined with the parameter estimates from the Weibull time to clearance model in order to estimate the adjusted duration for a generic patient who has a specific covariate pattern.

A similar model building process to that used throughout this chapter was applied in the derivation of the appropriate prevalence model. The results of the model building process are shown in Table 16. It can be seen that there is a small change in the AIC statistics across all models. Model $B4$ appeared

to be the most appropriate model as it was the model that consisted of only covariates that had a strong association with the prevalence of gametocytemia. In addition the deviance tests that were applied revealed that this model was not significantly different from the more complicated models that were fit in this analysis.

The model fitting results from Model $B4$ revealed that treatment, the first 24 hour parasite reduction ratio, baseline asexual parasite density and patient gender had strong associations with the prevalence of gametocytemia, as their 95% CIs did not include 0. Model $B4$ is defined as

$$\log(\pi_i) = \beta_{B0} + (trt_i \times \beta_{B1}) + (ratio_i \times \beta_{B2}) + (pzero_i \times \beta_{B6}) + (gender_i \times \beta_{B7}),$$

and it can be interpreted as follows

- Receiving a combination of artesunate and sulfadoxine-pyrimethamine treatment as compared to sulfadoxine-pyrimethamine treatment only, reduces the relative risk of gametocytemia by 52.5% $(= 1 - e^{-0.744})$.
- Clearing all baseline asexual parasites in the first 24 hours reduces the relative risk of gametocytemia by 38.0% $(= 1 - e^{-0.479})$.
- Every ten-fold increase in the baseline asexual parasite density increases the relative risk of gametocytemia by 6.2% $(= e^{0.060} - 1)$.
- Male patients have a 20.3% $(= e^{0.185} - 1)$ higher relative risk of gametocytemia as compared to female patients.

Table 16: Parameter estimates (95% CI) for the prevalence of gametocytemia models.

Definition	Parameter	B1	B2	B3	B4
intercept	β_{B0}	-1.597 (-2.159 ; -1.034)	-1.534 (-2.018 ; -1.051)	-1.508 (-1.988 ; -1.027)	-1.535 (-2.025 ; -1.045)
trt	β_{B1}	-0.766 (-1.134 ; -0.397)	-0.765 (-1.133 ; -0.397)	-0.758 (-1.127 ; -0.389)	-0.744 (-1.113 ; -0.375)
ratio	β_{B2}	-0.461 (-0.723 ; -0.199)	-0.462 (-0.723 ; -0.201)	-0.478 (-0.739 ; -0.216)	-0.479 (-0.742 ; -0.216)
anaemia	β_{B3}	0.171 (-0.015 ; 0.357)	0.158 (-0.016 ; 0.332)	0.154 (-0.021 ; 0.330)	
*mut*5	β_{B4}	0.128 (-0.070 ; 0.326)	0.126 (-0.071 ; 0.322)		
lage	β_{B5}	0.012 (-0.053 ; 0.077)			
pzero	β_{B6}	0.053 (0.020 ; 0.087)	0.052 (0.019 ; 0.085)	0.053 (0.020 ; 0.086)	0.060 (0.027 ; 0.093)
gender	β_{B7}	0.208 (0.030 ; 0.387)	0.205 (0.029 ; 0.38)	0.195 (0.019 ; 0.370)	0.185 (0.009 ; 0.361)
AIC		728	726	726	727
Deviance test statistic			(B2 vs B1) 0.124	(B3 vs B2) 1.447	(B4 vs B3) 2.910
Deviance test p-value			(B2 vs B1) 0.725	(B3 vs B2) 0.229	(B4 vs B3) 0.088

Equation 37 outlined the methodology required to estimate the population level duration of gametocytemia. This method involved using a patient's specific covariate pattern to combine the estimated prevalence of gametocytemia with the estimated duration gametocytemia. The confidence intervals associated with these duration estimates would subsequently be estimated using the delta method. These predicted durations are shown in Table 17 and they allow for the comparison of the following characteristics

- Sulfadoxine-pyrimethamine treatment only (SP) and a combination treatment of artesunate and sulfadoxine-pyrimethamine (ACT).
- "High" $\log_{10}$ baseline asexual parasite density of 17.754 relative to a "Low" baseline asexual parasite density of 5.
- "High" first 24 hour parasite reduction rate of 100% relative to a "Low" first 24 hour parasite reduction rate of -31.4%.
- Female patients compared to male patients.

The 2.5 and the 97.5 percentiles for the baseline asexual parasite density and the first 24 hour parasite reduction ratio were used as the respective low and high covariate values described above.

The following steps were taken to derive the predicted adjusted durations

1. A dataset with the levels given above (used for comparison) was created.
2. The parameter estimates from Model $W4$ were used to estimate the duration of gametocytemia on the log-scale.
3. The parameter estimates from Model $B4$ were used to estimate the prevalence of gametocytemia on the log-scale.
4. The estimates of duration and prevalence, on the log-scale given above, were combined as per Equation 39 in order to generate estimates for the adjusted duration on the log-scale.
5. The adjusted duration on the original scale was derived by taking the exponent of the log-scale adjusted durations. Subsequently the confidence intervals of the adjusted durations, on the original scale, were calculated using the delta method.

Table 17 reveals that the covariate patterns with the longest adjusted duration have the following characteristics

- Male patients

- Receive sulfadoxine-pyrimethamine treatment only
- Have a high baseline asexual parasite density
- Have a low first 24 hour parasite reduction ratio

Table 17: Estimated prevalence and duration of gametocytemia (95% CI), in days.

trt	*pzero*	*ratio*	*gender*	Estimated Prevalence	Estimated Duration	Adjusted Duration
SP	Low	low	F	0.34 (0.22 ; 0.46)	7.23 (4.91 ; 10.65)	2.45 (1.17 ; 5.40)
			M	0.41 (0.26 ; 0.55)	7.9 (5.30 ; 11.77)	3.22 (1.49 ; 7.21)
		High	F	0.18 (0.11 ; 0.25)	5.58 (3.59 ; 8.69)	1.01 (0.40 ; 2.44)
			M	0.22 (0.13 ; 0.31)	6.10 (3.92 ; 9.49)	1.32 (0.52 ; 3.24)
	High	low	F	0.73 (0.57 ; 0.88)	25.57 (20.23 ; 32.32)	18.64 (12.73 ; 31.48)
			M	0.88 (0.71 ; 1.05)	27.93 (22.04 ; 35.39)	24.49 (17.01 ; 40.71)
		High	F	0.39 (0.28 ; 0.50)	19.74 (14.81 ; 26.31)	7.67 (4.54 ; 14.68)
			M	0.47 (0.34 ; 0.60)	21.56 (16.39 ; 28.36)	10.08 (6.11 ; 18.93)
ACT	Low	low	F	0.16 (0.08 ; 0.24)	2.57 (1.68 ; 3.92)	0.41 (0.15 ; 1.06)
			M	0.19 (0.10 ; 0.29)	2.80 (1.82 ; 4.32)	0.54 (0.19 ; 1.40)
		High	F	0.09 (0.05 ; 0.12)	1.98 (1.39 ; 2.82)	0.17 (0.08 ; 0.39)
			M	0.10 (0.06 ; 0.15)	2.16 (1.52 ; 3.08)	0.22 (0.10 ; 0.52)
	High	low	F	0.35 (0.19 ; 0.51)	9.07 (5.43 ; 15.15)	3.14 (0.98 ; 8.42)
			M	0.42 (0.23 ; 0.60)	9.91 (5.93 ; 16.55)	4.13 (1.32 ; 10.98)
		High	F	0.18 (0.11 ; 0.26)	7.01 (4.54 ; 10.82)	1.29 (0.54 ; 3.07)
			M	0.22 (0.14 ; 0.31)	7.65 (5.00 ; 11.71)	1.70 (0.73 ; 3.98)

3.5.4 Time to early exit from the study

In the previous section the time to gametocyte clearance was investigated. This was done in the presence of censoring, which arose in the form of treatment failure and loss-to-follow-up. These two forms of censoring are examples of informative censoring. Informative censoring arises when the experience of patients who exit the study is different to the experience of patients who remain in the study.

It can be hypothesized that the patients who experience treatment failure have high densities of asexual parasites in their system for an extended period of time. These high densities of asexual parasites would be expected to develop into gametocytes, which would result in these patients having a higher prevalence of gametocytemia as compared to patients who stay in the study. It can possibly be assumed that patients who have successfully overcome their malaria infection are more likely to exit the study before it is completed, resulting in loss-to-follow-up. The implication of this assumption is that the prevalence of gametocytemia in patients who are lost-to-follow-up is less than the prevalence of gametocytemia in the patients who remain in the study. Alternatively, loss-to-follow-up my be due to reasons that are unrelated to the study like relocation for work or accidental death. The implication of loss-to-follow-up being not at random is that it becomes a competing cause of early exit from the study, to the treatment failure cause of exit.

This section will begin by investigating the risk factors for the hazard of early exit from the study, due to either treatment failure or loss-to-follow-up, using proportional hazard models. This involves combining treatment failure and loss-to-follow-up in order to create a single event classified as early exit from the study. The next step will involve extending the analysis into a cause-specific investigation, whereby treatment failure and loss-to-follow-up are treated as competing events.

Before beginning the time to early exit analysis, a test of the proportional hazards assumption was conducted. This test was performed graphically by assessing the relationship between the log cumulative hazard function and time, across different strata of the categorical variables used in this analysis. The results of this test are presented in Figure 17. It can be seen that the proportional hazards assumption for patient gender is not appropriate as the log cumulative hazard plots for males and females cross over after day 7. The log cumulative hazard functions for the different strata, of each of the other categorical variables, remain fairly parallel over time thus implying that the proportional hazards assumption if appropriate for them. Overall it appeared as though the

proportional hazards assumption was valid for this analysis.

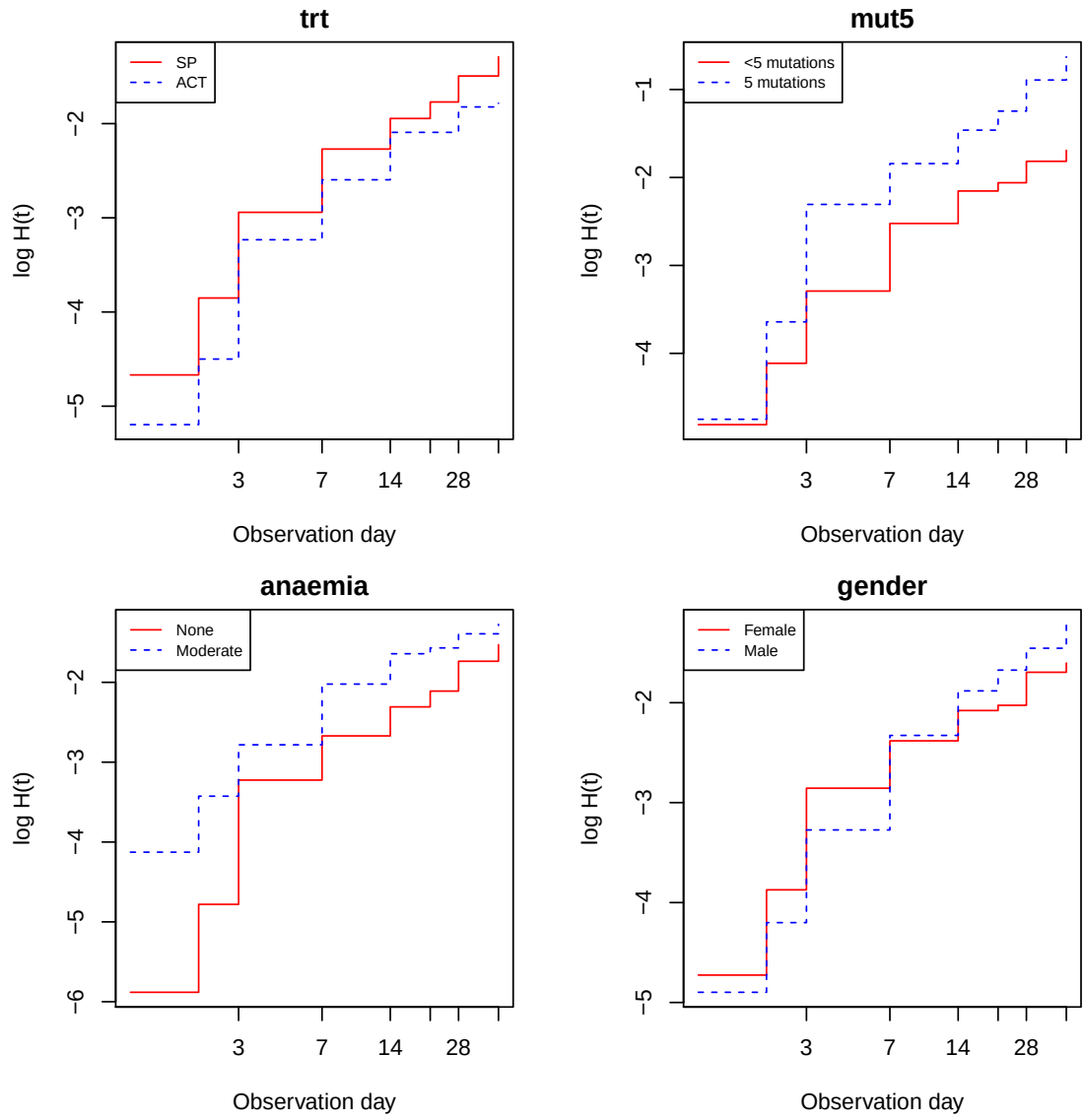

Figure 17: Plots of $\log H(t)$ against time for the time to early exit process, stratified by categorical risk factors.

The Cox, exponential and Weibull proportional hazard models were fit to the data and the estimated β coefficients, along with their confidence intervals, are shown in Table 18. It can be seen that the Weibull model has the lowest AIC statistic, with the exponential model having a moderately higher AIC statistic. It is also evident that there is a close similarity in the estimated coefficients

between the exponential and Weibull models. A comparison of the Cox-Snell residual plots for the fitted models is provided in Figure 18. This plot suggests that the exponential and Weibull models fit the data adequately. The exponential model is a special case of the Weibull model, where $\rho = 1$. It can be seen that the confidence interval for ρ, under the Weibull model, does not include 1. As a result is can be concluded that the ρ parameter is significantly different from 1, which implies that the Weibull model is the most appropriate model to consider for further investigation.

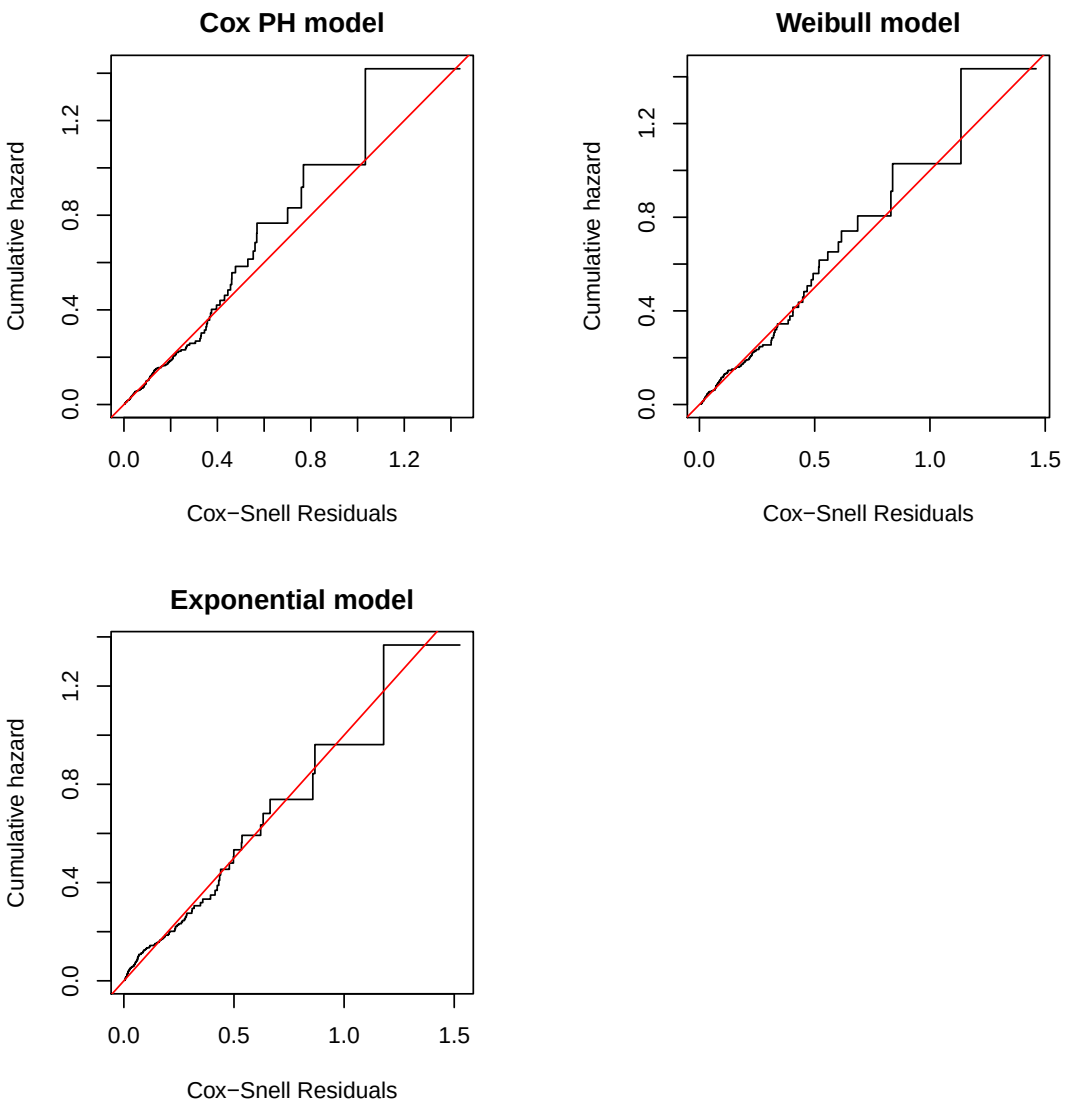

Figure 18: Cox-Snell residual plots for the time to early exit from the study

Table 18: Parameter estimates (95% CI) for the time to early exit Cox PH survival models.

Definition	Parameter	Cox	Exponential	Weibull
scale	λ		0.004 (0.000 ; 0.008)	0.008 (0.000 ; 0.018)
trt	β_1	-0.115 (-0.599 ; 0.369)	-0.118 (-0.601 ; 0.365)	-0.113 (-0.597 ; 0.371)
ratio	β_2	-0.696 (-1.181 ; -0.211)	-0.714 (-1.200 ; -0.228)	-0.702 (-1.188 ; -0.216)
anaemia	β_3	0.170 (-0.205 ; 0.545)	0.176 (-0.199 ; 0.551)	0.174 (-0.200 ; 0.548)
*mut*5	β_4	1.022 (0.662 ; 1.382)	1.041 (0.682 ; 1.400)	1.024 (0.665 ; 1.383)
lage	β_5	-0.110 (-0.242 ; 0.022)	-0.114 (-0.245 ; 0.017)	-0.111 (-0.242 ; 0.020)
pzero	β_6	0.036 (-0.022 ; 0.094)	0.036 (-0.022 ; 0.094)	0.036 (-0.023 ; 0.095)
gender	β_7	0.204 (-0.150 ; 0.558)	0.205 (-0.149 ; 0.559)	0.204 (-0.151 ; 0.559)
shape	ρ		1	0.825 (0.690 ; 0.960)
AIC		1608	1566	1562

Similarly to the previously defined model building processes outlined in this chapter, non-significant covariates were systematically removed from the Weibull model provided in Table 18. The impact on the change in the AIC statistic and the resultant deviance test, was used as the criteria in deciding the most appropriate fitted model. The results of the model building process are shown in Table 19. The results show that there is a small change in the AIC statistics, as non-significant covariates were being removed from the fitted models. In addition it can be seen that there is a small change in the parameter effect sizes of the covariates remaining in the fitted models, after the removal of non-significant covariates. The Cox-snell residuals, shown in Figure 19, were created to assess the adequacy of the individual models fit to the data. These plots revealed that all the fitted models were broadly appropriate.

Based on all the preceding information, it was decided that Model $W5$ was the most appropriate model to consider for further investigation. This was because all its covariates had a significant effect on the hazard of early exit from the study. In addition the results of the deviance tests, which compared Model $W4$ to Model $W5$, indicated that the presence of an additional non-significant covariate did not significantly improve the fit of Model $W4$ to the data as compared to Model $W5$.

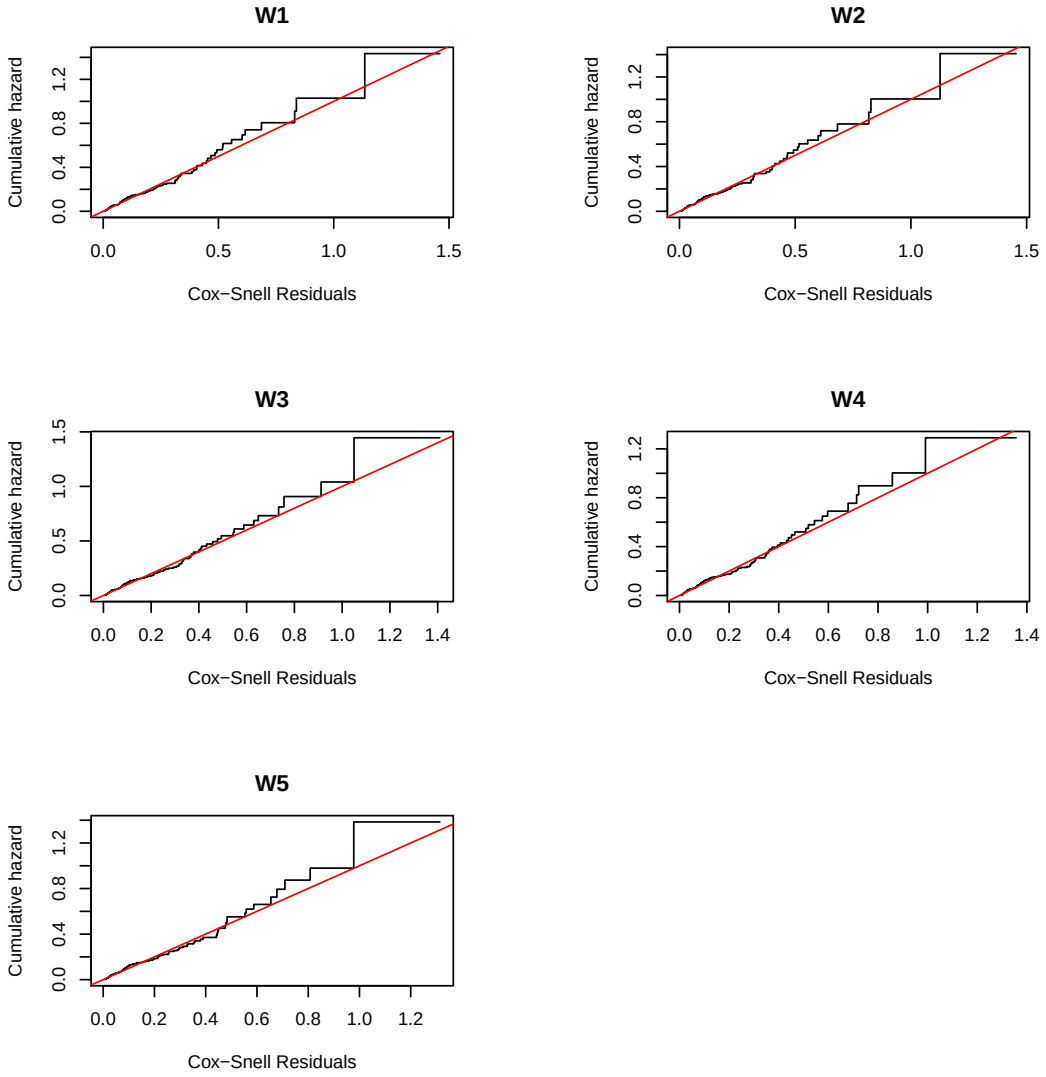

Figure 19: Cox-Snell residual plots for the time to early exit from the study

Table 19: Parameter estimates (95% CI) for the time to early exit Weibull PH models.

Definition	Parameter	W1	W2	W3	W4	W5
scale	λ	0.008 (0.000 ; 0.018)	0.008 (0.000 ; 0.018)	0.009 (0.000 ; 0.019)	0.010 (0.000 ; 0.020)	0.019 (0.007 ; 0.031)
trt	β_1	-0.113 (-0.597 ; 0.371)				
ratio	β_2	-0.702 (-1.188 ; -0.216)	-0.757 (-1.186 ; -0.328)	-0.744 (-1.171 ; -0.317)	-0.748 (-1.175 ; -0.321)	-0.800 (-1.204 ; -0.396)
anaemia	β_3	0.174 (-0.200 ; 0.548)	0.173 (-0.201 ; 0.547)			
*mut*5	β_4	1.024 (0.665 ; 1.383)	1.024 (0.663 ; 1.385)	1.020 (0.661 ; 1.379)	1.039 (0.680 ; 1.398)	1.068 (0.711 ; 1.425)
lage	β_5	-0.111 (-0.242 ; 0.020)	-0.110 (-0.241 ; 0.021)	-0.134 (-0.256 ; -0.012)	-0.145 (-0.265 ; -0.025)	-0.154 (-0.274 ; -0.034)
pzero	β_6	0.036 (-0.023 ; 0.095)	0.038 (-0.019 ; 0.095)	0.040 (-0.017 ; 0.097)	0.042 (-0.015 ; 0.099)	
gender	β_7	0.204 (-0.151 ; 0.559)	0.201 (-0.154 ; 0.556)	0.184 (-0.169 ; 0.537)		
shape	ρ	0.825 (0.690 ; 0.960)	0.825 (0.690 ; 0.960)	0.825 (0.690 ; 0.960)	0.824 (0.689 ; 0.959)	0.824 (0.689 ; 0.959)
AIC		1562	1561	1559	1558	1559
Deviance test statistic			(W2 vs W1) 0.210	(W3 vs W2) 0.813	(W4 vs W3) 1.050	(W5 vs W4) 2.238
Deviance test p-value			(W2 vs W1) 0.647	(W3 vs W2) 0.367	(W4 vs W3) 0.305	(W5 vs W4) 0.135

Deviance residual plots were created for Model $W5$ in order to assess its overall model fit. These plots are shown in Figure 20. There do not appear to be any significantly large outliers, with the range of the residuals being from -2.92 to 1.62. The residuals have a mean 0.16 and a standard deviation of 0.98, which is in line with expectations. However, it can be seen that the deviance residuals have a high proportion of small positive residuals that are offset by fewer relatively larger negative residuals. The implication is that the observed times to exit are larger than expected. These discrepancies were considered to be minor and not enough to disprove the conclusion that model $W5$ was an appropriate model.

The covariates in Model $W5$ all had strong associations with the hazard of early exit from the study, as their 95% CIs did not include 0. These covariates were the first 24 hour parasite reduction ratio, the prevalence of quintuple mutation and patient age. Model $W5$ is defined as

$$h_i(t) = \lambda\rho t^{\rho-1} \exp[(ratio_i \times \beta_2) + (mut5_i \times \beta_4) + (lage_i \times \beta_5)],$$

and it can be interpreted as follows,

- Clearing all baseline asexual parasites in the first 24 hours leads to a 55% $(= 1 - e^{-0.800})$ reduction in the hazard of early exit from the study.
- Patients with quintuple mutations present have a 2.91 $(= e^{1.068})$ fold higher hazard of early exit as compared to patients with less than 5 mutations.
- Doubling patient age leads to a 14.2% $(= 1 - e^{-0.154})$ reduction in the hazard of early exit.

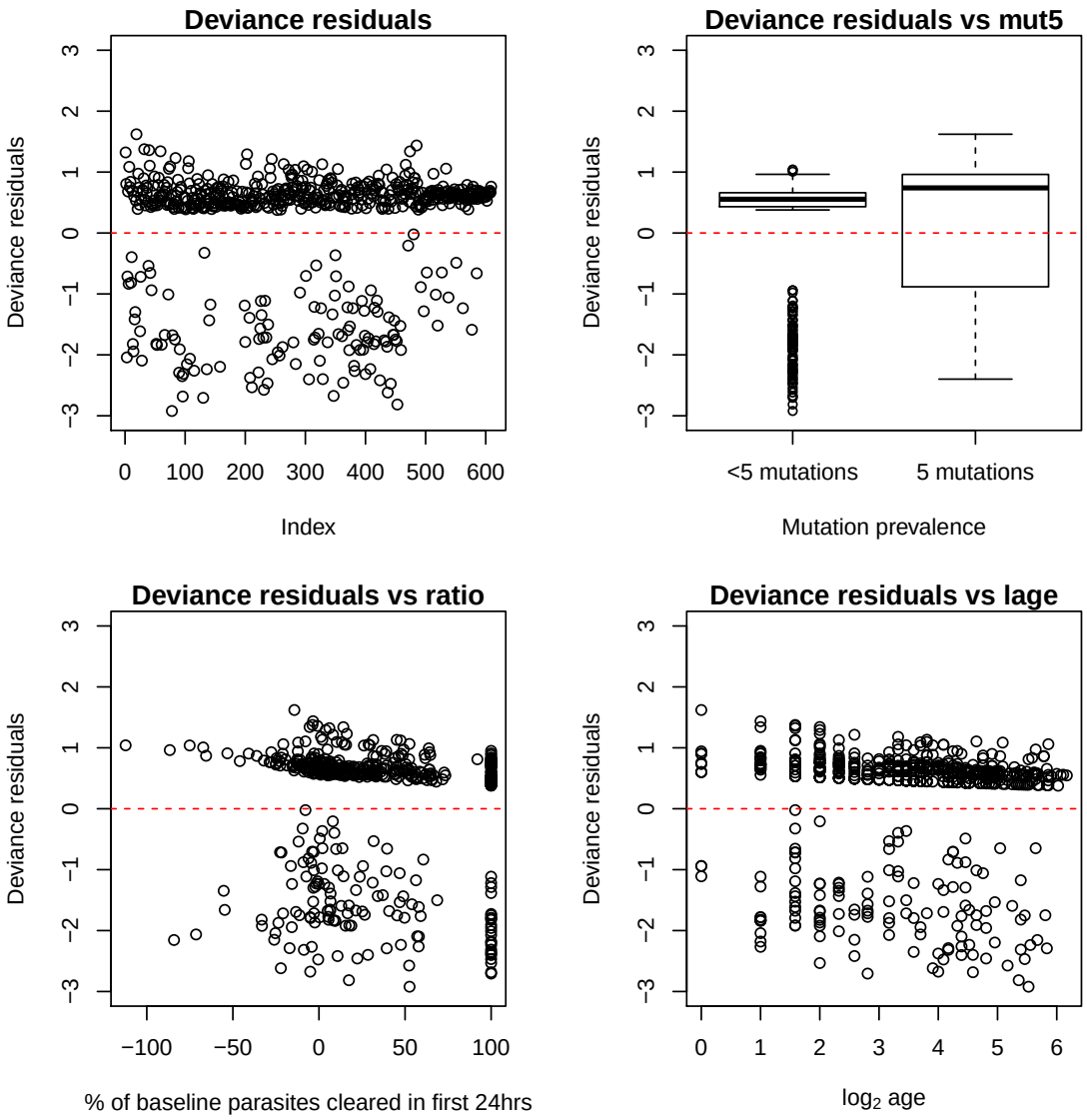

Figure 20: Deviance residual plots for the, Model $W5$, time to early exit from the study Weibull PH survival model

3.5.5 Cause-specific hazard analysis

As previously stated, the underlying characteristics of patients who exit the study can be very different depending on whether they were lost-to-follow-up or whether they experienced treatment failure. If the hypothesis that the cause of exit from the study is related to a patient's health status is correct, it would be expected that the risk factors used in this analysis would behave differently depending on the cause of exit being investigated. A cause-specific hazards analysis allows for the impacts of the risk factors, considered in this study, to be investigated with reference to each cause of early exit.

A cause-specific hazard model was fit to the data under the proportional hazards framework. Tests for the validity of the proportional hazards assumption are shown in Figure 21 and Figure 22 for the treatment failure and loss-to-follow-up causes of exit respectively. Considering the plots for the treatment failure cause of exit, it can be seen that the plots for patient gender cross over after 7 days, in addition the plots for the treatment effect cross over after day 3. When considering the plots for the LTFU cause of exit, the prevalence of mutation plots cross over at a couple of observation days, in addition treatment effect and gender plots can also be seen to cross over a certain points. These plots do raise some concerns around the validity of the proportional hazards assumption. However, since the plots are predominately parallel across the covariates considered, it was assumed that the proportional hazards assumption was valid.

A comparison of the cause-specific proportional hazard models is provided in Table 20. This table provides a comparison of the Cox, exponential and Weibull proportional hazard models. It can be seen that the parameter estimates, by cause of exit, were broadly the same across all the fitted models. The Cox proportional hazards model provides the lowest overall AIC statistic. However, Figure 23 reveals that the Weibull model provides a moderately better fit to both the treatment failure and loss-to-follow-up causes of exit. The Weibull model was thus selected as the final model to apply to the data. For ease of comparison between the two causes of exit no further model refinement, in respect of covariate selection, was applied.

The Weibull model is a combination of the treatment failure and the loss-to-follow-up (LTFU) hazard functions. The hazard functions are defined as

$$h_{Fi}(t) = \lambda_F \rho_F t^{\rho_F - 1} \exp[(trt_i \times \beta_{F1}) + (ratio_i \times \beta_{F2}) + (anaemia_i \times \beta_{F3}) + (mut5_i \times \beta_{F4}) + (lage_i \times \beta_{F5}) + (pzero_i \times \beta_{F6}) + (gender_i \times \beta_{F7})],$$

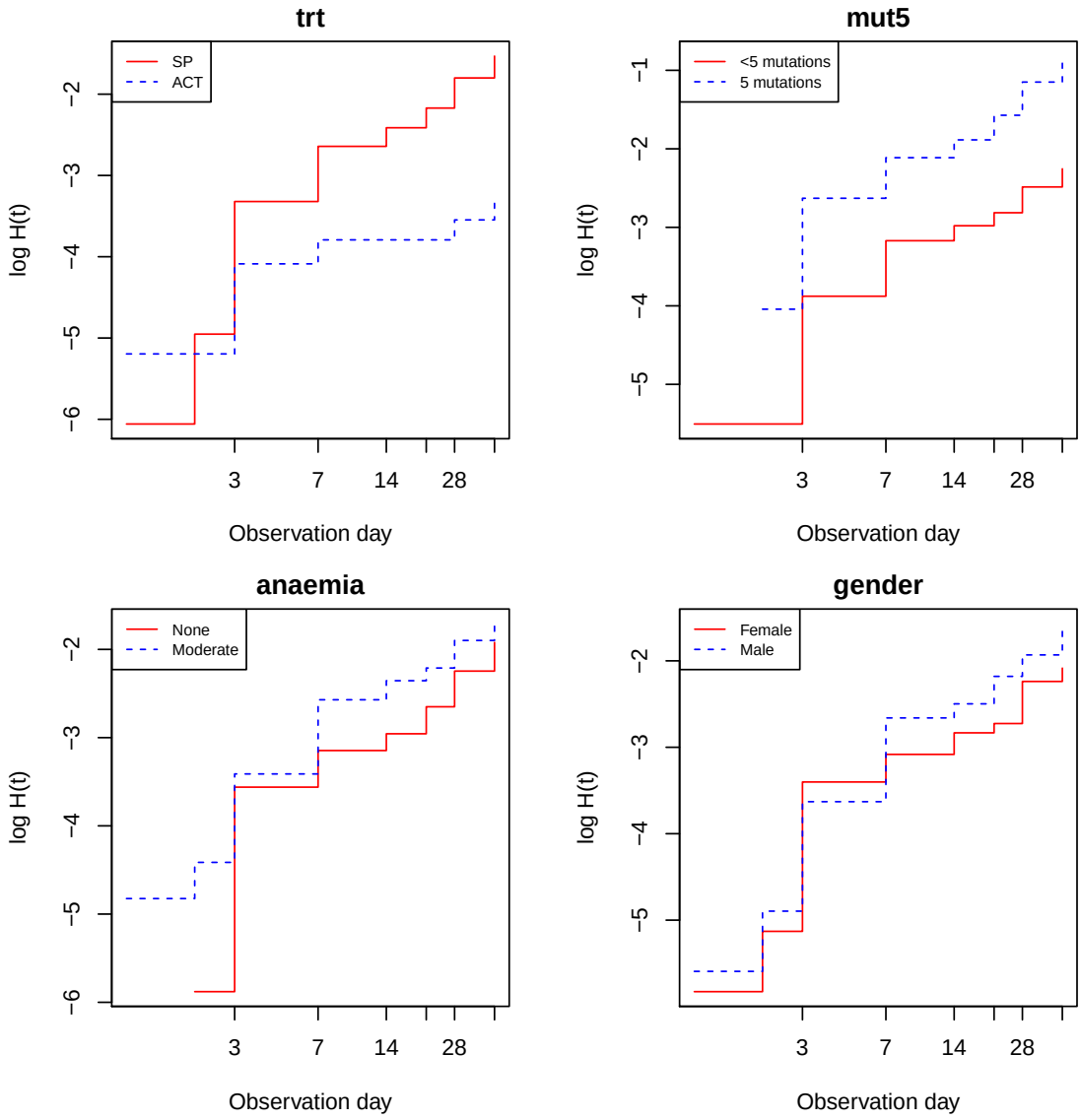

Figure 21: Plots of $\log H(t)$ against time, for the time to treatment failure cause of early exit

and

$$h_{Li}(t) = \lambda_L \rho_L t^{\rho_L - 1} \exp[(trt_i \times \beta_{L1}) + (ratio_i \times \beta_{L2}) + (anaemia_i \times \beta_{L3}) + (mut5_i \times \beta_{L4}) + (lage_i \times \beta_{L5}) + (pzero_i \times \beta_{L6}) + (gender_i \times \beta_{L7})],$$

for the treatment failure and LTFU hazard functions respectively.

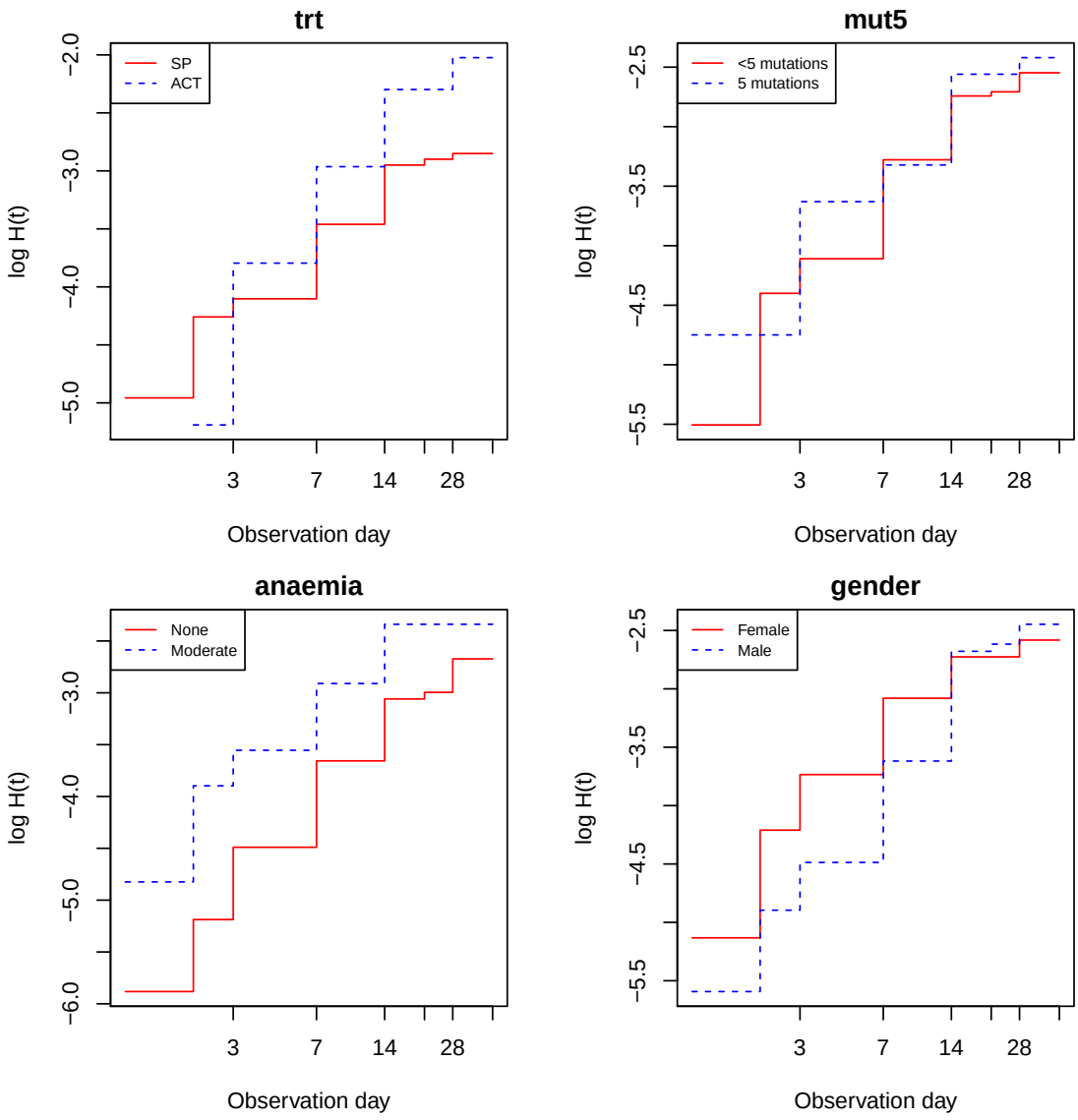

Figure 22: Plots of $\log H(t)$ against time for the time to LTFU cause of early exit

Table 20: Parameter estimates (95% CI) for the cause-specific PH models.

		Cox		Exponential		Weibull	
Definition	Parameter [a]	Failure	LTFU	Failure	LTFU	Failure	LTFU
scale	λ_K			0.003 (0.000 ; 0.007)	0.001 (0.000 ; 0.003)	0.003 (0.000 ; 0.007)	0.004 (0.000 ; 0.010)
trt	β_{K1}	-1.307 (-2.194 ; -0.420)	1.004 (0.288 ; 1.720)	-1.287 (-2.174 ; -0.400)	0.983 (0.271 ; 1.695)	-1.286 (-2.174 ; -0.398)	0.988 (0.275 ; 1.701)
ratio	β_{K2}	-0.843 (-1.476 ; -0.210)	-0.449 (-1.253 ; 0.355)	-0.864 (-1.498 ; -0.230)	-0.466 (-1.270 ; 0.338)	-0.861 (-1.496 ; -0.226)	-0.454 (-1.258 ; 0.350)
anaemia	β_{K3}	0.019 (-0.446 ; 0.484)	0.426 (-0.216 ; 1.068)	0.039 (-0.426 ; 0.504)	0.427 (-0.214 ; 1.068)	0.038 (-0.427 ; 0.503)	0.421 (-0.220 ; 1.062)
*mut*5	β_{K4}	1.392 (0.960 ; 1.824)	0.129 (-0.605 ; 0.863)	1.383 (0.951 ; 1.815)	0.186 (-0.548 ; 0.920)	1.379 (0.946 ; 1.812)	0.161 (-0.572 ; 0.894)
lage	β_{K5}	-0.179 (-0.338 ; -0.020)	0.068 (-0.169 ; 0.305)	-0.179 (-0.338 ; -0.020)	0.060 (-0.178 ; 0.298)	-0.178 (-0.337 ; -0.019)	0.061 (-0.176 ; 0.298)
pzero	β_{K6}	0.069 (-0.011 ; 0.149)	-0.002 (-0.090 ; 0.086)	0.067 (-0.013 ; 0.147)	0.000 (-0.088 ; 0.088)	0.067 (-0.013 ; 0.147)	-0.001 (-0.089 ; 0.087)
gender	β_{K7}	0.238 (-0.201 ; 0.677)	0.137 (-0.470 ; 0.744)	0.232 (-0.207 ; 0.671)	0.146 (-0.462 ; 0.754)	0.232 (-0.207 ; 0.671)	0.143 (-0.465 ; 0.751)
shape	ρ			1	1	0.965 (0.773 ; 1.157)	0.644 (0.462 ; 0.826)
AIC		1002	574	1051	655	1053	646
Total AIC		1576		1706		1699	

[a] $K = F$ when referencing treatment failure and L when referencing LTFU

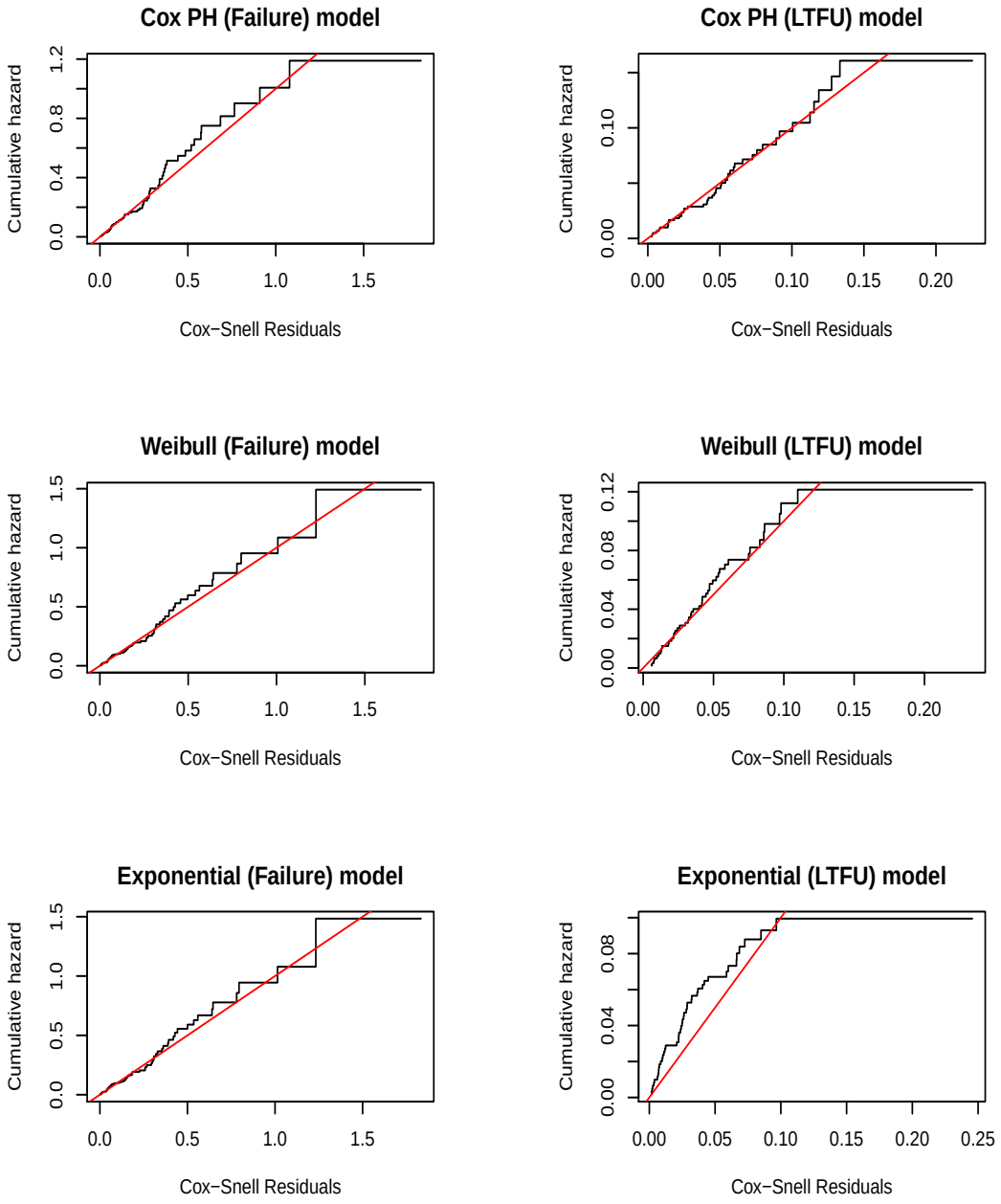

Figure 23: Cox-Snell residual plots for the cause specific time to early exit models

The interpretation of the Weibull cause-specific model is provided in Table 21. It can be seen from Table 20 that the treatment effect is highly significant once the two causes of early exit are isolated. The rationale behind artemisinin-based combination therapy is that adding an artemisinin derivative such as artesunate to sulfadoxine-pyrimethamine results in more rapid clearance of both asexual parasites and fever, a reduction in gametocyte carriage and a reduction in the risk of treatment failure. As artesunate and sulfadoxine-pyrimethamine have different mechanisms of action, the probabilities of a parasite being resistant to both antimalarials is markedly reduced. The interpretation of the effect of treatment, under the Weibull cause-specific hazards survival model, thus makes clinical sense. This is because the hazard of treatment failure for patients receiving ACT treatment would be expected to be significantly less than that of patients receiving SP treatment only. In addition since the signs and symptoms of malaria are expected to be relieved timeously in patients receiving ACT treatment, there would be less incentive for these patients to remain in the study once their health has improved. These findings add weight to the argument that both causes of early exit from the study are forms of informative censoring. It is important to note that treatment was the only covariate that had a strong association with the hazard of LTFU.

With regards to the hazard of treatment failure; patient gender and the prevalence of moderate anaemia did not have an association with the hazard of treatment failure. Baseline parasite density had a moderate association as its 95% CI was predominately positive, with the remaining covariates all having strong associations with the hazard of treatment failure.

Figure 24 provides a graphical illustration of the predicted hazard functions arising from the cause-specific hazards analysis, by treatment. These plots stack the predicted hazard functions arising from the LTFU and treatment failure, causes of early exit. As a result the highest curve on each of the plots is representative of the overall hazard of early exit arising from either LTFU or treatment failure. The predicted hazard rates were made on a male patient, with a haemoglobin density > 11g/dL, 5 mutations present, median $\log_{10}$ baseline asexual parasite density (13.7), median first 24 hour parasite clearance rate (28.3%) and median patient age of 12 years ($lage = 3.6$). It can be seen that the hazard of treatment failure is greater in patients receiving SP treatment only, as compared to ACT treatment. In addition it can be seen that the hazard of LTFU was greater in patients receiving ACT treatment as compared to SP treatment only.

Table 21: Interpretation of the Weibull Cause-specific PH model.

Definition	Treatment Failure	Loss-to-follow-up (LTFU)
trt	Receiving ACT treatment, as compared to SP treatment, reduces the hazard of treatment failure by 72%	Receiving ACT treatment, as compared to SP treatment, leads to a 2.7 fold increase in the hazard of LTFU
ratio	Clearing all baseline asexual parasites in the first 24 hour is associated with a 58% reduction in the hazard of treatment failure	The first 24 hour parasite reduction ratio does not have an association with the hazard of LTFU
anaemia	Anaemia status does not have an association with the hazard of treatment failure	Anaemia status does not have an association with the hazard of LTFU
*mut*5	Having 5 mutations leads to a 4 fold increase in the hazard of treatment failure	The prevalence of quintuple mutations does not have an association with the hazard of LTFU
lage	Doubling patient age reduces the hazard of treatment failure by 16%	Patient age does not have an association with the hazard of LTFU
pzero	Every ten-fold increase in the baseline asexual parasite density results in a 7% increase in the hazard of treatment failure	Baseline asexual parasite density does not have an association with the hazard of LTFU
gender	Patient gender does not have an association with the hazard of treatment failure	Patient gender does not have an association with the hazard of LTFU

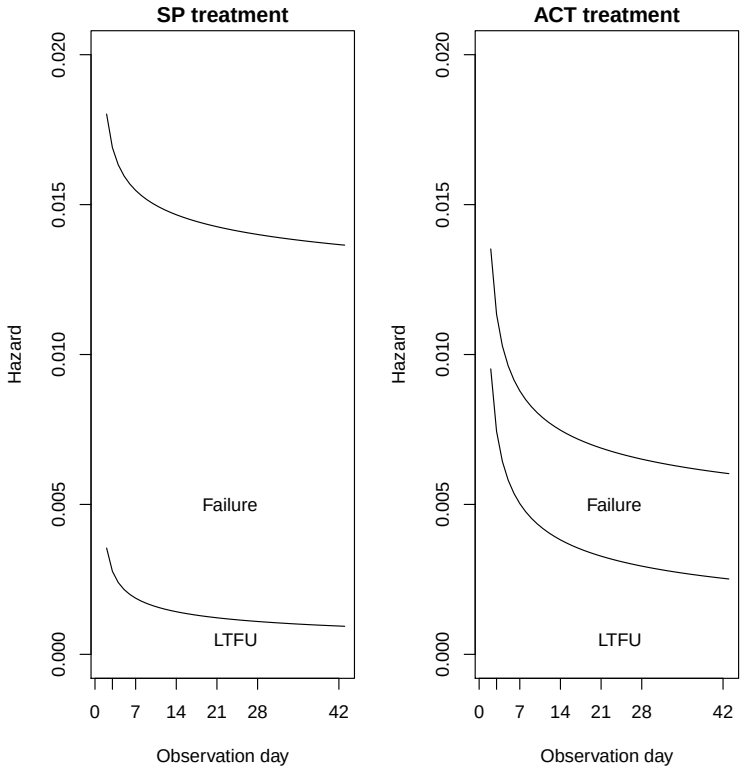

Figure 24: Stacked cause-specific hazard plots for LTFU and treatment failure causes of exit, by treatment

4 Joint Models

4.1 Overview of joint modeling techniques

Time-to-event and longitudinal (repeated measures) data were generated in this analysis. The modeling of the longitudinal gametocyte profile is a key area of focus in this thesis. It has been shown, in Chapter 2, that gametocyte profiles are zero-inflated and that they have a nonlinear relationship with time. Another feature of these profiles is that some patients have incomplete profiles, as they do not have gametocyte measurements at all observation times. Examples of individual gametocyte profiles are shown in Figure 25.

It was discussed in Chapter 3, that an informative censoring mechanism may be prevalent in this study. As a result standard longitudinal models fit to the observed data would be expected to give biased parameter estimates. Faucett and Thomas (1996) highlighted how non-random dropout led to biased estimates in their analysis of CD4 count progression over time. Another feature of longitudinal data is measurement error. Measurement error can arise, in longitudinal studies, due to erroneous measurement readings or due to short-term biological variability in patients (Wang and Taylor, 2001). Prentice (1982) showed that failure to account for measurement error, in longitudinal measurements included in a survival analysis model as time-dependent covariates, would lead to biased hazard estimates and incorrect variance estimates. Joint models can be used to correct for the bias in parameter estimation, which arises when modeling both the survival and longitudinal processes. The relationship between missing data and joint models will be expanded upon in the next section.

An overview of the joint modeling framework, that was applied in this analysis, is presented in this chapter. This overview considers generic longitudinal and survival processes in its development. The survival and longitudinal submodels were then subsequently investigated further and made more specific to this analysis. The methodology applied was an extension of the work by Henderson et al. (2000), which involved the use of latent random effects in the joint model formulation. The complex likelihood functions developed in this chapter were estimated under the Bayesian framework. The resulting joint posterior distributions of the parameters, to be estimated, were analytically intractable thus Markov Chain Monte Carlo (MCMC) methods were used to obtain the point and interval estimates of parameters (Ghosh et al., 2006). The survival process included in the joint model is usually reflective of the informative censoring mechanism taking place in the data. In this analysis the informative censoring mechanism was characterized by the time to early exit survival process. As a

result this process was included in the joint models that were applied in this chapter.

Two types of longitudinal submodels were considered in this analysis. These models can be characterized as those that ignore the zero-inflation in the data and those that account for it. Models that ignore zero-inflation are referred to as Nonlinear (NL) joint models while the models that accounted for the zero-inflation are referred to as zero-adjusted gamma (ZAG) joint models.

This chapter will proceed as follows, firstly notation that will be applied in this chapter will be presented. Secondly, the relationship between missing data and joint models will be discussed. Subsequently the joint models that were applied to the data generated by this study will be described. Finally these models will be applied to the data and the results will be presented.

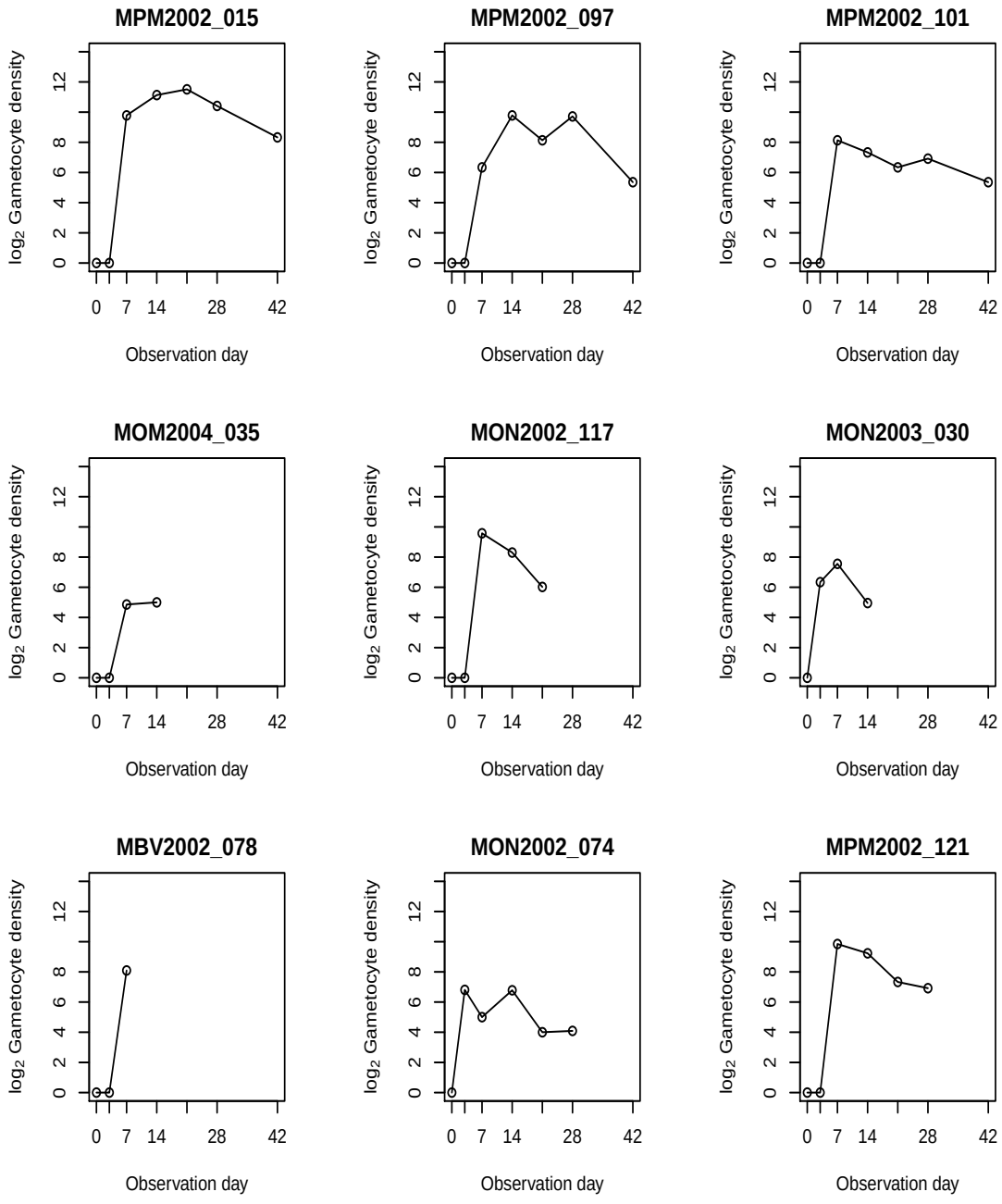

Figure 25: Observed Gametocyte profiles (black circles) for selected patients. The first row shows profiles for successful treatment outcomes, the second row shows profiles of patients lost-to-follow-up and the third row shows profiles of patients who experienced treatment failure

4.2 Notation

Given that there are m patients who are followed up for a period $[0, c]$, where in the case of this analysis $c \leq 42$, the number of observations collected for the i^{th} patient is n_i $(\forall i = 1, \ldots, m)$ with the total number of observations collected during the study given as $n = \sum_{i=1}^{m} n_i$. These observations are collected, for the i^{th} patient, at visitation times s_{ij} $(\forall j = 1, \ldots, n_i)$. The observed gametocyte density recorded for the i^{th} patient at the j^{th} visit is given as y_{ij}. The observed covariates are defined as $\mathbf{x}_{ij}$, where $\mathbf{x}_{ij}$ is a p-dimensional vector, with the same covariates being applied to both the survival and longitudinal submodels. In this analysis only baseline covariates were considered thus $\mathbf{x}_{ij}$ simplifies to $\mathbf{x}_i$. The corresponding p dimensional vector of fixed effect coefficients, for the i^{th} patient, is defined as $\boldsymbol{\theta}_i$. The covariates for the random effects are defined as $\mathbf{z}_i$ with associated random effect parameters of $\mathbf{b}_i$.

The time that the last gametocyte reading was taken is given as t_i and the reasons for that timepoint being the last observed time can either be a result of treatment failure, loss-to-follow-up or the end of the study. In addition consider that the response vector for a i^{th} patient, who is censored during the study, consists of an observed $(\mathbf{y}_i^o)$ and a missing component $(\mathbf{y}_i^m)$. The observed component consists of all observations collected up to the event time, whilst the missing component contains observations that would have been collected up to the end of the study if the patient had remained in the study. Interval censoring also gives rise to missing data, as the $\mathbf{y}_i^m$ observations can fall in between observed timepoints. The response vector for the i^{th} subject is thus given as $\mathbf{y}_i = (\mathbf{y}_i^o, \mathbf{y}_i^m)$. A missing data indicator will also be introduced as

$$r_{ij} = \begin{cases} 1 & \text{if } y_{ij} \text{ is observed} \\ 0 & \text{otherwise} \end{cases}.$$

The missing data indicator vector is thus $\mathbf{r}_i$ and the associated underlying missingness mechanism is defined as $\mathbf{w}_i$ with $\boldsymbol{\vartheta}$ as the set of missingness parameters.

4.3 Missing data mechanisms

In malaria intervention studies, gametocyte longitudinal profiles are generally incomplete. Early exit from the study, either due to loss-to-follow-up or treatment failure, is responsible for the incomplete profiles. Missing data arising from early exit from the study are referred to as monotone missingness. Intermittent missingness is an additional form of missingness, which arises in these studies,

whereby a patient misses some follow-up visits but returns to the study before it ends.

A key problem with missing data is a loss of efficiency, as additional patients need to be recruited into the study to achieve a specific level of power required to detect significant effects. An additional problem is bias in parameter estimates, in the event of informative censoring.

There are several approaches that can be used to handle missing data. A common approach used in most statistical packages is the deletion of information from patients who had missing values. An analysis that uses this approach is referred to as a "complete-case analysis" or "per protocol analysis". Estimates that are derived from this approach can be biased if the individuals excluded from the study are significantly different from those who remain. Additional approaches involve the imputation of the missing data. A simple imputation approach is referred to as the "last observation carried forward" approach. In this approach the missing responses are replaced by the last observed responses from the patient. A problem with this approach is that it underestimates the variability of the patient's responses, as it assumes that the missing responses are constant. Inverse probability weighting can also be used to deal with missing data. In this approach observed data is weighted by the inverse of the probability that the data was observed (Seaman and White, 2013). Little and Rubin (2002) proposed a more sophisticated multiple imputation approach. This approach involves the use of a statistical model to impute M values for each of the missing datapoints, thereafter M statistical analyses are conducted on the reconstituted datasets (i.e. datasets where the missing observations have been replaced). This is an example of explicit imputation.

The aim of the analysis conducted in this chapter is the imputation of incomplete gametocyte profiles. The method to be used depends on the underlying missingness mechanism. Authors like Fitzmaurice et al. (2004) and Molenberghs and Kenward (2007) outlined three types of missingness mechanisms. These mechanisms are Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing not at Random (MNAR).

A MCAR mechanism refers to the scenario where the outcomes are completely unrelated to the missingness process, as a result the full data density is given as

$$f(\mathbf{y}_i, \mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_i, \mathbf{w}_i, \boldsymbol{\theta}, \boldsymbol{\vartheta}) = f(\mathbf{y}_i|\mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\theta}) f(\mathbf{r}_i|\mathbf{w}_i, \boldsymbol{\vartheta}). \tag{41}$$

It is thus implied that the observed responses are independent of the missingness process, which means that the missingness mechanism does not need to be considered in the analysis of the observed responses. This allows Equation 41

to be simplified to

$$f(\mathbf{y}_i^o, \mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_i, \mathbf{w}_i, \boldsymbol{\theta}, \boldsymbol{\vartheta}) = f(\mathbf{y}_i^o|\mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\theta}) f(\mathbf{r}_i|\mathbf{w}_i, \boldsymbol{\vartheta}).$$

A MAR mechanism refers to the scenario where the probability of missingness is conditionally independent of the unobserved data ($\mathbf{y}_i^m$), given the observed data ($\mathbf{y}_i^o$). This implies that

$$f(\mathbf{r}_i|\mathbf{y}_i^o, \mathbf{y}_i^m, \mathbf{w}_i, \boldsymbol{\vartheta}) = f(\mathbf{r}_i|\mathbf{y}_i^o, \mathbf{w}_i, \boldsymbol{\vartheta}),$$

which leads to a density of

$$f(\mathbf{y}_i^o, \mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_i, \mathbf{w}_i, \boldsymbol{\theta}, \boldsymbol{\vartheta}) = f(\mathbf{y}_i^o|\mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\theta}) f(\mathbf{r}_i|\mathbf{y}_i^o, \mathbf{w}_i, \boldsymbol{\vartheta}). \tag{42}$$

The missingness process in the scenario above can be referred to as "ignorable", since the $\boldsymbol{\theta}$ and $\boldsymbol{\vartheta}$ are not related. As a result, likelihood based analysis conducted using only the observed data would give an appropriate result, provided that the model for the measurement process is correctly specified.

The MNAR mechanism is associated with informative censoring. Under this mechanism, the probability of a missing response is dependent on the unobserved data. As a result there is a convolved relationship between the observed data and the missingness process, which can be give rise to the following density

$$f(\mathbf{y}_i^o, \mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_i, \mathbf{w}_i, \boldsymbol{\theta}, \boldsymbol{\vartheta}) = \int f(\mathbf{y}_i|\mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\theta}) f(\mathbf{r}_i|\mathbf{y}_i, \mathbf{w}_i, \boldsymbol{\vartheta}) \mathrm{d}\mathbf{y}_i^m. \tag{43}$$

It is evident that the missingness process cannot be ignored and valid inferences can only be obtained when an analysis is conducted using a joint distribution of the measurement and missingness processes. Authors like Molenberghs and Kenward (2007), Little (1995) and Daniels and Hogan (2008) outlined three model families that can be used for the aforementioned joint distribution. These families are selection models, pattern mixture models and shared-parameter models.

The selection model defines a full dataset density of

$$f(\mathbf{y}_i, \mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_i, \mathbf{w}_i, \boldsymbol{\theta}, \boldsymbol{\vartheta}) = f(\mathbf{y}_i|\mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\theta}) f(\mathbf{r}_i|\mathbf{y}_i, \mathbf{w}_i, \boldsymbol{\vartheta}),$$

where $f(\mathbf{y}_i|\mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\theta})$ is the marginal density of the measurement process while $f(\mathbf{r}_i|\mathbf{y}_i, \mathbf{w}_i, \boldsymbol{\vartheta})$ is the density of the missingness process that is dependent on the responses ($\mathbf{y}_i$). The $f(\mathbf{r}_i|\mathbf{y}_i, \mathbf{w}_i, \boldsymbol{\vartheta})$ component can be considered as a probabilistic mechanism that describes a patient's self-selection with regards to either withdrawing or staying the study (Rizopoulos, 2012).

Pattern mixture models define a full dataset density of

$$f(\mathbf{y}_i, \mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_i, \mathbf{w}_i, \boldsymbol{\theta}, \boldsymbol{\vartheta}) = f(\mathbf{y}_i|\mathbf{r}_i, \mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\theta}) f(\mathbf{r}_i|\mathbf{w}_i, \boldsymbol{\vartheta}).$$

In this model the responses are modeled conditionally on the missingness process, while the missingness process can be modeled directly without considering the responses. As a result there is allowance for a different measurement model with each pattern of missing responses.

Shared-parameter models capture the association between measurement and missingness processes through the use of random effects. This leads to a full dataset density of

$$f(\mathbf{y}_i, \mathbf{r}_i|\mathbf{x}_i, \mathbf{z}_i, \mathbf{w}_i, \boldsymbol{\theta}, \boldsymbol{\vartheta}) = \int f(\mathbf{y}_i|\mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\theta}, \mathbf{b}_i) f(\mathbf{r}_i|\mathbf{w}_i, \boldsymbol{\vartheta}, \mathbf{b}_i) \mathrm{d}\mathbf{b}_i.$$

It is thus assumed that the measurement and missingness processes are independent given a latent random effect structure $\mathbf{b}_i$. Shared-parameter models will be used for the remainder of this analysis.

As discussed in the previous chapter informative censoring may have occurred in this study, which implies that a MNAR mechanism may apply to the data used in this thesis. Authors like Guo and Carlin (2004), Rizopoulos (2012) and Ibrahim et al. (2001) showed that joint models can be used to model informative dropout, where the survival time was taken as the time to dropout. The joint density function for the observed longitudinal and survival data is given as

$$\begin{aligned} f(\mathbf{y}_i^o, d_i|\mathbf{x}_i, \mathbf{z}_i, \mathbf{w}_i, \boldsymbol{\theta}, \boldsymbol{\vartheta}) &= \int \int f(\mathbf{y}_i|\mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\theta}, \mathbf{b}_i) f(d_i|\mathbf{w}_i, \boldsymbol{\vartheta}, \mathbf{b}_i) \mathrm{d}\mathbf{y}_i^m \mathrm{d}\mathbf{b}_i \\ &= \int f(\mathbf{y}_i^o|\mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\theta}, \mathbf{b}_i) f(d_i|\mathbf{w}_i, \boldsymbol{\vartheta}, \mathbf{b}_i) \mathrm{d}\mathbf{b}_i, \end{aligned} \tag{44}$$

where d_i is an indicator for the time to dropout. It is evident from Equation 44 that the measurement and dropout processes are conditionally independent given the random effects $\mathbf{b}_i$.

In this analysis joint models will be used to impute the incomplete gametocyte profiles. These joint models will combine the longitudinal gametocyte profiles with the survival model attributed to the time to early exit from the study. As shown above this method will accommodate the informative censoring associated with the study. The types of joint models that were used in this analysis will be described in the following sections.

4.4 Normally distributed nonlinear joint models

4.4.1 Nonlinear mixed effect models

Nonlinear mixed effect models are an extension of the widely used linear mixed effects models. They are used to model longitudinal (repeated measures) data. These models allow the conditional expectation of the response variable, given

random effects, to be a nonlinear function of model coefficients. The formulation of the nonlinear mixed effects model presented in this section was initially proposed by Lindstrom and Bates (1990). This formulation was generalized by Pinheiro and Bates (2000) in order to accommodate time varying covariates. The generalized formulation of this model, adapted for this analysis, can be described using the equation below

$$y_{ij} = f(\boldsymbol{\theta}_i, \mathbf{x}_i, s_{ij}) + e_{ij}, \tag{45}$$

where

- The mean response is given by $f(\boldsymbol{\theta}_i, \mathbf{x}_i, s_{ij})$.
- f is a nonlinear function of the observed covariates.
- s_{ij} is the patient observation day.
- e_{ij} is the within group error term that is assumed to follow a normal distribution given as $e_{ij} \sim \mathrm{N}(0, \sigma_e^2)$ with σ_e^2 being the variance of the error term.
- σ_e^2 is equivalent to τ_e^{-1}, where τ_e is the precision parameter.

The parameter vector $\boldsymbol{\theta}_i$ is defined as

$$\boldsymbol{\theta}_i = \mathbf{A}_i\boldsymbol{\beta} + \mathbf{B}_i\mathbf{b}_i,$$

where $\mathbf{b}_i$ is a m-dimensional vector of random effects linked to the i^{th} subject. These random effects are assumed to follow a normal distribution given by $\mathbf{b}_i \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\psi})$ with $\boldsymbol{\psi}$ being the variance-covariance matrix; $\boldsymbol{\beta}$ is a p-dimensional vector of fixed effects. The matrices $\mathbf{A}_i$ and $\mathbf{B}_i$ are design matrices for the fixed effects and the random effects respectively. It is assumed that the within-group errors (e_{ij}) are independently distributed and that they are independent of $\mathbf{b}_i$. A further assumption is that observations corresponding to different groups are independent of each other. A more detailed review of nonlinear mixed effect models is provided by Pinheiro and Bates (2000).

As illustrated in Chapter 1, the gametocyte longitudinal profile has a nonlinear relationship with time. In this thesis the modified critical exponential nonlinear mixed effect model was applied to capture the shape of the longitudinal profile. Distiller et al. (2010) identified this as an appropriate model to illustrate the gametocyte profile. Assuming that $y_{ij} \sim N(\mu_{ij}, \tau_e^{-1})$, the critical exponential model can be defined as

$$\begin{aligned} E[y_{ij}|\mathbf{b}_i] &= \mu_{ij} \\ &= A_{ij} + [C_{ij} \times s_{ij}] \times [R_{ij}^{s_{ij}}], \end{aligned} \tag{46}$$

where

$$A_{ij} = \mathbf{x}_{Ai}^T \boldsymbol{\beta_A} + b_{Ai},$$
$$C_{ij} = \mathbf{x}_{Ci}^T \boldsymbol{\beta_C} + b_{Ci},$$
$$R_{ij} = \mathbf{x}_{Ri}^T \boldsymbol{\beta_R} + b_{Ri},$$

with

- $\{\mathbf{x}_{Ai}^T, \mathbf{x}_{Ci}^T, \mathbf{x}_{Ri}^T\}$ being a set of observed explanatory variables for the components $\{A_{ij}, C_{ij}, R_{ij}\}$, where the different components are allowed to have different explanatory variables.
- $\{\boldsymbol{\beta_A}, \boldsymbol{\beta_C}, \boldsymbol{\beta_R}\}$ are the fixed effects for the different model components.
- $\{b_{Ai}, b_{Bi}, b_{Ci}\}$ are the random effects for the different model components.

Since only baseline covariates were considered in this analysis, it follows that $\{A_{ij}, C_{ij}, R_{ij}\}$ is equivalent to $\{A_i, C_i, R_i\}$.

In this analysis patients who entered the study had no gametocytes at day 0, as a result it can be seen that the A_i parameter would not be required. The modified critical exponential model is a simplification of Equation 46, whereby the A_i component is removed from the model. As a result the modified critical exponential model is given as

$$\mu_{ij} = [C_i \times s_{ij}] \times [R_i^{s_{ij}}]. \tag{47}$$

This formulation gives an expected value of 0, on day 0, which is consistent with the underlying data. Figure 26 illustrates the various shapes that the model can take, across varying values of the C and R components. These plots are superimposed onto the observed mean gametocyte profiles for patients receiving sulfadoxine-pyrimethamine (SP) treatment. Figure 27 illustrates how the C (set to 0.8) and R (set to 0.91) components combine to give the shape of the modified critical exponential model.

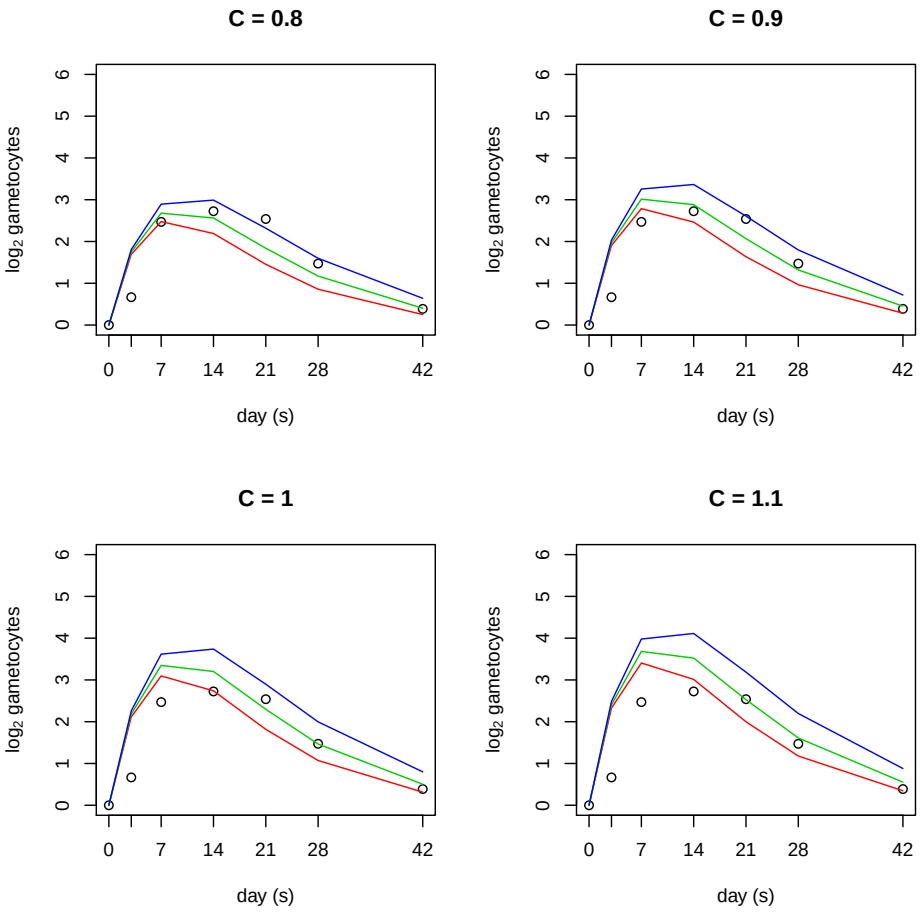

Figure 26: Shapes of the modified critical exponential Model by varying C and R parameters: $R = 0.89$ (Red line) , $R = 0.90$ (Green line), $R = 0.91$ (Blue line) and circles are the mean gametocyte profile for patients receiving SP treatment

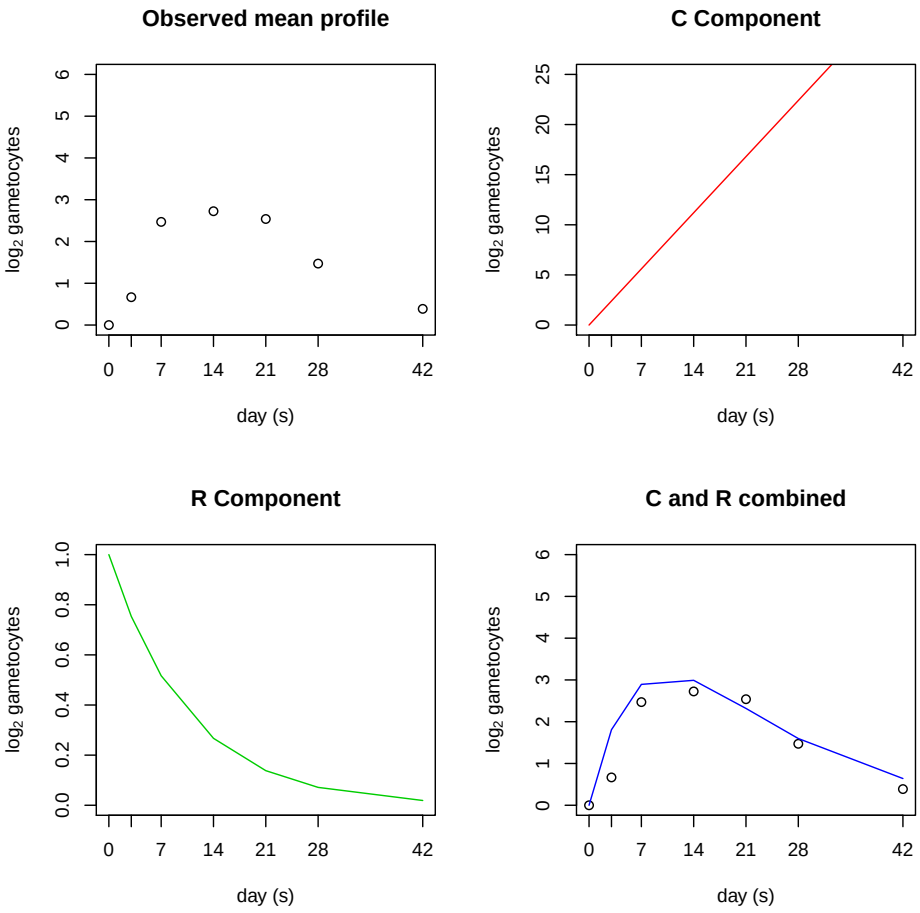

Figure 27: Build-up of the modified critical exponential Model for $C = 0.8$ and $R = 0.91$, with the circles representing the mean gametocyte profile for patients receiving SP treatment

4.4.2 Extension to a normally distributed nonlinear joint model

In this section the nonlinear mixed effect model, outlined above, was extended to a joint model that combined the longitudinal gametocyte profile to the time to early exit survival model. The longitudinal component of the model can be defined as

$$y_{ij}|\mathbf{b}_i \sim N(\mu_{ij}, \tau_e^{-1}),$$

where

$$\mu_{ij}|\mathbf{b}_i = [C_i \times s_{ij}] \times [R_i^{s_{ij}}]$$

with

$$C_i = \beta_{C0} + (trt_i \times \beta_{C1}) + (ratio_i \times \beta_{C2}) + (anaemia_i \times \beta_{C3}) + (mut5_i \times \beta_{C4}) + (lage_i \times \beta_{C5}) + (pzero_i \times \beta_{C6}) + (gender_i \times \beta_{C7}) + c_i$$

and

$$R_i = \beta_{R0} + (trt_i \times \beta_{R1}) + (ratio_i \times \beta_{R2}) + (anaemia_i \times \beta_{R3}) + (mut5_i \times \beta_{R4}) + (lage_i \times \beta_{R5}) + (pzero_i \times \beta_{R6}) + (gender_i \times \beta_{R7}) + r_i.$$

The random effects c_i and r_i (collectively referred to as $\mathbf{b}_i$) are assumed to follow normal distributions such that $c_i \sim N(0, \tau_c^{-1})$ and $r_i \sim N(0, \tau_r^{-1})$. Non-informative prior distributions were applied to both the fixed and random effects of the longitudinal component of the model. The fixed effects were assumed to follow normal distributions such that $\beta_{Cl} \sim N(0, 10000)$ and $\beta_{Rl} \sim N(0, 10000)$ $\forall l = 1, \ldots, 7$. The precision parameters and random error were all assumed to follow a gamma distribution such that $\tau_k \sim \Gamma(0.001, 0.001)$ $\forall k = c, r, e$.

The survival component of the model was assumed to be either a standard exponential or Weibull parametric model. These models were described in detail in Chapter 3.

Assuming that the time to early exit (t_i) follows a Weibull distribution, such that $t_i \sim W(\rho, \lambda_i)$, with ρ as the shape parameter that is greater than zero and that the same baseline covariates applied in the longitudinal submodel were applied to the survival submodel; the following parametric survival model applies

$$\begin{aligned}\log(\lambda_i) =& \beta_0 + (trt_i \times \beta_1) + (ratio_i \times \beta_2) + (anaemia_i \times \beta_3) + (mut5_i \times \beta_4) \\ &+ (lage_i \times \beta_5) + (pzero_i \times \beta_6) + (gender_i \times \beta_7) + (\omega_c \times c_i) + (\omega_r \times r_i),\end{aligned} \tag{48}$$

where c_i and r_i are the same random effects applied in the longitudinal model, with ω_c and ω_r as parameters that measure the strength of the association

between the survival and longitudinal components of the joint models. The same non-informative prior distributions, applied to the fixed effects used in the longitudinal model, were applied to the fixed effects given in Equation 48; including the ω_c and ω_r parameters. The shape parameter, ρ, was assumed to follow a uniform distribution such that $\rho \sim U(0, 100)$. The above result can be simplified to an exponential model by setting $\rho = 1$.

The two components, outlined above, are connected through the use of common random effects in both components of the model. The parameters, ω_c and ω_r, are used to scale the random effects in the survival model. The ω_c and ω_r parameters can thus be seen to be jointly estimated from both components. A further connection between the two components arises from the use of common covariates across both components.

The likelihood function for the joint model specified above can be expressed in the following manner. Consider it given that, $\mathbf{y}$ and $\mathbf{t}$ are representative of the observed longitudinal and survival data, whereby $\mathbf{t}$ also accommodates the indicator variable that specifies whether a patient exited the study or not. In addition it is given that $\boldsymbol{\beta}_y$ and $\boldsymbol{\beta}_t$ represent the fixed effect parameters for these components of the data respectively. Additionally $\mathbf{b}$ represents the random effect parameters applied in the model with $\boldsymbol{\Sigma}$ as the corresponding covariance matrix for these effects. The likelihood function for the joint model (ignoring covariates for simplicity) is given as

$$\begin{aligned}\pi(\boldsymbol{\beta}_y, \boldsymbol{\beta}_t, \mathbf{b}, \boldsymbol{\omega}, \boldsymbol{\Sigma}, \tau_e | \mathbf{y}, \mathbf{t}) \propto \pi(\mathbf{y}|\boldsymbol{\beta}_y, \mathbf{b}, \boldsymbol{\Sigma}, \tau_e) \times \pi(\mathbf{t}|\boldsymbol{\beta}_t, \boldsymbol{\omega}, \mathbf{b}, \boldsymbol{\Sigma}) \times \pi(\boldsymbol{\beta}_y) \\ \times \pi(\boldsymbol{\beta}_t) \times \pi(\mathbf{b}|\boldsymbol{\Sigma}) \times \pi(\boldsymbol{\Sigma}) \times \pi(\boldsymbol{\omega}) \times \pi(\tau_e).\end{aligned} \tag{49}$$

The longitudinal submodel contribution to the likelihood given above simplifies to

$$\pi(\mathbf{y}|\boldsymbol{\beta}_y, \mathbf{b}, \boldsymbol{\Sigma}, \tau_e) = \prod_{i=1}^{m} \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\tau_e^{-1}}} \exp\Big(- \frac{\tau_e (y_{ij} - \mu_{ij})^2}{2} \Big), \tag{50}$$

where μ_{ij} is as defined in Equation 47 and $\mathbf{y}$ is assumed to be normally distributed.

Assuming that the Weibull survival model is applied, the corresponding survival submodel contribution is given as

$$\pi(\mathbf{t}|\boldsymbol{\beta}_t, \mathbf{b}, \boldsymbol{\omega}, \Sigma) = \prod_{i=1}^{m} \{\lambda_i \rho t_i^{\rho-1}\}^{\delta_i} \exp\{-\lambda_i t_i^{\rho}\} \tag{51}$$

where δ_i is an indicator value that takes a value of 1 when a patient exits the study and 0 otherwise and λ_i is as defined in Equation 48. This contribution can be simplified to give the contribution for the exponential model by setting $\rho = 1$. In addition this contribution can be extended to accommodate

cause-specific risks using the methodology outlined in Chapter 3.3. The resulting cause-specific risks survival submodel contribution is thus an extension of Equation 51. The model is given as

$$\pi(\mathbf{t}|\boldsymbol{\beta}_t, \mathbf{b}, \boldsymbol{\omega}, \Sigma) = \prod_{i=1}^{m} \prod_{k=1}^{h} \{\lambda_{ki}\rho_k t_i^{\rho_k - 1}\}^{\delta_{ki}} \exp\{-\lambda_{ki} t_i^{\rho_k}\}, \tag{52}$$

where h represents the set of distinct causes of early exit, that is loss-to-follow-up or treatment failure, with δ_{ki} as an indicator variable that gives a value of 1 when the k^{th} cause of exit occurs and 0 otherwise. In addition $\boldsymbol{\beta}_t$ is extended to accommodate the fixed effect parameters for both causes of early exit from the study.

The models outlined above are able to account for the overarching shape of the gametocyte profile. However, they do not account for the large number of zero values in the data. In the proceeding section a model that is able account for the zero-inflation in the data will be outlined.

4.5 Zero-adjusted gamma nonlinear joint models

4.5.1 Zero-adjusted gamma nonlinear mixed effects model

The gametocyte data used in this analysis was found to be zero-inflated, as the majority of observed responses had zero values. This can be seen in Figure 28, where the distribution of the gametocyte density is illustrated for patients who received sulfadoxine-pyrimethamine (SP) treatment only and for patients who received a combination of artesunate and sulfadoxine-pyrimethamine (ACT) treatment. It can be seen that more than 80% of the observed responses were zeros, across both treatments, with 97% of the observations collected from patients receiving ACT having zero values.

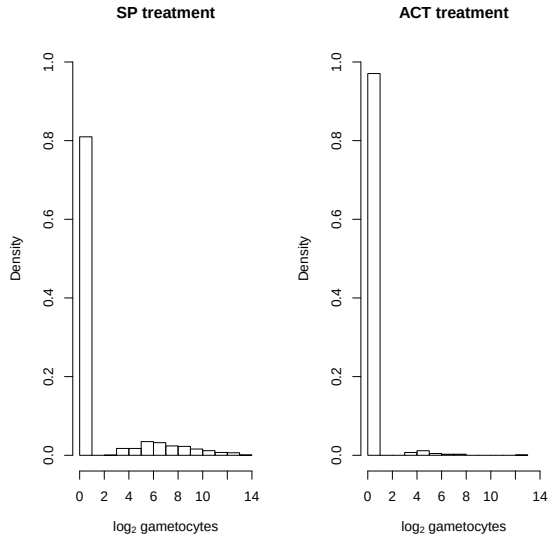

Figure 28: Distribution of $\log_2$ gametocyte density by treatment

Zero inflated data, of this nature, is fairly common in both industrial data (Lambert, 1992) and epidemiological data (Hall, 2000). Extensive work has been done in modeling zero inflated count data using discrete mixture models. Examples of these models include hurdle models (Mullahy, 1986, Heilbron, 1989), which comprise of a mixture of a point mass at zero combined with a truncated discrete count model for values greater than zero.

The zero inflated model (Heilbron, 1994, Lambert, 1992) is an alternative model that can used. It comprises of a mixture of a point mass at zero and an untruncated discrete count model. A key characteristic of this type of model is that zero values are included in both parts of the mixture model. This allows

for zeros to be categorized as "structural" (zeros that are imposed due to the design of the experiment) and "chance" (zeros that occur randomly) (Neelon et al., 2010). Zero deflation cannot be accommodated by such models as the probability of structural zeros is non-negative. Examples of such models are the zero inflated Poisson model and the zero inflated negative binomial model. Ridout et al. (1998) provides an in-depth overview of zero inflated models.

Bayesian approaches to fitting zero-inflated data have beeen well documented in literature. Ghosh et al. (2006) developed Bayesian zero-inflated Poisson models for cross-sectional data. This approach involved the use of Markov chain Monte-Carlo simulations with a "data augmentation" step (Tanner and Wong, 1987), which allowed the authors to obtain posterior samples. Neelon et al. (2010) presented a Bayesian modeling approach for repeated measures zero-inflated count data, which allowed for correlated random effects in both parts of the mixture model.

In this study the non-zero data was considered as being continuous. As a result, a "zero-adjusted" model was applied to the data as opposed to a "zero-inflated" model. A zero-adjusted model combines discrete and continuous distributions. The discrete component of the model applies the Bernoulli distribution to account for the probability of gametocytemia. The continuous component of the model accounts for the non-zero values in the data, which were assumed to follow a gamma distribution. This type of model is referred to as a zero-adjusted gamma model (ZAG). A variation of the three parameter ZAG, defined by Rigby and Stasinopoulos (2010), is outlined below

$$f(y_{ij}|\mu_{ij},\tau,p_{ij}) = \begin{cases} 1-p_{ij} & \text{if } y_{ij}=0 \\ p_{ij}\times\left\{\dfrac{y_{ij}^{\tau-1}\times\exp\left[-\left(\dfrac{y_{ij}\times\tau}{\mu_{ij}}\right)\right]}{\Gamma(\tau)\times\left(\dfrac{\mu_{ij}}{\tau}\right)^{\tau}}\right\} & \text{if } y_{ij}>0, \end{cases}$$

where p_{ij} is the probability that gametocytes are present. It follows that $0 < p_{ij} < 1$, $\mu_{ij} > 0$ and $\tau > 0$; where $\tau = \frac{1}{\sigma^2}$, with σ as the dispersion parameter. The properties of this distribution are

- $\mathrm{E}(Y_{ij}|\mathbf{b}_i) = p_{ij}\mu_{ij}$
- $\mathrm{Var}(Y_{ij}|\mathbf{b}_i) = p_{ij}\mu_{ij}^2\{(1-p_{ij}) + \frac{1}{\tau}\}$.

This model can be applied using the Generalized Additive Models for Location, Scale and Shape (GAMLSS) framework proposed by Rigby and Stasinopoulos (2005). This framework is very flexible and it can be applied to a wide range

of distributions, including distributions that are not part of the exponential family of distributions.

The ZAG model outlined above can be extended to allow for covariates and random effects under a mixed effect regression framework. Using the predefined notation given at the start of this chapter, the two components of the ZAG regression model are defined as

$$\text{logit}(p_{ij}|\mathbf{b}_{1i}) = g(\mathbf{x}_{1i}, \mathbf{b}_{1i}, \boldsymbol{\theta}_1) \tag{53}$$

and

$$\log(\mu_{ij}|\mathbf{b}_{2i}) = h(\mathbf{x}_{2i}, \mathbf{b}_{2i}, \boldsymbol{\theta}_2), \tag{54}$$

where p_{ij} is the probability that gametocytes are present, with g and h as nonlinear functions, $\mathbf{x}_{ki}$ $(\forall k = 1, 2)$ are covariate vectors, $\boldsymbol{\theta}_k$ $(\forall k = 1, 2)$ are vectors of fixed effect coefficients and $\mathbf{b}_{ki}$ $(\forall k = 1, 2)$ denote random effect parameters. In this investigation the same set of covariates were applied in the Bernoulli and gamma components of the model, thus $\mathbf{x}_{ki}$ can be replaced with $\mathbf{x}_i$ in the above formulation. The covariates in the two aforementioned components can be allowed to differ when there is a prior scientific reason to believe that the covariates are different or if the aim of the investigation is to find the most parsimonious model, which would arise due to extensive model building (Neelon et al., 2010). In this investigation, no model building will be applied as the aim is to compare different joint models using the same set of covariates. It will also be assumed that only random intercepts are required to capture the heterogeneity between patients in the study. The implication of this assumptions is that $\mathbf{b}_{1i} = b_{1i}$ and $\mathbf{b}_{2i} = b_{2i}$.

Prior distributions for parameters are a requirement when undertaking a Bayesian analysis. The fixed effects in this analysis are assumed to have an independent, multivariate normal prior distribution such that

$$\boldsymbol{\theta}_k \sim N_p(\boldsymbol{\theta}_0, \sigma_\theta^2 \mathbf{I}_s),$$

where $\boldsymbol{\theta}_0$ and σ_θ^2 are known hyper-parameters with $\mathbf{I}_s$ as a $s \times s$ identity matrix. The prior distribution applied to the random effects was a bivariate normal distribution with independent random effects that are represented as below

$$\mathbf{b}_i = \begin{pmatrix} b_{1i} \\ b_{2i} \end{pmatrix} \sim N_2(\mathbf{0}, \boldsymbol{\Sigma}),$$

where

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}.$$

Alternative prior distributions can be applied to the random effects. These include the diffuse inverse-Wishart distribution (Neelon et al., 2010) and the product normal parameterization for the random effects (Spiegelhalter, 1998, Cooper et al., 2007). In this analysis the simplified covariance structure, given above, was applied as it allowed for easy comparison across the models fit in this analysis.

Assuming prior independence, the joint posterior distribution of the model parameters can be defined as

$$\pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \mathbf{b}, \boldsymbol{\Sigma}|\mathbf{y}) \propto \pi(\mathbf{y}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \mathbf{b}, \boldsymbol{\Sigma})\pi(\boldsymbol{\theta}_1)\pi(\boldsymbol{\theta}_2)\pi(\mathbf{b}|\boldsymbol{\Sigma})\pi(\boldsymbol{\Sigma})$$

$$\propto \prod_{i=1}^{m}\prod_{j=1}^{n_i}(1-p_{ij})^{(1-d_{ij})}\left\{p_{ij} \times \frac{y_{ij}^{\tau-1}\exp\left[-\left(\frac{y_{ij}\times\tau}{\mu_{ij}}\right)\right]}{\Gamma(\tau)\times\left(\frac{\mu_{ij}}{\tau}\right)^{\tau}}\right\}^{d_{ij}}\pi(\boldsymbol{\theta}_1)\pi(\boldsymbol{\theta}_2)\pi(\mathbf{b}|\boldsymbol{\Sigma})\pi(\boldsymbol{\Sigma}), \tag{55}$$

where $\pi(\mathbf{y}|.)$ is the likelihood for the data ($\mathbf{y}$) given the model parameters, with d_{ij} as a indicator variable with a value of 1 when gametocytes are present and 0 otherwise; $\pi(\boldsymbol{\theta}_1)$,$\pi(\boldsymbol{\theta}_2)$,$\pi(\Sigma)$ are prior densities for $\boldsymbol{\theta}_1$,$\boldsymbol{\theta}_2$ and Σ.

Equations 53 and 54 introduced two nonlinear functions, g and h, which would be required to fit the zero-adjusted gamma (ZAG) model. These functions, $g(.)$ and $h(.)$, would affect the prevalence and gamma distributed components of the model respectively. These functions would have to accommodate the nonlinear profiles of the prevalence of gametocytemia as well as the nonlinear nature of the observed gametocyte profiles. It has already been shown that the modified critical exponential model provides an appropriate representation of the nonlinear relationship between gametocyte density and time. As a result it would be expected that the functions to apply in the zero-adjusted gamma analysis would be based on the modified critical exponential model. Unfortunately the modified critical exponential model cannot be used when modeling the prevalence and gamma components of the ZAG model. This is because at day 0 the prevalence model would provide a predicted probability of 0.5 ($= \frac{\exp([C_i\times 0]\times[R_i^0])}{1+\exp([C_i\times 0]\times[R_i^0])}$) while the gamma component would give an expected density of 1 ($= \exp([C_i \times 0] \times [R_i^0])$. This is not a desirable result as the patients analyzed in this study did not have gametocytes on day 0. It is clear that in both cases there is a need for an additional component to be included in the model, which can make an adjustment to the predicted profiles at day 0. This requirement led to the use of the previously described critical exponential model

(Equation 46). As a result the prevalence model that was applied is defined as

$$\begin{aligned}\text{logit}(p_{ij}|\mathbf{b}_{1i}) &= g(\mathbf{x}_{1ij}, \mathbf{b}_{1i}, \boldsymbol{\beta}_1) \\ &= BA_i + [(BC_i \times s_{ij}) \times BR_i^{s_{ij}}],\end{aligned} \tag{56}$$

this leads to

$$p_{ij}|\mathbf{b}_{1i} = \frac{\exp\{BA_i + [(BC_i \times s_{ij}) \times BR_i^{s_{ij}}]\}}{1 + \exp\{BA_i + [(BC_i \times s_{ij}) \times BR_i^{s_{ij}}]\}}. \tag{57}$$

where

$$\begin{aligned}BA_i = \beta_{BA0} &+ (trt_i \times \beta_{BA1}) + (ratio_i \times \beta_{BA2}) + (anaemia_i \times \beta_{BA3}) + (mut5_i \times \beta_{BA4}) \\ &+ (lage_i \times \beta_{BA5}) + (pzero_i \times \beta_{BA6}) + (gender_i \times \beta_{BA7}),\end{aligned}$$

with

$$\begin{aligned}BC_i = \beta_{BC0} &+ (trt_i \times \beta_{BC1}) + (ratio_i \times \beta_{BC2}) + (anaemia_i \times \beta_{BC3}) + (mut5_i \times \beta_{BC4}) \\ &+ (lage_i \times \beta_{BC5}) + (pzero_i \times \beta_{BC6}) + (gender_i \times \beta_{BC7}) + bc_i,\end{aligned}$$

and

$$\begin{aligned}BR_i = \beta_{BR0} &+ (trt_i \times \beta_{BR1}) + (ratio_i \times \beta_{BR2}) + (anaemia_i \times \beta_{BR3}) + (mut5_i \times \beta_{BR4}) \\ &+ (lage_i \times \beta_{BR5}) + (pzero_i \times \beta_{BR6}) + (gender_i \times \beta_{BR7}) + br_i.\end{aligned}$$

It follows that $\mathbf{b}_{1i}$ refers to the random effects bc_i and br_i. It is evident that no random effects were placed on the BA component of the model. This is because any random patient level variation in the upward trajectory of the gametocyte profile, can be sufficiently accommodated by the BC component of the model. At the same time the patient level variability in the downward trajectory is accounted for by the BR component.

The different shapes of the critical exponential model, for the prevalence model, are shown in Figure 29. In addition Figure 30 gives a graphical representation of the transformation steps that give the predicted probability of gametocytemia. This figure gives the profile of $BA_i + [(BC_i \times s_{ij}) \times BR_i^{s_{ij}}]$ and subsequently the impact of taking the exponent of that profile before finally converting it into a probability.

The critical exponential model was also applied to the gamma distributed component of the ZAG model. The resulting model is given as

$$\log(\mu_{ij}|\mathbf{b}_{2i}) = A_i + [(C_i \times s_{ij}) \times R_i^{s_{ij}}]. \tag{58}$$

where

$$\begin{aligned}A_i = \beta_{A0} &+ (trt_i \times \beta_{A1}) + (ratio_i \times \beta_{A2}) + (anaemia_i \times \beta_{A3}) + (mut5_i \times \beta_{A4}) \\ &+ (lage_i \times \beta_{A5}) + (pzero_i \times \beta_{A6}) + (gender_i \times \beta_{A7})\end{aligned}$$

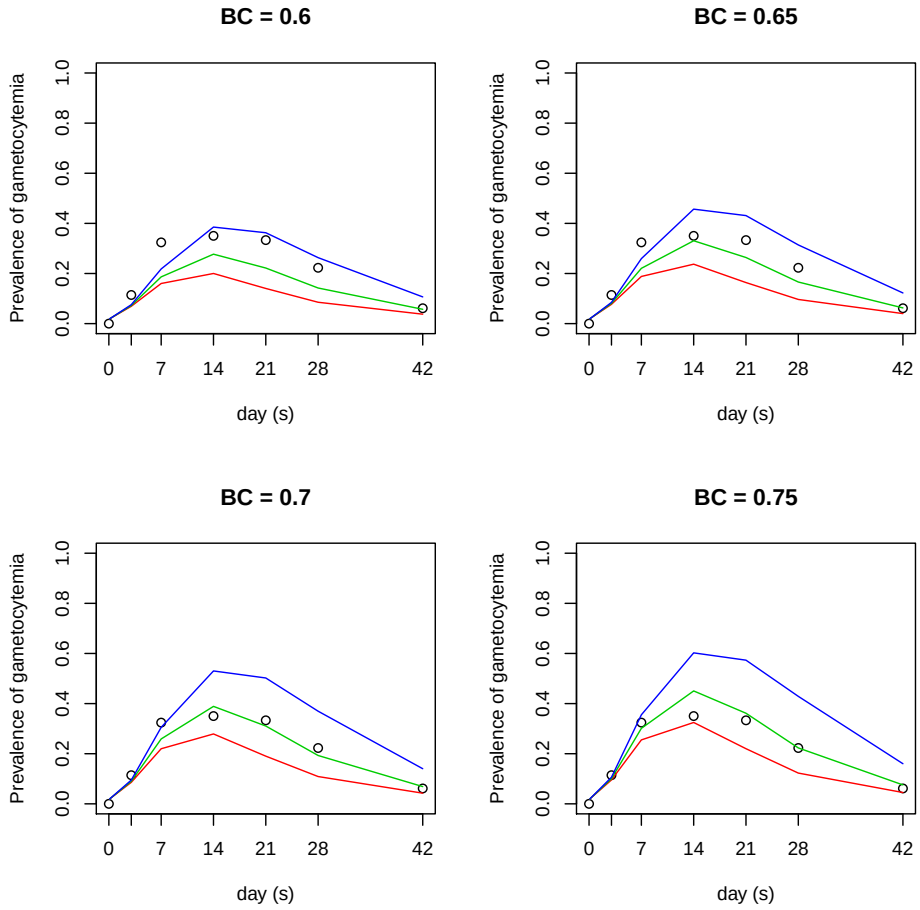

Figure 29: Shapes of the critical exponential model for the prevalence of gametocytemia by varying BC and BR parameters ($BR = 0.92$ (Red line), $BR = 0.93$ (Green line), $BR = 0.94$ (Blue line)), with BA set to -4; where the circles represent the observed prevalence of gametocytemia for patients receiving SP treatment.

with

$$C_i = \beta_{C0} + (trt_i \times \beta_{C1}) + (ratio_i \times \beta_{C2}) + (anaemia_i \times \beta_{C3}) + (mut5_i \times \beta_{C4}) + (lage_i \times \beta_{C5}) + (pzero_i \times \beta_{C6}) + (gender_i \times \beta_{C7}) + c_i$$

and

$$R_i = \beta_{R0} + (trt_i \times \beta_{R1}) + (ratio_i \times \beta_{R2}) + (anaemia_i \times \beta_{R3}) + (mut5_i \times \beta_{R4}) + (lage_i \times \beta_{R5}) + (pzero_i \times \beta_{R6}) + (gender_i \times \beta_{R7}) + r_i,$$

where $\mathbf{b}_{2i}$ refers to the random effects c_i and r_i. Similarly to the prevalence

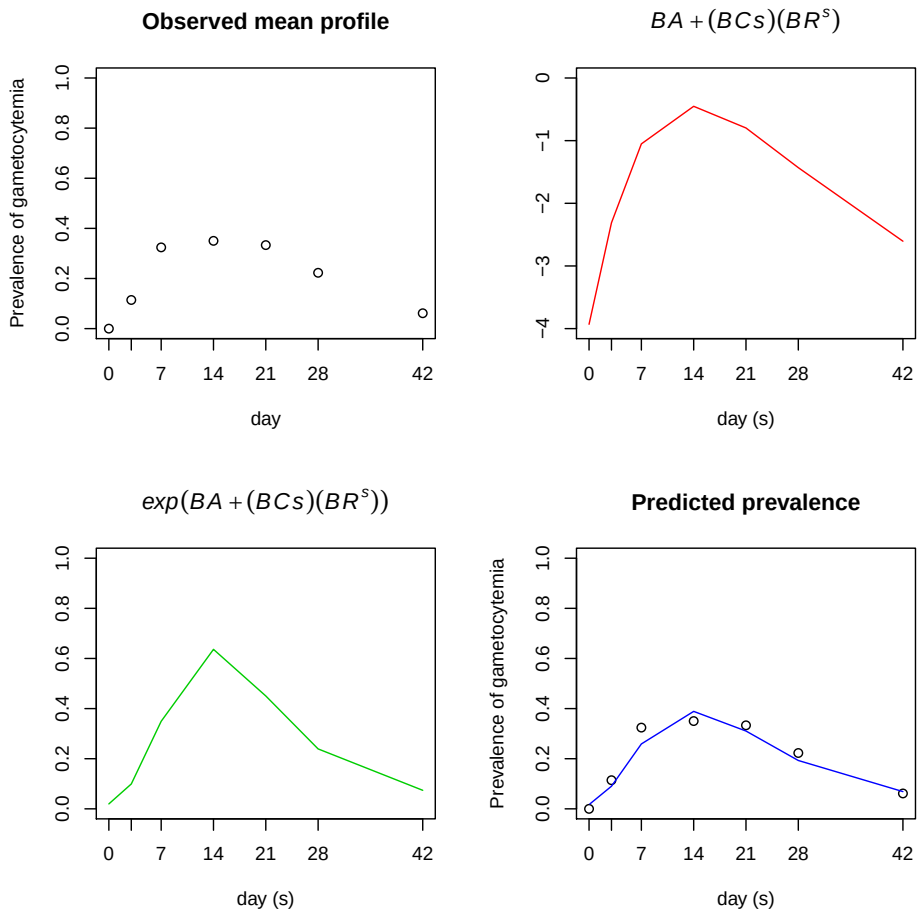

Figure 30: Build up of the critical exponential model for prevalence of gametocytemia where BA = -4, $BC = 0.7$ and $BR = 0.93$ with the circles representing the observed prevalence of gametocytemia for patients receiving SP treatment.

model, no random effects were placed on the A component of the model.

The various shapes of the critical exponential model for the gamma component of the zero-adjusted gamma (ZAG) model are shown in Figure 31, with the impact of the transformation steps used to get the expected response shown in Figure 32.

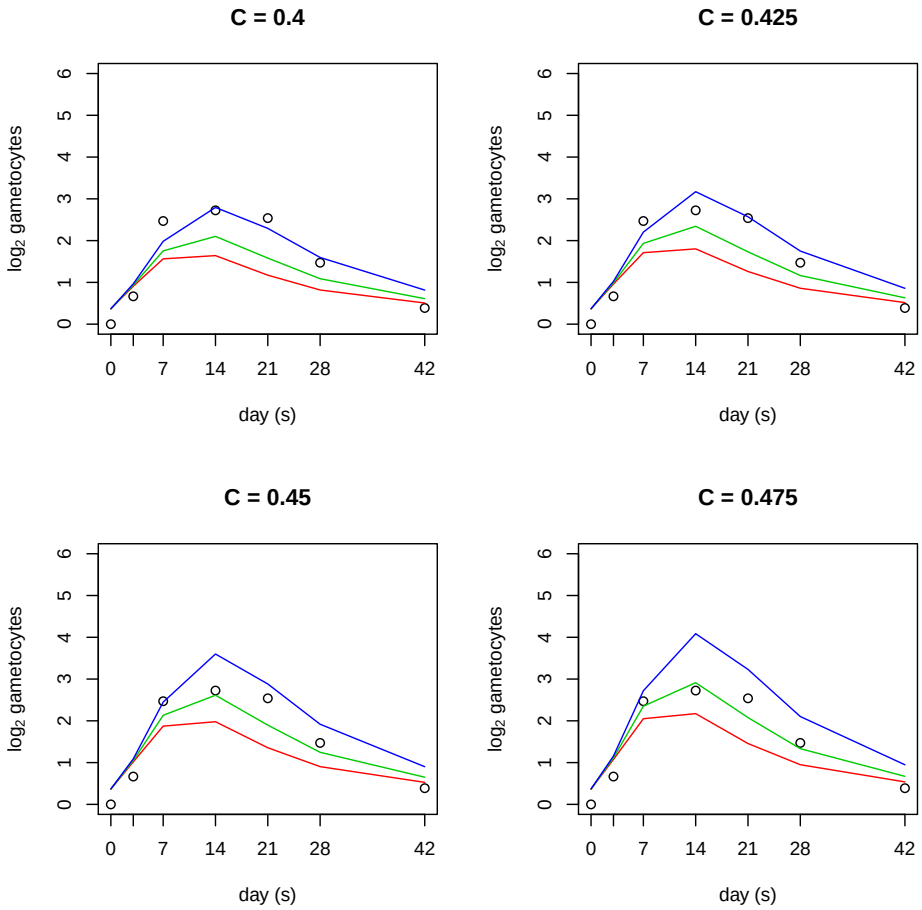

Figure 31: Shapes of the critical exponential model for the gamma distributed continuous component of the ZAG model by varying C and R parameters: $R =$ 0.91 (Red line), $R = 0.92$ (Green line), $R = 0.93$ (Blue line) with $A =$ -1 and the circles representing the mean gametocyte profile for patients receiving SP treatment.

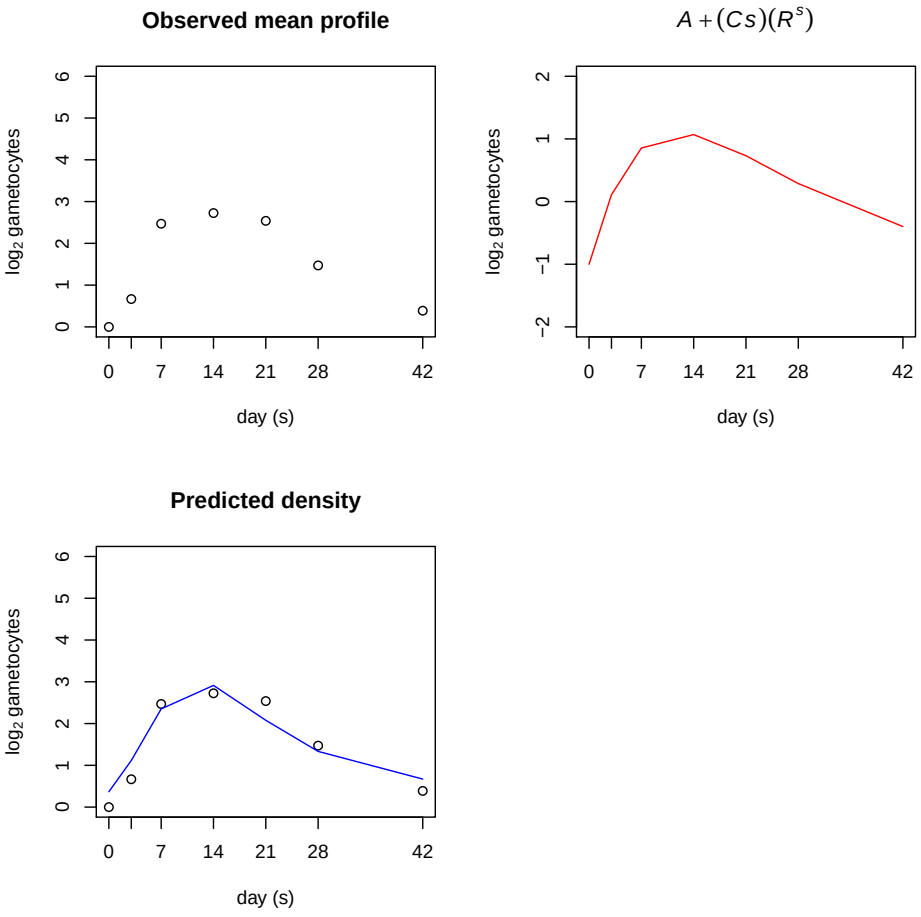

Figure 32: Build up of the critical exponential model for the gamma distributed continuous component of the ZAG model where A = -1, C = 0.475 and R = 0.92 with the circles representing the mean gametocyte profile for patients receiving SP treatment.

4.5.2 Extension to a zero-adjusted gamma nonlinear joint model

The zero-adjusted gamma (ZAG) model outlined above can be extended to a joint model that combines the longitudinal gametocyte profile with the time to early exit survival model. The longitudinal component of the model is split into a prevalence component and a continuous component as previously shown by Equations 57 and 58 respectively.

The random effects bc_i and br_i (collectively referred to as $\mathbf{b}_{1i}$), along with c_i and r_i (collectively referred to as $\mathbf{b}_{2i}$) are assumed to follow normal distributions. This implies that $bc_i \sim N(0, \tau_{bc}^{-1})$, $br_i \sim N(0, \tau_{br}^{-1})$, $c_i \sim N(0, \tau_c^{-1})$ and $r_i \sim N(0, \tau_r^{-1})$. Non-informative prior distributions were applied to both the fixed and random effects of the longitudinal component of the model. The fixed effects were assumed to follow normal distributions such that $\beta_{Cl} \sim N(0, 10000)$, $\beta_{Rl} \sim N(0, 10000)$, $\beta_{BCl} \sim N(0, 10000)$ and $\beta_{BRl} \sim N(0, 10000)$ $\forall l = 1, \ldots, 7$. The random effects and random error were all assumed to follow a gamma distribution such that $\tau_k \sim \Gamma(0.001, 0.001)$ $\forall k = c, r, bc, br, e$.

The survival model, shown in Equation 48, was extended to accommodate the prevalence random effects such that it became

$$\begin{aligned}\log(\lambda_i) =& \beta_0 + (trt_i \times \beta_1) + (ratio_i \times \beta_2) + (anaemia_i \times \beta_3) + (mut5_i \times \beta_4) \\ &+ (lage_i \times \beta_5) + (pzero_i \times \beta_6) + (gender_i \times \beta_7) \\ &+ (\omega_c \times c_i) + (\omega_r \times r_i) + (\omega_{bc} \times bc_i) + (\omega_{br} \times br_i),\end{aligned} \tag{59}$$

where c_i, r_i, bc_i and br_i are the same random effects applied in the longitudinal model, with ω_c, ω_r, ω_{bc} and ω_{br} as parameters that measure the strength of the association between the survival and longitudinal components of the fitted joint models. The same non-informative prior distributions, applied to the fixed effects used in the longitudinal model, were applied to the fixed effects given in Equation 59. The shape parameter, ρ, is assumed to follow a uniform distribution such that $\rho \sim U(0, 100)$. The Weibull survival model, given above, can be simplified to an exponential model by setting $\rho = 1$.

By applying the same notation as that used in the previous section it follows that, the likelihood function for the joint model (ignoring covariates for simplicity) specified above is given as

$$\begin{aligned}\pi(\boldsymbol{\beta}_y, \boldsymbol{\beta}_t, \mathbf{b}, \boldsymbol{\omega}, \boldsymbol{\Sigma}, \tau | \mathbf{y}, \mathbf{t}) \propto& \pi(\mathbf{y}|\boldsymbol{\beta}_y, \mathbf{b}, \boldsymbol{\Sigma}, \tau) \times \pi(\mathbf{t}|\boldsymbol{\beta}_t, \boldsymbol{\omega}, \mathbf{b}, \boldsymbol{\Sigma}) \times \pi(\boldsymbol{\beta}_y) \\ &\times \pi(\boldsymbol{\beta}_t) \times \pi(\mathbf{b}|\boldsymbol{\Sigma}) \times \pi(\boldsymbol{\Sigma}) \times \pi(\boldsymbol{\omega}) \times \pi(\tau).\end{aligned} \tag{60}$$

The contribution of the longitudinal component to the likelihood is given as

$$\pi(\mathbf{y}|\boldsymbol{\beta}_y, \mathbf{b}, \boldsymbol{\Sigma}, \tau) = \prod_{i=1}^{m}\prod_{j=1}^{n_i}(1-p_{ij})^{(1-d_{ij})}\Big\{\frac{y_{ij}^{\tau-1}\exp[-(\frac{y_{ij}\tau}{\mu_{ij}})]}{(\frac{\mu_{ij}}{\tau})^{\tau}\Gamma(\tau)}\Big\}^{d_{ij}}, \tag{61}$$

where μ_{ij} and p_{ij} are as previously defined in Equations 57 and 58.

Assuming that the Weibull model is applied, the corresponding survival submodel contribution is given as

$$\pi(\mathbf{t}|\boldsymbol{\beta}_t, \mathbf{b}, \boldsymbol{\omega}, \Sigma) = \prod_{i=1}^{m}\{\lambda_i\rho t_i^{\rho-1}\}^{\delta_i}\exp\{-\lambda_i t_i^{\rho}\}, \tag{62}$$

where δ_i is an indicator variable having a value of 1 if exit occurs and 0 otherwise. The survival submodel provided above can also be extended to accommodate cause-specific risks resulting in a survival submodel contribution given by

$$\pi(\mathbf{t}|\boldsymbol{\beta}_t, \mathbf{b}, \boldsymbol{\omega}, \Sigma) = \prod_{i=1}^{m}\prod_{k=1}^{h}\{\lambda_{ki}\rho_k t_i^{\rho_k-1}\}^{\delta_{ki}}\exp\{-\lambda_{ki} t_i^{\rho_k}\},$$

where h represents the set of distinct causes of early exit, that is loss-to-follow-up or treatment failure, with δ_{ki} as an indicator variable that gives a value of 1 when the k^{th} cause of exit occurs and 0 otherwise; in addition $\boldsymbol{\beta}_t$ is extended to accommodate the fixed effect parameters for both causes of early exit from the study.

4.6 Model estimation

The models outlined above were computed using the Gibbs sampling MCMC algorithm. This algorithm iteratively samples from the full conditional distributions of the parameters used in the models. The Gibbs sampling algorithm was implemented using JAGS (Plummer, 2003), through the *jagsUI* (Kellner, 2016) package in R (R Core Team, 2015).

The zero-adjusted gamma (ZAG) distribution is not a pre-designated distribution in JAGS, thus a likelihood for the distribution needed to be defined as part of the fitting process. This was done by applying the "zeros trick" that is defined in Spiegelhalter et al. (2003). This trick converts the likelihood contributions for the observed data into a format that can be processed by JAGS. Given that responses (y_{ij}) follow a ZAG distribution, with each observation contributing a likelihood term L_{ij}, the "zeros trick" would involve the creation of a dummy set of observations (z_{ij}) with values of zero, which are assumed to follow a Poisson distribution with mean ψ_{ij}. The associated likelihood contributions for these dummy values would be $\exp(-\psi_{ij})$. Setting ψ_{ij} equal to $-\log L_{ij}$ would result in JAGS processing the correct likelihood contribution for y_{ij}. Since ψ_{ij} should always be greater than 0, a large positive value is usually included in the formulation of ψ_{ij} such that

$$\psi_{ij} = -\log L_{ij} + C,$$

where C was considered as 1000 in this analysis.

The same modeling strategy was applied to all fitted models in this analysis, to allow for consistency in approach. This strategy involved initializing three chains with 30 000 iterations for each chain, with 15 000 of these iterations being considered as burn-in iterations. In order to reduce autocorrelation "thinning" was applied. This involved retaining only the 10^{th} value of each chain as part of the estimation procedure. The median of the posterior samples was used as the point estimate of the parameters fitted in this chapter, with the 2.5 and 97.5 percentiles of the posterior samples used as the 95% credibility interval.

4.7 Model selection

In this analysis no variable selection was conducted, as this would have made it difficult to compare the different model structures fit to the data. The only selection process implied in this analysis was based on determining the most appropriate underlying model structure, based on a fixed set of covariates. The covariates used in this analysis were

- $\log_{10}$ baseline asexual parasite density (*pzero*).
- Prevalence of quintuple mutations (*mut*5), defined as 1 for presence and 0 for absence of quintuple mutations.
- Treatment, defined as sulfadoxine-pyrimethamine treatment only (SP) or a combination of artesunate and sulfadoxine-pyrimethamine (ACT) (*trt*). SP was given a value of 0 and ACT was given a value of 1.
- Parasite reduction ratio at 24 hours (*ratio*).
- Gender of patient, with males given a value of 1 and females given a value of 0 (*gender*).
- Prevalence of moderate anaemia (*anaemia*), where moderate anaemia was defined as having a haemoglobin density less than 11g/dL. Patients with moderate anaemia were given a value of 1 while patients with a haemoglobin density greater than 11g/dL were given a value of 0.
- $\log_2$ of patient age (in years) (*lage*).

There are several procedures that can be used to determine the optimal model, from a set of models developed as part of the model building stage of an analysis. One such procedure involves taking an average of all posterior probabilities that each one of the fitted models is the true model and then subsequently obtaining posterior parameters (Hoeting et al., 1999). This type of method is computationally intensive as it requires the use of procedures that can provide coverage for the entire model space, an example being the reversible jump MCMC (Green, 1995).

Another method involves assessing the ratios of posterior probabilities from competing models. These ratios are called Bayes factors (Kass and Raftery, 1995). Given two models M_0 and M_1, the Bayes factors in favor of model M_0 is defined as

$$B_{01} = \frac{\pi(\mathbf{y}|M_0)}{\pi(\mathbf{y}|M_1)},$$

where $\mathbf{y}$ is a vector of observed data, with $\pi(\mathbf{y}|M_0)$ and $\pi(\mathbf{y}|M_1)$ as the marginal likelihoods of $\mathbf{y}$ under the respective models. It is important to note that Bayes factors require the use of proper priors (Neelon et al., 2010).

In this analysis the deviance information criterion (DIC) was used as a model adequacy and goodness of fit measure. It was proposed by Spiegelhalter et al. (2002) as a Bayesian model selection tool. This measure is similar to the Akaike information criterion (AIC) and Bayesian information criterion (BIC), which

are both measures commonly used under the frequentist framework. The DIC applies a penalty to offset the gains in a model fit that arise due to an increase in the complexity of a model, this is similar to the approach applied by the AIC and BIC. It is fairly straight-forward to determine the complexity of the fixed effects in a model as this is defined as the number of model parameters. The same cannot be said when trying to assess the complexity of the random effects. The impact of random effects, on the number of model parameters, is determined by the variance of these effects. Random effects with a large variance contribute approximately one parameter while random effects with a very small variance contribute approximately zero parameters (Elliott et al., 2005). The DIC is able to estimate the effective number of parameters in a fitted mixed effect model. The DIC for a model M is defined as

$$\begin{aligned} \text{DIC}(M) &= 2\overline{D(\boldsymbol{\theta}_m, M)} - D(\overline{\boldsymbol{\theta}}_m, M) \\ &= D(\overline{\boldsymbol{\theta}}_m, M) + 2p_m \end{aligned} \tag{63}$$

where $\boldsymbol{\theta}_m$ is a vector of model parameters used in model M, $\mathbf{y}$ is a vector of observed data, $D(\boldsymbol{\theta}_m, M)$ is the Bayesian deviance defined as

$$D(\boldsymbol{\theta}_m, M) = -2\log[\pi(y|\boldsymbol{\theta}, M)],$$

$\overline{\boldsymbol{\theta}}_m$ and $\overline{D(\boldsymbol{\theta}_m, M)}$ are the means of the posterior distributions of $\boldsymbol{\theta}_m$ and $D(\boldsymbol{\theta}_m, M)$ respectively, p_m is the number of effective parameters defined as

$$p_m = \overline{D(\boldsymbol{\theta}_m, M)} - D(\overline{\boldsymbol{\theta}}_m, M).$$

The effective number of parameters (p_m) is a measure of model complexity, which penalizes the measure of the goodness of fit. A model with the lowest DIC is considered the best fitting model.

In this analysis the "zeros trick" was applied to fit the ZAG model using JAGS. This approach derives the DIC on a different scale to that used by built-in distributions in JAGS. Therefore, the DIC statistics produced from fitting the ZAG models would need to be adjusted in order to compare them to those derived from models that were fit whilst ignoring the zero-inflation in the data. The methodology used to make this adjustment was outlined by Lunn et al. (2012) and it is shown below.

Given that the Bayesian deviance, derived using the "zeros trick" (D_{zero}) refers to a Poisson model with dummy data made up of zero values, that is $z_{ij} = 0$, the sampling distribution for z_{ij} is defined as

$$
\begin{aligned}
f(z_{ij}|\psi_{ij}) &= \frac{e^{-\psi_{ij}}\psi_{ij}^{z_{ij}}}{z_{ij}!} \\
&= e^{-\psi_{ij}}.
\end{aligned}
\tag{64}
$$

As part of the model fitting process, ψ_{ij} is defined as

$$
\psi_{ij} = C - \log[g(y_{ij}|\boldsymbol{\theta})],
$$

with $g(y_{ij}|\boldsymbol{\theta})$ as the sampling distribution for the ZAG response and C as a large constant value (set as 1000 in this analysis) that ensures that ψ_{ij} remains positive. It follows that Equation 64 can be simplified to

$$
\begin{aligned}
f(z_{ij}|\psi_{ij}) &= e^{-C+\log[g(y_{ij}|\boldsymbol{\theta})]} \\
&= e^{-C} g(y_{ij}|\boldsymbol{\theta}).
\end{aligned}
\tag{65}
$$

Both sides of Equation 65 can be converted into Bayesian deviances by taking the sum, over all observations, of -2 times the logarithm of both sides of the equation leading to

$$
\begin{aligned}
\sum_{i=1}^{m}\sum_{j=1}^{n_i}\{-2\log[f(z_{ij}|\psi_{ij})]\} &= 2mn_iC + \sum_{i=1}^{n}\sum_{j=1}^{n_i}\{-2\log[g(y_{ij}|\boldsymbol{\theta})]\} \\
D_{zero} &= 2mn_iC + D,
\end{aligned}
\tag{66}
$$

where D is the Bayesian deviance measured on the scale of the observed data y_{ij}. It thus follows that the DIC on the zero scale is different to the DIC on the observed data scale by a value of $2mn_iC$. This value will be removed from the DIC extracted from JAGS, when comparing the goodness of fit of the models applied.

An additional goodness of fit test is based on the predictive loss function. This function compares the observed with the predicted response. Laud and Ibrahim (1995) and Gelfand and Ghosh (1998) proposed the use of the posterior predictive loss (PPL) and mean square predictive error (MSPE). In this analysis the MSPE will be used. Given an observed response y_{ij} and its corresponding predicted value $\widetilde{y_{ij}}$, which is a simulation from the posterior predictive distribution during simulation k ($\forall k = 1, \ldots, K$), the MSPE is defined as

$$
\text{MSPE}_k = \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \widetilde{y_{ij}})^2.
$$

The median of the generated MSPE statistics was used as a point estimate, with the 2.5 and 97.5 percentiles used as the 95% credibility interval.

4.8 Model checking

Once the optimal model was selected, the fit of the model to the observed data was assessed. This was done by assessing the posterior distribution of standardized Pearson residuals, from the selected model. These residuals were used to assess the deviation between observed and predicted values. They are defined as

$$r_{ij} = \frac{y_{ij} - E(y_{ij}|\boldsymbol{\theta})}{\sqrt{Var(y_{ij}|\boldsymbol{\theta})}},$$

with large values of r_{ij} indicating a poor model fit. It is assumed that

$$r_{ij} \sim N(0, 1),$$

thus r_{ij} values would be expected to lie between -2 and 2 if a fitted model is appropriate. The median of the simulated distribution of residuals can be plotted against the median fitted values in order to assess model adequacy. Systematic trends in the plot would indicate that the plot was not appropriate. Pearson residuals can also be derived for the missing observations. These residuals are derived as

$$r_{ij}^m = \frac{y_{ij}^m - E(y_{ij}^m|\boldsymbol{\theta})}{\sqrt{Var(y_{ij}^m|\boldsymbol{\theta})}},$$

where r_{ij}^m are the standardized Pearson residuals for the missing data, y_{ij}^m are the imputed responses produced as part of the MCMC estimation procedure with $E(y_{ij}^m|\boldsymbol{\theta})$ as the fitted values attributed to the missing observations. Assuming that y_{ij}^m follows a normal distribution with $E(y_{ij}^m|\boldsymbol{\theta}) = f(\boldsymbol{\theta}_i, \mathbf{x}_i, s_{ij}^m)$, the response y_{ij}^m is defined as

$$y_{ij}^m = f(\boldsymbol{\theta}_i, \mathbf{x}_i, s_{ij}^m) + e_{ij}$$

where e_{ij} is the within group error term as defined in Equation 45 and s_{ij}^m is the observation day associated with the missing value.

4.9 Imputation of incomplete gametocyte profiles

The main aim of fitting the joint models, derived earlier in this chapter, was to impute the incomplete gametocyte profiles for patients who exited the study early. This was achieved through the use of the posterior samples generated as part of the Bayesian MCMC procedure. Once complete gametocyte profiles have been developed, they can be used to investigate clinical research questions like the estimation of the duration of gametocytemia or the area under the curve (AUC) of the gametocyte profile. It was highlighted in Chapter 3 that the data used in this analysis was affected by interval censoring. The imputation

of incomplete gametocyte profiles mitigates the effect of interval censoring as predicted responses can be generated for the timepoints that lie in between the study's predefined observation dates. In this thesis the imputed gametocyte profiles were used to extend the analysis into the duration of gametocytemia, which was previously discussed in Chapter 3.

The methodology used to impute the incomplete gametocyte profiles and subsequently estimate the duration of gametocytemia is outlined below. The methodology will be outlined for a ZAG joint model with a time to early exit survival component that assumes that time follows an exponential distribution.

The methodology applied proceeded as follows, given that the survival component of the joint model was defined as

$$\begin{aligned}\log(\lambda_i) =&\beta_0 + (trt_i \times \beta_1) + (ratio_i \times \beta_2) + (anaemia_i \times \beta_3) + (mut5_i \times \beta_4)\\ &+ (lage_i \times \beta_5) + (pzero_i \times \beta_6) + (gender_i \times \beta_7)\\ &+ (\omega_c \times c_i) + (\omega_r \times r_i) + (\omega_{bc} \times bc_i) + (\omega_{br} \times br_i),\end{aligned}$$

with the prevalence component of the longitudinal model defined as

$$\text{logit}(p_{ij}|\mathbf{b}_{1i}) = BA_i + [(BC_i \times s_{ij}) \times BR_i^{s_{ij}}],$$

where

$$\begin{aligned}BA_i =& \beta_{BA0} + (trt_i \times \beta_{BA1}) + (ratio_i \times \beta_{BA2}) + (anaemia_i \times \beta_{BA3}) + (mut5_i \times \beta_{BA4})\\ &+ (lage_i \times \beta_{BA5}) + (pzero_i \times \beta_{BA6}) + (gender_i \times \beta_{BA7}),\end{aligned}$$

with

$$\begin{aligned}BC_i =& \beta_{BC0} + (trt_i \times \beta_{BC1}) + (ratio_i \times \beta_{BC2}) + (anaemia_i \times \beta_{BC3}) + (mut5_i \times \beta_{BC4})\\ &+ (lage_i \times \beta_{BC5}) + (pzero_i \times \beta_{BC6}) + (gender_i \times \beta_{BC7}) + bc_i\end{aligned}$$

and

$$\begin{aligned}BR_i =& \beta_{BR0} + (trt_i \times \beta_{BR1}) + (ratio_i \times \beta_{BR2}) + (anaemia_i \times \beta_{BR3}) + (mut5_i \times \beta_{BR4})\\ &+ (lage_i \times \beta_{BR5}) + (pzero_i \times \beta_{BR6}) + (gender_i \times \beta_{BR7}) + br_i.\end{aligned}$$

In addition the gamma component was defined as

$$\log(\mu_{ij}|\mathbf{b}_{2i}) = A_i + [(C_i \times s_{ij}) \times R_i^{s_{ij}}],$$

where

$$\begin{aligned}A_i =& \beta_{A0} + (trt_i \times \beta_{A1}) + (ratio_i \times \beta_{A2}) + (anaemia_i \times \beta_{A3}) + (mut5_i \times \beta_{A4})\\ &+ (lage_i \times \beta_{A5}) + (pzero_i \times \beta_{A6}) + (gender_i \times \beta_{A7})\end{aligned}$$

with

$$C_i = \beta_{C0} + (trt_i \times \beta_{C1}) + (ratio_i \times \beta_{C2}) + (anaemia_i \times \beta_{C3}) + (mut5_i \times \beta_{C4}) + (lage_i \times \beta_{C5}) + (pzero_i \times \beta_{C6}) + (gender_i \times \beta_{C7}) + c_i$$

and

$$R_i = \beta_{R0} + (trt_i \times \beta_{R1}) + (ratio_i \times \beta_{R2}) + (anaemia_i \times \beta_{R3}) + (mut5_i \times \beta_{R4}) + (lage_i \times \beta_{R5}) + (pzero_i \times \beta_{R6}) + (gender_i \times \beta_{R7}) + r_i.$$

After a sufficient burn-in period, L iterations were used to generate the posterior samples used in this analysis. After each iteration l ($\forall l = 1, \ldots, L$) the following occurred.

1. Posterior draws for $\beta_{hk}^{(l)}$ ($\forall h = A, C, R, BA, BC, BR$ and $\forall k = 1, \ldots 7$) along with the random effects $c_i^{(l)}$, $r_i^{(l)}$,$bc_i^{(l)}$ and $br_i^{(l)}$ were generated.

2. The fixed and random effect posterior draws given above were subsequently used to derive $A_i^{(l)}$, $C_i^{(l)}$, $R_i^{(l)}$, $BA_i^{(l)}$, $BC_i^{(l)}$ and $BR_i^{(l)}$, using the respective linear predictors given as part of the joint model formulation

3. The posteriors draws for $A_i^{(l)}$, $C_i^{(l)}$, $R_i^{(l)}$, $BA_i^{(l)}$, $BC_i^{(l)}$ and $BR_i^{(l)}$ were then used to impute values for $p_{ij}^{(l)}$ and $\mu_{ij}^{(l)}$ for any timepoint (s_{ij}). This was achieved using the relationship

$$p_{ij}|\mathbf{b}_{1i} = \frac{\exp\{BA_i + [(BC_i \times s_{ij}) \times BR_i^{s_{ij}}]\}}{1 + \exp\{BA_i + [(BC_i \times s_{ij}) \times BR_i^{s_{ij}}]\}}.$$

   and

$$(\mu_{ij}|\mathbf{b}_{2i}) = \exp\{A_i + [(C_i \times s_{ij}) \times R_i^{s_{ij}}]\}.$$

4. The $p_{ij}^{(l)}$ and $\mu_{ij}^{(l)}$ estimates were then be used to derive the expected gametocyte density over time for the i^{th} patient in the study at the jth time, as $E(y_{ij})^{(l)} = p_{ij}^{(l)} \times \mu_{ij}^{(l)}$. In this section the j was considered to range from day 0 to 100, with daily increments.

The process above was repeated L times until L complete datasets were generated for the patients in the study. In each iteration a subset of patients who were considered to have gametocytemia was taken. Since male and female gametocytes are required for the lifecyle of the parasite to continue, it was assumed in this thesis that gametocytemia would occur if a count of more than two gametocytes was predicted; this is equivalent to having an expected logarithm to the base two gametocyte density of at least 1. Subsequently L Weibull AFT models

for the time to gametocyte clearance were fit to those subsets with the model for the l^{th} iteration defined as

$$\log T_i^{(l)} = \alpha_0^{(l)} + (trt_i \times \alpha_1^{(l)}) + (ratio_i \times \alpha_2^{(l)}) + (anaemia_i \times \alpha_3^{(l)}) + (mut5_i \times \alpha_4^{(l)}) + (lage_i \times \alpha_5^{(l)}) + (pzero_i \times \alpha_6^{(l)}) + (gender_i \times \alpha_7^{(l)}),$$

where $T_i^{(l)}$ is the estimated time to gametocyte clearance for the i^{th} patient during the l^{th} iteration with $\alpha_k^{(l)}$ $(\forall k = 1, \ldots, 7)$ as the fixed effect AFT parameters calculated during the l^{th} iteration . These generated parameters were considered as a pseudo posterior distribution for the resulting Weibull AFT parameters arising from the joint modeling procedure. At this stage these generated parameters could be used to make predictions on the duration of gametocytemia. The medians of each of these generated AFT parameter samples were considered as the point estimates for the parameters. The 2.5 and 97.5 percentiles, from the generated samples, were used as the 95% credibility interval for the AFT parameters. The same approach would be taken to derive the 95% credibility interval for the estimated durations, derived for a generic group of patients with specific covariate patterns, generated at each iteration.

The methodology outlined above would be able to provide duration estimates for a subset of the population that would have experienced gametocytemia. In order to extend these findings to provide population level estimates, based on patient covariate patterns, the methodology associated with Equation 37 (discussed in Chapter 3) was applied. This approach required the development of a baseline prevalence model for gametocytemia. A similar approach to the one used to derive the Weibull AFT model point estimate parameters was taken to derive the parameters for the baseline prevalence model. This approach involved fitting a baseline gametocyte prevalence model during each iteration. The baseline prevalence model for the l^{th} iteration is defined as

$$\log \pi_i^{(l)} = \beta_{B0}^{(l)} + (trt_i \times \beta_{B1}^{(l)}) + (ratio_i \times \beta_{B2}^{(l)}) + (anaemia_i \times \beta_{B3}^{(l)}) + (mut5_i \times \beta_{B4}^{(l)}) + (lage_i \times \beta_{B5}^{(l)}) + (pzero_i \times \beta_{B6}^{(l)}) + (gender_i \times \beta_{B7}^{(l)}),$$

where $\pi_i^{(l)}$ is the estimated prevalence to gametocytemia the i^{th} patient during the l^{th} iteration and $\beta_{Bk}^{(l)}$ $(\forall k = 1, \ldots, 7)$ are the fixed effect baseline prevalence parameters calculated during the l^{th} iteration.

The point estimates from the prevalence model and the survival model were then used to derive an adjusted duration for generic patients with specific covariate patterns, as per Equation 37.

4.10 Model fitting results

A range of joint models were fit to the data recorded in the study. These models were split into those that ignored zero-inflation and those that accounted for it. The differences, within each of these classes of models, was determined by the definition of the survival component of the models. The types of survival components used in this analysis were the exponential survival model, Weibull survival model, exponential cause-specific model and the Weibull cause-specific model. The definitions of each of the fitted models used in this analysis are shown in Table 22.

Table 22: Definitions of the fitted Joint models.

Class of Joint models	Label	Description of survival component
Nonlinear joint that assumes that the longitudinal response follows a normal distribution (NL)	NLJ	Exponential survival model
	NLJW	Weibull survival model
	NLJC	Exponential cause-specific model
	NLJWC	Weibull cause-specific model
Nonlinear joint models that assumes that the longitudinal response follows a zero-adjusted gamma distribution (ZAG)	ZAGJ	Exponential survival model
	ZAGJW	Weibull survival model
	ZAGJC	Exponential cause-specific model
	ZAGJWC	Weibull cause-specific model

The model fitting results of the survival components from the nonlinear joint models that assume that the longitudinal response follows a normal distribution (NL), are shown in Tables 23 and 24.

It can be seen from Table 23 that there is a strong association between the survival and longitudinal components, across all the fitted joint models. This is shown by the 95% CI of the association parameters not including 0. When considering the NLJW (normally distributed nonlinear joint model with a Weibull survival component) model, it is evident that the hypothesis that the Weibull shape parameter is equal to 1 (implying an exponential distribution for the survival component of the joint model) cannot be rejected as the 95% CI for this parameter includes 1. Another interesting result is that the treatment effect does not have an association with the hazard of time to early exit from the study. This result may be due to combining the treatment failure and LTFU outcomes, into a single outcome, as part of the formulation of the survival component of the model.

Table 23: Model estimates (95% CI) for the survival component of the fitted normally distributed nonlinear joint models.

Definition	Parameter	NLJ	NLJW
Intercept	β_0	-8.702 (-10.471 ; -6.731)	-9.123 (-13.491 ; -5.500)
trt	β_1	0.071 (-0.672 ; 0.786)	0.111 (-0.631 ; 0.957)
ratio	β_2	-1.182 (-1.951 ; -0.534)	-1.325 (-2.431 ; -0.507)
anaemia	β_3	0.962 (0.260 ; 1.660)	0.777 (0.136 ; 1.572)
*mut*5	β_4	1.383 (0.819 ; 1.999)	1.474 (0.843 ; 2.690)
lage	β_5	0.389 (0.022 ; 0.757)	0.287 (-0.053 ; 0.595)
pzero	β_6	0.057 (-0.012 ; 0.140)	0.064 (-0.024 ; 0.194)
gender	β_7	0.542 (0.012 ; 1.091)	0.459 (-0.075 ; 1.063)
Association parameters	ω_c	-0.863 (-1.204 ; -0.541)	-1.102 (-1.843 ; -0.235)
	ω_r	3.786 (2.682 ; 4.976)	4.948 (1.349 ; 8.701)
Precision parameters	τ_c	0.577 (0.382 ; 0.806)	0.827 (0.538 ; 1.295)
	τ_e	0.447 (0.415 ; 0.517)	0.493 (0.460 ; 0.536)
	τ_r	5.178 (3.855 ; 8.110)	8.721 (6.344 ; 15.038)
shape	ρ	1	1.144 (0.850 ; 1.702)

In Table 24 the treatment failure and LTFU outcomes, were analyzed separately as part of the cause-specific survival model. It can be seen that separating these outcomes resulted in a strong association between treatment and the relative hazards of both treatment failure and LTFU. The results from both the NLJC (normally distributed nonlinear joint model with an exponential cause-specific survival component) and NLJWC (normally distributed nonlinear joint model with a Weibull cause-specific survival component) models imply that receiving a combination of artesunate and sulfadoxine-pyrimethamine treatment significantly decreases the hazard of treatment failure whilst significantly increasing the hazard of LTFU. Barnes and White (2005) highlighted that $Plasmodium falciparum$ has developed clinically significant resistance to most antimalarials when they are given alone (as monotherapy) and not combined with an artemisinin derivative. As a result patients who receive a combination of artesunate and sulfadoxine-pyrimethamine treatment would be expected to clear the signs and symptoms associated with malaria infection at a faster rate than those who receive sulfadoxine-pyrimethamine treatment only. Based on the treatment effects highlighted above it appears as though patients who re-

cover quickly are more likely to choose to exit the study, through LTFU, instead of waiting for the completion of the study. The implication is that patients who are LTFU are generally healthier than those who have to be rescued from the study due to treatment failure.

Table 24 also reveals that the 95% credibility intervals for some of the association parameters include 0, however in these cases the credibility interval is predominately greater than or less than 0. This implies that there is evidence of an association between the survival and longitudinal components of the fitted cause-specific joint models. This in turn implies that there is a missing not at random (MNAR) dropout mechanism associated with the data used in this thesis.

The results of the longitudinal component of the fitted joint models are provided in Table 25, along with their associated DIC and MSPE statistics. It can be seen that model NLJWC (normally distributed nonlinear joint model with a Weibull cause-specific survival component) had the lowest DIC statistic, which would imply that it was the most appropriate model to apply. However Figure 33, which provides a plot of residuals against fitted values, indicates that there is a distinct systematic trend in the residuals of the fitted models. This trend is due to the excess number of zero values in the data that are not being accounted for. The implication of this result is that none of the normally distributed joint models are appropriate for further consideration.

Table 24: Model estimates (95% CI) for the survival component of the fitted cause-specific normally distributed nonlinear joint models.

Cause of exit	Definition	Parameter	NLJC	NLJWC
Treatment failure	*Intercept*	β_{F0}	-6.347 (-7.757 ; -4.934)	-5.876 (-7.830 ; -4.470)
	trt	β_{F1}	-1.317 (-2.340 ; -0.442)	-1.411 (-2.471 ; -0.456)
	ratio	β_{F2}	-1.037 (-1.713 ; -0.363)	-1.196 (-2.000 ; -0.503)
	anaemia	β_{F3}	0.141 (-0.351 ; 0.666)	-0.045 (-0.585 ; 0.475)
	*mut*5	β_{F4}	1.499 (1.034 ; 2.003)	1.506 (1.013 ; 2.014)
	lage	β_{F5}	-0.176 (-0.347 ; -0.005)	-0.232 (-0.447 ; -0.050)
	pzero	β_{F6}	0.075 (-0.005 ; 0.172)	0.066 (-0.008 ; 0.161)
	gender	β_{F7}	0.248 (-0.208 ; 0.730)	0.313 (-0.164 ; 0.852)
	Association parameters	ω_{cf}	-0.172 (-0.639 ; 0.262)	-0.360 (-0.806 ; 0.133)
		ω_{rf}	6.813 (0.000 ; 12.958)	6.064 (2.155 ; 12.173)
	shape	ρ_f		0.957 (0.779 ; 1.161)
LTFU	*Intercept*	β_{L0}	-9.386 (-12.277 ; -6.795)	-27.521 (-40.447 ; -17.289)
	trt	β_{L1}	1.268 (0.072 ; 2.724)	7.599 (1.189 ; 11.836)
	ratio	β_{L2}	-1.337 (-2.932 ; 0.085)	-6.644 (-17.775 ; -0.827)
	anaemia	β_{L3}	0.588 (-0.532 ; 1.651)	0.624 (-4.259 ; 8.935)
	*mut*5	β_{L4}	0.481 (-0.998 ; 1.630)	1.190 (-10.693 ; 6.616)
	lage	β_{L5}	0.071 (-0.289 ; 0.434)	-0.462 (-1.395 ; 1.094)
	pzero	β_{L6}	-0.025 (-0.147 ; 0.120)	-0.183 (-4.397 ; 0.602)
	gender	β_{L7}	0.319 (-0.702 ; 1.328)	-0.956 (-14.044 ; 3.740)
	Association parameters	ω_{cl}	-2.636 (-3.913 ; -1.463)	11.606 (7.769 ; 35.704)
		ω_{rl}	23.203 (12.371 ; 41.956)	-74.080 (-113.138 ; -53.153)
	shape	ρ_l		4.067 (2.670 ; 5.831)
Precision parameters		τ_c	1.176 (0.948 ; 1.459)	1.198 (0.755 ; 1.486)
		τ_r	109.738 (66.259 ; 208.054)	73.118 (20.107 ; 126.889)
		τ_e	0.508 (0.482 ; 0.535)	0.502 (0.476 ; 0.528)

Table 25: Model estimates (95% CI) for the longitudinal component of Nonlinear Joint Models, with no allowance for Zero Inflation.

Definition	Parameter	NLJ	NLJW	NLJC	NLJWC
intercept	β_{C0}	0.297 (-1.171 ; 1.995)	0.452 (-1.022 ; 2.259)	0.154 (-0.513 ; 0.832)	0.463 (-0.426 ; 1.378)
trt	β_{C1}	-0.388 (-1.150 ; 0.392)	-0.304 (-1.181 ; 0.600)	0.360 (-0.376 ; 2.544)	-0.055 (-0.655 ; 2.524)
ratio	β_{C2}	0.041 (-0.618 ; 0.652)	0.274 (-0.485 ; 0.870)	-0.034 (-0.338 ; 0.281)	0.083 (-0.242 ; 0.399)
anaemia	β_{C3}	-0.176 (-0.631 ; 0.247)	-0.032 (-0.412 ; 0.365)	0.152 (-0.098 ; 0.397)	0.369 (0.123 ; 0.686)
*mut*5	β_{C4}	0.062 (-0.380 ; 0.499)	-0.062 (-0.449 ; 0.338)	0.027 (-0.248 ; 0.292)	0.034 (-0.272 ; 0.308)
lage	β_{C5}	0.003 (-0.145 ; 0.158)	-0.043 (-0.201 ; 0.095)	-0.024 (-0.107 ; 0.060)	0.035 (-0.063 ; 0.126)
pzero	β_{C6}	0.055 (-0.037 ; 0.137)	0.061 (-0.035 ; 0.141)	0.051 (0.010 ; 0.091)	0.023 (-0.037 ; 0.075)
gender	β_{C7}	0.084 (-0.363 ; 0.477)	0.026 (-0.282 ; 0.354)	0.107 (-0.126 ; 0.325)	0.011 (-0.216 ; 0.270)
Intercept	β_{R0}	-0.866 (-1.118 ; -0.500)	-0.507 (-0.711 ; -0.259)	0.607 (0.509 ; 0.687)	0.549 (0.454 ; 0.719)
trt	β_{R1}	-0.269 (-0.412 ; -0.129)	-0.304 (-0.435 ; -0.177)	-0.290 (-0.434 ; -0.193)	-0.224 (-0.444 ; -0.139)
ratio	β_{R2}	-0.301 (-0.396 ; -0.209)	-0.262 (-0.387 ; -0.183)	-0.110 (-0.154 ; -0.069)	-0.143 (-0.257 ; -0.093)
anaemia	β_{R3}	0.245 (0.102 ; 0.393)	0.177 (0.113 ; 0.247)	0.029 (0.008 ; 0.059)	0.004 (-0.026 ; 0.038)
*mut*5	β_{R4}	0.089 (-0.039 ; 0.182)	0.066 (-0.005 ; 0.150)	0.025 (-0.009 ; 0.064)	0.023 (-0.023 ; 0.063)
lage	β_{R5}	0.226 (0.107 ; 0.307)	0.132 (0.095 ; 0.209)	0.007 (-0.004 ; 0.019)	0.000 (-0.020 ; 0.013)
pzero	β_{R6}	0.037 (0.023 ; 0.048)	0.037 (0.027 ; 0.056)	0.014 (0.010 ; 0.018)	0.016 (0.009 ; 0.026)
gender	β_{R7}	0.155 (0.049 ; 0.261)	0.108 (0.047 ; 0.178)	0.024 (0.003 ; 0.052)	0.049 (0.016 ; 0.087)
DIC		16450	16752	15892	15767
MSPE		4.475 (3.875 ; 4.817)	4.054 (3.739 ; 4.342)	3.938 (3.742 ; 4.144)	3.988 (3.788 ; 4.204)

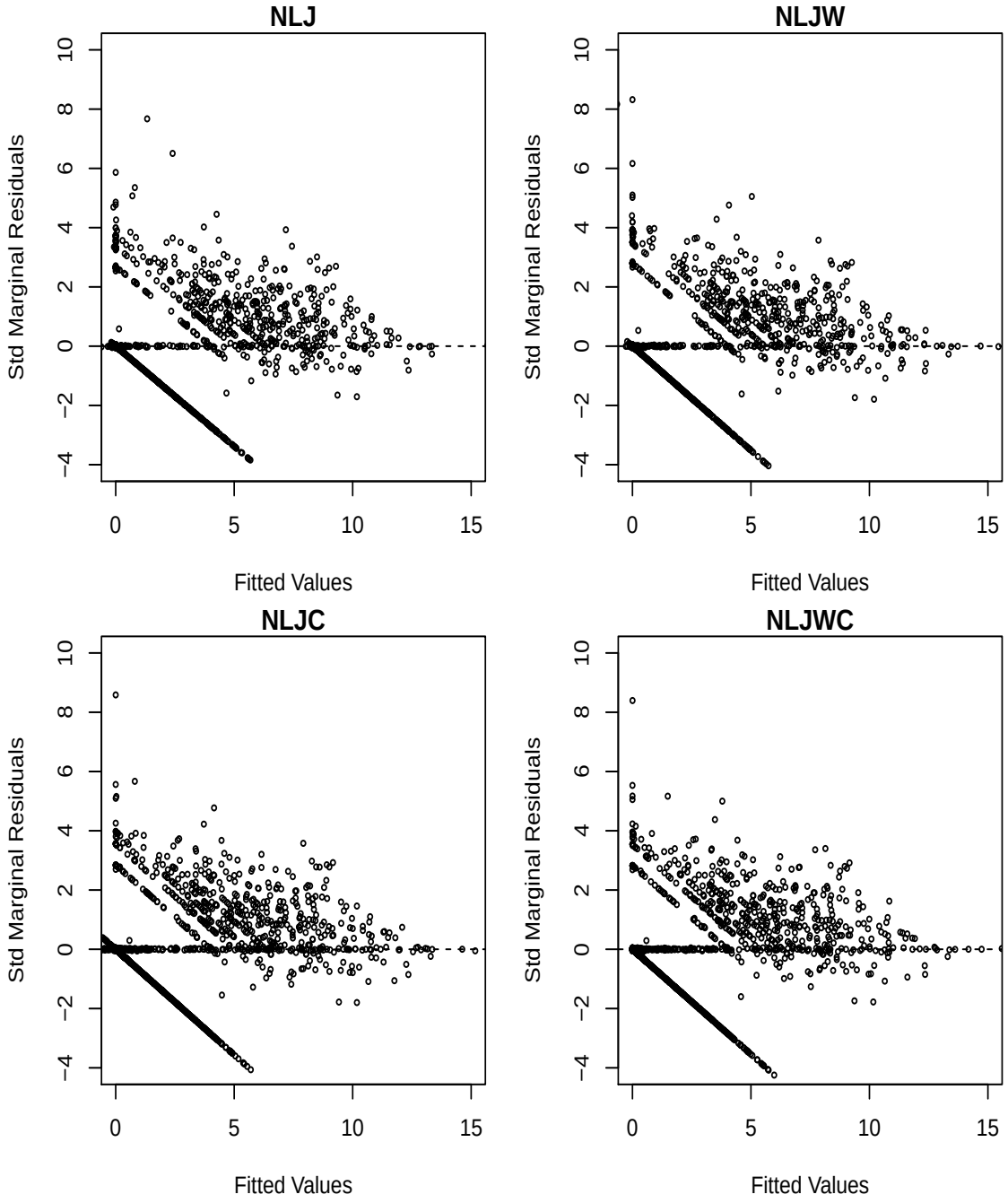

Figure 33: Standardized residual plots for the fitted normally distributed non-linear joint models

Zero-adjusted gamma joint models were fit to the data to account for the observed zero-inflation. Table 26 provides the survival component parameter estimates for the ZAGJ (zero-adjusted gamma nonlinear joint model with an exponential survival component) and ZAGJW (zero-adjusted gamma nonlinear joint model with a Weibull survival component) models.

It can be seen that the longitudinal component, of the ZAGJ model, has a marginal association with the survival component through the ω_{br} parameter. This is because the 95% CI for the ω_{br} parameter is be predominately negative. All other association parameters attributed to this model indicate that there is no association between the survival and longitudinal components, as shown by 0 being included the 95% CI for these parameters.

When considering the ZAGJW model, it is evident that there is a strong association between the survival and longitudinal components of the joint model. This is because the ω_{br}, ω_c and ω_r parameters all have 95% CIs that do not include 0.

The strength of the association between the survival and longitudinal components in the fitted ZAGJ and ZAGJW models imply that there is a missing not at random (MNAR) dropout mechanism associated with the data used in this thesis. The use of joint modeling techniques adjusts for this type of dropout, leading to less biased parameter estimates.

The results from Table 26 indicate that combining the treatment failures with the loss-to-follow-up outcomes, leads to treatment not having an association with the hazard of early exit from the study. In addition it can be seen that the hypothesis of the shape parameter being equal to 1, implying that an exponential distribution should be assumed for the survival model, can be rejected as the 95% CI for the parameter does not contain 1.

Table 26: Model estimates (95% CI) for the survival component of the fitted zero-adjusted gamma (ZAG) Joint Models.

Definition	Parameter	ZAGJ	ZAGJW
intercept	β_0	-6.535 (-7.920 ; -5.148)	-18.637 (-25.284 ; -12.658)
trt	β_1	-0.008 (-0.639 ; 0.673)	0.063 (-1.819 ; 1.957)
ratio	β_2	-1.023 (-1.731 ; -0.377)	-2.887 (-5.408 ; -1.029)
anaemia	β_3	0.346 (-0.230 ; 0.926)	1.079 (-0.247 ; 2.742)
mut5	β_4	1.430 (0.867 ; 2.032)	3.950 (2.053 ; 6.511)
lage	β_5	-0.132 (-0.328 ; 0.070)	-0.302 (-1.060 ; 0.145)
pzero	β_6	0.055 (-0.024 ; 0.128)	0.113 (-0.090 ; 0.341)
gender	β_7	0.302 (-0.179 ; 0.820)	0.696 (-0.992 ; 2.160)
Association parameters	ω_{bc}	0.169 (-0.132 ; 0.356)	-0.058 (-0.583 ; 0.408)
	ω_{br}	-1.298 (-26.384 ; 0.119)	-68.810 (-98.430 ; -39.042)
	ω_c	7.088 (-29.838 ; 24.780)	63.351 (25.770 ; 135.539)
	ω_r	-3.643 (-63.900 ; 52.632)	-93.745 (-236.250 ; -50.471)
Precision parameters	τ_{bc}	0.120 (0.073 ; 0.310)	0.235 (0.155 ; 0.347)
	τ_{br}	12.98 (9.01 ; 559.37)	460.13 (335.22 ; 626.49)
	τ_c	283.82 (40.01 ; 2209.10)	232.03 (112.70 ; 650.04)
	τ_r	816.63 (201.88 ; 2861.71)	848.26 (408.50 ; 1506.54)
	τ_e	12.624 (11.130 ; 14.183)	22.524 (19.086 ; 26.517)
Shape	ρ		3.104 (2.060 ; 4.443)

The results of applying a cause-specific survival component, to the fitted joint models, are shown in Table 27. This table compares the results of the ZAGJC (zero-adjusted gamma nonlinear joint model with an exponential cause-specific survival component) and ZAGJWC (zero-adjusted gamma nonlinear joint model with a Weibull cause-specific survival component) models.

Table 27 reveals that treatment is strongly associated with both the hazard of treatment failure and the hazard of loss-to-follow-up, with the 95% CI for the treatment effect not including 0. The treatment effect indicates that receiving a combination treatment of artesunate and sulfadoxine-pyrimethamine (ACT) considerably decreases the hazard of treatment failure (78% reduction for the ZAGJC model and 94% reduction for the ZAGJWC model) whilst increasing the hazard of loss-to-follow-up (9 fold increase for the ZAGJC model and 289 fold increase for the ZAGJWC model). It can also be seen that resistance to treatment, in the form of the presence of quintuple mutations ($mut5$), has a strong association with the hazard of treatment failure. However, the $mut5$ effect did not have an association with the hazard of LTFU. An interesting finding is that apart from patient gender, all the covariates used in this analysis had a strong association (or marginal association in the case of the patient age when considering the ZAGJWC model and baseline asexual parasite density when considering the ZAGJC model) with the hazard of treatment failure. On the other hand treatment (trt) was the only covariate that had a strong association with the hazard of loss-to-follow-up. These findings apply to both the ZAGJC and ZAGJWC models.

Table 27 provides an assessment of the association between the survival and longitudinal components of the fitted ZAGJC and ZAGJWC models. When considering the ZAGJC (zero-adjusted gamma nonlinear joint model with an exponential cause-specific survival component) model, it can be seen that there is a strong association between the survival component and the longitudinal component of the joint model through the ω_{brf}, ω_{rf}, ω_{brl}, ω_{cl} and ω_{rl} (marginal association) parameters. The strong association between the survival component and the longitudinal component of the ZAGJWC (zero-adjusted gamma nonlinear joint model with a Weibull cause-specific survival component) model arose through the ω_{brf}, ω_{brl} (marginal association), ω_{cl} (marginal association) and ω_{rl} parameters. Once again it must be highlighted that this association indicates that dropout was MNAR. Since this association arises when considering both the treatment failure and LTFU causes of exit, it can be concluded that both causes of exit are forms of informative censoring.

It can be seen, when considering the results of the ZAGJWC model fit shown

in Table 27, that the 95% CI for the shape parameters associated with both causes of early exit do not include 1. As a result it can be concluded that a Weibull survival component is more appropriate as compared to an exponential survival component. These findings support the use of the ZAGJWC model over the ZAGJC model. The shape parameters for both causes of exit can be seen to be greater than 1, thus implying that the hazards of treatment failure and loss-to-follow-up increase over time.

Table 27: Model estimates (95% CI) for the survival component of cause-specific ZAG Joint Models.

Cause of exit	Definition	Parameter	ZAGJC	ZAGJWC
Treatment failure	*intercept*	β_{F0}	-7.082 (-9.050 ; -5.078)	-18.277 (-25.277 ; -8.709)
	trt	β_{F1}	-1.504 (-2.858 ; -0.491)	-2.801 (-6.169 ; -0.855)
	ratio	β_{F2}	-1.150 (-1.960 ; -0.395)	-2.918 (-5.594 ; -1.005)
	anaemia	β_{F3}	0.072 (-0.563 ; 0.674)	0.298 (-1.140 ; 1.822)
	*mut*5	β_{F4}	1.742 (1.137 ; 2.459)	4.287 (1.884 ; 7.215)
	lage	β_{F5}	-0.220 (-0.429 ; -0.009)	-0.423 (-1.060 ; 0.022)
	pzero	β_{F6}	0.107 (-0.008 ; 0.240)	0.196 (0.021 ; 0.413)
	gender	β_{F7}	0.315 (-0.227 ; 0.891)	0.860 (-0.349 ; 2.919)
	Association parameters	ω_{bcf}	0.049 (-0.127 ; 0.221)	0.176 (-0.202 ; 0.633)
		ω_{brf}	-16.320 (-27.158 ; -7.039)	-42.735 (-89.230 ; -20.081)
		ω_{cf}	-0.406 (-5.509 ; 3.126)	-1.984 (-21.051 ; 2.263)
		ω_{rf}	60.247 (14.866 ; 85.394)	43.650 (-48.313 ; 127.865)
	Shape	ρ_f		2.565 (1.453 ; 3.674)
LTFU	*intercept*	β_{L0}	-11.226 (-16.336 ; -7.707)	-22.736 (-31.378 ; -14.754)
	trt	β_{L1}	2.235 (0.724 ; 4.255)	5.666 (1.986 ; 9.199)
	ratio	β_{L2}	-1.466 (-3.188 ; 0.705)	-3.778 (-6.789 ; 0.937)
	anaemia	β_{L3}	0.924 (-0.269 ; 2.525)	0.837 (-1.844 ; 5.349)
	*mut*5	β_{L4}	0.344 (-1.435 ; 1.814)	0.606 (-3.693 ; 4.302)
	lage	β_{L5}	0.082 (-0.280 ; 0.592)	-0.286 (-1.440 ; 0.766)
	pzero	β_{L6}	-0.020 (-0.202 ; 0.213)	-0.194 (-0.534 ; 0.109)
	gender	β_{L7}	0.270 (-1.087 ; 1.552)	0.205 (-2.314 ; 3.434)
	Association parameters	ω_{bcl}	-0.261 (-0.757 ; 0.246)	-0.670 (-2.246 ; 0.243)
		ω_{brl}	-32.414 (-59.791 ; -6.659)	-70.772 (-140.790 ; 8.471)
		ω_{cl}	10.491 (2.130 ; 27.380)	5.177 (-6.307 ; 80.827)
		ω_{rl}	-24.272 (-101.908 ; 11.539)	-36.369 (-80.049 ; -3.846)
	Shape	ρ_l		2.227 (1.670 ; 2.812)
Precision parameters		τ_{bc}	0.228 (0.151 ; 0.347)	0.221 (0.148 ; 0.332)
		τ_{br}	468.0 (340.0 ; 646.5)	435.2 (318.2 ; 626.1)
		τ_c	12.341 (0.288 ; 96.764)	0.314 (0.106 ; 101.943)
		τ_r	4431.7 (575.6 ; 25361.1)	514.5 (138.8 ; 9178.0)
		τ_e	12.470 (10.991 ; 14.077)	12.586 (11.137 ; 14.193)

The model fitting results of the longitudinal components of the fitted models, along with their respective DIC statistics are provided in Table 28. It can be seen that the ZAGJ and ZAGJC models had the lowest DIC statistics. The MSPE was also used as an additional goodness of fit test, to determine that zero-adjusted gamma joint model was more appropriate. The median MSPE statistics derived for all the fitted models are fairly similar. However, the ZAGJW model can be seen to have the largest credibility range for the MSPE statistic that would imply that this model is not appropriate.

Figure 34 provides a comparison of the plots of residuals against fitted values, for the zero-adjusted gamma joint models fitted in this analysis. This figure shows that, across all models, there were a small number of fairly large positive residuals attributed to small fitted values. It is also evident from this figure that the standardized residuals, across all the fitted models predominately ranged between -2 and 2. In addition the systematic trend associated with the large number of zero values in the data, which was previously an issue under the normally distributed joint models, is no longer present. As a result it can be concluded that all the models performed appropriately.

It was decided that the model that would be suitable for further investigation was the ZAGJWC (zero-adjusted gamma nonlinear joint model with a Weibull cause-specific survival component) model. The reasons for this were that firstly the Weibull shape parameters were found to significantly different from 1, as the 95% CI for these parameters (when considering the treatment failure and loss-to-follow-up causes of exit) did not include 1. In addition the effect of the treatment allocated to patients under this model provided clinically meaningful results that support the validity of this model.

Table 28: Model estimates (95% CI) for the longitudinal component of the fitted zero-adjusted gamma (ZAG) Joint Models.

Component	Definition	Parameter	ZAGJ	ZAGJW	ZAGJC	ZAGJWC
Prevalence	*Intercept*	β_{BA0}	-4.587 (-8.953 ; -0.342)	-6.161 (-8.913 ; -3.515)	-6.010 (-8.905 ; -3.059)	-4.844 (-9.170 ; -1.923)
	trt	β_{BA1}	0.347 (-3.398 ; 3.270)	1.950 (-2.063 ; 4.134)	2.066 (-1.218 ; 4.199)	1.658 (-3.097 ; 4.571)
	ratio	β_{BA2}	2.155 (-0.237 ; 4.469)	1.793 (-0.048 ; 3.742)	1.718 (-0.584 ; 4.005)	1.610 (-0.830 ; 3.489)
	anaemia	β_{BA3}	-1.246 (-3.656 ; 1.292)	-0.435 (-2.217 ; 1.829)	-0.688 (-2.432 ; 1.515)	-0.423 (-2.028 ; 1.415)
	mut5	β_{BA4}	0.251 (-2.284 ; 2.318)	0.029 (-2.066 ; 2.152)	-0.002 (-2.016 ; 1.910)	0.275 (-1.691 ; 2.139)
	lage	β_{BA5}	-1.090 (-1.875 ; -0.155)	-0.856 (-1.435 ; -0.316)	-0.948 (-1.517 ; -0.360)	-0.998 (-1.534 ; -0.417)
	pzero	β_{BA6}	-0.086 (-0.375 ; 0.180)	-0.053 (-0.259 ; 0.172)	-0.029 (-0.240 ; 0.206)	-0.123 (-0.357 ; 0.166)
	gender	β_{BA7}	-3.027 (-5.433 ; -0.564)	-2.152 (-4.706 ; -0.504)	-2.711 (-5.319 ; -0.595)	-2.446 (-4.534 ; -0.638)
	Intercept	β_{BC0}	-0.045 (-1.491 ; 2.466)	0.581 (-1.311 ; 1.770)	0.335 (-1.220 ; 1.221)	-0.074 (-1.866 ; 2.275)
	trt	β_{BC1}	1.183 (-1.687 ; 4.361)	-1.058 (-2.305 ; 0.929)	-1.154 (-2.361 ; 0.448)	-0.964 (-2.780 ; 1.406)
	ratio	β_{BC2}	-1.026 (-2.666 ; 0.495)	-1.077 (-2.064 ; -0.195)	-1.020 (-2.207 ; 0.162)	-0.944 (-2.055 ; 0.375)
	anaemia	β_{BC3}	1.263 (0.067 ; 2.653)	0.826 (-0.275 ; 1.745)	0.908 (-0.041 ; 1.702)	0.836 (0.032 ; 1.549)
	mut5	β_{BC4}	0.648 (-0.588 ; 2.133)	0.441 (-0.580 ; 1.462)	0.477 (-0.493 ; 1.483)	0.356 (-0.606 ; 1.302)
	lage	β_{BC5}	0.267 (-0.083 ; 0.877)	0.226 (-0.036 ; 0.565)	0.277 (0.020 ; 0.513)	0.316 (0.050 ; 0.511)
	pzero	β_{BC6}	0.037 (-0.038 ; 0.174)	0.042 (-0.095 ; 0.141)	0.022 (-0.053 ; 0.117)	0.054 (-0.083 ; 0.143)
	gender	β_{BC7}	1.013 (0.004 ; 2.059)	0.935 (0.255 ; 1.966)	1.186 (0.299 ; 2.339)	1.067 (0.357 ; 1.935)
	Intercept	β_{BR0}	0.692 (0.630 ; 0.828)	0.830 (0.800 ; 0.861)	0.838 (0.806 ; 0.890)	0.821 (0.788 ; 0.847)
	trt	β_{BR1}	-0.190 (-0.416 ; -0.031)	-0.050 (-0.077 ; -0.026)	-0.050 (-0.075 ; -0.026)	-0.048 (-0.074 ; -0.022)
	ratio	β_{BR2}	-0.083 (-0.195 ; -0.008)	-0.019 (-0.036 ; -0.003)	-0.019 (-0.036 ; -0.000)	-0.020 (-0.037 ; -0.003)
	anaemia	β_{BR3}	0.039 (-0.021 ; 0.081)	-0.013 (-0.026 ; 0.000)	-0.012 (-0.024 ; 0.001)	-0.012 (-0.024 ; 0.000)
	mut5	β_{BR4}	-0.005 (-0.074 ; 0.066)	-0.001 (-0.015 ; 0.014)	-0.003 (-0.019 ; 0.011)	-0.002 (-0.017 ; 0.013)
	lage	β_{BR5}	0.003 (-0.018 ; 0.016)	0.002 (-0.003 ; 0.007)	0.001 (-0.003 ; 0.006)	0.002 (-0.003 ; 0.006)
	pzero	β_{BR6}	0.004 (0.000 ; 0.007)	0.005 (0.002 ; 0.006)	0.004 (0.001 ; 0.006)	0.005 (0.004 ; 0.007)
	gender	β_{BR7}	0.042 (0.000 ; 0.089)	0.008 (-0.003 ; 0.021)	0.008 (-0.004 ; 0.021)	0.009 (-0.003 ; 0.020)
Continuous	*Intercept*	β_{A0}	1.231 (1.061 ; 1.394)	1.118 (0.888 ; 1.393)	1.239 (1.076 ; 1.410)	1.233 (1.076 ; 1.398)
	trt	β_{A1}	-0.105 (-0.216 ; 0.020)	-0.186 (-0.357 ; 0.004)	-0.108 (-0.218 ; 0.012)	-0.104 (-0.219 ; 0.019)
	ratio	β_{A2}	-0.130 (-0.198 ; -0.061)	-0.134 (-0.223 ; -0.038)	-0.128 (-0.197 ; -0.060)	-0.130 (-0.200 ; -0.059)
	anaemia	β_{A3}	0.012 (-0.040 ; 0.064)	-0.050 (-0.135 ; 0.039)	0.012 (-0.040 ; 0.065)	0.013 (-0.039 ; 0.066)
	mut5	β_{A4}	0.011 (-0.052 ; 0.074)	0.317 (0.180 ; 0.420)	0.016 (-0.048 ; 0.081)	0.012 (-0.049 ; 0.076)
	lage	β_{A5}	0.041 (0.022 ; 0.060)	0.011 (-0.017 ; 0.042)	0.040 (0.022 ; 0.059)	0.042 (0.022 ; 0.061)
	pzero	β_{A6}	0.042 (0.032 ; 0.053)	0.039 (0.021 ; 0.053)	0.042 (0.031 ; 0.052)	0.042 (0.031 ; 0.052)
	gender	β_{A7}	0.007 (-0.042 ; 0.057)	0.003 (-0.073 ; 0.074)	0.002 (-0.048 ; 0.054)	0.006 (-0.045 ; 0.055)
	Intercept	β_{C0}	-8.527 (-33.933 ; 36.426)	0.129 (-0.047 ; 0.284)	0.837 (-0.204 ; 9.432)	-1.356 (-10.889 ; 47.035)
	trt	β_{C1}	-1.756 (-121.744 ; 147.249)	-0.008 (-0.109 ; 0.085)	-0.121 (-60.383 ; 39.356)	7.066 (-133.329 ; 54.903)
	ratio	β_{C2}	-37.107 (-99.475 ; 48.753)	-0.001 (-0.064 ; 0.093)	4.899 (-1.818 ; 41.183)	-26.943 (-117.538 ; 6.573)
	anaemia	β_{C3}	10.895 (-32.655 ; 54.580)	0.025 (-0.037 ; 0.075)	-7.138 (-37.897 ; 23.174)	19.448 (-10.676 ; 118.204)
	mut5	β_{C4}	7.834 (-78.139 ; 66.588)	68.115 (-131.875 ; 137.078)	-16.657 (-64.132 ; 2.267)	33.681 (-18.710 ; 183.493)
	lage	β_{C5}	4.046 (-0.872 ; 9.309)	-0.001 (-0.018 ; 0.018)	0.111 (-6.137 ; 0.236)	1.407 (-1.502 ; 6.800)
	pzero	β_{C6}	-0.557 (-1.393 ; 1.993)	-0.002 (-0.014 ; 0.008)	0.005 (-0.069 ; 0.290)	-0.998 (-3.811 ; 0.297)
	gender	β_{C7}	33.634 (-45.637 ; 147.892)	0.015 (-0.034 ; 0.066)	10.654 (0.377 ; 89.949)	29.689 (-2.630 ; 68.580)
	Intercept	β_{R0}	0.072 (-0.353 ; 0.322)	0.808 (0.674 ; 0.888)	-0.074 (-0.443 ; 0.433)	0.072 (-0.309 ; 0.316)
	trt	β_{R1}	-0.007 (-0.139 ; 0.111)	0.042 (-0.012 ; 0.091)	0.014 (-0.163 ; 0.225)	-0.023 (-0.177 ; 0.224)
	ratio	β_{R2}	0.061 (-0.152 ; 0.150)	0.017 (-0.069 ; 0.054)	-0.104 (-0.250 ; 0.311)	0.052 (-0.226 ; 0.158)
	anaemia	β_{R3}	-0.008 (-0.098 ; 0.076)	-0.000 (-0.026 ; 0.030)	0.036 (-0.118 ; 0.147)	-0.004 (-0.093 ; 0.091)
	mut5	β_{R4}	-0.050 (-0.151 ; 0.118)	-0.960 (-1.006 ; -0.709)	0.065 (-0.244 ; 0.260)	-0.034 (-0.143 ; 0.190)
	lage	β_{R5}	-0.007 (-0.038 ; 0.031)	0.010 (-0.000 ; 0.021)	0.005 (-0.047 ; 0.042)	-0.008 (-0.038 ; 0.037)
	pzero	β_{R6}	-0.006 (-0.021 ; 0.022)	0.002 (-0.003 ; 0.011)	0.002 (-0.030 ; 0.023)	-0.004 (-0.018 ; 0.020)
	gender	β_{R7}	0.007 (-0.092 ; 0.089)	-0.002 (-0.025 ; 0.022)	0.058 (-0.107 ; 0.207)	0.015 (-0.077 ; 0.108)
DIC			10084	12030	10739	11767
MSPE			3.254 (2.802 ; 3.914)	3.009 (2.612 ; 5.589)	3.545 (3.128 ; 4.074)	3.503 (3.092 ; 3.997)

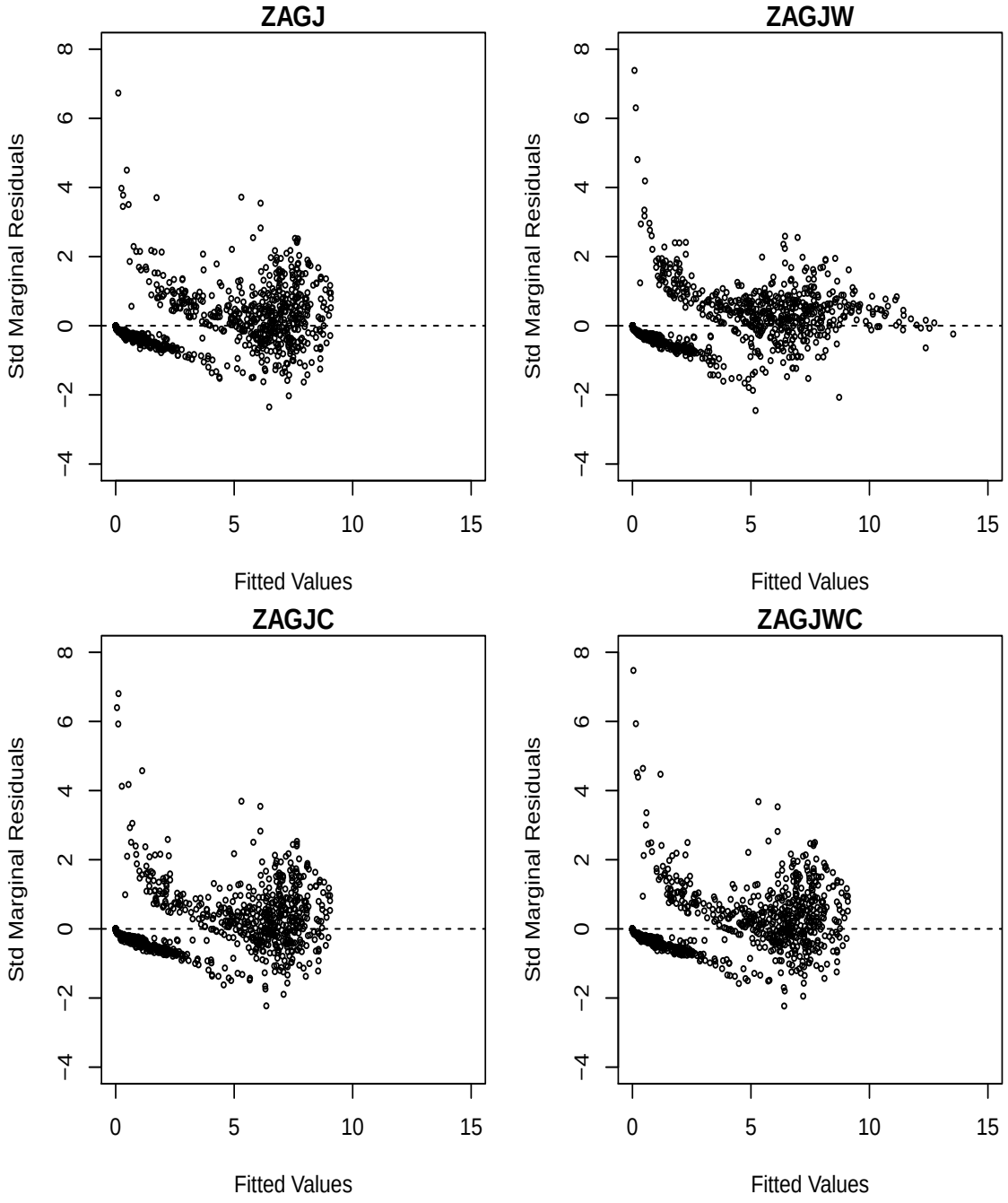

Figure 34: Standardized residual plots for the fitted zero-adjusted gamma (ZAG) joint models

In order to graphically illustrate the predicted patient gametocyte profiles, a generic patient dataset was created. This dataset allowed for the comparison of specific covariates patterns. The covariate patterns were based on the following covariate levels

- SP and ACT treatments
- "High" first 24 hour parasite reduction rate of 100% relative to a "Low" first 24 hour parasite reduction rate of -31.4%
- Presence and absence of moderate anaemia
- Presence and absence of quintuple mutations
- "Old" patient with an age of 55.0 ($lage = 5.781$) and "Young" patient with an age of 2 ($lage = 1$)
- "High" $\log_{10}$ baseline parasite density of 17.754 and "Low" baseline parasite density of 5
- Male and female patients

Based on the above characteristics a dataset of 128 distinct covariate patterns was created. Subsequently the median fixed effect parameters from the fitted ZAGJWC model (shown in Table 28) were used to create estimated gametocyte profiles for these patients. This dataset allows for categories within covariates to be compared in isolation, that it categories with-in a covariate can be allowed to vary whilst all other covariates are held constant.

Figure 35 was created based on this approach and it allowed for the comparison of the predicted profiles for a patient receiving a combination treatment of artesunate and sulfadoxine-pyrimethamine and a patient receiving sulfadoxine-pyrimethamine, whilst holding all other covariates at a constant value. It can be seen from this plot that the predicted gametocyte profiles for patients receiving SP treatment only are significantly higher than those for patients who received a combination treatment of artesunate and sulfadoxine-pyrimethamine. It is of interest to stratify the predicted profiles, of the other covariates considered, with treatment to assess the level of impact that treatment has on the gametocyte profiles relative to the other covariates. This is illustrated in Figure 36. It can be seen that the application of artesunate sulfadoxine-pyrimethamine treatment generally suppresses the magnitude of the predicted gametocyte density, regardless of the level of the associated covariate. The implication of these findings is that regardless of the levels of the other covariates considered in this study, the

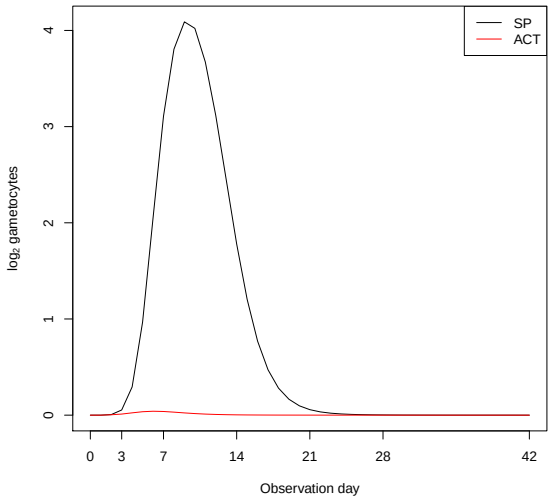

Figure 35: Predicted gametocyte patient profiles by treatment, for the ZAGJWC model

application of artesunate sulfadoxine-pyrimethamine treatment will result in a suppressed gametocyte density.

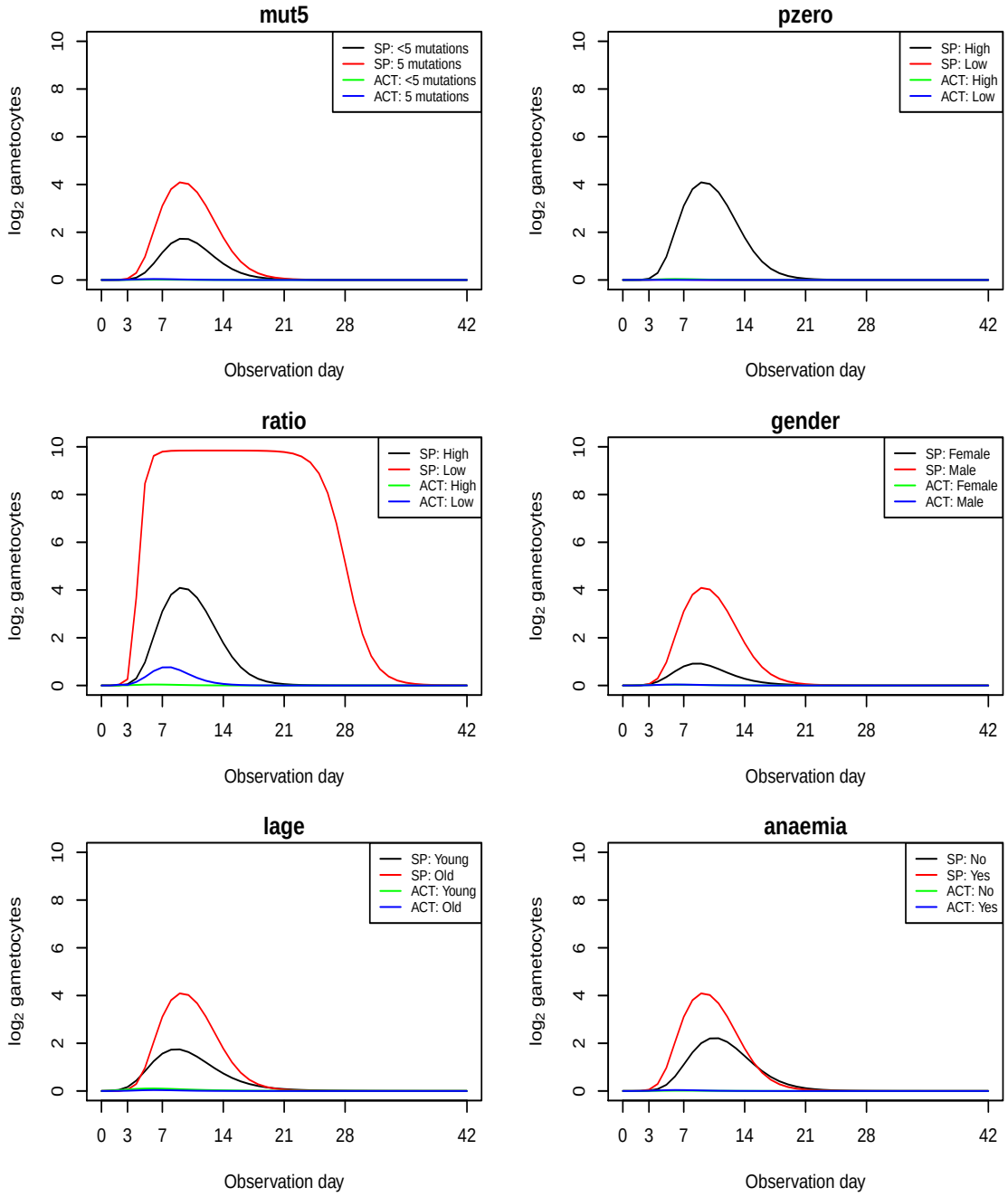

Figure 36: Predicted patient profiles by categorical covariates, stratified by treatment, for the ZAGJWC model

In order to assess the impact of informative censoring on parameter estimates, a zero-adjusted gamma mixed effect model (ZAG) was fit to the longitudinal data. The results of this model fit are shown in Table 29. This table also provides a comparison with the parameters from the longitudinal component of the ZAGJWC (zero-adjusted gamma nonlinear joint model with a Weibull cause-specific survival component) model. It can be seen that there are noticeable differences in the magnitude of the fixed effects across the ZAG and ZAGJWC models. In some cases parameters that were found to have no association with the longitudinal gametocyte profile under the ZAG model, were found to have a strong association with the gametocyte profile once a ZAGJWC model was applied. This finding applied to the β_{BC3} (moderate anaemia) , β_{BC5} (age) and β_{BC7} (gender) parameters.

A comparison of the predicted profiles arising from the ZAG and ZAGJWC models, using the generic dataset previously created are shown in Figure 37. This figure compares the profiles from the two models using set values for the first 24 hour parasite reduction ratio grouped by the treatment administered. It can be seen that the ZAG model generally predicted higher gametocyte densities as compared to the ZAGJWC model.

The fitted values from the ZAG and ZAGJWC model fits were superimposed onto the observed gametocyte profiles for select patients from the study (Figure 38). The patients in the top row of the plot were observed at all observation points and thus had complete profiles. It is evident that the ZAGJWC predicted profiles outperformed the ZAG across this group of patients. The remaining rows provide examples of incomplete profiles (second row consisted of patients who experienced treatment failure while the third row consisted of patients lost-to-follow-up). The incomplete gametocyte profiles are superimposed with complete predicted profiles from the ZAG and the ZAGJWC models. The profiles from the two models can be seen to be fairly similar for this subset of patients.

Table 29: Comparison of Fixed effects Model estimates (95% CI) for the longitudinal component of a fitted standard ZAG and a ZAGJWC Joint Model.

Component	Definition	Parameter	ZAG	ZAGJWC
Prevalence	*Intercept*	β_{BC0}	0.682 (-1.130 ; 2.302)	-0.074 (-1.866 ; 2.275)
	trt	β_{BC1}	-0.367 (-2.342 ; 1.260)	-0.964 (-2.780 ; 1.406)
	ratio	β_{BC2}	-0.742 (-2.370 ; 0.728)	-0.944 (-2.055 ; 0.375)
	anaemia	β_{BC3}	0.659 (-0.234 ; 2.186)	0.836 (0.032 ; 1.549)
	mut5	β_{BC4}	0.302 (-0.702 ; 1.557)	0.356 (-0.606 ; 1.302)
	lage	β_{BC5}	0.153 (-0.118 ; 0.643)	0.316 (0.050 ; 0.511)
	pzero	β_{BC6}	0.085 (-0.097 ; 0.162)	0.054 (-0.083 ; 0.143)
	gender	β_{BC7}	1.166 (-0.028 ; 2.539)	1.067 (0.357 ; 1.935)
	Intercept	β_{BR0}	0.810 (0.681 ; 0.858)	0.821 (0.788 ; 0.847)
	trt	β_{BR1}	-0.055 (-0.197 ; -0.027)	-0.048 (-0.074 ; -0.022)
	ratio	β_{BR2}	-0.029 (-0.138 ; -0.005)	-0.020 (-0.037 ; -0.003)
	anaemia	β_{BR3}	-0.009 (-0.028 ; 0.015)	-0.012 (-0.024 ; 0.000)
	mut5	β_{BR4}	0.000 (-0.018 ; 0.066)	-0.002 (-0.017 ; 0.013)
	lage	β_{BR5}	0.002 (-0.007 ; 0.007)	0.002 (-0.003 ; 0.006)
	pzero	β_{BR6}	0.005 (0.003 ; 0.008)	0.005 (0.004 ; 0.007)
	gender	β_{BR7}	0.014 (-0.001 ; 0.085)	0.009 (-0.003 ; 0.020)
Continuous	*Intercept*	β_{C0}	0.127 (-0.140 ; 7.008)	-1.356 (-10.889 ; 47.035)
	trt	β_{C1}	18.203 (-68.669 ; 47.597)	7.066 (-133.329 ; 54.903)
	ratio	β_{C2}	-0.081 (-57.425 ; 11.600)	-26.943 (-117.54 ; 6.573)
	anaemia	β_{C3}	-0.031 (-21.679 ; 42.475)	19.448 (-10.676 ; 118.204)
	mut5	β_{C4}	-0.019 (-31.576 ; 90.112)	33.681 (-18.710 ; 183.493)
	lage	β_{C5}	0.002 (-0.438 ; 2.272)	1.407 (-1.502 ; 6.800)
	pzero	β_{C6}	0.001 (-0.872 ; 0.020)	-0.998 (-3.811 ; 0.297)
	gender	β_{C7}	0.039 (-0.100 ; 42.445)	29.689 (-2.630 ; 68.580)
	Intercept	β_{R0}	0.738 (-0.080 ; 0.886)	0.072 (-0.309 ; 0.316)
	trt	β_{R1}	-0.721 (-0.968 ; 0.089)	-0.023 (-0.177 ; 0.224)
	ratio	β_{R2}	-0.030 (-0.108 ; 0.147)	0.052 (-0.226 ; 0.158)
	anaemia	β_{R3}	-0.018 (-0.748 ; 0.047)	-0.004 (-0.093 ; 0.091)
	mut5	β_{R4}	-0.069 (-0.715 ; 0.077)	-0.034 (-0.143 ; 0.190)
	lage	β_{R5}	0.005 (-0.036 ; 0.020)	-0.008 (-0.038 ; 0.037)
	pzero	β_{R6}	0.002 (-0.019 ; 0.010)	-0.004 (-0.018 ; 0.020)
	gender	β_{R7}	0.001 (-0.078 ; 0.051)	0.015 (-0.077 ; 0.108)
Precision parameters		τ_{bc}	0.206 (0.080 ; 0.349)	0.221 (0.148 ; 0.332)
		τ_{br}	397.36 (17.40 ; 600.34)	435.22 (318.23 ; 626.17)
		τ_c	156.95 (0.03 ; 837.21)	0.31 (0.11 ; 101.94)
		τ_r	693.98 (200.06 ; 2496.40)	514.48 (138.79 ; 9177.98)
		τ_e	16.817 (11.445 ; 22.616)	12.586 (11.137 ; 14.193)

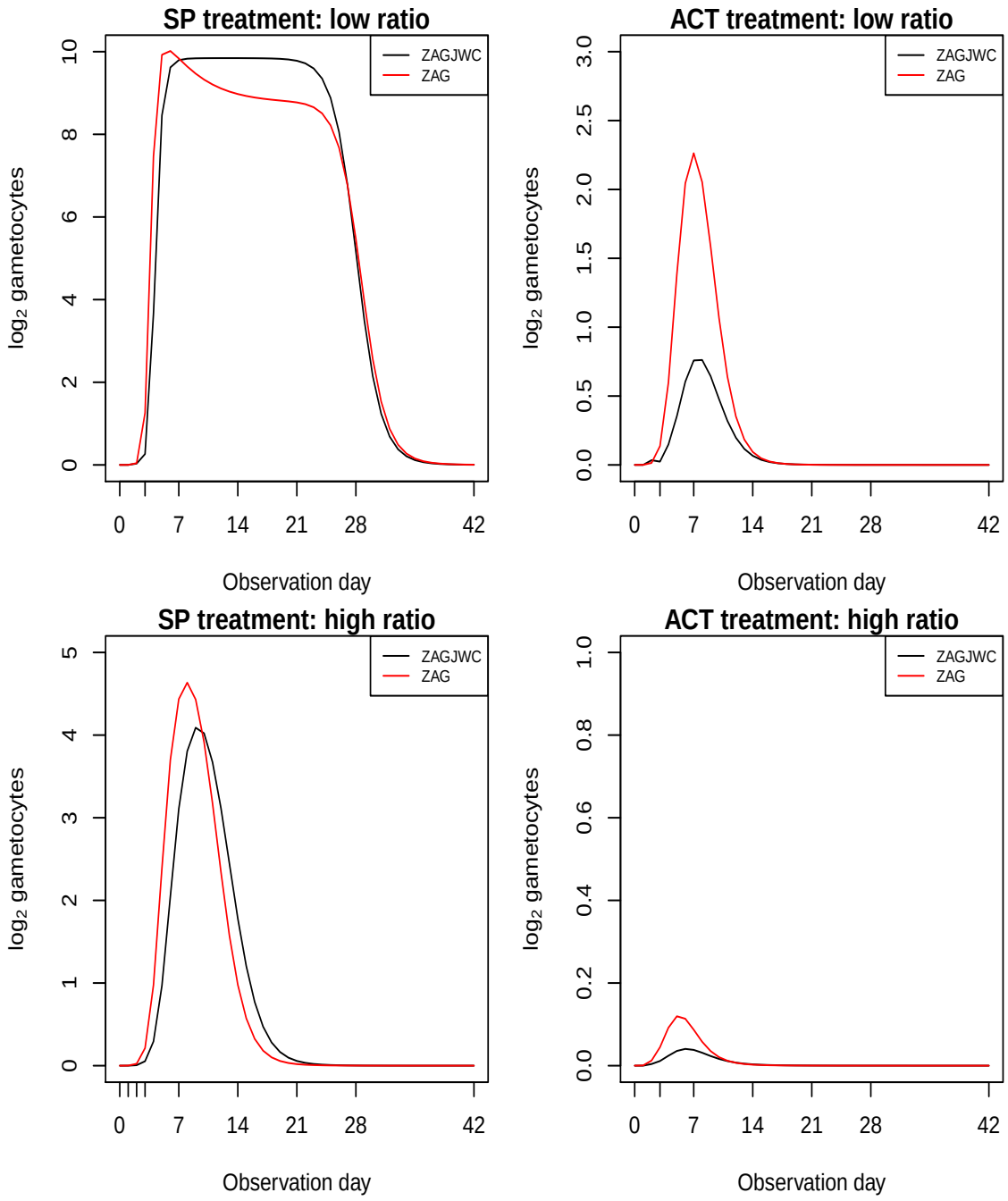

Figure 37: Predicted patient profiles by first 24 hour parasite reduction ratios (low vs high), stratified by treatment, for the ZAG and ZAGJWC models

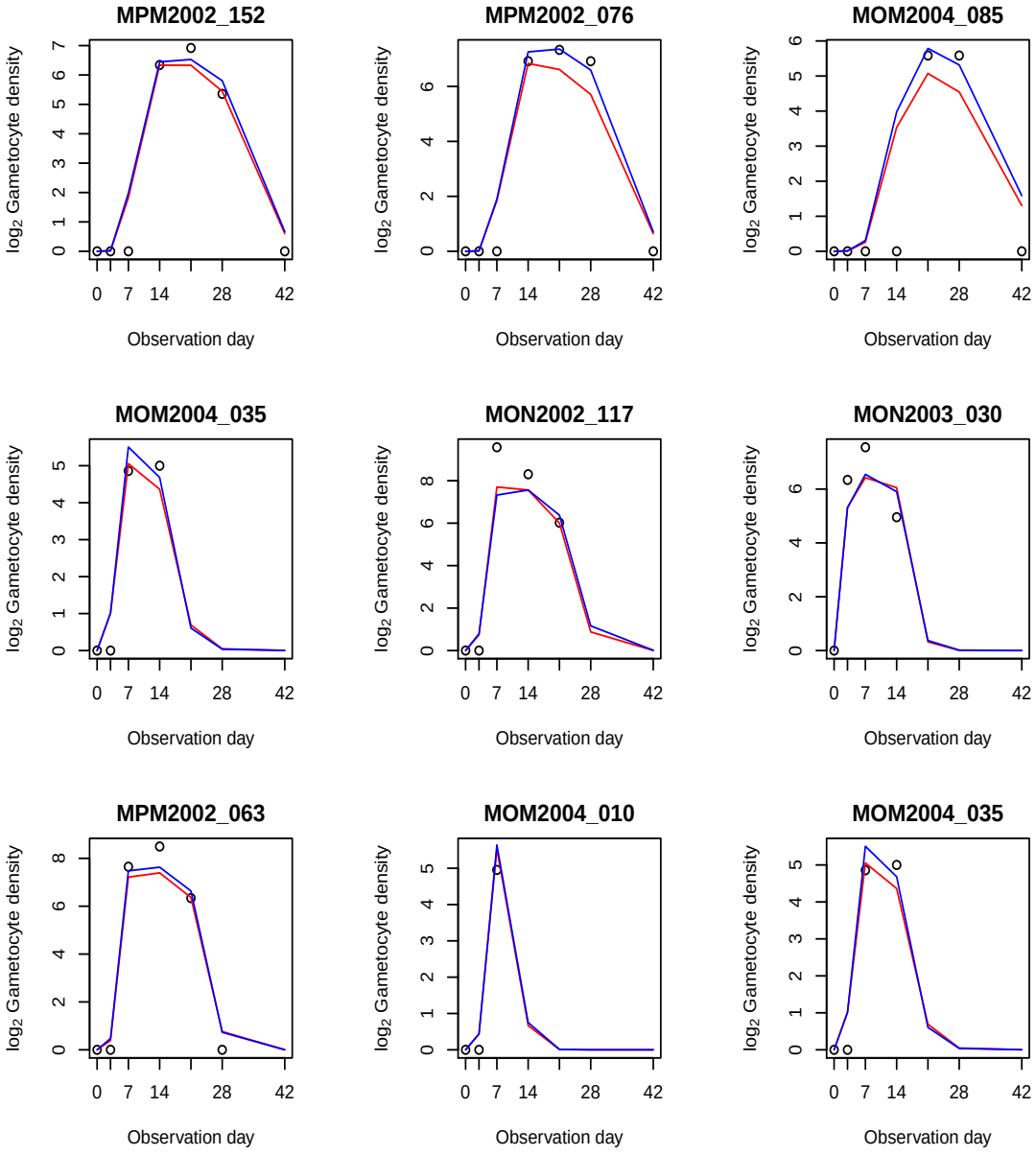

Figure 38: Observed Gametocyte mean profiles (black circles) for selected patients superimposed with predicted mean profiles using the ZAG model (red line) and the ZAGJWC model (blue line). The first row consists of patients who were observed at all observations days. The remaining rows are examples of incomplete gametocyte profiles with the second row consists of patients who experienced treatment failure with the third row consisting of patients lost-to-follow-up

4.11 Estimation of the duration of gametocytemia using imputed gametocyte profiles

In Chapter 3 an analysis into the time to gametocyte clearance was conducted. In that analysis it was assumed that gametocyte clearance occurred in the middle of adjacent observation days. This was a simplistic assumption used to account for interval censoring. In this Chapter a ZAGJWC (zero-adjusted gamma nonlinear joint model with a Weibull cause-specific survival component) model was developed. This model can be used to impute incomplete gametocyte profiles whilst accounting for informative censoring. These imputed profiles would provide estimates for all missing observations, including observations that would lie in-between patient follow up days. These complete gametocyte profiles can thus be used to investigate the risk factors that influence a more precisely defined time to gametocyte clearance and subsequently be used to estimate the duration of gametocytemia. The use of these imputed profiles, in an investigation to estimate the duration of gametocytemia, would mitigate the effect of interval censoring on the model estimation procedure.

The procedure outlined in Chapter 4.9 was used to derive estimates for a Weibull AFT model, applied to imputed profiles within the iterative imputation algorithm, designed to estimate the duration of gametocytemia in patients. The survival model results are presented in Table 30. This table also provides a comparison with the fitted Weibull AFT model from Chapter 3, which is referred to as the "Original" model.

Table 30: Comparison of Weibull AFT models (95% CI).

Definition	Parameter	Original	Imputed
intercept	α_0	1.281 (0.661 ; 1.902)	1.817 (0.959 ; 2.370)
trt	α_1	-1.022 (-1.401 ; -0.643)	-0.778 (-1.710 ; -0.286)
ratio	α_2	-0.224 (-0.484 ; 0.037)	-0.183 (-0.471 ; 0.131)
anaemia	α_3	0.005 (-0.214 ; 0.223)	-0.109 (-0.477 ; 0.096)
*mut*5	α_4	0.041 (-0.229 ; 0.311)	-0.151 (-0.756 ; 0.139)
lage	α_5	0.046 (-0.034 ; 0.125)	0.065 (-0.009 ; 0.151)
pzero	α_6	0.096 (0.058 ; 0.134)	0.092 (0.053 ; 0.141)
gender	α_7	0.120 (-0.091 ; 0.330)	0.022 (-0.237 ; 0.314)
scale	σ	0.711 (0.631 ; 0.790)	0.804 (0.632 ; 1.130)

Table 30 reveals that, under the imputed model, patient age has a marginal association with the time to gametocyte clearance (as shown by the predomi-

nately positive 95% CI for this parameter). The implication is that doubling patient age leads to a 6.8% increase in the time to gametocyte clearance, under the imputed model. The magnitude of the treatment effect was found to have changed considerably between the two models. Receiving a combination treatment of artesunate and sulfadoxine-pyrimethamine was found to reduce the time to gametocyte clearance by 54.1%, under the imputed model, whilst it reduced the time to gametocyte clearance by 64% under the "original" model.

The magnitude of the effect associated with the first 24 hour parasite reduction ratio and the baseline asexual parasite density, did not change considerably between the two models. Under the imputed model, clearing all baseline asexual parasites in the first 24 hours reduces the time to gametocyte clearance by 16.7% as compared to the "original" model, where clearing all baseline asexual parasites in the first 24 hours reduces the time to gametocyte clearance by 20%. Every ten fold increase in the baseline asexual parasite density leads to a 9.6% increase in the time to gametocyte clearance, when considering the imputed model, as compared to a 10.1% increase under the "original" model. It is also evident from Table 30 that the effects of the prevalence of moderate anaemia, prevalence of quintuple mutations and patient gender, do not have an association with the time to gametocyte clearance.

The approach taken to derive the time to gametocyte clearance parameter estimates, from the ZAGJWC model, was also applied to derive parameter estimates for the prevalence of gametocytemia model. The results are shown in Table 31. It can be seen that there is a marginal difference, in the parameter estimates, between the model fit in Chapter 3 and the model resulting from the imputed gametocyte profiles. It can also be seen that the strength of the association between treatment and the prevalence of gametocytemia is reduced, when comparing the treatment effect in the imputed model to the original model. An additional finding is that the prevalence of mutation can be seen to have a strong association with the prevalence of gametocytemia in the imputed model (the presence of quintuple mutations increases the relative risk of gametocytemia by 18.3%), whilst it previously had no association under the original model.

As previously discussed, estimates for the duration of gametocytemia that were derived from the Weibull AFT model would only be valid for the cohort of patients who experienced gametocytemia. An adjustment would be required to derive population level duration estimates for generic patients with specific covariate patterns. This adjustment would require that the estimated prevalence of gametocytemia be combined with the estimated duration of gametocytemia, for a generic patient with a specific covariate pattern. Given that the duration of

Table 31: Comparison of gametocyte prevalence models (95% CI).

Definition	Parameter	Original	Imputed
intercept	β_{B0}	-1.597 (-2.159 ; -1.034)	-1.500 (-1.893 ; -0.821)
trt	β_{B1}	-0.766 (-1.134 ; -0.397)	-0.644 (-1.024 ; 0.040)
ratio	β_{B2}	-0.461 (-0.723 ; -0.199)	-0.416 (-0.585 ; -0.204)
anaemia	β_{B3}	0.171 (-0.015 ; 0.357)	0.181 (0.062 ; 0.346)
*mut*5	β_{B4}	0.128 (-0.070 ; 0.326)	0.168 (0.045 ; 0.397)
lage	β_{B5}	0.012 (-0.053 ; 0.077)	0.017 (-0.025 ; 0.054)
pzero	β_{B6}	0.053 (0.020 ; 0.087)	0.049 (0.003 ; 0.073)
gender	β_{B7}	0.208 (0.030 ; 0.387)	0.181 (0.033 ; 0.302)

gametocytemia for a generic patient with a specific covariate pattern is defined as s_i with the prevalence of gametocytemia for the same patient given π_i, the resulting adjusted duration (d_i) would be given as

$$d_i = s_i \times \pi_i.$$

A dataset of generic patients was created to illustrate the impact of specific covariate patterns on the estimated duration and adjusted duration of gametocytemia. The dataset that was created was similar to the one used to derive Table 17 in Chapter 3. The dataset applied in this section allows for comparisons to be made on the covariates that were found to have a strong association with the time to gametocyte clearance, under the imputed survival model. The following characteristics were compared in this table

- Sulfadoxine-pyrimethamine treatment only (SP) and a combination treatment of artesunate and sulfadoxine-pyrimethamine (ACT)
- "High" $\log_{10}$ baseline parasite density of 17.754 relative to a "Low" baseline parasite density of 5
- "High" first 24 hour parasite reduction ratio of 100% relative to a "Low" first 24 hour parasite reduction ratio of -31.4%
- "Old" patient with an age of 55.0 ($lage = 5.781$) relative to "Young" patient with age of 2 ($lage = 1$)

These results are presented, in Table 32, for a female with a haemoglobin density greater than 11g/dL and less than 5 mutations present. It can be seen that patients who would have the longest duration have the following characteristics

- Receive sulfadoxine-pyrimethamine (SP) treatment only
- Have a high baseline asexual parasite density
- Have a low first 24 hour parasite reduction ratio
- Are generally older

Table 32: Estimated duration of gametocytemia, in days, derived using imputed gametocyte profiles for a a female with a haemoglobin density greater than 11g/dL and less than 5 mutations present.

trt	*pzero*	*ratio*	*lage*	Duration	Adjusted Duration
SP	Low	low	Young	11.05 (5.82 ; 15.91)	3.62 (2.33 ; 5.2)
			Old	14.97 (8.58 ; 21.66)	5.38 (3.6 ; 7.55)
		High	Young	8.52 (4.18 ; 14.55)	1.66 (0.98 ; 2.82)
			Old	11.71 (6.30 ; 19.29)	2.44 (1.58 ; 4.08)
	High	low	Young	35.70 (20.83 ; 50.06)	21.69 (13.20 ; 29.29)
			Old	48.70 (32.71 ; 63.76)	32.01 (21.75 ; 41.75)
		High	Young	28.04 (15.56 ; 43.33)	9.78 (6.43 ; 14.33)
			Old	38.02 (25.81 ; 53.98)	14.51 (10.70 ; 19.68)
ACT	Low	low	Young	5.03 (1.60 ; 9.39)	0.85 (0.45 ; 1.50)
			Old	6.92 (2.24 ; 12.40)	1.29 (0.70 ; 2.08)
		High	Young	3.81 (1.29 ; 7.75)	0.40 (0.19 ; 0.74)
			Old	5.29 (1.90 ; 9.80)	0.60 (0.30 ; 1.04)
	High	low	Young	16.46 (4.64 ; 30.53)	5.01 (2.64 ; 8.82)
			Old	22.76 (6.69 ; 39.88)	7.57 (4.11 ; 12.15)
		High	Young	12.65 (3.83 ; 23.20)	2.31 (1.27 ; 4.06)
			Old	17.55 (5.55 ; 29.74)	3.43 (2.08 ; 5.56)

It has been shown in this thesis, when considering the results of the ZAGJWC (zero-adjusted gamma nonlinear joint model with a Weibull cause-specific survival component) model, that treatment has a strong association with the hazard of treatment failure and the hazard of loss-to-follow-up (Table 27). It was found that receiving a combination treatment of artesunate and sulfadoxine-pyrimethamine (ACT) decreased the hazard of treatment failure by 94%, whilst increasing the hazard of loss-to-follow-up by 289 fold. The implication of these findings is that the treatment administered to a patient is strongly associated with treatment outcome. Patients who experienced treatment failure are thus closely associated with patients who received SP treatment whilst patients who were lost-to-follow-up are associated with patients who received a combination treatment of artesunate and sulfadoxine-pyrimethamine (ACT). Based on these findings it can be concluded that patients who experience treatment failure are expected to have longer durations of gametocytemia as compared to patients who are lost-to-follow-up.

Baseline asexual parasite density was found to have strong association with the hazard of treatment failure with a ten-fold increase in the baseline asexual parasite density resulting in a 21.7% increase in the hazard of treatment failure (Table 27). Baseline asexual parasite density was also found to have a strong association with the prevalence of and duration of gametocytemia analyses conducted using the imputed gametocyte profiles from the ZAGJWC model. It was found that a ten fold increase in the baseline asexual parasite density resulted in a 5.0% increase in the relative risk of gametocytemia and a 9.6% increase in the time to gametocyte clearance.

The prevalence of quintuple mutations was used as a measure of resistance to treatment in this thesis. It was found that the prevalence of quintuple mutations had a strong association with the hazard of treatment failure, with the presence of quintuple mutations leading to a 21.7% increase in the hazard of treatment failure. There was no association between the prevalence of quintuple mutations and the hazard of loss-to-follow-up. It was found that the prevalence of quintuple mutations had a strong association with the prevalence of gametocytemia (the presence of quintuple mutations leads to a 18.3% increase in the relative risk of gametocytemia), when considering the analysis conducted using the imputed gametocyte profiles from the ZAGJWC model. The impact of the prevalence of quintuple mutations is as expected in literature and this result signifies an improvement from the prevalence of gametocytemia model derived in Chapter 3. It was also highlighted that the prevalence of quintuple mutations did not have an association with the duration of gametocytemia, in either the original

model or the imputed model.

The first 24 hour parasite reduction ratio was also found to have a strong association with the hazard of treatment failure, whilst not having an association with the hazard of loss-to-follow-up. Clearing all baseline asexual parasites in the first 24 hours was found to lead to a 94.6% reduction in the hazard of treatment failure, under the ZAGJWC model. The first 24 hour parasite reduction ratio was found to have a strong association with the prevalence of gametocytemia (clearing all baseline asexual parasites in the first 24 hours leads to a 34.0% reduction in the relative risk of gametocytemia), when considering the imputed model. The first 24 hour parasite reduction ratio was found to have no association with the duration of gametocytemia.

5 Discussion

Gametocytes are the sexual form of the *Plasmodium falciparum* parasite and they have been found to have a significant influence on the infectivity of a host patient. Most malaria intervention trials focus on the clearance of asexual parasites, which are responsible for the signs and symptoms associated with malaria infection. An effective treatment is one that can quickly and efficiently clear both forms of the *Plasmodium falciparum* parasite.

The main aim of this thesis was the imputation of the incomplete gametocyte profile in the presence of censoring. The clinical motivation for this analysis was that research into the risk factors that influence the emergence of gametocytes, duration of gametocytemia and density of gametocytes is of vital public health importance as the findings from such work can be used to reduce the transmission of malaria and subsequently the spread of antimalarial drug resistance.

There is also a methodological motivation for this analysis. Gametocyte data is characterized by being zero inflated, skew with a long-tail to the right and nonlinear with regards to the relationship between the gametocyte density profile and time. There is also a competing risk nature associated with the different reasons for early exit from the study. These features of gametocyte data make it a statistically appealing dataset to analyze. In order to answer the clinical questions raised above, joint modeling techniques based on cause-specific Weibull models for the times to early exit and zero-adjusted gamma models for the longitudinal profiles had to be developed.

This chapter will be split into methodological and clinical discussion sections.

5.1 Methodological discussion

The gametocyte data used in this investigation was found to be zero-inflated with a long-tail to the right. The observed longitudinal gametocyte profiles, of patients included in this analysis, were found to have a nonlinear relationship with time. Another feature of this data was the presence of censoring. Some patients exited the study before its completion either due to treatment failure or loss-to-follow-up. It was found that the experience of patients who exited the study early was different to that of patients who remained, thus implying that this exit was a form of informative censoring. Additionally it was found that there was a competing risks nature to the forms of early exit from the study. Patients who withdrew from the study either due to treatment failure or loss-to-follow-up had incomplete gametocyte profiles. It can reasonably be assumed

that the gametocyte profiles of patients who were rescued from the study or who were lost-to-follow-up (assuming informative censoring) would differ from the observed profiles of patients who were successfully treated and observed to the end of the gametocyte cycle.

Longitudinal nonlinear profiles are typically modeled using nonlinear mixed effect models. These models typically assume an underlying normal distribution for the response. Since gametocyte data are zero-inflated with a long-tail to the right, the assumption of normality is not be valid. This distributional assumption can be changed, with the response assumed to follow a zero-adjusted gamma distribution. Making this change would allow for a standard nonlinear mixed effect model to be fit. However, standard nonlinear mixed effect models rely on model based imputation due to maximum likelihood estimation in the presence of missing data. Hence these models do not account for informative censoring, which can lead to biased parameter estimates.

These aforementioned problems were overcome by using joint modeling techniques based on cause-specific Weibull models for the times to early exit and zero-adjusted gamma models for the longitudinal profiles. The resulting joint model used shared random effects to combine a Weibull survival model, describing the cause-specific hazards of patient exit from the study, with a nonlinear zero-adjusted gamma mixed effect model for the longitudinal gametocyte profile (ZAGJWC model). Table 29 compared the longitudinal parameter estimates of the zero-adjusted gamma mixed effect (ZAG) and the ZAGJWC models. It was shown that there were noticeable differences in the magnitude of the fixed effects across the ZAG and ZAGJWC models. In some cases parameters that were found to have no association with the longitudinal gametocyte profile, under the ZAG model, were found to have a strong association with the gametocyte profile once a ZAGJWC model was applied. A comparison of the observed gametocyte profiles for select patients from the study, along with the fitted values from both the ZAG and ZAGJWC models, was illustrated graphically in Figure 38. This graphic revealed that the ZAGJWC predicted profiles outperformed the ZAG across the group of patients who were observed at all timepoints. However, it was observed that there was a small difference between the models when predictions were made on incomplete profiles. The development of additional model comparison techniques is thus an area of further research that would enrich this analysis and provide further backing to the assertion that the ZAGJWC model is a better model as compared to the ZAG model.

It was shown in Table 26 and Table 27 that several precision parameters could arguably be omitted from the fitted joint models. However, these random

effects were retained in the models as they captured the within subject correlated structure. These random effects were also retained as they allowed for an easy comparison of the different model structures fit to the data.

Complex likelihood functions were developed in order to fit the joint models described above. These functions were estimated under the Bayesian framework. The resulting joint posterior distributions for the parameters were analytically intractable thus Markov Chain Monte Carlo (MCMC) methods were used to obtain the point and interval estimates of parameters. The imputed gametocyte profiles were derived from posterior samples generated as part of the Bayesian MCMC procedure used in the development of the applicable joint model. These completed profiles would provide estimates for all missing data, including missing data arising in-between the predefined study observation days.

The complete gametocyte profiles were subsequently used in an analysis to investigate the risk factors that affect the prevalence of gametocytemia and the time to gametocyte clearance. The advantages of using imputed gametocyte profiles are that they account for informative censoring and they mitigate the effect of interval censoring as predicted responses can be generated for the time-points that lie in-between the study's predefined observation dates. It was found that the use of the imputed data improved the results of the analysis. This was highlighted by the prevalence of quintuple mutations (used as a measure of resistance to treatment in this thesis) which was found to have no association with the prevalence of gametocytemia, when considering the observed data. However, when the imputed data was analyzed a strong association was found between the prevalence of quintuple mutations and the prevalence of gametocytemia. This association is as expected in literature and this result highlights the improvement in analysis arising from the use of imputed data derived from the ZAGJWC model.

An unfortunate drawback of the methodology developed above is that it is computationally intensive due to the complexity of the fitted models. It took 2 days to fit the ZAGJWC model using a laptop with an i7 processor, 64-bit operating system and 8GB RAM. It is hoped that with future advancements in computational power, the running time for these models would decrease.

The treatment administered to a patient was found to have a strong association with the patient's treatment outcome. Treatment outcome provided information used in the imputation of the longitudinal gametocyte profile through the cause-specific survival component of the ZAGJWC model. Generally a patient's response to treatment depends on the immunity level of the host, the level of $Plasmodium\ falciparum$ parasite resistance to treatment and the phar-

macokinetic properties of the antimalarial treatment. The models considered in this thesis did not account for the pharmacokinetic properties of the antimalarial treatment. The work conducted in this thesis can thus be extended to include the evolution of patient drug concentrations in the longitudinal gametocyte model. The expectation is that adding the pharmacokinetics of the drugs administered to the patient would improve the performance of the fitted models.

An additional methodological innovation that was discussed in this thesis was the development of a framework for the estimation of population level gametocyte duration estimates, for patients with specific covariate patterns. This methodology combined estimates for the prevalence of gametocytemia with estimates for the duration of gametocytemia to give an expected population level estimate for the duration of gametocytemia, with an associated confidence interval derived using the delta method. A simplifying assumption of independence between the duration of gametocytemia and the prevalence of gametocytemia was applied in the derivation of the variance structure for the estimated population level duration estimates. It is acknowledged that there could possibly be a correlation between the duration of gametocytemia and the prevalence of gametocytemia. An investigation into the appropriateness of the independence assumption and the resulting change in the variance structure of the population level duration estimates, assuming that the independence assumption is not appropriate, is an area of further work.

An additional area of further work, is the design of an optimal cost effective study that is able to capture data at key points of the gametocyte life cycle. Currently the majority of malaria intervention studies are designed to capture the evolution of the asexual parasites, resulting in sparse observations in the periods when gametocyte are expected to be prevalent. The methodology used to impute incomplete gametocyte profiles, which was outlined in this study, can be used to estimate time periods of high gametocyte prevalence. Subsequently studies can be designed that capture gametocyte data at these key points.

5.2 Clinical discussion

Since 2000 there has been a considerable reduction in the global malaria burden. This has been associated with the wide-scale deployment of artemisinin based combination therapies in malaria endemic countries. Artemisinin based treatments work quickly and efficiently against asexual parasites (White, 1997) and immature gametocytes (Adjalley et al., 2011, Price et al., 1996). As a result these treatments are associated with high patient cure rates and low post-

treatment malaria transmission (Group et al., 2016). Unfortunately over the last few years artemisinin resistance has been confirmed in the Greater Mekong Subregion, in countries like Cambodia, the Lao Peoples Democratic Republic, Myanmar, Thailand and Vietnam. In most cases, patients with artemisinin-resistant parasites are still able to achieve an adequate clinical and parasitological response if the longer-acting partner drug used in conjunction with the artemisinin derivate is still effective in that geographical area. However, it has been found that *Plasmodium falciparum* has become resistant to almost all available antimalarial treatments in areas along the Cambodia-Thailand border (WHO, 2013). It is thus imperative that this treatment resistant strain is stopped from spreading. Currently a single dose of primaquine is recommended, by the WHO (WHO, 2017b), to be used in conjunction with artemisinin in order to prevent transmission of *Plasmodium falciparum* to mosquitoes in areas threatened by artemisinin resistance. Research into the risk factors that affect the prevalence gametocytemia, duration of gametocytemia and the density of gametocytes is of vital importance to public health. This is because a decrease gametocyte carriage does not just reduce ongoing malaria transmission, but it can also slow down the spread of antimalarial drug resistance that is threatening current malaria control and elimination efforts.

The modeling of the key features of gametocytemia, was an area of focus in this thesis. Particular emphasis was placed on the modeling of the time to gametocyte emergence, prevalence of gametocytemia, duration of gametocytemia and the density of gametocytes. It was also of interest to investigate the impact of current and novel antimalarials under development on the aforementioned features.

The data used in this analysis was from a series of clinical trials conducted between 2002 and 2004 in southern Mozambique and the Mpumulanga province of South Africa. The aim of these studies was primarily to measure the efficacy of two treatments, in eliminating asexual parasites in patients. The two treatment procedures used were sulfadoxine-pyrimethamine (SP) only and a combination of artesunate and sulfadoxine-pyrimethamine (ACT). The patients enrolled in the trials had moderate uncomplicated malaria, in a period of increasing resistance to sulfadoxine-pyrimethamine (SP) treatment.

The gametocyte data collected in this study was initially analyzed using standard survival modeling techniques. Non-parametric Cox regression models and parametric survival models were applied to the data as part of this initial investigation. These models were used to investigate the factors that affected the time to gametocyte emergence. Subsequently, using the subset of the pop-

ulation that experienced gametocytemia, accelerated failure time models were applied to investigate the factors that affected the duration of gametocytemia. It is evident that the findings from the aforementioned duration investigation would only be able to provide valid duration estimates for patients who experienced gametocytemia. This work was extended to allow for population level duration estimates by incorporating the prevalence of gametocytemia into the estimation of duration. The prevalence of gametocytemia was modeled using an underlying binomial distribution. The delta method was subsequently used to derive confidence intervals for the population level duration estimates that were associated with specific covariate patterns.

The gametocyte data used in this analysis was collected over a 42 day follow up period with observations occurring on days 0, 3, 7, 14, 21, 28 and 42. Due to the design of the study, it is evident that in addition to informative censoring, interval censoring would also be expected to arise. The previously discussed accelerated failure time models, which were used to investigate the factors that affected the duration of gametocytemia, applied a simplistic approach to the analysis of the gametocyte clearance data. It was assumed that gametocyte clearance occurred in the middle of adjacent observation days, with this midpoint used as a point estimate for the gametocyte clearance time. The findings from the investigation into the duration of gametocytemia were updated by taking the imputed profiles, derived as part of the fitting process for the ZAGJWC model and then subsequently using them to estimate the duration of gametocytemia. The use of these imputed profiles mitigated the effect of informative censoring and interval censoring on the resulting analysis.

The results of the updated time to gametocyte clearance analysis (results shown in Table 30) highlighted that a patient's treatment had a strong association with the duration of gametocytemia. It was found that receiving a combination of artesunate and sulfadoxine-pyrimethamine treatment as compared to sulfadoxine-pyrimethamine treatment only, decreases the time to gametocyte clearance by 54.1%. The density of baseline asexual parasites was also found to have a strong effect on the duration of gametocytemia, with a ten-fold increase in the baseline asexual parasite density leading to a 9.6% increase in the time to gametocyte clearance. It was also found that patient age has a marginally strong association with the time to gametocyte clearance, as doubling patient age leads to a 6.8% increase in the time to gametocyte clearance. In literature, increased age is usually associated with lower gametocyte carriage (in areas of moderate to intense malaria transmission such as Mozambique where partial immunity is acquired with age), or has no effect on gametocyte carriage (in

non-immune patients such as those living in areas of low intensity transmission, e.g. Mpumalanga). It is thus unusual that the results of this analysis imply that increasing age leads to an increase in the duration of gametocytemia. A possible reason for this result is that patients from different study sites were combined for the purposes of this investigation. As previously discussed the impact of age is dependent on the malaria transmission rates associated with the site of the study, as a result an improvement in the interpretation of the results might occur if study site is incorporated into the model.

The imputed gametocyte profiles from the ZAGJWC model were used to derive a prevalence of gametocytemia model (results shown in Table 31). It was found that treatment had a marginal association with the prevalence of gametocytemia, as the 95% CI for the parameter estimate was predominately negative. The first 24 hour parasite reduction ratio, the prevalence of anaemia, the prevalence of quintuple mutations, baseline asexual parasite density and gender were all found to have strong associations with the relative risk of gametocytemia. It is interesting to note that the prevalence of gametocytemia model derived in Chapter 3 found that the prevalence of quintuple mutations had no association with the prevalence of gametocytemia. The use of imputed gametocyte profiles increases the strength of the association between the prevalence of quintuple mutations and the prevalence of gametocytemia. The association between the prevalence of quintuple mutations and the prevalence of gametocytemia makes clinical sense and it is as expected from literature. The same applies for the prevalence of moderate anaemia, as this covariate was found to have no association with the prevalence of gametocytemia in the model fit in Chapter 3. However, the use of imputed gametocyte profiles increased the strength of the association between the prevalence of moderate anemia and the prevalence of gametocytemia. The model derived from the imputed gametocyte profiles can be intercepted as

- Receiving a combination of artesunate and sulfadoxine-pyrimethamine (ACT) reduces the relative risk of gametocytemia by 47.4% as compared to receiving sulfadoxine-pyrimethamine treatment only
- Clearing all baseline asexual parasites in the first 24 hours reduces the relative hazard of gametocytemia by 34.0%
- The presence of moderate anaemia increases in the relative risk of gametocytemia by 19.8%
- The presence of quintuple mutations increases the relative risk of gametocytemia by 18.3%

- A ten-fold increase in the baseline asexual parasite density increases the relative risk of gametocytemia by 5.0%
- Being a male patient increases the relative risk of gametocytemia by 19.8%

The prevalence of quintuple mutations and the first 24 hour parasite reduction ratios did not have an association with the duration of gametocytemia, in either of the original model (derived from Chapter 3) or the imputed (derived using imputed gametocyte profiles from the ZAGJWC model) models. These results are unexpected as the initial assumption around the impact of resistance to treatment and the first 24 hour parasite reduction ratio was that these covariates would affect both the prevalence and duration of gametocytemia. Further work is required in order to understand the aforementioned findings.

In summary, the key findings from this investigation were that treatment had a highly significant effect on the time to gametocyte emergence, extent of gametocytemia and the duration of gametocytemia. Patients who received a combination of artesunate and sulfadoxine-pyrimethamine treatment were found to have significantly lower hazards of gametocyte emergence, lower predicted durations of gametocytemia and lower predicted longitudinal gametocyte densities, as compared to patients who received sulfadoxine-pyrimethamine treatment only. It was also shown that the impact of a combination of artesunate and sulfadoxine-pyrimethamine treatment on the gametocyte density was stronger than the impact of all the other covariates considered in this study. This result implies that regardless of a patient's covariate pattern, the application of a combination artesunate and sulfadoxine-pyrimethamine treatment will suppress the gametocyte density. This is a powerful result that highlights the importance of artemisinin based treatment to the fight against malaria, as this treatment can be seen to have a significant effect on the key driver of malaria transmission. A reduction in the transmission rate of malaria in a community can also lead to a reduction in the spread of antimalarial drug resistance, that is threatening current malaria control and elimination efforts.

5.3 Concluding remarks

The overall contribution of this piece of work was that a methodology was developed that allowed for the imputation of incomplete gametocyte profiles, whilst taking into account informative censoring and the underlying structure of the longitudinal gametocyte patient profile, through the use joint modeling techniques. The imputed gametocyte profiles were derived from posterior samples generated as part of the Bayesian MCMC procedure used in the development

of the applicable joint model. The resulting joint model used shared random effects to combine a Weibull survival model, describing the cause-specific hazards of patient exit from the study, with a nonlinear zero-adjusted gamma mixed effect model for the longitudinal gametocyte profile. This model was used to impute the incomplete gametocyte profiles, after adjusting for informative censoring. These imputed profiles were then used to identify the risk factors that influence the duration and prevalence of gametocytemia. An additional contribution of this work was the development of a methodology for the estimation of population level gametocyte duration estimates. This methodology combined estimates for the prevalence of gametocytemia with estimates for the duration of gametocytemia to give an expected population level estimate for the duration of gametocytemia with an associated confidence interval, derived using the delta method.

It is hoped that the findings from this research will be incorporated into the continuous fight against malaria infection and transmission.

References

Aalen, O. (1976). Nonparametric inference in connection with multiple decrement models, *Scandinavian Journal of Statistics* **3**: 15–27.

Adjalley, S. H., Johnston, G. L., Li, T., Eastman, R. T., Ekland, E. H., Eappen, A. G., Richman, A., Sim, B. K. L., Lee, M. C., Hoffman, S. L. et al. (2011). Quantitative assessment of plasmodium falciparum sexual development reveals potent transmission-blocking activity by methylene blue, *Proceedings of the National Academy of Sciences* **108**(47): E1214–E1223.

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control* **19**: 716–723.

Allen, E. N., Little, F., Camba, T., Cassam, Y., Raman, J., Boulle, A. and Barnes, K. I. (2009). Efficacy of sulphadoxine-pyrimethamine with or without artesunate for the treatment of uncomplicated plasmodium falciparum malaria in southern mozambique: a randomized controlled trial, *Malaria Journal* **8**: 141.
URL: *https://doi.org/10.1186/1475-2875-8-141*

Babiker, H. A., Satti, G., Ferguson, H., Bayoumi, R. and Walliker, D. (2005). Drug resistant plasmodium falciparum in an area of seasonal transmission, *Acta Tropica* **94**(3): 260268.

Barnes, K. I. and White, N. J. (2005). Population biology and antimalarial resistance: The transmission of antimalarial drug resistance in plasmodium falciparum, *Acta tropica* **94**(3): 230–240.

Barnes, K., Little, F., Smith, P., Evans, A., Watkins, W. and J White, N. (2006). Sulfadoxine-pyrimethamine pharmacokinetics in malaria: Pediatric dosing implications, *Clinical pharmacology and therapeutics* **80**: 582–596.

Collett, D. (1994). *Modelling Survival Data in Medical Research*, London: Chapman and Hall.

Cooper, N., Lambert, P., Abrams, K. and Sutton, A. (2007). Predicting costs over time using Bayesian Markov chain Monte Carlo methods: an application to early inflammatory polyarthritis, *Health Economics* **16**: 37–56.

Cox, D, R. (1972). Regression models and life-tables (with discussion), *Journal of the Royal Statistical Society* **34**: 187 – 220.

Daniels, M. and Hogan, J. (2008). *Missing Data in Longitudinal Studies. Strategies for Bayesian Modeling and Sensitivity Analysis*, Chapman and Hall/CRC, Boca Raton.

Delignette-Muller, M., Dutang, C., Pouillot, R. and Denis, J. (2014). Help to fit of a parametric distribution to non-censored or censored data.

Despommier, D., Gwadz, R. and Hotez, P. (1994). *Parasitic Diseases*, 3rd edn, Springer, Berlin.

Diebner, H., Eichner, M., Molineaux, L., Collins, W., Jeffery, G. and Dietz, K. (2000). Modelling the transition of asexual blood stages of plasmodium falciparum to gametocytes, *Journal of Theoretical Biology* **202**: 113–127.

Distiller, G., B., Little, F. and Barnes, K., I. (2010). Nonlinear mixed effects modeling of gametocyte, *Malaria Journal* **9**(60).

Draper, C. (1953). Observations on the infectiousness of gametocytes in hyperendemic malaira, *Transactions of the Royal Society of Tropical Medicine and Hygiene* **47**: 160–165.

Elliott, M., Gallo, J., Ten Have, T., Bogner, H. and Katz, I. (2005). Using a Bayesian latent growth curve model to identify trajectories of positive affect and negative events following myocardial infarction, *Biostatistics* **6**: 119–43.

Faucett, C. J. and Thomas, D. C. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: A Gibbs sampling approach, *Statistics in Medicine* **15**: 1663 – 1685.

Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk, *Journal of the American statistical association* **94**(446): 496–509.

Fitzmaurice, G., Laird, N. and Ware, J. (2004). *Applied Longitudinal Analysis*, Wiley, Hoboken.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis*, New York: John Wiley and Sons.

Gelfand, A. and Ghosh, S. (1998). Model choice: A minimum posterior predictive loss approach, *Biometrika* **85**: 1–11.

Ghosh, S., Mukhopadhyay, P. and Lu, J. (2006). Bayesian analysis of zero-inflated regression models, *Journal of Statistical Planning and Inference* **136**: 1360–1375.

Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika* **82**: 711–732.

Group, W. G. S. et al. (2016). Gametocyte carriage in uncomplicated plasmodium falciparum malaria following treatment with artemisinin combination therapy: a systematic review and meta-analysis of individual patient data, *BMC medicine* **14**(1): 79.

Guo, X. and Carlin, B. (2004). Separate and joint modeling of longitudinal and event time data using standard computer packages, *The American Statistician* **58**: 1–9.

Hall, D. (2000). Zero-inflated poisson and binomial regression with random effects: a case study., *Biometrics* **56**: 1030–1039.

Heilbron, D. (1989). Generalized linear models for altered zero probabilities and overdispersion in count data., *Technical report*, Department of Epidemiology and Biostatistics, University of California; San Francisco.

Heilbron, D. (1994). Zero-altered and other regression models for count data with added zeros., *Biometrical Journal* **36**: 531–547.

Henderson, R., Diggle, P. and Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data, *Biostatistics* **1**(4): 465–480.

Hoeting, J.A.and Madigan, D., Raftery, A. and Volinsky, C. (1999). Bayesian model averaging: a tutorial, *Statistical Science* **14**: 382–417.

Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring, *Annals of Statistics* **24**: 540–568.

Ibrahim, J., Chu, H. and Chen, L. (2001). Basic concepts and methods for joint models of longitudinal and survival data, *Journal of Clinical Oncology* **28**: 2796–2801.

Jackson, C. (2016). *flexsurv: Flexible Parametric Survival and Multi-State Models*. R package version 0.7.1.
URL: *http://CRAN.R-project.org/package=flexsurv*

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, John Wiley and Sons, Inc.

Kaplan, E. and Meier, P. (1958). Nonparametric estimation for incomplete observations, *Journal of the American Statistical Association* **93**: 457–481.

Kass, R. and Raftery, A. (1995). Bayes factors, *Journal of American Statistical Association* **90**: 773–795.

Kellner, K. (2016). *jagsUI: A Wrapper Around 'rjags' to Streamline 'JAGS' Analyses*. R package version 1.4.4.
URL: *https://CRAN.R-project.org/package=jagsUI*

Killeen, G., Ross, A. and Smith, T. (2006). Infectiousness of malaria-endemic human populations to vectors, *American Journal of Tropical Medical Hygiene* **75**: 38–45.

Klein, J. P. and Moeschberger, M. L. (2005). *Survival analysis: techniques for censored and truncated data*, Springer Science & Business Media.

Kooperberg, C. and Clarkson, D. B. (1997). Hazard regression with interval-censored data, *Biometrics* **53**: 1485–1494.

Korenromp, E., Williams, B., Gouws, E., Dye, C. and Snow, R. (2003). Measurement of trends in childhood malaria mortality in africa: an assessment of progress towards targets based on verbal autopsy, *The Lancet Infectious Diseases* **3**: 349–358.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing, *Technometrics* **34**: 1–14.

Lange, K. (2004). *Optimization*, Springer-Verlag New York.

Lau, B., Cole, S. R. and Gange, S. J. (2009). Competing risk regression models for epidemiologic data, *American journal of epidemiology* **170**(2): 244–256.

Laud, P. and Ibrahim, J. (1995). Predictive model selection, *Journal of the Royal Statistical Society* **57**: 247–262.

Lina, H., Turnbulla, B. W., McCullocha, C. E. and Slatea, E. H. (2002). Latent class models for joint analysis of longitudinal biomarker and event process data, *Journal of the American Statistical Association* **97**: 53–65.

Lindsey, J. C. and Ryan, L. M. (1998). Methods for interval-censored data, *Statistics in Medicine* **17**: 219–238.

Lindsey, J. K. (1998). A study of interval censoring in parametric regression models, *Lifetime Data Analysis* **4**: 329–354.

Lindstrom, M, J. and Bates, D, M. (1990). Nonlinear mixed-effects models for repeated measures data, *Biometrics* **46**: 673–687.

Little, R. (1995). Modeling the drop-out mechanism in repeated-measures studies, *Journal of the American Statistical Association* **90**: 1112–1121.

Little, R. and Rubin, D. (2002). *Statistical Analysis with Missing Data*, 2nd edn, John Wiley and Sons, New York.

Lunn, D., Jackson, C., Best, N., Thomas, A. and Spiegelhalter, D. (2012). *The BUGS Book: A Practical Introduction to Bayesian Analysis*, Chapman & Hall/CRC Texts in Statistical Science, Taylor & Francis.
URL: *https://books.google.co.za/books?id=Cthz3XMa_VQC*

Michalakis, Y. and Renaud, F. (2009). Malaria evolution in vector control, **462**: 298–300.

Molenberghs, G. and Kenward, M. (2007). *Missing Data in Clinical Studies*, Wiley, New York.

Mullahy, J. (1986). Specification and testing of some modified count data models., *Journal of Econometrics* **33**: 341–365.

Murawska, M., Rizopoulos, D. and Lesaffre, E. (2012). A two-stage joint model for nonlinear longitudinal response and a time-to-event with application in transplantation studies, **2012**.

Nacher, M., Singhasivanon, P., Silachamroon, U., Treeprasertsuk, S., Tosukhowong, T., Vannaphan, S., Gay, F., Mazier, D. and Looareesuwan, S. (2002). Decreased hemoglobin concentrations, hyperparasitemia, and severe malaria are associated with increased plasmodium falciparum gametocyte carriage, *The Journal of Parasitology* **88**(1): 97–101.
URL: *http://www.jstor.org/stable/3285398*

Neelon, B., O'Malley, A. and Normand, S. (2010). A Bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use, *Statistical Modelling* **10**: 421–439.

Nosten, F., van Vugt, M., Luxemburger, C., Thway, K. L., Brockman, A., McGready, R., ter Kuile, F. Looareesuwan, S. and White, N. J. (2000). Effects of artesunate-mefloquine combination on incidence of plasmodium falciparum malaria and mefloquine resistance in western thailand: a prospective study, *The Lancet* **356**: 297–302.

Oehlert, G. (1992). A note on the delta method, *American Statistician* **46**.

Pinheiro, Jos, C. and Bates, Douglas, M. (2000). *Mixed-Effects Models in S and S-PLUS*, Statistics and Computing, Springer New York.
URL: *https://books.google.co.za/books?id=N3WeyHFbHLQC*

Pintilie, M. (2007). Analysing and interpreting competing risk data, *Statistics in medicine* **26**: 1360–1367.

Plummer, M. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling.

Pongtavornpinyo, W. (2006). *Mathematical modelling of antimalarial drug resistance*, PhD thesis, School of Tropical Medicine, University of Liverpool.

Prentice, R, L. (1974). A log gamma model and its maximum likelihood estimation, *Biometrika* **61**(3): 539–544.

Prentice, R. L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model, *Biometrika* **69**: 331–342.

Price, R. N., Nosten, F., Luxemburger, C., Ter Kuile, F., Paiphun, L., Chongsuphajaisiddhi, T. and White, N. (1996). Effects of artemisinin derivatives on malaria transmissibility, *The Lancet* **347**(9016): 1654–1658.

Price, R., Nosten, F., Simpson, J. A., Luxemburger, C., Phaipun, L., ter Kuile, F., van Vugt, M., Chongsuphajaisiddhi, T. and White, N. J. (1999). Risk factors for gametocyte carriage in uncomplicated falciparum malaria., *The American Journal of Tropical Medicine and Hygiene* **60**(6): 1019–1023.
URL: *http://www.ajtmh.org/content/journals/10.4269/ajtmh.1999.60.1019*

R Core Team (2015). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Ridout, M., Demtrio, C. and Hinde, J. (1998). Models for count data with many zeros., *International Biometric Conference Cape Town* .

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape., *Journal of the Royal Statistical Society* **54**: 507554.

Rigby, R. A. and Stasinopoulos, D. M. (2010). A flexible regression approach using GAMLSS in R, *London Metropolitan University, London* .

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data with Applications in R*, Chapman and Hall.

Roper, C., Pearce, R., Bredenkamp, B., Gumede, J., Drakeley, C., Mosha, F., Chandramohan, D. and Sharp, B. (2003). Antifolate antimalarial resistance in southeast africa: a population-based analysis, *The Lancet* **361**(9364): 1174 – 1181.
URL: *http://www.sciencedirect.com/science/article/pii/S0140673603129510*

Ross, A., Killeen, G. and Smith, T. (2006). Relationships between host infectivity to mosquitoes and asexual parasite density in plasmodium falciparum, *AM J Trop Med Hyg* **75 (2 Suppl)**.

Samuelson, S. O. and Kongerud, J. (1994). Interval censoring in longitudinal data of respiratory symptoms in aluminium potroom workers: a comparison of methods, *Statistics in Medicine* **13**: 1771–1780.

Seaman, S. R. and White, I. R. (2013). Review of inverse probability weighting for dealing with missing data, *Statistical Methods in Medical Research* **22**(3): 278–295.
URL: *https://doi.org/10.1177/0962280210395740*

Snow, R. W., Trape, J. F. and Marsh, K. (2001). The past, present and future, *TRENDS in Parasitology* **17**(17): 593–597.

Spiegelhalter, D. (1998). Bayesian graphical modelling: a case study in monitoring health outcomes, *Applied Statistics* **47**: 115–133.

Spiegelhalter, D., Best, N., Carlin, B. and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion), *Journal of the Royal Statistical Society* **64**: 583–639.

Spiegelhalter, D., Thomas, A., Best, N. and Lunn, D. (2003). *WinBugs Version 1.4: User Manual*, Cambridge: Medical Research Council Biostatistics Unit.
URL: *http://www.mrc-bsu.cam.ac.uk/bugs*

Tanner, M. and Wong, W. (1987). The calculation of posterior distributions by data augmentation (with discussion)., *American Statistical Association* **82**: 528–550.

Targett, G., Drakeley, C., Jawara, M., von Seidlein, L., Coleman, R., Deen, J., Pinder, M., Doherty, T., Sutherland, C., Walraven, G. and Milligan, P. (2001). Artesunate reduces but does not prevent posttreatment transmission of plasmodium falciparum to anopheles gambiae, *Journal of Infectious Diseases* **183**: 1254–1259.

Therneau, T. M., Grambsch, P. M. and Fleming, T. (1990). Martingale-based residuals for survival models, *Biometrika* **77**: 147–160.

Trape, J. F. (2001). The public health impact of chloroquine resistance in africa, *American Journal of Tropical Medicine and Hygiene* **64**(Supplement): 1217.

von Seidlein, L., Drakeley, C., Greenwood, B., Walraven, G. and Targett, G. (2001). Risk factors for gametocyte carriage in Gambian children., *The American Journal of Tropical Medicine and Hygiene* **65**(5): 523–527.
URL: *http://www.ajtmh.org/content/journals/10.4269/ajtmh.2001.65.523*

Wang, Y. and Taylor, J. M. G. (2001). Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome, *Journal of the American Statistical Association* **96**(455): 895–905.

White, N. (1997). Assessment of the pharmacodynamic properties of antimalarial drugs in vivo., *Antimicrobial agents and chemotherapy* **41**(7): 1413.

White, N. J. (2004). Antimalarial drug resistance, *The Journal of Clinical Investigation* **113**(8): 10841092.

WHO (2011). *Haemoglobin concentrations for the diagnosis of anaemia and assessment of severity*, Geneva, Switzerland.
URL: *http://www.who.int/vmnis/indicators/haemoglobin.pdf.*

WHO (2013). *Emergency response to artemisinin resistance in the Greater Mekong subregion: regional framework for action 20132015*, Geneva: World Health Organization.

WHO (2015). *Guidelines for the treatment of malaria.*
URL: *www.who.int/malaria/publications/atoz/9789241549127/en/*

WHO (2017a). Antimalarial drug resistance in the Greater Mekong subregion: How concerned should we be?
URL: *http://www.who.int/malaria/media/drug-resistance-greater-mekong-qa/en/*

WHO (2017b). *World malaria report 2017*, World Health Organization.

Younes, N. and Lachin, J. (1997). Link-based models for survival data with interval and continuous time censoring, *Biometrics* **53**: 1199–1211.

6 Appendix

6.1 Modified critical exponential nonlinear mixed effects model used for modeling the longitudinal gametocyte profile: JAGS code

```
model{
for(i in 1:N)
{
c[i] ~dnorm(0.0,tauc)
r[i] ~dnorm(0.0,taur)

C[i]<- betaC0+ (betaC1 * trt[i]) + (betaC2 * ratio[i]) + (betaC3 * anaemia[i])
+(betaC4 * mut5[i]) + (betaC5 * lage[i]) +(pzero[i]*betaC6)+ (sex[i]*betaC7)   + c[i]

R[i]<- betaR0+ (betaR1 * trt[i]) + (betaR2 * ratio[i]) + (betaR3 * anaemia[i])
+(betaR4 * mut5[i]) + (betaR5 * lage[i]) + (pzero[i]*betaR6) + (sex[i]*betaR7) + r[i]

   for(j in 1:M)
{
y[i,j] ~dnorm(eta[i,j],taue)

eta[i,j] <- ((C[i]* pday[i,j] ) * pow(R[i],pday[i,j]) )
}
}
tauc ~dgamma(0.001,0.001)
taur ~dgamma(0.001,0.001)
taue~dgamma(0.001,0.001)

betaR0 ~dnorm(0.0,1.0E-4)
betaR1 ~dnorm(0.0,1.0E-4)
betaR2 ~dnorm(0.0,1.0E-4)
betaR3 ~dnorm(0.0,1.0E-4)
betaR4 ~dnorm(0.0,1.0E-4)
betaR5 ~dnorm(0.0,1.0E-4)
betaR6 ~dnorm(0.0,1.0E-4)
betaR7 ~dnorm(0.0,1.0E-4)

betaC0 ~dnorm(0.0,1.0E-4)
```

```
betaC1 ~dnorm(0.0,1.0E-4)
betaC2 ~dnorm(0.0,1.0E-4)
betaC3 ~dnorm(0.0,1.0E-4)
betaC4 ~dnorm(0.0,1.0E-4)
betaC5 ~dnorm(0.0,1.0E-4)
betaC6 ~dnorm(0.0,1.0E-4)
betaC7 ~dnorm(0.0,1.0E-4)
}
```

6.2 Weibull cause-specific hazards survival model: JAGS code

```
model{
for(i in 1:N)
{

######
#########Survival
######
#Failure
tF[i]~dweib(wF,muF[i])

censoredF[i] ~ dinterval(tF[i], cenF[i])

log(muF[i])<-betaF0+(trt[i]*betaF1)+(ratio[i]*betaF2)+(anaemia[i]*betaF3)
+(mut5[i]*betaF4)+(lage[i]*betaF5)+ (pzero[i]*betaF6)+ (sex[i]*betaF7)

#LTFU
tL[i]~dweib(wL,muL[i])

censoredL[i] ~ dinterval(tL[i], cenL[i])

log(muL[i])<-betaL0+(trt[i]*betaL1)+(ratio[i]*betaL2)+(anaemia[i]*betaL3)
+(mut5[i]*betaL4)+(lage[i]*betaL5)+ (pzero[i]*betaL6)+ (sex[i]*betaL7)

}

wL~dunif(0,100)
wF~dunif(0,100)
```

```
betaF0 ~dnorm(0.0,1.0E-4)
betaF1 ~dnorm(0.0,1.0E-4)
betaF2 ~dnorm(0.0,1.0E-4)
betaF3 ~dnorm(0.0,1.0E-4)
betaF4 ~dnorm(0.0,1.0E-4)
betaF5 ~dnorm(0.0,1.0E-4)
betaF6 ~dnorm(0.0,1.0E-4)
betaF7 ~dnorm(0.0,1.0E-4)

betaL0 ~dnorm(0.0,1.0E-4)
betaL1 ~dnorm(0.0,1.0E-4)
betaL2 ~dnorm(0.0,1.0E-4)
betaL3 ~dnorm(0.0,1.0E-4)
betaL4 ~dnorm(0.0,1.0E-4)
betaL5 ~dnorm(0.0,1.0E-4)
betaL6 ~dnorm(0.0,1.0E-4)
betaL7 ~dnorm(0.0,1.0E-4)
}
```

6.3 Normally distributed Joint model, with Weibull cause-specific hazard survival component model, used for modeling the longitudinal gametocyte profile: JAGS code

```
model{
for(i in 1:N)
{
################################
#########Survival component
################################
#Failure
tF[i]~dweib(wF,muF[i])

censoredF[i] ~ dinterval(tF[i], cenF[i])

log(muF[i])<-betaF0+(trt[i]*betaF1)+(ratio[i]*betaF2)+(anaemia[i]*betaF3)+(mut5[i]*betaF4)
+(lage[i]*betaF5)+ (pzero[i]*betaF6)+ (sex[i]*betaF7)  +(cf*c[i]) + (rf* r[i])
```

```
#LTFU
tL[i]~dweib(wL,muL[i])

censoredL[i] ~ dinterval(tL[i], cenL[i])

log(muL[i])<-betaL0+(trt[i]*betaL1)+(ratio[i]*betaL2)+(anaemia[i]*betaL3)+(mut5[i]*betaL4)
+(lage[i]*betaL5)+ (pzero[i]*betaL6)+ (sex[i]*betaL7)  +(cl*c[i]) + (rl* r[i])
################################
#########Longitudinal component
################################
c[i] ~dnorm(0.0,tauc)
r[i] ~dnorm(0.0,taur)

C[i]<- betaC0+ (betaC1 * trt[i]) + (betaC2 * ratio[i]) + (betaC3 * anaemia[i]) +
(betaC4 * mut5[i]) + (betaC5 * lage[i]) +(pzero[i]*betaC6)+ (sex[i]*betaC7)   + c[i]

R[i]<- betaR0+ (betaR1 * trt[i]) + (betaR2 * ratio[i]) + (betaR3 * anaemia[i]) +
(betaR4 * mut5[i]) + (betaR5 * lage[i]) + (pzero[i]*betaR6) + (sex[i]*betaR7) + r[i]

   for(j in 1:M)
{
y[i,j] ~dnorm(eta[i,j],taue)

eta[i,j] <- ((C[i]* pday[i,j] ) * pow(R[i],pday[i,j]) )
}
}
tauc ~dgamma(0.001,0.001)
taur~dgamma(0.001,0.001)

taue~dgamma(0.001,0.001)

cf~dnorm(0.0,1.0E-4)
rf~dnorm(0.0,1.0E-4)
cl~dnorm(0.0,1.0E-4)
rl~dnorm(0.0,1.0E-4)

wL~dunif(0,100)
```

```
wF~dunif(0,100)

betaF0 ~dnorm(0.0,1.0E-4)
betaF1 ~dnorm(0.0,1.0E-4)
betaF2 ~dnorm(0.0,1.0E-4)
betaF3 ~dnorm(0.0,1.0E-4)
betaF4 ~dnorm(0.0,1.0E-4)
betaF5 ~dnorm(0.0,1.0E-4)
betaF6 ~dnorm(0.0,1.0E-4)
betaF7 ~dnorm(0.0,1.0E-4)

betaL0 ~dnorm(0.0,1.0E-4)
betaL1 ~dnorm(0.0,1.0E-4)
betaL2 ~dnorm(0.0,1.0E-4)
betaL3 ~dnorm(0.0,1.0E-4)
betaL4 ~dnorm(0.0,1.0E-4)
betaL5 ~dnorm(0.0,1.0E-4)
betaL6 ~dnorm(0.0,1.0E-4)
betaL7 ~dnorm(0.0,1.0E-4)

betaR0 ~dnorm(0.0,1.0E-4)
betaR1 ~dnorm(0.0,1.0E-4)
betaR2 ~dnorm(0.0,1.0E-4)
betaR3 ~dnorm(0.0,1.0E-4)
betaR4 ~dnorm(0.0,1.0E-4)
betaR5 ~dnorm(0.0,1.0E-4)
betaR6 ~dnorm(0.0,1.0E-4)
betaR7 ~dnorm(0.0,1.0E-4)

betaC0 ~dnorm(0.0,1.0E-4)
betaC1 ~dnorm(0.0,1.0E-4)
betaC2 ~dnorm(0.0,1.0E-4)
betaC3 ~dnorm(0.0,1.0E-4)
betaC4 ~dnorm(0.0,1.0E-4)
betaC5 ~dnorm(0.0,1.0E-4)
betaC6 ~dnorm(0.0,1.0E-4)
betaC7 ~dnorm(0.0,1.0E-4)
}
```

6.4 Zero-adjusted gamma Joint model, with Weibull cause-specific hazard survival component model, used for modeling the longitudinal gametocyte profile: JAGS code

```
model{K<-1000
for(i in 1:N)
{
################################
#########Survival component
################################

#Treatment Failure cause of exit
tF[i]~dweib(wF,muF[i])
censoredF[i] ~ dinterval(tF[i], cenF[i])
log(muF[i])<-betaF0+(trt[i]*betaF1)+(ratio[i]*betaF2)+(anaemia[i]*betaF3)
+(mut5[i]*betaF4)+(lage[i]*betaF5)+  (pzero[i]*betaF6)+ (sex[i]*betaF7)
+ (cf*c[i])+ (rf*r[i]) + (brf* br[i]) + (bcf* bc[i])

#LTFU cause of exit
tL[i]~dweib(wL,muL[i])
censoredL[i] ~ dinterval(tL[i], cenL[i])
log(muL[i])<-betaL0+(trt[i]*betaL1)+(ratio[i]*betaL2)+(anaemia[i]*betaL3)
+(mut5[i]*betaL4)+(lage[i]*betaL5) + (pzero[i]*betaL6)
+ (sex[i]*betaL7)+ (cl*c[i])+(rl*r[i]) + (brl* br[i]) + (bcl* bc[i])

################################
#########Longitudinal component
################################

####gamma component
c[i] ~dnorm(0.0,tauc)
r[i] ~dnorm(0.0,taur)

A[i]<- betaA0  + (betaA1 * trt[i]) + (betaA2 * ratio[i]) + (betaA3 * anaemia[i])
```

```
+ (betaA4 * mut5[i]) + (betaA5 * lage[i])+ (betaA6 * pzero[i])+ (sex[i]*betaA7)

C[i]<- betaC0  + (betaC1 * trt[i]) + (betaC2 * ratio[i]) + (betaC3 * anaemia[i])
+ (betaC4 * mut5[i]) + (betaC5 * lage[i])+ (betaC6 * pzero[i])+ (sex[i]*betaC7)  + c[i]

R[i]<- betaR0  + (betaR1 * trt[i]) + (betaR2 * ratio[i]) + (betaR3 * anaemia[i])
+ (betaR4 * mut5[i]) + (betaR5 * lage[i]) + (betaR6 * pzero[i])+ (sex[i]*betaR7) + r[i]

####prevalence component
br[i] ~dnorm(0.0,taubr)
bc[i] ~dnorm(0.0,taubc)

bA[i]<- betaBA0  + (betaBA1 * trt[i]) + (betaBA2 * ratio[i]) + (betaBA3 * anaemia[i])
 + (betaBA4 * mut5[i]) + (betaBA5 * lage[i])+ (betaBA6 * pzero[i])+ (sex[i]*betaBA7)

bC[i]<- betaBC0  + (betaBC1 * trt[i]) + (betaBC2 * ratio[i]) + (betaBC3 * anaemia[i])
+ (betaBC4 * mut5[i]) + (betaBC5 * lage[i])+ (betaBC6 * pzero[i])+(sex[i]*betaBC7)+ bc[i]

bR[i]<- betaBR0  + (betaBR1 * trt[i]) + (betaBR2 * ratio[i]) + (betaBR3 * anaemia[i])
+ (betaBR4 * mut5[i]) + (betaBR5 * lage[i])+(betaBR6 * pzero[i])+(sex[i]*betaBR7)+br[i]

################################
#########Likelihood function
################################

   for(j in 1:M)
{
#prevalence
pres[i,j] ~dbern(PI[i,j])

logit(PI[i,j]) <- bA[i] + ((bC[i]* pday[i,j] ) * pow(bR[i],pday[i,j]) )

#gamma
log(eta[i,j]) <- A[i]+ ((C[i]* pday[i,j] ) * pow(R[i],pday[i,j]) )

logGamma[i,j]<-log(dgamma(y[i,j],taue, (taue/eta[i,j])))

d[i,j] <- pres[i,j]
```

```
#mixture likelihood
l[i,j] <- ((1-d[i,j])* log(1-PI[i,j])) + (d[i,j] * (log(PI[i,j]) + logGamma[i,j]))

phi[i,j]<-  K - l[i,j]

zeros[i,j]~dpois(phi[i,j])
}
}
wL~dunif(0,100)
wF~dunif(0,100)

tauc ~dgamma(0.0001,0.0001)
taur ~dgamma(0.0001,0.0001)

taue~dgamma(0.0001,0.0001)

taubr ~dgamma(0.0001,0.0001)
taubc ~dgamma(0.0001,0.0001)

rf~dnorm(0.0,1.0E-4)
cf~dnorm(0.0,1.0E-4)
brf~dnorm(0.0,1.0E-4)
bcf~dnorm(0.0,1.0E-4)

rl~dnorm(0.0,1.0E-4)
cl~dnorm(0.0,1.0E-4)
brl~dnorm(0.0,1.0E-4)
bcl~dnorm(0.0,1.0E-4)

betaL0 ~dnorm(0.0,1.0E-4)
betaL1 ~dnorm(0.0,1.0E-4)
betaL2 ~dnorm(0.0,1.0E-4)
betaL3 ~dnorm(0.0,1.0E-4)
betaL4 ~dnorm(0.0,1.0E-4)
betaL5 ~dnorm(0.0,1.0E-4)
betaL6 ~dnorm(0.0,1.0E-4)
betaL7 ~dnorm(0.0,1.0E-4)
```

```
betaF0 ~dnorm(0.0,1.0E-4)
betaF1 ~dnorm(0.0,1.0E-4)
betaF2 ~dnorm(0.0,1.0E-4)
betaF3 ~dnorm(0.0,1.0E-4)
betaF4 ~dnorm(0.0,1.0E-4)
betaF5 ~dnorm(0.0,1.0E-4)
betaF6 ~dnorm(0.0,1.0E-4)
betaF7 ~dnorm(0.0,1.0E-4)


betaA0 ~dnorm(0.0,1.0E-4)
betaA1 ~dnorm(0.0,1.0E-4)
betaA2 ~dnorm(0.0,1.0E-4)
betaA3 ~dnorm(0.0,1.0E-4)
betaA4 ~dnorm(0.0,1.0E-4)
betaA5 ~dnorm(0.0,1.0E-4)
betaA6 ~dnorm(0.0,1.0E-4)
betaA7 ~dnorm(0.0,1.0E-4)

betaC0 ~dnorm(0.0,1.0E-4)
betaC1 ~dnorm(0.0,1.0E-4)
betaC2 ~dnorm(0.0,1.0E-4)
betaC3 ~dnorm(0.0,1.0E-4)
betaC4 ~dnorm(0.0,1.0E-4)
betaC5 ~dnorm(0.0,1.0E-4)
betaC6 ~dnorm(0.0,1.0E-4)
betaC7 ~dnorm(0.0,1.0E-4)

betaR0 ~dnorm(0.0,1.0E-4)
betaR1 ~dnorm(0.0,1.0E-4)
betaR2 ~dnorm(0.0,1.0E-4)
betaR3 ~dnorm(0.0,1.0E-4)
betaR4 ~dnorm(0.0,1.0E-4)
betaR5 ~dnorm(0.0,1.0E-4)
betaR6 ~dnorm(0.0,1.0E-4)
betaR7 ~dnorm(0.0,1.0E-4)
```

```
betaBA0 ~dnorm(0.0,1.0E-4)
betaBA1 ~dnorm(0.0,1.0E-4)
betaBA2 ~dnorm(0.0,1.0E-4)
betaBA3 ~dnorm(0.0,1.0E-4)
betaBA4 ~dnorm(0.0,1.0E-4)
betaBA5 ~dnorm(0.0,1.0E-4)
betaBA6 ~dnorm(0.0,1.0E-4)
betaBA7 ~dnorm(0.0,1.0E-4)


betaBC0 ~dnorm(0.0,1.0E-4)
betaBC1 ~dnorm(0.0,1.0E-4)
betaBC2 ~dnorm(0.0,1.0E-4)
betaBC3 ~dnorm(0.0,1.0E-4)
betaBC4 ~dnorm(0.0,1.0E-4)
betaBC5 ~dnorm(0.0,1.0E-4)
betaBC6 ~dnorm(0.0,1.0E-4)
betaBC7 ~dnorm(0.0,1.0E-4)

betaBR0 ~dnorm(0.0,1.0E-4)
betaBR1 ~dnorm(0.0,1.0E-4)
betaBR2 ~dnorm(0.0,1.0E-4)
betaBR3 ~dnorm(0.0,1.0E-4)
betaBR4 ~dnorm(0.0,1.0E-4)
betaBR5 ~dnorm(0.0,1.0E-4)
betaBR6 ~dnorm(0.0,1.0E-4)
betaBR7 ~dnorm(0.0,1.0E-4)


#prior distributions for the missing response values excluded in the above extract
}
```