



UNIVERSITY OF CAPE TOWN

STA5079W

MINOR DISSERTATION

Modelling Gametocytes in the Presence of Interval-Censoring

Author:
Michaela Faro C. Takawira

Student Number:
TKWMIC001

January 20, 2022

The copyright of this thesis vests in the author. No quotation from it or information derived from it is to be published without full acknowledgement of the source. The thesis is to be used for private study or non-commercial research purposes only.

Published by the University of Cape Town (UCT) in terms of the non-exclusive license granted to UCT by the author.

Abstract

Malaria is a parasitic disease that has afflicted many over the years, with *Plasmodium falciparum* malaria accounting for many deaths. Gametocytes are the sexual form of the parasite responsible for transmitting and spreading malaria from the human host to the mosquito vector. Most studies are designed to measure asexual parasites and hence the measurement intervals are not optimal for measuring gametocytes.

The data analysed was obtained from a series of clinical trials conducted between 2002 and 2004 in Mpumalanga, South Africa and Mozambique. As part of the South-East African Combination Antimalarial Therapy (SEACAT) evaluation of the phased introduction of combination antimalarial drug under the Lubombo Spatial Development Initiative. Patients were observed on days 0, 3, 7, 14, 21, 28 and 42, where blood samples were collected and were analysed, providing gametocyte densities and other information.

Due to the study design, observing gametocyte profiles was complex due to censoring. Censoring occurs when an event has not been observed, and hence the event time is only partially known. Censoring occurs for various reasons, which include the following. In some instances, a patient may enter a study with gametocytes circulating in their system before enrolling into the study. The time in which gametocytes have emerged is therefore unknown. Under such circumstances, this type of censoring is known as left censoring. Another example of censoring occurs when a patient is lost-to-follow-up or leaves the study before the end of the study period due to treatment failure leaving the gametocyte profile of the patient incomplete, resulting in right censoring. Lastly, besides left and right censoring, the gametocyte data has been affected by interval-censoring. Patients are observed and monitored on specific days; thus, the actual moment of observation of gametocyte emergence or clearance is estimated to have occurred between two observation days. The gametocyte data is thus also characterized by interval-censoring.

Given these interesting characteristics of the gametocyte data, this research aims to directly model gametocytes while taking into account censoring. Researchers often opt to assume the times to events of interest are at the moment of observation (right endpoint of an interval) or midpoint of the interval in which the event is assumed to have occurred. In this research, we also investigate and discuss the impact of ignoring the interval-censored mechanism present in the data and how parameter estimates based on these ad-hoc approaches might differ from interval-censored results. Following the above discussion, the research will then apply interval-censored techniques to the data.

Several survival analysis models were applied to the gametocyte data. These models included the Cox Proportional Hazards (PH) model, parametric PH model and accelerated failure time models, which were used to illustrate how results may differ based on whether interval-censoring was taken into account or not. From the analysis, it was observed that the midpoint imputation was a better proxy for interval-censoring compared to the right-imputation method and that ignoring interval censoring had more of an impact on events that occurred during the wider intervals towards the latter part of the observation period. From the clinical perspective, it was found that younger patients who had high levels of

baseline asexual parasitemia, quintuple mutations and used sulfadoxine-pyrimethamine were estimated to experience the most prolonged duration of gametocytemia. It was also found that age, treatment and baseline asexual parasitemia influenced whether a patient would develop gametocytes.

Acknowledgements

I would like to thank Professor Francesca Little for being awesome! I will always appreciate you for being so kind and patient with me. You held my hand and guided me (sometimes pushed me) through the course of this dissertation while we were in the middle of a pandemic and because of that I am truly grateful. I would also like to thank the entire staff in the Statistical Sciences department you all played some role in my master's journey whether it was a smile, a "Hi" or a conversation in the tea room. To the future Dr. Stefan Britz, thank you for being my therapist, when the world got a bit too loud, I knew I could count on you to help me refocus and find my inner strength. Prof. Scott I wish words were enough to express how much I appreciate you but for now they will have to do. Dr. Er thank you for believing in me and boosting my confidence when you saw I needed it. Kimmy Kim Kim, you were our resident fashion icon in the department thank you for the laughs, fashion tips and being a big sister when I was in need of one! Beverley King thank you for being my female knight in shining armor whenever I needed assistance you were right there. Kevin, thank you for all coffees and teas you made for me just to cheer me up, the running around that you did for me in the background and for introducing me to the RB group The Whispers! Celene since my first year at UCT until now you've always had your arms wide open to carry me when I thought I couldn't go any further. I cannot thank you enough for being by my side throughout all these years you will always hold a special place in my heart!

To all my friends, I made it this far because of you all. You were always a phone call, an uber or a message away. If there are people who truly witnessed all the tears and joy that came with this journey it was you all. Thank you for being there. Lily and Zan my two sisters from other misters my God-sent angels on earth! Even though we meet years ago, God knew I would need you for such a time as this thank you. Lihle my safe-house, when I got overwhelmed by the journey, I knew it was ok to lean/fall on you because you were right there to catch me and help me back up. Tanaka, I do not know where to begin to thank you but thank you for reminding me what it means to be alive. Hassan at first we did not get along but now I cannot imagine life without oga. Kuda and Tavonga you two taught me how to push through the pain thank you for never letting me get "too comfortable" or stagnant.

To my biological and adopted family thank you for holding me together and carrying me over the finish line. Carrie a.k.a Sweet Caroline thank you for taking your position as my second mummy I appreciate you dearly. Olufunke! My answered prayer, thank you, just thank you for accepting me as one of your own, you did not have to but because of your big heart you did and I will forever be grateful. Ore a.k.a Auntie Biscuit may you continue to "Bounce in the Lord" always, and thank you for keeping me laughing, I have a bit of ab definition because of you. Ife my fellow partner in almost everything, thank you for the free counselling sessions also, please do not forget to credit me on Apple music as a writer even if I contributed two words to your song (a contribution is a contribution)! Tofunmi a.k.a my twin you always saw the best in me thank you for never leaving my

side, next stop the moon! To my mom, dad and my precious sisters. Thank you for all the sacrifices, prayers (Laughing out loud - I know there was a lot of prayer involved) and hard decisions you made to make sure I got to this point! Where would I be if it were not for you all. God truly outdid Himself when He blessed me with all four of you. I love you guys dearly.

Last but certainly not least I want to thank God for bringing me this far, for bringing all these wonderful people onto my journey, making sure every need I had was supplied for and keeping Your promises to me! Years ago You told me that You had great plans for me, plans to prosper me and not to harm, so I just want to thank you for not disappointing me and staying true to Your word!

Contents

Abstract	i
Acknowledgements	iii
List of Tables	vii
List of Figures	viii
1 Introduction	1
1.1 Background	1
1.2 Study design	3
1.3 Past contributions	4
1.4 Objectives	6
1.5 Data overview	7
1.5.1 Prevalence of gametocytes	9
1.5.2 Association with treatment arm	15
1.5.3 Individual gametocyte profiles	17
2 Censoring	19
2.1 Right censoring	19
2.2 Left censoring	20
2.3 Interval-censoring	22
2.4 Doubly interval-censoring	23
2.5 Left truncation	24
2.6 Assumptions of censoring	25
2.7 Mechanisms for dealing with censoring	26
2.8 Notation and transforming the data	27
3 Survival curves	28
3.1 Methodology	29
3.1.1 Kaplan-Meier curve	29
3.1.2 Non-parametric maximum likelihood estimator	30
3.2 Results	33
3.2.1 Time to emergence	33
3.2.2 Time to elimination	36
3.3 Comparison of survival distributions	39
3.3.1 Logrank tests for right-censored data	39
3.3.2 Weighted logrank tests for interval-censored data	41
3.3.3 Weighted logrank significance testing	43
3.4 Results	44
3.4.1 Time to emergence	44
3.4.2 Time to elimination	47

4	Models for survival data	51
4.1	Semi-parametric Proportional Hazard models	51
4.1.1	The Cox Proportional Hazards model	51
4.1.2	Interval-censored semi-parametric PH model	52
4.1.3	Results	53
4.2	Parametric PH models	55
4.2.1	Regression models	56
4.2.2	Results	58
4.3	Accelerated failure time models	60
4.3.1	Right censored accelerated failure time models	60
4.3.2	Accelerated failure time models for interval-censored data	61
4.3.3	Results	62
5	Building survival models for interval-censored data	64
5.1	Model fitting approach	64
5.1.1	Time to gametocyte emergence	66
5.1.2	Time to gametocyte clearance	73
5.1.3	Duration	77
5.2	Doubly interval-censored data	79
5.2.1	Methodology	79
5.2.2	Results	80
6	Discussion	86
6.1	Methodological discussion	86
6.2	Clinical discussion	88
6.3	Concluding remarks	90
	References	92
A	Appendices	97
A.1	Libraries and data used	97
A.2	Data imported	98
A.3	Descriptive statistics	99
A.4	Standard survival analysis	106
A.5	Midpoint censoring	109
A.6	Interval-censored approach	112
A.7	Building up the full model	117
A.8	Doubly censored	125

List of Tables

1	Covariates	8
2	Treatment Outcomes	11
3	Gametocyte prevalence descriptive statistics	12
4	Descriptive statistics of covariates per treatment group	15
5	Censoring notation and interval specifications	27
6	Regions of mass support for time to emergence	34
7	Survival rate estimates for gametocyte emergence	34
8	Regions of mass support for time to elimination	36
9	Survival rate estimates for gametocyte elimination	37
10	Observed data at time t_j to compute the logrank test	39
11	Regions of mass support for gametocyte emergence per treatment group . .	45
12	Survival rate estimates for patients based on treatment	45
13	Logrank and Wilcoxon Tests and Statistics for gametocyte emergence by treatment arm	47
14	Regions of mass support for gametocyte elimination per treatment group .	48
15	Survival rate estimates for gametocyte elimination for patients based on treatment	48
16	Logrank and Wilcoxon Tests and Statistics for gametocyte elimination by treatment arm	50
17	Gametocyte Emergence semi-parametric PH Models	54
18	Gametocyte Elimination semi-parametric PH Models	54
19	Exponential and Weibull underlying survival functions	56
20	Gametocyte Emergence Parametric PH Models	59
21	Gametocyte Elimination AFT Models	62
22	Duration estimates and standard errors based on treatment arm	63
23	Parameter estimates and 95% C.I. for time to emergence PH models	68
24	Parameter estimates and 95% C.I. for nested time to emergence Cox PH models	69
25	Revised Parameter estimates and 95% C.I. for time to emergence PH models models	73
26	Parameter estimates and 95% C.I. for time to clearance AFT models	74
27	Parameter estimates and 95% C.I. for nested time to clearance AFT models	76
28	Estimated Duration of gametocytemia and Standard Errors	78
29	Comparison of Parameter estimates of the Log-normal AFT models	81
30	Estimated Duration of gametocytemia and Standard Errors based on the Doubly interval-censored, Midpoint and Singly interval-censored approaches	84

List of Figures

1	Plasmodium falciparum life cycle as depicted by Bousema and Drakeley (2011)	2
2	Gametocyte prevalence by day by treatment group	10
3	Boxplots for <i>logage</i> and <i>pzero</i> for by gametocyte prevalence	13
4	Mean gametocyte profiles across the categorical covariates	14
5	Boxplots for <i>logage</i> and <i>pzero</i> covariates by treatment arm	16
6	Gametocyte Profile spaghetti plots	18
7	Right censoring examples	20
8	Left censoring examples	21
9	Examples of point censored observations	22
10	Interval-censoring examples	23
11	Doubly interval-censored example	24
12	Left Truncation example	25
13	Example of Turnbull's intervals: Suppose events take place within the intervals (0,7],(3,21] and (14,28]. The Turnbull intervals in this case would be (3,7] and (14,21].	32
14	K-M Emergence Survival Curve	35
15	K-M Elimination Survival Curve	38
16	Comparison of gametocyte emergence survival curves based on treatment group	46
17	Comparison of gametocyte elimination survival curves based on treatment group	49
18	Examining Baseline Weibull and Exponential Distribution Fits	67
19	PH Assumption Validity	71
20	PH Assumption Validity with interaction	72
21	All Gametocyte Profile Spaghetti Plots	131

1 Introduction

1.1 Background

Malaria is a parasitic disease that has been responsible for the deaths of many over the past decades and remains a burden to the public population to this day. The malaria scourge continues to afflict many of those, particularly in Africa, Asia and parts of Latin America. In 2018 alone, 228 million cases of malaria were reported (WHO, 2019). About 405 000 of those cases resulted in death (WHO, 2019) and 94 per cent of those deaths came from the WHO African Region (WHO, 2019). Malaria is arguably still one of the largest contributors to the death toll in Africa. With that being said, the fight against malaria over the past few years has seen annual malaria-related deaths since 2000 being halved (WHO, 2015*b*). In the year 2000, 839,000 deaths were reported which reduced substantially to 438,000 deaths in 2015 (Roser and Ritchie, 2019). However, in the past few years, the global incidence rate of malaria, which refers to the number of malaria cases per 1000 population, has reduced from 71 to 57 between 2010 and 2018 (WHO, 2019). Given these numbers, it appears that the reduction in malaria cases is no longer as drastic as it once was. However, this does not mean complete global eradication is not possible. But it does call our attention to focus on improving malaria elimination and strategies that prevent malaria transmission with the hopes of controlling malaria better. One of those strategies involves interfering with the parasite transmission cycle between the mosquito and human host (Karunajeewa and Mueller, 2016). To undermine this transmission cycle, it would require a better awareness and understanding of how the malaria transmission cycle works.

Gametocytes are the sexual form of the *Plasmodium falciparum* parasite and are responsible for the transmission of the malaria parasite (Bousema and Drakeley, 2011). Gametocytes are different from their asexual precursor of the parasite (Bousema and Drakeley, 2011), which is responsible for the outward signs and symptoms of malaria. The malaria parasite is introduced into the human bloodstream by the female *Anopheles* mosquito as they take their blood meal (Bousema and Drakeley, 2011). As the female mosquito withdraws blood, she simultaneously injects sporozoites (this is the form of the malaria parasite that invades the liver) into the human host (Bousema and Drakeley (2011), Michalakis and Renaud (2009)). These sporozoites then travel to the liver, where they multiply asexually before becoming merozoites (Michalakis and Renaud, 2009). The merozoites (which are also a form of the malaria parasite) then enter into the bloodstream, where they infect the red blood cells, this process becomes a repetitive asexual cycle (Bousema and Drakeley, 2011). Eventually, the host starts to experience symptoms of malaria, such as chills and fever (Michalakis and Renaud, 2009). While some of the merozoites continue in the asexual cycle, a portion of the merozoites released from the red blood cells they infected form male or female gametocytes (Barnes and White, 2005).

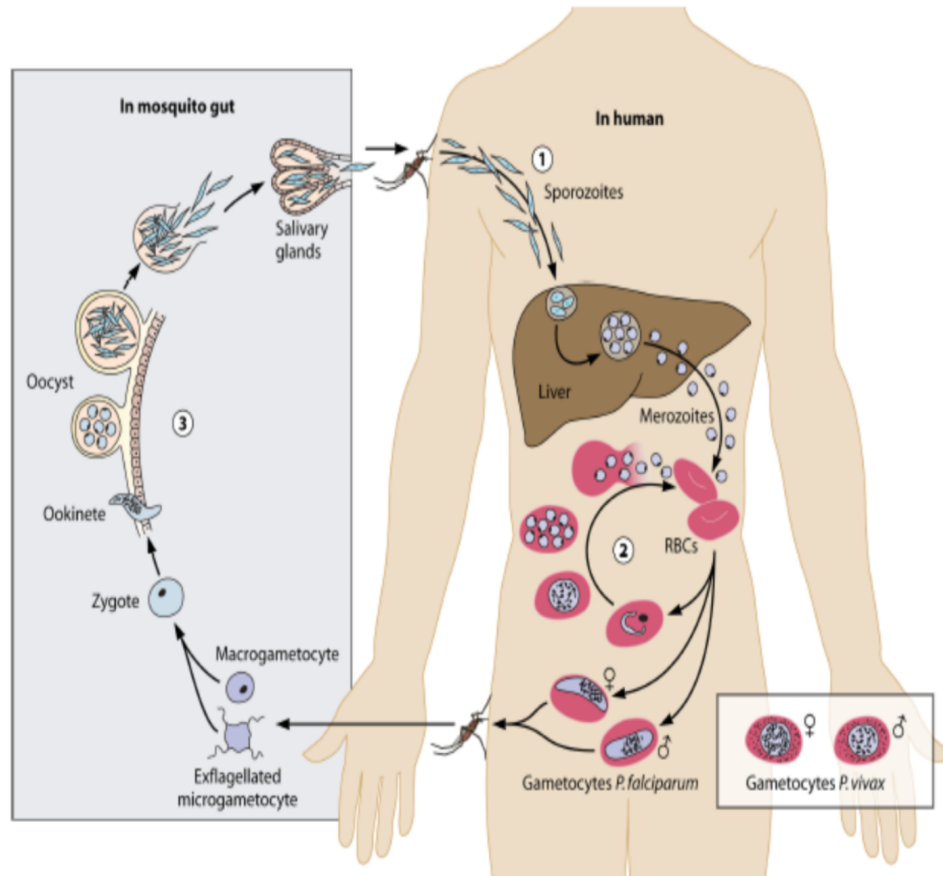


Figure 1: *Plasmodium falciparum* life cycle as depicted by Bousema and Drakeley (2011)

The gametocyte development and maturation phase are divided into five morphologically distinct stages (Bousema and Drakeley, 2011). Using the text from Bousema and Drakeley (2011) the development of gametocytes and *Plasmodium falciparum* life cycle are summarized as follows. The first stage of gametocyte development begins with the sequestration of gametocytes, the stages that follow see gametocytes developing till they reach maturity, which is the final stage. Only then are gametocytes transmissible to the mosquitoes who ingest these gametocytes that circulate within the host's peripheral blood. Each gametocyte is transmitted back to the Anopheles mosquito during their feeding time in the form of one female macrogamete or not more than eight male microgametes. Within the mosquito's midgut, the gametes fuse together, resulting in the formation of a zygote. The zygote then matures into a motile ookinete that can break through the midgut wall to form oocysts. Over time, these oocysts divide and burst out of the mosquito's midgut to create sporozoites that then enter the mosquito's salivary gland. The mosquito is now a carrier of the malaria parasite, making it infectious to the human hosts and starting the infection cycle again. This cycle is illustrated in Figure 1.

To disrupt the parasite transmission cycle as suggested by Karunajeewa and Mueller (2016), we could focus on targeting gametocytes (the sexual parasites responsible for transmission) which will help in the control and elimination of malaria (Bousema et al., 2010). However, for us to achieve this, methods of analyzing data related to gametocytes must be improved to define risk factors that determine the extent of gametocyte carriage better; host susceptibility; and develop treatments/measures to eradicate gametocytes. This dissertation focuses on analyzing longitudinal data, which includes gametocyte densities from multi-centre clinical trials, using survival analysis techniques. The rest of this chapter is structured as follows, first, a small discussion on study design and some of the problems that come with this particular design - followed by a discussion of previous works done in this field of research. The chapter then closes with a description and exploration of the data used.

1.2 Study design

Before analyzing gametocytemia (the presence of gametocytes) distributions, it is important to consider the study design. How a study is designed potentially impacts the quality of results obtained and usefulness of the study. A study designed poorly cannot be fixed but a study analysed poorly can be reanalyzed (Campbell and Machin, 1999). And in this case, the study design is just as important as in any other research. The interesting feature of some malaria intervention trials is that they are designed to observe and focus on eliminating the asexual parasite as opposed to gametocyte profiles (Distiller et al., 2010). Based on such a design, measurement of gametocytemia is therefore not optimal, seeing as the gametocyte emergence lags that of the asexual parasite emergence (Distiller et al., 2010).

The present longitudinal studies collected asexual and sexual parasite density data for patients at ever-increasing intervals up until the last day of the trial or day of withdrawal (due to treatment failure or loss to follow up). However, observing gametocyte profiles over time may not always be possible due to censoring.

Censoring in the context of survival analysis is when an event time is only partially unknown. Censoring happens for various reasons. For one, Distiller et al. (2010) gave an example where a patient could be withdrawn from the study because of treatment failure due to asexual parasites failing to clear or parasites re-emerging. The presence of asexual parasites that have not cleared yet end up becoming a competing event to measuring gametocytemia (Distiller et al., 2010). This results in incomplete gametocyte distributions for patients who cannot succumb to anti-malarial drugs and are thus withdrawn from the study due to treatment failure (Distiller et al., 2010). This is known as informative censoring. Other reasons for censoring due to study design within the present context include: left censoring where some patients enter into the study with gametocytes; right censoring, which occurs when patients withdraw from the study before gametocytes clear and lastly, interval-censoring, which occurs when the exact date on which gametocytes enter the patient or clear out of the patient's system is only partially known, in that the time interval when the event happens is known, but the exact date is not. More details of

the study design underlying the data will be discussed in section 1.4.

1.3 Past contributions

Previous researchers have contributed greatly towards modelling gametocytes, and two such researchers include Distiller et al. (2010) and Chatora (2018) who have previously analyzed the gametocyte profiles from this data. Both researchers focus on the imputation of incomplete gametocyte profiles through nonlinear mixed effect modelling. Distiller et al. (2010) found that the gametocyte density time profiles were generally nonlinear and characterized by intra-individual variability and inter-individual variability. For these reasons, nonlinear mixed-effects modelling was the appropriate statistical methodology (Distiller et al., 2010). Distiller et al. (2010) then went on to compare six distinguishable nonlinear functions, one and two-compartment models, a critical exponential model, a reduced critical exponential model, and two wave-like functions, Fourier and double Fourier. In the end, the reduced form of the critical exponential function was chosen as the most appropriate underlying nonlinear function to help in the modelling of gametocytes. Although their paper took an innovative approach to model gametocytes directly, it did not account for informative censoring.

Another approach on the same data used by Distiller et al. (2010) was proposed by Chatora (2018). The author's focus was turned towards modelling the time to gametocyte emergence, the density of gametocytes, and the duration of gametocytemia (the presence of gametocytes). It was also of interest to impute gametocyte profiles for reasons that included accounting for the presence of informative censoring and mitigating the effect of interval-censoring. Chatora (2018) combined nonlinear mixed effects and time to event models that take into account informative censoring in a joint modelling framework, following on from the work proposed by Henderson et al. (2000). The author then analyzed the prevalence of gametocytes on the imputed data through these joint models. This current dissertation presents the reanalysis of the same data used by Distiller et al. (2010) and Chatora (2018). However, this study will not impute gametocyte profiles. Instead, the focus will be on the use of time to event models that account for censoring, particularly the presence of interval-censoring since the study has been designed in such a way that data is collected over ever-increasing intervals.

The use of methods that account for interval-censoring has, over recent years, become more and more common, with more literature around the statistical analysis of interval-censored data being developed over the past two decades (Singh and Totawattage, 2013). Several authors have developed many methods to deal with interval-censored data. Finkelstein and Wolfe (1985) presented a semi-parametric model for regression analysis that allows for the analysis of interval-censored data. Turnbull (1976) put forward a method that estimates the non-parametric distribution function for interval-censored data using the idea of self-consistency equations. Peto (1973) proposed a method that estimates the cumulative distribution function for interval-censored data, which corresponds to the life table technique used for right-censored data. Finkelstein (1986) also proposed a method for

fitting proportional hazard regression models for interval-censored data using the Newton-Rhapson iteration algorithm.

Gomez et al. (2009) presented a tutorial on the methods for interval-censored data using the then available software for analysis. Their paper discussed the common methods that take into account interval-censoring such as the parametric and semi-parametric regression models, estimating the non-parametric survival function, comparison of survival curves and tests, which the authors then applied to interval-censored data.

Lindsey and Ryan (1998) also provided a tutorial using interval-censored methods. Their paper describes parametric and non-parametric methods for estimating the survival curves, regression models and compare interval-censored results with the results from approaches that assume event times would have taken place at the start, midpoint, and end of an interval.

This dissertation will focus on applying the appropriate interval-censored methods to a real data set related to gametocytemia to demonstrate the impact of ignoring interval-censoring. As previously mentioned, by improving our methods of analyzing gametocyte data this will help better answer some of the clinical questions. Hence, interval-censored methods will then be applied to investigate the time to gametocytemia, the relative hazard of carrying gametocytes and the duration of gametocytemia.

The latter will help determine the infectivity period of the host whilst, the former will inform on the factors that influence the host's susceptibility. These findings will help advise us of the impact on transmission, which could help in the combat against malaria and drug resistance.

1.4 Objectives

The main focus of this dissertation is to model gametocytes in the presence of interval-censoring. In clinical trials, patients are examined at follow-up visits, which implies that an event's exact time is only known to be between two follow-up visits. The survival data is then said to be interval-censored. However, some authors may choose to ignore this and treat the interval-censored data as exact time data.

In this dissertation, different approaches to modelling the data are used to assess the impact of ignoring the interval-censored mechanism. Regarding the clinical objectives, the aim is to determine (1) the time to gametocytemia and the relative hazard of carrying gametocytes and (2) gametocytemia duration. The statistical approach outlined in this study takes into account:

- patients entering the study with gametocytes thus left censoring.
- patients who withdraw before clearance of gametocytes, as such the elimination phase is incomplete, leading to right censoring.
- measurements at time points taken at ever-increasing intervals, hence the exact date of either gametocyte emergence or elimination is unknown, thus interval-censoring.

In this dissertation, three approaches will be used to analyze the data. The one approach correctly assumes data is interval-censored. With this approach, data will be analyzed using interval-censored methods. The second approach assumes the time of observation as the event time (the day the event is first observed to have occurred). The last approach assumes the event time occurs at the midpoint of the interval in which the event has occurred.

These approaches are applied to analyze the current data set, to illustrate the impact on results when different statistical methods are applied to interval-censored data.

1.5 Data overview

The data used in this analysis were obtained from a series of randomized control trials conducted between 2002 and 2004 in Mpumalanga, South Africa and southern Mozambique, (Barnes et al. (2006), Allen et al. (2009)) as part of the South-East African Combination Antimalarial Therapy (SEACAT) Evaluation of the phased introduction of combination anti-malarial therapy (Allen et al., 2009) under the Lubombo Spatial Development Initiative. These studies' primary aim was to measure the effectiveness of two treatments, sulfadoxine-pyrimethamine (SP) and a combination of artesunate and sulfadoxine-pyrimethamine (ACT), in clearing asexual parasites in patients.

Patients included in the study were male and female (non-pregnant and non-breastfeeding) subjects who showed signs and symptoms of malaria, were older than a year and weighed more than 10 kilograms (Allen et al., 2009). Patients who either showed signs of complicated malaria, were allergic to the drugs included in the study or were diagnosed with an underlying disease were excluded from the study (Carter et al., 2005). Blood samples from 784 patients were drawn on prescheduled visitation days. Patients were required to come for clinical and parasitological assessments on days 0,1,2,3,7,14,21,28 and 42 after being enrolled, where day 0 was enrollment day. At each follow-up, thick blood smears samples were taken. With these collected blood samples, further analysis was carried out and provided the following information:

- Asexual Parasite Density over time
- Gametocyte Density over time
- Sulfadoxine drug concentration (ug/ml)
- Pyrimethamine drug concentration (ng/ml)

In addition to the blood samples collected, several baseline covariates were recorded. A select few of these covariates were included in this analysis due to their potential significance and effect on the emergence, elimination and duration of gametocytemia (the presence of gametocytes). The selected covariates are presented below in Table 1.

Table 1: Covariates

Variable	Covariate
anemia	Absence (0) or presence (1) of moderate anaemia
treatment	Treatment received by patient, with a value of 0 for sulfadoxine-pyrimethamine treatment only (SP) and 1 for the combination of artesunate and sulfadoxine-pyrimethamine treatment (ACT)
gender	Gender of a patient with female given by 1 and male 0
logage	Age of patient in years to the logarithm base of 2 ($\log_2$)
mutat	The presence of quintuple mutations with presence denoted by 1 or absence represented by 0
pzero	Baseline asexual parasite density per microliter to the logarithm of base 10 ($\log_{10}$)

Anemia was included in the analysis as a covariate since authors Price et al. (1999) and von Seidlein et al. (2001) identified anaemia as a risk factor of gametocyte carriage. Sowunmi et al. (2004) suggests that there is a relationship between gametocytemia and anaemia. This may be because low levels of haemoglobin reticulocytosis trigger gametocyte production (Trager et al., 1999). For this analysis, a patient is considered moderately anaemic if the haemoglobin concentration level is less than 11g/dl. This definition aligns with the guidelines given by the World Health Organization (WHO, 2011b).

In addition, patient age is included in the analysis as the covariate age is said to be a proxy for immunity against gametocyte carriage (WWARN Gametocyte Study Group, 2016). Many authors (see for example the WWARN Gametocyte Study Group (2016)) have identified a negative relationship between age and the prevalence of gametocytemia. With the much younger demographic, particularly children frequently observed to have a higher level of gametocyte prevalence.

Treatment (*arm*) is included in the analysis as a covariate since the aim of the trials was to compare the effectiveness of Artemisinin-based combination therapies to the sulfadoxine-pyrimethamine treatment only (SP) to reduce the development of gametocytemia and consequently suppress uncomplicated *falciparum* malaria.

The gender of a patient has been previously found by some authors to be associated with gametocyte carriage. For instance, the WWARN Gametocyte Study Group (2016), found that the male gender to be a risk factor of gametocytemia. Gender is therefore included as a covariate since it is related to gametocyte carriage.

Ross et al. (2006) suggested that gametocyte densities lag behind asexual parasite densities. In other words, the baseline asexual parasite preceded *Plasmodium falciparum* gametocytemia. Price et al. (1999) even states that the peak of gametocytemia appears 7-10 days after the peak of the asexual parasite. Therefore, asexual densities can inform on gametocytemia.

According to Barnes et al. (2008), there is a correlation between dihydrofolate reductase (dhfr) and dihydropteroate synthase (dhps) mutations and sulfadoxine-pyrimethamine resistance, thus affecting gametocyte carriage. For this reason, the categorical variable *mutat* is included in the study. The value 1 represents a patient with five mutations (3 dhfr and 2 dhps), and 0 represents the patient with fewer than five mutations.

From the patient data of the 784 patients enrolled in these studies, only a subset of 391 patients had records of the above covariates and did not withdraw on the day of enrollment because of treatment failure or loss to follow up are included for further analysis.

1.5.1 Prevalence of gametocytes

Gametocyte prevalence refers to patients who have at least one nonzero gametocyte density at any time throughout the study window period. Of the 391 individuals, 37.1% (145 out of 391) of the participants experienced gametocytaemia. The remaining patients had zero gametocytes throughout the study. Figure 2 shows the number of individuals with a positive gametocyte infection per day by treatment group. For example, on day 0, of those using the ACT treatment, 16 individuals were recorded to have gametocytes present in their blood systems. From the same figure, it also appears that under the ACT treatment, there are more active cases (patients presently experiencing gametocytaemia) in the earlier days, especially on day 3 than in the later days. The number of active cases quickly drops by day 14, which could suggest quicker recovery.

Whereas the number of patients experiencing gametocytaemia under the standard treatment, SP, is much larger in the later periods than in the earlier days. We only see a drastic drop in the number of patients who actively have gametocytes by day 28. And with the ever-increasing interval, this could potentially suggest a slower recovery rate. Overall, it is evident that more patients under the SP regimen who experience gametocytemia than those who are using ACT. Which suggests that ACT is the superior treatment.

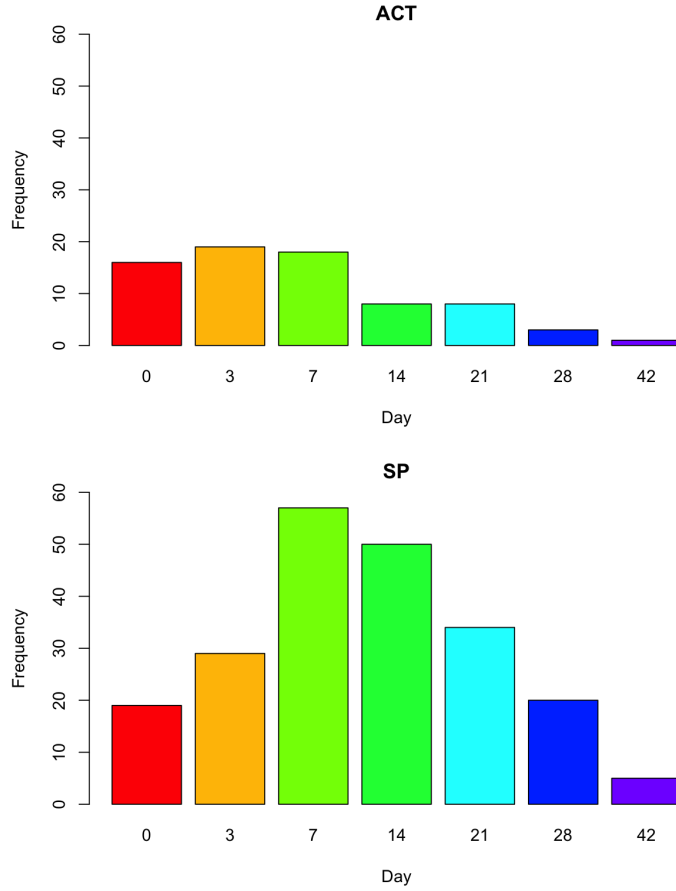


Figure 2: Gametocyte prevalence by day by treatment group

Table 3 shows the descriptive statistics of the covariates included in the analysis together with therapeutic responses against gametocyte prevalence. The therapeutic responses (treatment outcomes based on asexual parasite clearance) categories from the data collected are summarized below in Table 2 and their corresponding definitions guided by the World Health Organization. However, in this paper, re-infection, recrudescence and early treatment failure are categorised into one group: Failure.

Table 2: Treatment Outcomes

Outcome	Definition	N
Success	Patients in this category see a clearance of asexual parasites from the patient's system with no re-infection.	302
Failure	Patients who are recorded to have this outcome experience one of the following : • Re-infection • Recrudescence • Early Treatment Failure • Death	50
Loss-to-follow-up	This category represents once active participants who no longer present themselves for observation.	39

Table 3 reveals a strong association between gametocyte prevalence and the covariate *arm*. It appears that of those patients taking ACT, 77% of them do not experience gametocytemia throughout the study. In contrast, of those using the SP treatment, only 48% did not develop gametocytes. Suggesting that ACT is the superior treatment.

There also appears to be a strong association between anaemia and the presence of gametocytes. Patients who are not moderately anaemic have a lower prevalence of gametocytes, with 67% of patients not recording any gametocytes throughout the study. Table 3 also reveals that the gametocyte prevalence is strongly associated with treatment outcome/therapeutic responses. 62% of those who experienced treatment failure experienced gametocytemia. And of those who had a successful treatment outcome 65% did not experience gametocytemia throughout the study. While 79% of patients lost-to-follow-up do not experience gametocytemia. This may, of course, suggest bias since patients lost-to-follow-up are not observed after they have withdrawn and neither can we observe the presence of gametocytes should they emerge.

From the same table, it is clear that *gender* and *mutat* do not have any association with gametocyte prevalence.

The Kruskal-Wallis test reveals that neither distribution of *logage* or *pzero* are significantly different for the presence or absence of gametocytes.

Table 3: Gametocyte prevalence descriptive statistics

Variable		Present	Absent	P-value
Treatment (arm)	SP	98 (52%)	91 (48%)	< 0.001
	ACT	47 (23%)	155 (77%)	
Gender (gender)	M	64 (40%)	98 (60%)	0.457
	F	81 (34%)	148 (66%)	
Quintuple Mutation (mutat)	0	109 (35%)	199 (65%)	0.201
	1	36 (43%)	47 (57%)	
Anemia (anemia)	0	66 (31%)	144 (69%)	0.016
	1	78 (43%)	102 (57%)	
Treatment Outcome (Based on asexual parasite clearance)	Success	106 (35%)	196 (65%)	< 0.001
	Failure	31 (62%)	19 (38%)	
	LTFU	8 (21%)	31 (79%)	
$\log_2$ age in years (logage)	Median	1.097	0.845	0.139
	IQ Range	0.778-1.415	0.602-1.255	
Baseline asexual parasitaemia (pzero)	Median	3.543	4.031	0.398
	IQ Range	2.441-4.463	2.918-4.591	

¹ Fisher's Exact tests were applied to assess dependence between categorical variables

² Kruskal Wallis tests were applied to test differences between continuous variables

³ Row percentages are in brackets

The boxplots in Figure 3 display the distribution of the data for the continuous covariates based on gametocyte prevalence. It appears that patients who develop gametocytes:

- are generally younger
- have higher baseline asexual parasite densities

compared to those who do not develop gametocytes.

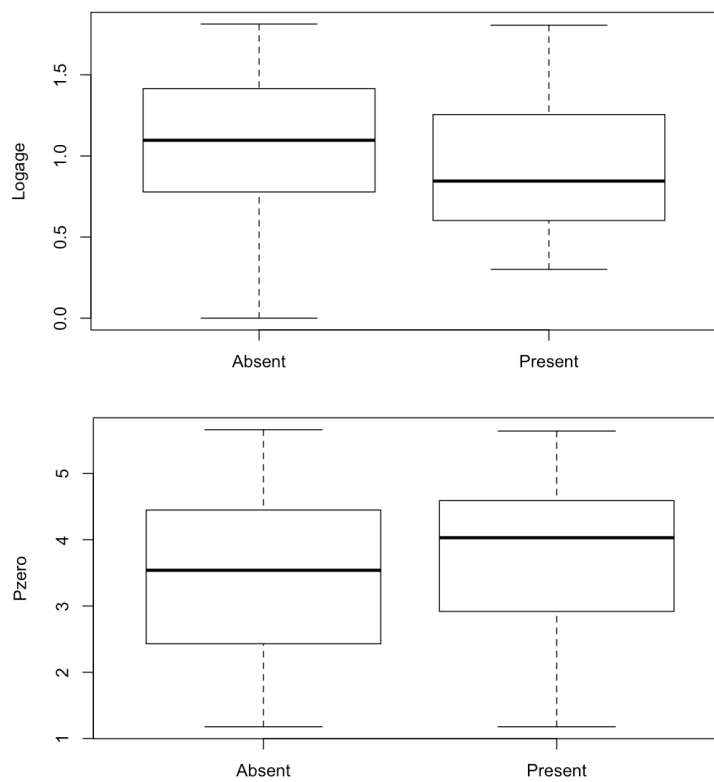


Figure 3: Boxplots for *logage* and *pzero* for by gametocyte prevalence

Figure 4 shows the mean profiles of gametocytemia by categorical covariates used in this analysis. The profiles below show that the highest gametocyte densities are associated with patients who experience treatment failure, are male, have five mutations are anaemic and use the standard treatment SP. Furthermore, gametocyte density profiles generally peak between days 7 and 21, where day 7 records the highest average across all covariates.

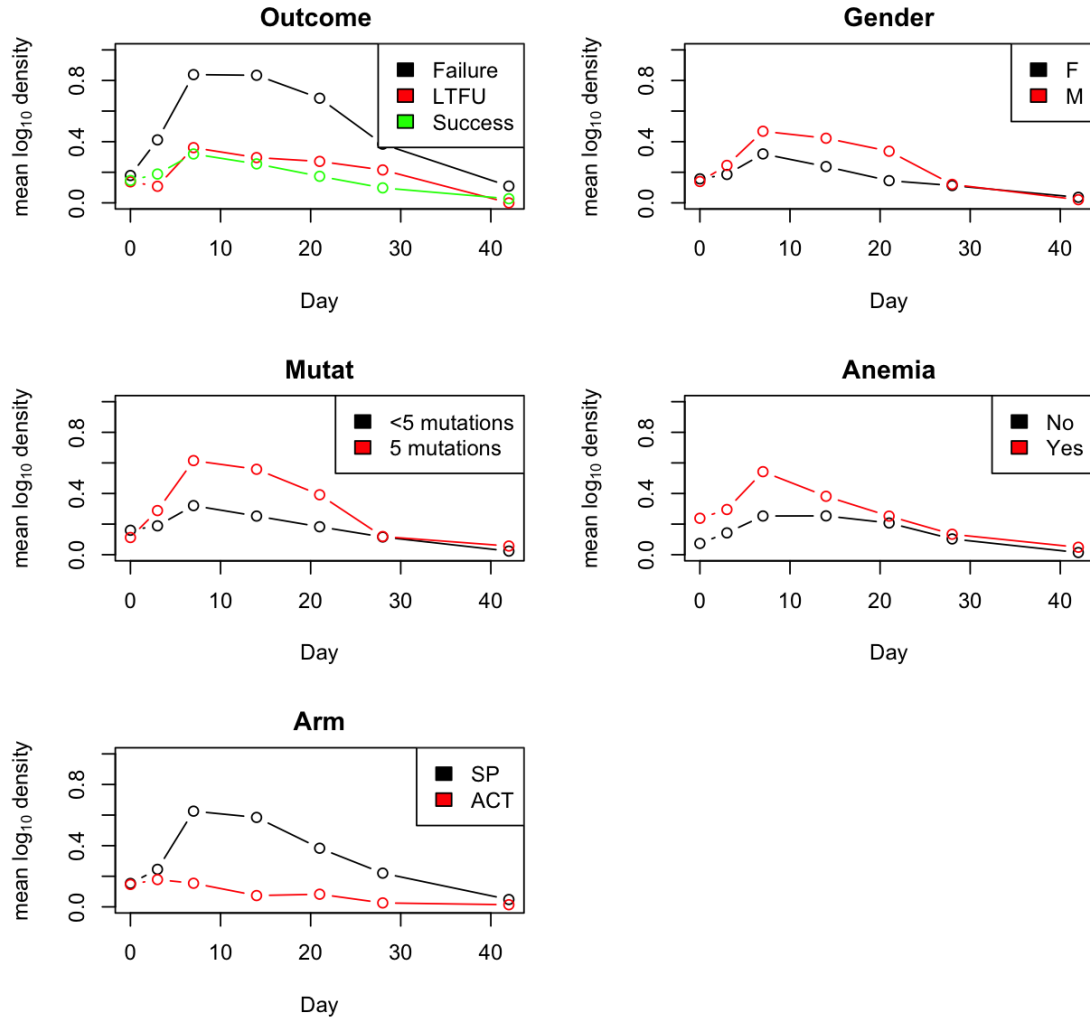


Figure 4: Mean gametocyte profiles across the categorical covariates

1.5.2 Association with treatment arm

Table 4 shows the descriptive statistics for the covariates considered in this analysis across the treatment arms. To assess if there was any association between treatment arms and the covariates. Once again, Fisher's exact test was used in the case of categorical variables and the Kruskal-Wallis test was used to test whether any differences between groups for continuous variables were significant.

From Table 4 it is shown that of the patients whose outcomes result in treatment failure in this sample, 28% were receiving the combination of artesunate and sulfadoxine-pyrimethamine (ACT) treatment, while, 72% of those who experienced treatment failure were the patients receiving the standard treatment sulfadoxine-pyrimethamine (SP). Suggesting treatment failure is more common under SP. Fisher's Exact test showed a strong association between therapeutic response and treatment arm. However, based on the Kruskal and Fisher's test, there appears to be no association between treatment and the other categorical variables.

Table 4: Descriptive statistics of covariates per treatment group

Variable		SP	ACT	P-value
Gender (gender)	M	73 (45%)	89 (55%)	0.305
	F	116 (51%)	113 (49%)	
Quintuple Mutation (mutat)	0	145 (47%)	163 (53%)	0.387
	1	44 (53%)	39 (47%)	
Anemia (anemia)	0	100 (48%)	110 (52%)	0.839
	1	88 (49%)	92 (51%)	
Treatment Outcome (Based on asexual parasite clearance)	Success	137 (45%)	165 (55%)	0.001
	Failure	36 (72%)	14 (28%)	
	LTFU	16 (41%)	23 (59%)	
$\log_2$ age in years (logage)	Median	1.000	1.079	0.502
	IQ Range	0.699-1.301	0.699-1.398	
Baseline asexual parasitaemia (pzero)	Median	3.926	3.655	0.571
	IQ Range	2.577-4.550	2.458-4.463	

¹ Fisher's Exact tests were applied to assess dependence between categorical variables

² Kruskal Wallis tests were applied to test differences between continuous variables

³ Row percentages are in brackets

Figure 5 shows the boxplots for the distribution of the continuous covariates *logage* and *pzero* across the treatment arms. From the boxplots in Figure 5 it is evident that there are only marginal differences between boxplots of each treatment arm.

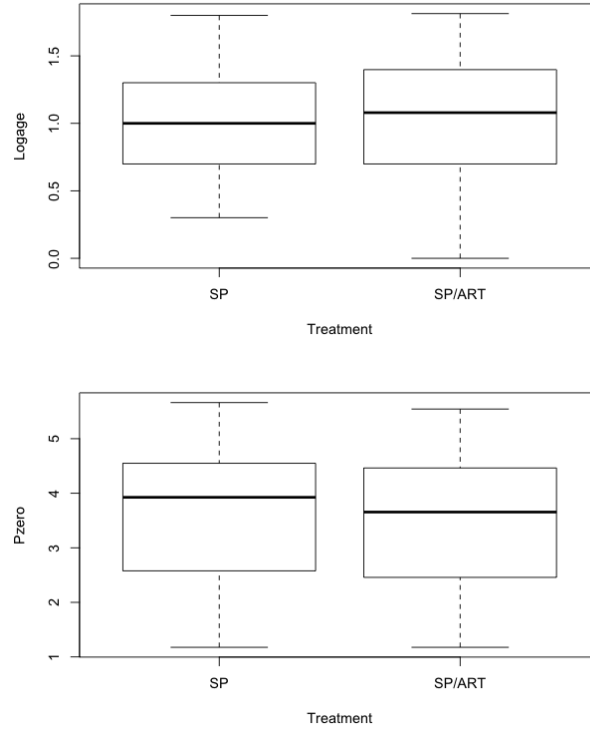


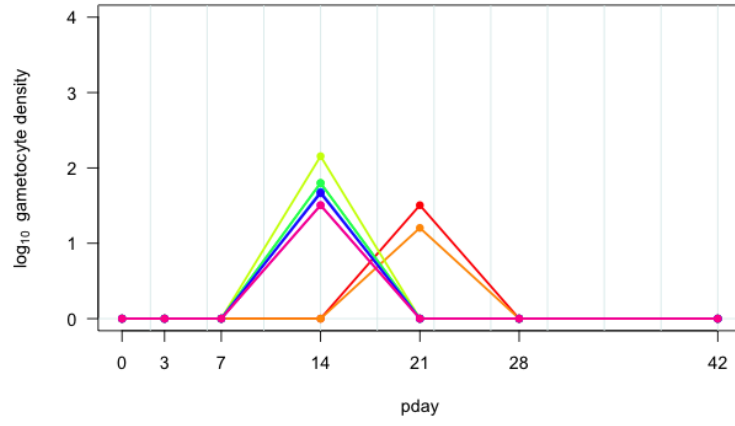
Figure 5: Boxplots for *logage* and *pzero* covariates by treatment arm

1.5.3 Individual gametocyte profiles

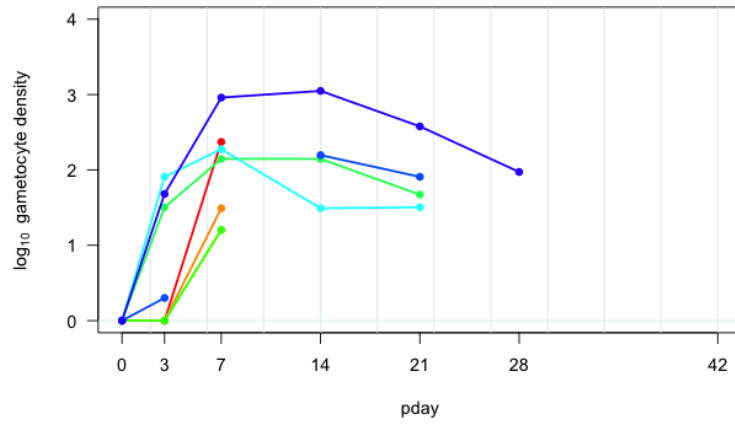
Figure 6 shows the spaghetti plots of a subset of patients who fall into one of the therapeutic response categories and their recorded gametocyte densities. The overall spaghetti plots that contains all the patient profiles based on treatment outcome are provided in the Appendix and it can be generally observed that

- Patients who experience successful treatment generally have relatively low gametocyte densities; have shorter periods of gametocytemia and quite a number of patients observing a peak on day 14;
- Patients who experience treatment failure appear to have moderately high gametocyte densities; their gametocyte density profiles seem to peak on day 7; tend to leave the study from day 21 hence their complete gametocyte profiles are not observed; more interestingly patients have longer periods of gametocytemia
- Patients who are lost-to-follow-up also have relatively low gametocyte densities and incomplete profiles with some patients entering the study with gametocytes
- We observe the widening interval lengths at the later time periods.

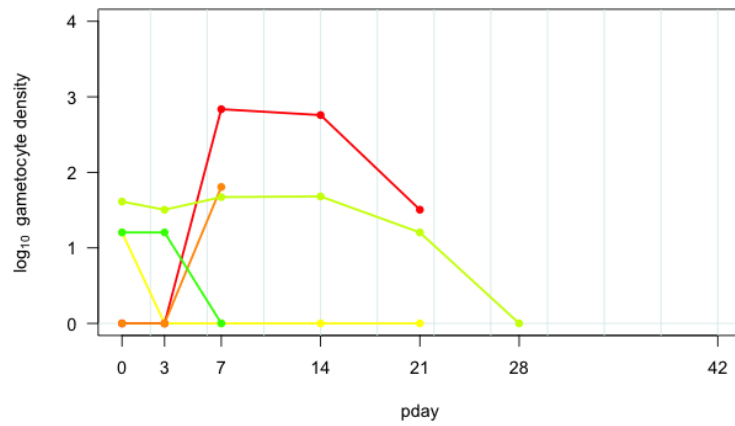
The outline of this dissertation is as follows: Chapter 2 discusses censoring mechanisms and gives a brief introduction to survival analysis methods. Chapter 3 presents the theory and application of survival curves and in Chapter 4, several survival functions and models will be fitted using the approaches mentioned above. These results will be compared to investigate and assess the significance of overlooking the interval-censoring mechanism. Finally, Chapter 5 will focus on the model-building approach, discussing the models fitted using the appropriate interval-censored methods and answering the clinical questions mentioned above.



(a) Success



(b) Failure



(c) LTFU

Figure 6: Gametocyte Profile spaghetti plots

2 Censoring

It is common that some information about an individual's survival time is incomplete or partially known in time to event data. The actual occurrence of an event happening is thus unknown, which results in data being censored. Censoring is an inherent characteristic of survival data that makes typical statistical procedures inappropriate (Little, 2013). In Chapter 1, the concept of censoring was introduced, and it was briefly mentioned that censoring within this data is a result of the study design. In this chapter, an in-depth look at the different forms of censoring are discussed.

There are generally two forms of censoring. The one form of censoring occurs when, for example, a study is terminated at a set point of time, such that those remaining observations are known to have survived up to that time. In this instance, time is fixed, and the number of failures is the random variable. This form of censoring is known as *generalized Type I censoring* (David and Moeschberger, 1978).

Type II censoring refers to when the study or test continues up until a pre-specified number of r items have failed ($r < n$) (Klein and Moeschberger, 2003). So we see that with Type II censoring, items failing is fixed, and time is the random variable. Generally, Type II censoring occurs in fields such as engineering when testing equipment for example (Klein and Moeschberger, 2003). Here all objects are tested simultaneously, and the test runs up until r out of n items have failed. Type I censoring is more common in the epidemiology and clinical field and will thus be this dissertation's main focus. This chapter will look at the different censoring types and how such scenarios are related to the present data. Furthermore, the assumptions of censoring and, finally, the mechanisms used to handle censored survival data are discussed.

2.1 Right censoring

Right censoring is the most well known and common type of censoring. It occurs when the event of interest is unobserved by the end of the study period. Right censoring might occur for various reasons.

Firstly, survival times might be censored when a competing event has occurred, which results in an individual exiting the study before the event of interest has happened. An example of this is when a patient withdraws from a clinical trial due to treatment failure. Secondly, the subject may leave the study before the event of interest (for example, the elimination of gametocytes) has occurred. This may be because, during the follow-up period, a patient may have stopped going to the clinic or may have moved to another city. When this happens, the patient is lost to follow up and is thus right-censored.

Lastly, if the event of interest has not occurred during the entire period under review, the right-censored observations arise due to administrative censoring.

In all the scenarios described above, the patient's information is known up until their last known observation time. Time to the event is, therefore, larger than the time under observation. Figure 7 shows gametocyte profiles of patients who were right-censored. These are examples of patients who dropped out of the clinical trial or were lost to follow

up or are right-censored due to administrative censoring. It would be possible with more information to determine if the patients dropped out because of reasons associated with the study. However, this is beyond the scope of this dissertation and will not be explored in this paper.

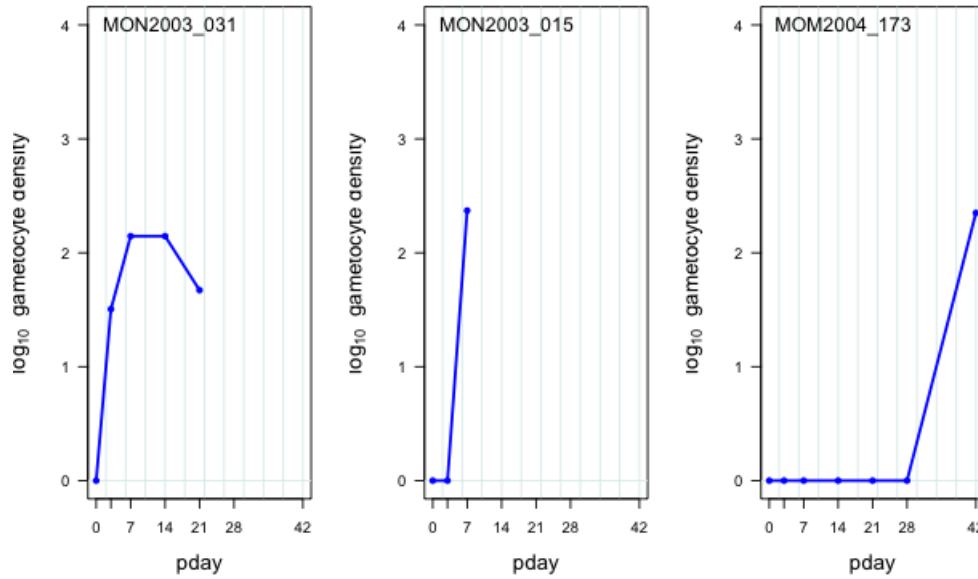


Figure 7: Right censoring examples

2.2 Left censoring

Left censoring occurs when the event of interest has already passed for an individual before they are under observation. In other words, the individual has already experienced the event before they have been enrolled in the study. In the context of this study, a patient is left-censored when a patient is recorded to have a gametocyte density greater than zero upon entry of the study. The particular date on which the gametocyte emergence first occurred is unknown since the patient had not been included in the study. Figure 8 are some examples of patients who entered into the study with gametocytes. The absorption phase is missing; hence the exact time to emergence is unobserved.

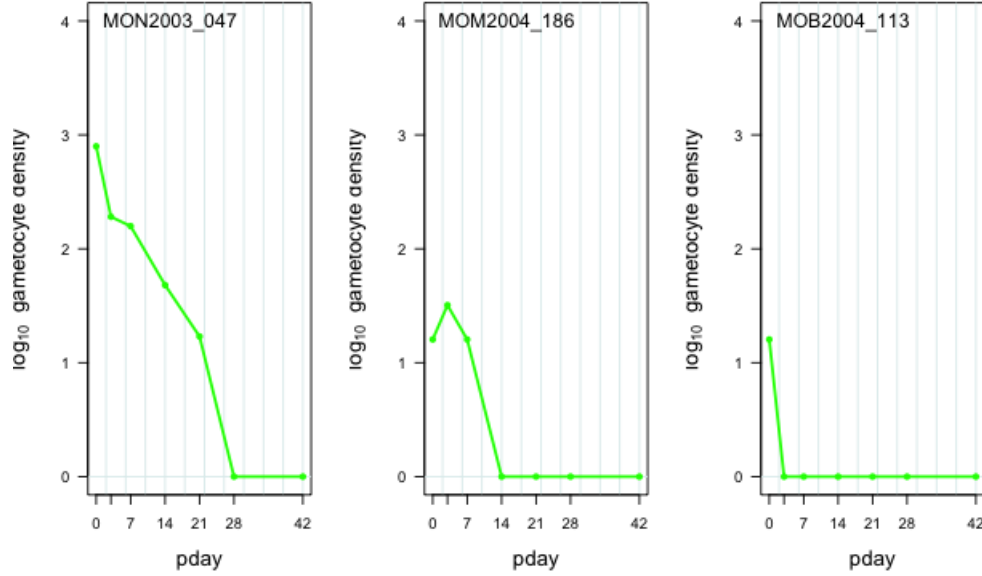


Figure 8: Left censoring examples

An illustration of both left and right censoring is discussed below in Figure 9. The example in the diagram relates to a hypothetical situation where the parameter of interest is time to gametocyte emergence. The * represents censored patients. Patient B's time to relapse is observed within the observational window period. Therefore, they are not an example of a censored patient. Patient D went throughout the study without experiencing gametocytemia (the presence of gametocytes) and is, therefore, right-censored. Patient A is also censored due to loss of follow up, this may be because the patient moved away, and their records are no longer available. They cannot be observed throughout the rest of the study. Patient C had gametocytes in their system before they had enrolled in the study. Patient C is thus an example of a left-censored subject.

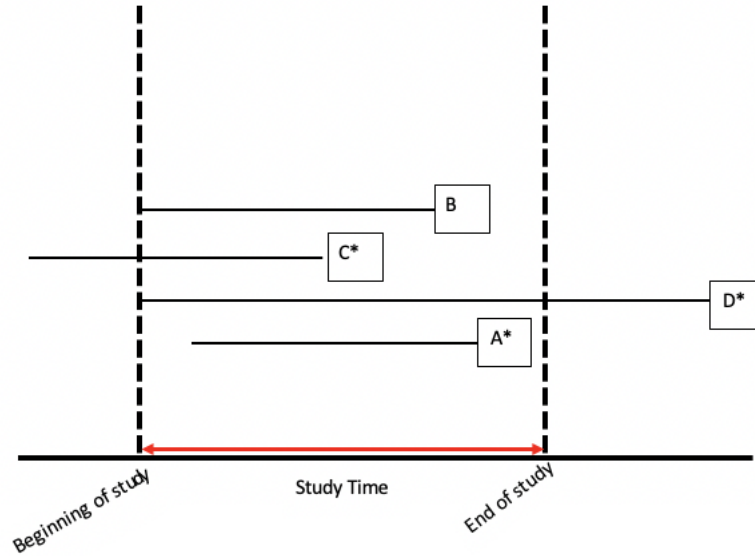


Figure 9: Examples of point censored observations

2.3 Interval-censoring

Interval-censoring occurs when an event happens in between two consecutive doctor's visits for example. Interval-censoring typically occurs in medical and longitudinal studies designed to have periodic follow-up dates. In between these prescheduled visitations, a patient could experience a change in their health status. In other words, during a patient's current examination, the patient's circumstances or diagnosis is different from their previous examination. Consequently, the exact time to event is unobserved, and all that is known is the interval in which the event of interest occurred. As previously mentioned, our data is collected from the blood samples of individuals were collected at specific time points, which means that patients are observed at specific periodic follow-up times, and further implies that the event times are interval-censored.

For example, say a patient at day 3 shows no signs of gametocytemia, but on day 7, the patient has a nonzero gametocyte density. Then, the researcher can only conclude that the gametocytes appeared some time greater than the last observation day but on or before the change in status was observed (less than or equal to the following observation time at which the change of status is first recorded). In other words, the time to emergence had occurred somewhere between days 3 and 7. The exact day on which the gametocytes first appeared is partially unknown, but the interval in which the event of interest occurs is known.

The profiles in Figure 10 represents patients with completed profiles; however, the exact dates of time to emergence and time to elimination is unknown. For example, patient

MOB2004_058 on day 21 showed relatively high gametocyte levels. However, at their next visit on day 28, there were no signs of gametocytes. Their exact time of elimination is unobserved; there is therefore only partial information available. That information being the interval window in which elimination happened. Moreover, with the clinical trials' intervals ever-increasing, estimating time to elimination is much more difficult. It is worth mentioning that it is possible for a patient to experience a combination of interval, left and right censoring (Rodrigues et al., 2018).

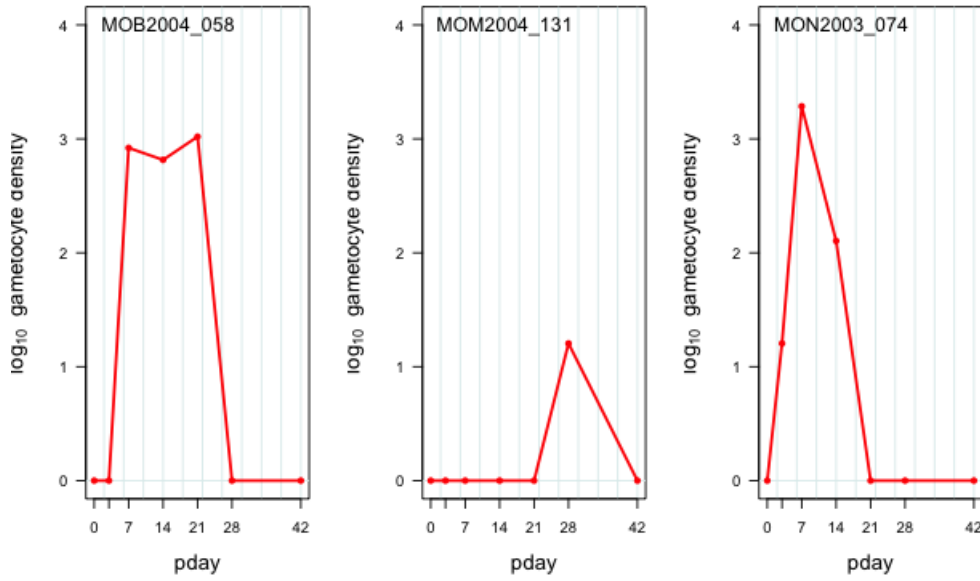


Figure 10: Interval-censoring examples

2.4 Doubly interval-censoring

The focus of this dissertation is on the modelling of crucial aspects of gametocytemia. This includes modelling the duration of gametocytemia. The duration of gametocytes is the difference between the time of gametocyte emergence and the time of gametocyte clearance. Simply put, this is the length of time that a patient has gametocytes in their system. This estimation would require the knowledge of the time to gametocyte emergence and time to gametocyte clearance. However, in our data, both these times are subject to interval-censoring, which means we only have partial information. In such instances where both time to emergence and time to elimination occur within an interval, this is known as doubly interval-censoring.

Figure 11 shows gametocytemia progression from time to emergence until time to elimination. In the diagram, the observed time to gametocyte emergence occurs between

day 3 and day 14 and the time to elimination is also interval-censored in the interval day 28 and day 42. Because of the interval-censoring mechanism, the true event times are unknown. Therefore, to calculate duration, both emergence and clearance times would need to take into account interval-censoring - hence we would need to apply doubly interval-censored methods.

The analysis of doubly interval-censored data will be discussed further in Chapter 5.

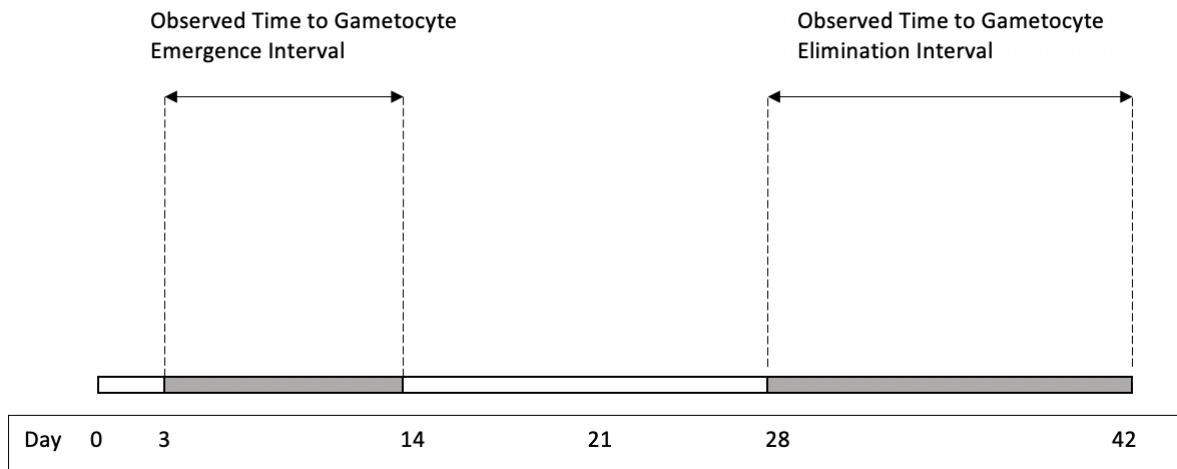


Figure 11: Doubly interval-censored example

2.5 Left truncation

An additional feature of survival data often confused with censoring is *truncation* (Klein and Moeschberger, 2003). Truncation is the exclusion of individuals based on some condition that determines whether the researcher is aware of their existence. That is different from censoring, as censored observations in the study are still included in the analysis. In some instances, the event of interest can only occur if an individual has experienced some other event prior. For example, for us to observe time to elimination (the main event of interest for this example), a patient must have observed the emergence of gametocytes. In Figure 12 patient *MOC2003_023* has no gametocytes recorded at any of the follow-up times. When investigating time to gametocyte elimination, patients with similar gametocyte profiles to patient *MOC2003_023* will be excluded since patients did not experience gametocytemia. In other words, for the event - gametocyte elimination, such patients will be left truncated.

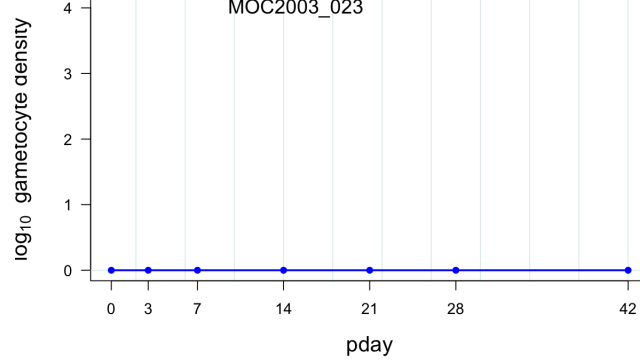


Figure 12: Left Truncation example

2.6 Assumptions of censoring

In cases where censoring is a present feature in the data, a general assumption underlying censoring requires censoring to occur independently of the outcome of interest. Should this not be the case bias may arise. In our data, when a patient experiences treatment failure, the patient is withdrawn from the study. Implying that withdrawn patients are, in fact, in a worse condition compared to those who remain in the study and may thus have an increased probability of carrying gametocytemia, resulting in informative censoring which in turn results in bias.

Unfortunately, standard survival analysis methods do not consider the presence of informative censoring (Chatora, 2018). To counter this Chatora (2018) dealt with this through joint modelling. However, Chatora (2018) did not find strong evidence of bias due to informative censoring.

This dissertation will not focus on informative censoring. Instead, the primary assumption used throughout this paper will be non-informative censoring, also known as the constant-sum condition (Leung et al., 1997). The non-informative interval-censoring assumption as specified by Sun (2006) is as follows:

$$Pr(T \leq t | L = l, R = r, L < T \leq R) = Pr(T \leq t | l < T \leq r),$$

where L and R are the endpoints of the interval such that $L < T \leq R$. This assumption implies that the only information provided by the censored interval $(l, r]$ with regards to survival time t is that no extra information is provided except that time t is found within interval $(l, r]$ (Zhang and Sun, 2010). In other words, the likelihood function does not consider the censoring mechanism (Zhang and Sun, 2010). All methods discussed below will be based on this assumption.

2.7 Mechanisms for dealing with censoring

As previously highlighted in section 2.3 the present data is interval-censored as a consequence of the study design. Traditional statistical approaches cannot be applied to interval-censored data. Described briefly below are standard practices for analyzing interval-censored data.

One common way of dealing with censored data is ignoring the interval-censoring mechanism completely.

This is done using imputation methods. The two more commonly used methods of imputation include: (1) assuming that the event time occurs at the moment the event is first observed to have occurred - right imputation and; (2) assuming that the event has occurred in the middle of the interval in which the event occurred - midpoint imputation. Authors tend to favour these two approaches due to the lack and availability of statistical software for the analysis of interval-censored data (Leung et al., 1997).

This study analysis will be based on three alternative approaches, which assume the survival data is (1) interval-censored (IC); (2) right imputation (RI), or (3) midpoint imputation (MP).

To illustrate the above approaches, consider the following scenario and examples below. Suppose a patient on day 14 has gametocytes, but on the subsequent visit, day 21, a patient has zero gametocytes. In that case, it is assumed that gametocyte elimination has occurred on day 21, even though elimination may have occurred before day 21 but after day 14. The three approaches will deal with the data in the following ways.

- Interval-censored - the event is assumed to have occurred within the interval $(l, u]$, thus $(14, 21]$
- Right imputation - the event is assumed to have occurred at the end of an interval such that $(t) = u$. In the above example $t = 21$.
- Midpoint imputation - the event is assumed to have occurred at the midpoint of the interval in which the event took place. In this case, the time to gametocyte elimination has “occurred” on day 17.5.

The above approaches will be applied to analyze time to gametocyte emergence and time to gametocyte elimination. It should be noted that time to gametocyte emergence is defined as the time from enrollment day (day 0), the starting point, until the first nonzero gametocyte density is observed. Time to elimination is the time when gametocytes were first observed to be eliminated following a previous visit where a nonzero gametocyte density was observed, less the time to emergence, which is considered as the starting point, assuming that the starting point (time to emergence) is an exact time. Finally, duration of gametocytemia is defined as the difference between time to emergence and time to elimination, assuming both starting point (time to emergence) and event time (time to elimination) are interval-censored - data is doubly censored.

The subsequent sections will compare results assuming exact times using the right and midpoint imputation methods against interval-censored to analyze time to emergence and time to gametocyte elimination. Chapter 5 will introduce and analyze the duration of gametocytemia.

2.8 Notation and transforming the data

Presently, there are survival statistical software programs, for example, in R, STATA and SAS that provide tools and functions to analyze right and interval-censored data. For this dissertation, the analysis uses specific packages in the R statistical program (R Core Team, 2019). Some of the software implementations in R (R Core Team, 2019) require the specification of a survival object. The inputs required include (1) time and event indicator or (2) the lower and upper limits of the interval in which the event is observed to have occurred.

The specification of the survival objects in this dissertation is guided by Bogaerts et al. (2017). In general, for interval-censoring, it is assumed that $l < T < u$, where l and u are the lower and upper limits, respectively of an interval. If an event is left censored then the lower limit is specified as, $l = 0$. For right censored observations, the upper limit is set to infinity, so $u = \infty$. In practice, the appropriate convention is to set the upper limit for right-censored observations to a value greater than the most extreme event time. The notation is summarised below,

Table 5: Censoring notation and interval specifications

Type of Censoring	Event time	Interval Specification	Indicator
Right	time l	$0 < l < u = \infty$	$\delta = 0$
Midpoint	time $\frac{l_i + u_i}{2}$	$0 < l = u = t < \infty$	$\delta = 1$
Left	time u	$0 = l < u < \infty$	$\delta = 2$
Interval	time in $[l_i, u_i]$	$0 < l < u < \infty$	$\delta = 3$

It should be noted that right and left censoring are also examples of interval-censoring (Lindsey and Ryan, 1998). With this notation in mind, a few adjustments have been made. Some functions in R (R Core Team, 2019) have a hard requirement (or a strict rule) that left-censored observations have a lower limit equal to 0. However, some observations that are not left-censored might also have a lower limit equal to 0. Therefore, to distinguish between left-censored observations from those not left-censored, the lower limit for the observations, not left-censored but had a lower limit of 0, was changed to 1. For the sake of consistency, “+1” was added to all event times in the data.

These adjustments have been applied throughout the data used in this analysis and should be kept in mind when reading.

3 Survival curves

Several functions could be used to summarize time-to-event outcomes (Collett, 1994). Four functions in particular are discussed below and drawing on the work from Collett (1994), we are able to see how these functions are all related. For the time to event response, T , the *probability density* function is defined as

$$\begin{aligned} f(t) &= F'(t) \\ &= \lim_{\delta t \rightarrow 0} \left\{ \frac{F(t + \delta t) - F(t)}{\delta t} \right\}, \end{aligned} \quad (1)$$

where $F(t) = Pr(T < t)$ is the probability that the event has occurred prior to time t . The second function measures the probability of surviving up until or beyond time t ,

$$\begin{aligned} S(t) &= Pr(T \geq t) \\ &= 1 - \int_0^t f(u) du \\ &= 1 - F(t). \end{aligned} \quad (2)$$

The third function that may be used to summarize survival data is the hazard function, which is defined as the probability that an event occurs in the interval $[t, t + \delta]$. The function is defined as

$$\begin{aligned} h(t) &= \lim_{\delta t \rightarrow 0} \left\{ \frac{Pr(T \leq t + \delta t | T \geq t)}{\delta t} \right\} \\ &= \lim_{\delta t \rightarrow 0} \left\{ \frac{1}{\delta t} \frac{Pr(t \leq t + \delta t)}{Pr(T \geq t)} \right\} \\ &= \lim_{\delta t \rightarrow 0} \left\{ \frac{1}{\delta t} \frac{F(t + \delta t) - F(t)}{S(t)} \right\} \\ &= \lim_{\delta t \rightarrow 0} \left\{ \frac{F(t + \delta t) - F(t)}{\delta t} \right\} / S(t) \\ &= \frac{f(t)}{S(t)}. \end{aligned} \quad (3)$$

It estimates the instantaneous risk of an event occurring within the small interval $[t, t + \delta]$ given that the person has survived up until time t . The hazard function and survival function are related in the following way.

Given that $S(t) = 1 - F(t)$ then $\log(S(t)) = \log(1 - F(t))$. It then follows that the derivatives of logs is given by

$$\begin{aligned}\frac{d}{dt}\log(1 - F(t)) &= \frac{\frac{d}{dt}(1 - F(t))}{1 - F(t)} \\ &= -\frac{f(t)}{S(t)},\end{aligned}\tag{4}$$

so the link between the hazard function and survival function is

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}\{\log S(t)\}.$$

The final function is the cumulative hazard function which is expressed as

$$\begin{aligned}H(t) &= \int_0^t h(u)du \\ &= -\log S(t),\end{aligned}\tag{5}$$

which can be rewritten as

$$\exp(-H(t)) = S(t),$$

therefore showing the relation between the hazard and survival models.

The survival functions estimate the percentage of individuals who survive up to some point and beyond. Methods used to do this could be non-parametric, parametric and semi-parametric. An example of a non-parametric estimator is the Kaplan-Meier method (Kaplan and Meier, 1958) which is discussed further below.

3.1 Methodology

3.1.1 Kaplan-Meier curve

The Kaplan-Meier estimator is a popular non-parametric maximum likelihood estimator of a survival function (Kaplan and Meier, 1958). It is non-parametric as it is distribution-free and therefore does not require any specific assumptions to be made with regards to the underlying distribution of the survival data. This function is used to assess the probability of occurrences of events over a period of time.

To obtain the Kaplan-Meier estimate of the survivor function for a sample of survival data, infinitesimally small time intervals are constructed. This is done by dividing the observation period into a series of organized intervals as follows.

Let survival time be t_x for $x = 1, 2, \dots, K$ survival times. Suppose there are n individuals in

a study with observed event times $t_1, t_2, t_3, \dots, t_K$ where after arranging the survival times in ascending order the survival times are such that $t_1 < t_2 < t_3 \dots < t_K$. At each event time let there be n_x number of subjects at risk including those who are about to experience the event of interest at time t_x . The number of subjects that experience the event of interest at time t_x will be denoted as d_x . Consider the time interval between $t_x - \delta$ and t_x , where δ represents a minute interval (Collett, 1994). Since there are n_x individuals alive at the start of time t_x and d_x events at time t_x , then the estimated probability of the event of interest occurring between time $t_x - \delta$ and t_x is given by $q(t_x) = \frac{d_x}{n_x}$ (Collett, 1994). Consequently, the probability of survival between the interval $t_x - \delta$ and t_x can be estimated by $1 - q(t_x) = \frac{(n_x - d_x)}{n_x}$.

When some subjects are censored it can happen that censored and death times occur at the same time. If there are any subjects that are censored at time t_x , these subjects are assumed to have survived longer than the event time at time t_x . Therefore, no adjustment is necessary for the number at risk n_x at time t_x . However, these censored subjects will be excluded from the number at risk at time t_{x+1} , the next time point. Furthermore, it can be assumed that events of interest of individuals in a sample occur independently of each other.

Previously, it was mentioned that the probability of survival through the time interval $t_x - \delta$ to t_x is given by $1 - q(t_x) = \frac{(n_x - d_x)}{n_x}$. However, Little (2013) demonstrated that calculating survival rates using $1 - q(t_x)$ could lead to inconsistent estimates and using individuals who are present at each survival time to estimate survival is wasteful. Instead the probability of survival is given as the probability of survival through all the intervals up to time t_x conditional upon the survival in previous time intervals. Therefore, the survival probability at time t_x is estimated using the Kaplan-Meier estimator which is defined by Kaplan and Meier (1958) as:

$$S(t_x) = \prod_x (1 - q(t_x)). \quad (6)$$

This estimator is a decreasing step function that remains constant between event times and 'jumps' at each event time. The product is cumulative over all intervals up to and including t_x . For this reason, the Kaplan-Meier survival function is also known as the product-limit estimator. However, this function has its limitations.

In a scenario where the largest observation time is censored, $\hat{S}(t)$ is undefined beyond that censored time (Collett, 1994). This then implies that the mean lifetime cannot be estimated (Rodrigues et al., 2018). Seeing as the mean survival time is calculated as the area under the survival curve from time 0 till the end of the observational period (Klein and Moeschberger, 2003).

3.1.2 Non-parametric maximum likelihood estimator

The Kaplan-Meier estimate is used to estimate the survival curves with data that is right-censored. This section focuses on how to estimate the survival distribution in the presence

of interval-censoring using the non-parametric maximum likelihood (NPMLE) approach.

The survivor function for interval-censored data may be estimated non parametrically. These estimated curves are interpreted in the same manner as the traditional Kaplan-Meier survival curves (Gomez et al., 2009). However, unlike the NPMLE used for right-censored data, the survival function used for interval-censored data is obtained using an iterative algorithm. The more common algorithms used to construct the NPMLE of an interval-censored survivor function include the self-consistency algorithm (Turnbull, 1976), the ICM algorithm proposed by Groeneboom and Wellner (1992) which uses a more efficient technique and the hybrid algorithm which was suggested for consideration by Wellner and Zhan (1997) which uses a composite algorithmic mapping combining the self-consistency algorithm with the ICM algorithm. Of these methods, the self-consistency algorithm proposed by Turnbull (1976) will be our focus.

The self-consistency algorithm is one of the more commonly used methods developed by Turnbull (1976) who extended the ideas of Peto (1973) by suggesting the use of self-consistency equations to estimate the non-parametric estimator of the survivor function. In the case of interval-censored data, subjects are evaluated at specific time points rather than continuously. For each interval, there is an upper boundary, u_i (observation time at which change in status of the subject is first noted) and a lower boundary, l_i (which is the preceding observation time before the change in status is first recorded). The estimation of the NPMLE of an interval-censored survival function requires a set of intervals known as Turnbull's intervals, which are constructed as follows.

Let $[l_i, u_i]$ ($i = 1, 2, \dots, n$) be the observed intervals for n independent subjects, which contain the unknown event times. Peto (1973) and Turnbull (1976) proposed a reduction algorithm to assist in the identification of regions of possible support as follows. Where regions of possible support are the intervals in which an event could have occurred.

For each interval of the observed intervals $[l_i, u_i]$, rank time points and indicate if it is a left end point (lower boundary) or a right end point (upper boundary). Next step, identify regions of possible mass.

This is done by pairing endpoints such that a left endpoint q_i is followed immediately by a right endpoint p_i . These set of so-called Turnbull intervals used are denoted by $I = (q_1, p_1], (q_2, p_2], (q_3, p_3], \dots, (q_m, p_m]$ (Gomez et al., 2009). Refer to Figure 13 for example of how this might be done.

The previous step is particularly important as it effectively reduces the size of the interval and thus narrowing the search area to a more specific time interval or possibly an exact time. These intervals are also known as regions of possible support because the algorithm aims to identify regions in which events may have possibly occurred (Bogaerts et al., 2017).

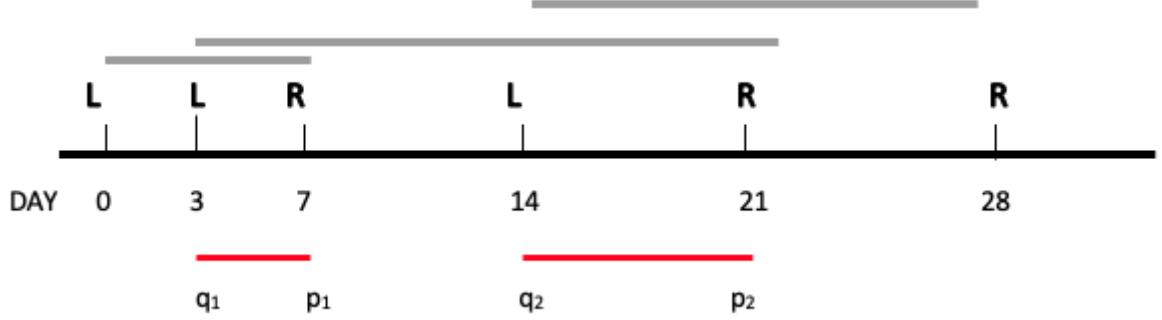


Figure 13: Example of Turnbull's intervals: Suppose events take place within the intervals $(0,7]$, $(3,21]$ and $(14,28]$. The Turnbull intervals in this case would be $(3,7]$ and $(14,21]$.

With these intervals the maximization for obtaining the NPMLE is simplified to the following likelihood function:

$$L_T(w_1, \dots, w_m) = \prod_{i=1}^n [S(q_j) - S(p_j)] = \prod_{i=1}^n \left(\sum_{j=1}^m \alpha_{ij} w_j \right), \quad (7)$$

where α_{ij} is an indicator variable equal to 1 if an event that has occurred in the interval $[l_i, u_i]$ could have also happened in the Turnbull interval $(q_j, p_j]$. In other words $\alpha_{ij} = 1$ if $(q_j, p_j] \subseteq [l_i, u_i]$ and is 0 otherwise. Probability of an event occurring at time p_j or weight is denoted by w_j and is calculated as

$$w_j = S(q_j) - S(p_j) \quad (j = 1, \dots, m).$$

After calculating the probability for each interval, we now have a vector containing these probabilities given by $\mathbf{w}$ such that w_j is subject to the constraints $w_j \geq 0$ and $\sum_{j=1}^m w_j = 1$. In order to estimate the NPMLE of the survival function given the above constraints Turnbull (1976) proposes using the self-consistency algorithm. The algorithm begins by making an initial estimate for w_j . An example of such an estimate would be to assume equal probability for all by assuming $w_j = 1/m$ for all j (Bogaerts et al., 2017). We estimate the NPMLE using the following algorithm:

$$\begin{aligned} \hat{S}(p_0) &= 1 \\ \hat{S}(p_j) &= \hat{\pi}_j \hat{S}(p_{j-1}) \\ \hat{\pi}_j &= \frac{n'_j - d'_j}{n'_j} \end{aligned} \quad (8)$$

Where the estimated number of events is given as

$$d'_j = \sum_{i=1}^n \frac{\alpha_{ij} w_j}{\sum_{k=1}^m \alpha_{ik} w_k}, \quad j = 1, \dots, m$$

and the number of individuals at risk, n'_j

$$n'_j = \sum_{k=j}^m \left(\sum_{i=1}^n \frac{\alpha_{ij} w_j}{\sum_{k=1}^m \alpha_{ik} w_k} \right).$$

occurring at time p_j .

The vector w_j is then updated using the newly estimated values from n'_j and d'_j . This iterative process is repeated until there is very little change between the new and previous estimates or until a level of accuracy is achieved.

3.2 Results

The methodology outlined above was applied to the longitudinal data discussed in Chapter 1. This section illustrates the impact of ignoring the presence of interval-censoring in data. Several survival functions were generated using the following approaches: (1) applying the Kaplan-Meier estimator assuming that the event time is an exact time equal to the time of observation - right imputation (RI) (2) applying the Kaplan- Meier estimator assuming that the event time is the midpoint of the interval of observation (MP) and (3) applying the Turnbull estimator after correctly recognizing the interval-censoring mechanism (IC). These functions were generated to measure the survival time for our events of interest-time to gametocyte emergence and time to gametocyte elimination.

3.2.1 Time to emergence

Time to gametocyte emergence is the time until the first nonzero gametocyte density is recorded. Patients who did not record any gametocytes throughout the 42 day study period or were lost to follow up were classified as right censored patients. It should be noted that as previously discussed in the section 2.7 to accommodate functions in the *Icens* package (Gentleman and Vandal, 2019) require left censored observations to have a lower limit of 0, all event times were shifted by 1. This was to help distinguish between left censored observations and observations that were not left censored but happened to have a lower limit $l_i = 0$. In addition to the above, some functions in the same package had a 'hard' requirement that specified right censored observations to have an upper limit equal to some arbitrary value. It should also be highlighted that the upper limit for right censored observations was set to 45 which is beyond the maximum values for both the lower and upper limit variables .

To generate the interval-censored distribution function using NPMLE, the function EMICM is used which yields the possible regions of support together with their respective

weights shown below in Table 6. Recall that these weights are the maximising solution to the likelihood function given in equation 7. These weights are interpreted as the estimated probability of an event occurring within the interval $(q_j, p_j]$. From the table below, the probability that gametocytes have emerged between the enrollment day and day 1 $(0, 1]$ is 0.090. Given the adjustments made to event times this interval represents the left censored observations. It can also be observed that Turnbull interval $(43, 45]$ has a probability of 0.611. However, it should not be concluded that the probability of gametocyte emergence within the interval $(43, 45]$ is equal to 0.611 as these are in-fact the right censored observations.

Table 6: Regions of mass support for time to emergence

Regions of Support	Probability
(0,1]	0.090
(3,4]	0.086
(4,8]	0.122
(8,15]	0.058
(15,22]	0.021
(22,29]	0.009
(29,43]	0.003
(43,45]	0.611

In Table 7 are the survival rate (probabilities of reaching a specific day without gametocyte emergence) estimates for days 4, 8, 15, 22, 29 and 43 based on the RI, MP and IC approaches. From this table it can be observed that there is not a relatively large difference among the three approaches.

Table 7: Survival rate estimates for gametocyte emergence

Analysis Method	Time (days)					
	4	8	15	22	29	43
Right	0.828	0.703	0.644	0.624	0.612	0.609
Midpoint	0.829	0.702	0.646	0.623	0.614	0.611
Interval	0.825	0.702	0.644	0.623	0.613	0.610

The survival curves for gametocyte emergence from all three approaches using the Kaplan-Meier and Turnbull estimators are shown in Figure 14. The black curve is the survival curve estimated from the Turnbull estimator, the red curve is the survival curve from the Kaplan-Meier estimator using the right end point (RI) and the blue curve is the

Kaplan-Meier estimator when midpoint imputation (MP) is used.

The survival curves appear to track each other very closely. The survival rates in the earlier days (when intervals are shorter) for all three approaches are very similar to each other. As the intervals increase in width the Kaplan-Meier estimator based on the RI method, generally estimates higher survival estimates in comparison to the curves where data is IC and MP.

Although the differences are very minor. The three curves become indistinguishable as time goes on. This is a result of all the data being right censored. The only difference is, the interval-censored survival curve eventually declines to 0 by day 45. This however, should not be taken to mean that by day 45 all patients would have experienced gametocytemia. Rather, this descend to 0 is a statistical artefact, a result of replacing the right end point of an interval of right censored patients with 45.

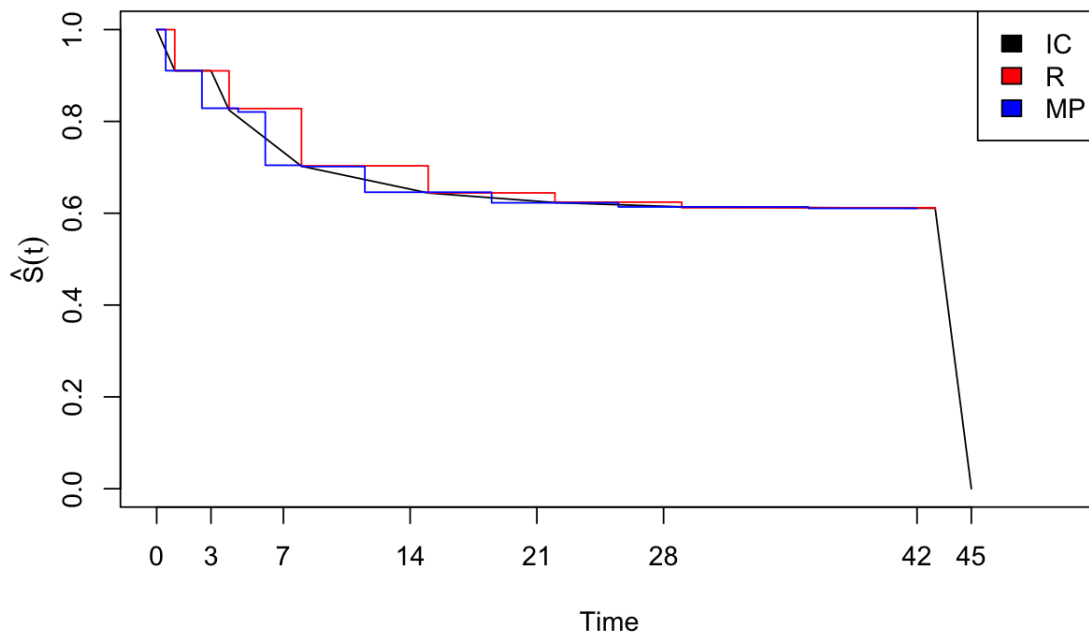


Figure 14: K-M Emergence Survival Curve

The inferences made above are a result of the following:

Firstly, with right censored data, it is assumed that the events of interest (in this case emergence of gametocytes) occurs at the end of the time point or time interval. By contrast, the NPMLE for interval-censored data employs an iterative reduction algorithm that seeks out to identify regions of possible support (time intervals in which an event may have occurred) and thus tapers to smaller time intervals (possibly a specific time point) in which the event could have occurred.

The event time under the interval-censored method (as well as the midpoint approach) could therefore occur earlier in comparison to the right imputation method. Which then explains why the right imputation method overestimates survival probabilities.

Secondly, looking at the above tables, much of the time to gametocyte emergence occurs within the smaller, earlier intervals. Interval-censoring would not have as great an impact which could explain why the three curves are similar. It should also be noted that the NPMLE assumes a linear decrease in survival over each region of support, in contrast to the KM estimates that assume constant survival until the next event time observed.

3.2.2 Time to elimination

Time to gametocyte elimination is the day of the first zero gametocyte density reading, following a nonzero gametocyte reading at the previous clinical visit. We further adjust this by subtracting a “start-time”, which is the first day we observed a nonzero gametocyte density. This is done because the time to gametocyte elimination can only begin after gametocytes have emerged. Elimination times thus represent duration of gametocytemia. Subjects who were still experiencing gametocytemia (where gametocytes had not been eliminated from their system) by the end of their last visit were classified as right-censored observations. While patients who did not experience gametocytemia were left truncated. Therefore, only 145 patients were included in this analysis. It should also be noted that although there are no left-censored observations, the event times were shifted by 1 for the sake of consistency. In Table 8 are the possible regions of support and respective weights. It can be observed that the probability of gametocyte elimination in interval (1,4] is 0.379. It can also be seen that the probability of gametocyte elimination between days (15,19] is 0.104 and we can deduce from the table that the estimated median time of gametocyte elimination is between (4,5] days from emergence.

Table 8: Regions of mass support for time to elimination

Regions of Support	Probability
(1,4]	0.379
(4,5]	0.154
(5,8]	0.000
(8,12]	0.191
(12,15]	0.011
(15,19]	0.104
(19,22]	0.000
(22,26]	0.000
(26,29]	0.119
(29,36]	0.000
(36,40]	0.042

Table 9 shows the gametocyte elimination survival rate estimates. In general, the RI survival curve estimates higher survival rates in comparison to both the MP and IC approaches.

This again is a result of assuming that the event time has occurred at the extreme end of the interval, which implies that for the same survival probability the event time under this approach occurs at a later time.

Table 9: Survival rate estimates for gametocyte elimination

Analysis Method	Time (days)					
	4	8	15	22	29	43
Right	0.912	0.489	0.299	0.190	0.114	0.043
Midpoint	0.788	0.467	0.262	0.131	0.042	0.000
Interval	0.621	0.467	0.264	0.161	0.042	0.000

Figure 15 shows the survival curves for all three approaches.

The difference in the survival curves is more pronounced compared to the curves in Figure 14. In the Figure below we see that in the narrower intervals both the RI and MP approaches overtake the IC survival curve. In the later and more wider intervals, the survival curves from the Kaplan-Meier estimates “fence in” the Turnbull method. As the RI overestimates the survival times while the MP approach generally underestimates the survival times compared to the survival estimates from the IC approach. This was also observed by Lindsey and Ryan (1998).

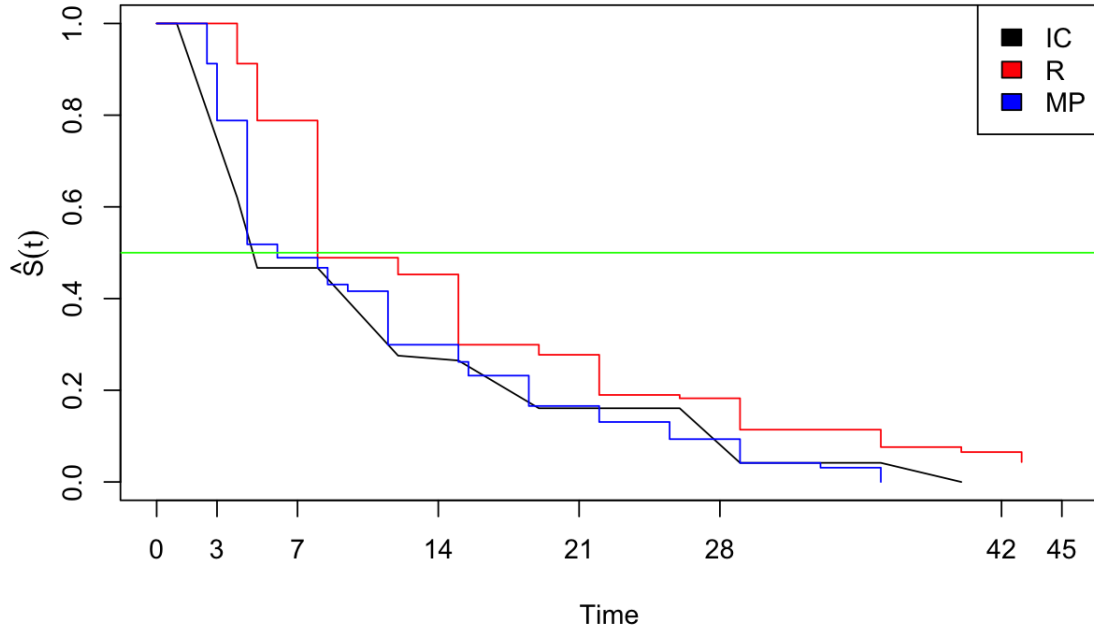


Figure 15: K-M Elimination Survival Curve

From the above figure we can observe that the median event times (which are estimated using the green horizontal line) for the RI, MP and IC approaches are 8, 5.5 and 5 respectively.

Overall it is apparent that where event times occur in smaller intervals ignoring the interval-censoring mechanism does not impact results as much, as is the case in gametocyte emergence where events occurred during the narrower windows. However, in instances where observations have occurred in wider intervals, RI and MP imputation may not be appropriate substitutes for interval-censored method due to relatively larger differences in approaches as was observed in table 9.

Secondly, when using the the RI and MP approaches, the the probability of survival is constant until the subsequent event time. Whereas, the interval-censored approach assumes a constant negative slope within each region of support.

3.3 Comparison of survival distributions

It is not uncommon for the data that is collected to include categorical covariates that stratify the sample observations into two or more groups. One may wish to observe if there is a significant difference between these groups, for example, between patients using standard treatment SP and those using the alternative ACT. This comparison would be of interest to help determine if, indeed, one treatment is more effective than the other. One way of determining this is through the use of plotted survival functions stratified according to treatment. However, conclusions made from such illustrations are not always reliable due to subjectivity.

Therefore, in addition to the graphical comparisons of each group's survival distributions, formal hypothesis tests may be conducted. Available methods used for right-censored data include the classical non-parametric test and the log-rank test that is based on a comparison of the observed events and the expected events overall event time intervals. For interval-censored data, we use a generalized weighted log-rank test. The details of these right-censored and interval-censored tests are discussed below.

3.3.1 Logrank tests for right-censored data

For right censored data, testing if there is a difference between two survival curves or more uses the logrank test derived by Mantel (1966). This is the classical non-parametric method for comparing survival distributions. Construction of the logrank test summarizing the text from Collett (1994) is as follows:

Begin by considering the different groups separately, for example Group I and Group II. Suppose there are r distinct event times such that $t_1 < t_2 < t_3, \dots, t_r$, across the two groups. Let n_{j1} and n_{j2} represent the number of individuals at risk in groups I and II respectively, just before time t_j and let d_{j1} and d_{j2} be the number of events. The values of d_{j1} and d_{j2} are equal to 0 or 1 except in cases where there are two or more individuals who experience the event at the same survival time.

Further, let $d_j = d_{j1} + d_{j2}$ and $n_j = n_{j1} + n_{j2}$ for $j = 1, 2, \dots, r$.

This data may be summarised into a contingency table such as Table 10 (adapted from Collett (1994)), which shows the number of events at the j^{th} time as well as the number of subjects surviving beyond t_j as shown below.

Table 10: Observed data at time t_j to compute the logrank test

Group	Number of events at t_j	Number at risk at t_j	Number surviving beyond t_j
I	d_{j1}	n_{j1}	$n_{j1} - d_{j1}$
II	d_{j2}	n_{j2}	$n_{j2} - d_{j2}$
Total	d_j	n_j	$n_j - d_j$

Summarizing data this way over 2x2 tables was proposed by Mantel and Haenszel (1959).

Now, consider the null hypothesis

$$H_0 : S_1 = S_2,$$

which states there is no real difference between two groups in terms of overall survival. To validate this claim, this would require assessing the extent to which the number of subjects who experience the event at a given time differ between the two groups and differ from the number that is expected under the null hypothesis of no difference.

In fact, the number of those expected to fail in Group I under the null hypothesis is given by

$$e(d_{j1}) = n_{j1} \times \frac{d_j}{n_j},$$

which is the mean of the hyper-geometric random variable d_{j1} and the variance of d_{j1} is given by:

$$\text{Var}(d_{j1}) = V_{j1} = d_j \frac{n_{j1}n_{j2}}{n_j^2} \frac{n_j - d_j}{n_j - 1}.$$

Together with the data in the contingency table, the difference between the total observed death d_{j1} and expected death $e(d_{j1})$ in Group I, leads to the following statistic which summed over all event times to give the following:

$$\chi_1^2 \sim \frac{U_1^2}{V_1} = \frac{(\sum_{j=1}^r (d_{j1} - e_{j1}))^2}{(\sum_{j=1}^r V_{j1})}. \quad (9)$$

The larger the differences between what was observed and what was expected under the null hypothesis, the larger the test statistic χ_1^2 will also be which will lead to the rejection of the null hypothesis, suggesting there is sufficient evidence to conclude that there is no difference between the survivor functions of the two groups.

This can be extended to a weighted logrank statistic

$$\chi_W^2 = \frac{(\sum_{j=1}^r w_j (d_{j1} - e_{j1}))^2}{\sum_{j=1}^r w_j^2 V_{j1}}, \quad (10)$$

allowing for different weights denoted by w_j at different event times $j = 1, 2, \dots, r$. Fleming and Harrington (1980) suggest using the weights to emphasize either early, middle or late differences using the $G^{\rho, \lambda}$ class of weighted logrank tests where the weighted function at time t_j is given as:

$$w_j = \hat{S}(t_j)^\rho (1 - \hat{S}(t_j))^\lambda \quad \rho \geq 0, \lambda \geq 0.$$

In instances where $\rho = 0$ and $\lambda = 0$ and then equal weight is given to all data therefore the function does not emphasize differences, this simplifies to the classical logrank test.

On the other hand when $\rho = 1$ and $\lambda = 0$ this reduces to the generalized Wilcoxon test statistic known as the Peto-Prentice test modified by Peto and Peto (1972) and Prentice (1978), emphasizing early differences. And when $\rho = 0$ and $\lambda = 1$ this emphasizes late differences.

3.3.2 Weighted logrank tests for interval-censored data

The methodology in this section is explained as a summary of the text from Bogaerts et al. (2017) and Gomez et al. (2009). As with right censored data, the test statistic of weighted logrank test (WRLT) is based on the difference between the observed and expected number of events. Equation 10 shows how the weighted logrank statistic for right-censored data is calculated, where d_{jk} is the observed number of events and e_{jk} the expected number of events for the k^{th} group of the j^{th} event time. This idea can be extended to interval-censored data by once again, comparing expected observation to observed at each interval of the NPMLE. This section shows how equation 10 has been extended to interval-censoring setting.

Estimation of the WLRT for interval-censored data requires the expected and observed number of events and the weights. Expected number of events are the number of events in the j^{th} region of support of $\hat{S}$. While, the observed number of events is the estimated number of events occurring in the k^{th} sample. To estimate these components Fay and Shih (1998) define the survival function of the i^{th} subject given as $\hat{S}^{(i)}$ which is truncated or cut-off at the i^{th} observed interval $(l_i, u_i]$.

$\hat{S}^{(i)}$ is defined as follows:

$$\hat{S}^i(t) = Pr_{\hat{S}}(T_i > t | T \in (l_i, r_i]) = \frac{\hat{S}(t) - \hat{S}(u_i)}{\hat{S}(l_i) - \hat{S}(u_i)}.$$

The survival function for those in the k^{th} group is considered to be the average of $\hat{S}^i$ values of individuals in the group, this can be written as:

$$\hat{S}_k(t) = \frac{1}{N_k} \sum_{i=1}^n \alpha_i \hat{S}^i(t),$$

where α_i is the indicator variables for group membership.

Just as Turnbull's intervals were identified when constructing the NPMLE, the same is also done here. So each interval is defined once again as $(q_j, p_j]$. Next, the probability of an event happening within the Turnbull interval $(q_j, p_j]$ is assigned to the right end point of the interval p_j . Note that $p_0 = 0$ (Gomez et al., 2009). Let the estimated total number of events d_j within the interval $(q_j, p_j]$ be calculated by:

$$D_j = n[\hat{S}(p_{j-1}) - \hat{S}(p_j)],$$

where p_j is the right end point that the probability of an event happening in $(q_j, p_j]$ is assigned to and p_{j-1} is the previous right end point and the estimated number of

individuals at risk prior to p_j is given by

$$n_j = n\widehat{S}(p_{j-1}).$$

While the expected number of individuals at risk prior to p_j in group k is

$$n_{jk} = N_k \widehat{S}(p_j) \quad (k = 1, 2, \dots, K, \quad j = 1, 2, \dots, J).$$

The 'observed' number of events in the k^{th} group that have occurred in the interval $(q_j, p_j]$ is estimated by

$$D_{jk} = N_k \{\widehat{S}_k(p_{j-1}) - \widehat{S}_k(p_j)\}$$

and the expected number of events under the null hypothesis is

$$E_{jk} = \frac{n_{jk}}{n_j} D_j \quad (k = 1, 2, \dots, K, \quad j = 1, 2, \dots, J).$$

The test statistic for the WLRT is given by:

$$U_K = \sum_{j=1}^J w_j (D_{jk} - E_{jk}) \quad (k = 1, \dots, K), \quad (11)$$

which may be rewritten as

$$U_K = \sum_{j=1}^J w_j (D_{jk} - \frac{n_{jk}}{n_j} D_j).$$

Using expression 11, Fay (1999) defines the vector of WLRT statistics for interval-censored data as

$$U = (U_1, U_2, \dots, U_K)'$$

Using a linear transformation model Gomez and Oller (2008) suggest extending the ideas of Fleming and Harrington (1980) $G^{\rho, \lambda}$ to an interval-censored setting. The weight functions are given as:

$$\begin{aligned} w_j &= w_i^{\lambda, \rho} \\ &= \widehat{S}(p_{j-1}) \frac{B(1 - \widehat{S}(p_j); \lambda + 1, \rho) - \widehat{S}(p_{j-1}) B(1 - \widehat{S}(p_j); \lambda + 1, \rho)}{\widehat{S}(p_{j-1}) - \widehat{S}(p_j)}, \end{aligned} \quad (12)$$

where $B(t, a, b) = \int_0^t x^{a-1}(1-x)^{b-1}dx$ is an incomplete beta function and $[p_{j-1}, p_j]$ are the observed intervals. Alternatively Gomez and Oller (2008) also demonstrate that the score tests may be expressed as:

$$c_j = c_i^{\lambda, \rho} = \frac{\widehat{S}(r_j)B(1 - \widehat{S}(r_j); \lambda + 1, \rho) - \widehat{S}(l_{j-1})B(1 - \widehat{S}(p_j); \lambda + 1, \rho)}{\widehat{S}(l_j) - \widehat{S}(r_j)}. \quad (13)$$

The sensitivity of the test are represented by ρ and λ . Different forms for c_i and w_i allows for emphasis on either early, middle or late intervals of time. Just as with the right-censored WLRT when values $\rho = 1$ and $\lambda = 0$ are used, the emphasis is on the early differences. This is the Wilcoxon-Peto test statistics suggested by Peto and Peto (1972). The combination of $\rho = 0$ and $\lambda = 1$ leads to late differences being emphasized and when $\rho = 1$ and $\lambda = 1$ middle differences are emphasized and where $\rho = 0$ and $\lambda = 0$ then equal weight is given to all data therefore the function does not emphasize differences, this simplifies to the classical logrank test.

Gomez and Oller (2008) also highlight that the weights in equation 12 do simplify to the original Fleming and Harrington family of $G^{\rho, \lambda}$ tests since the weight function, $w_i^{\lambda, \rho}$ tends towards $\widehat{S}(p_{j-1})^\rho(1 - \widehat{S}(p_{j-1}))^\lambda$ as $\frac{\widehat{S}(p_{j-1})}{\widehat{S}(p_j)}$ approaches 1.

3.3.3 Weighted logrank significance testing

In this section we once again summarize the text of Bogaerts et al. (2017) and Gomez et al. (2009). To determine the significance of the weighted logrank statistic based on the score vector U with weight or scores, the asymptotic distribution must be established. This can be done using the permutation principle (see e.g., Fay (1996), Fay and Shih (1998) and Gomez and Oller (2008)). To perform this approach rewrite U (see e.g., Gomez and Oller (2008), Fay (1999)) as

$$U = \sum_{i=1}^n z_i c_i,$$

where c_i is given by equation 13 and $z_i = (\alpha_i^{(1)}, \alpha_i^{(2)}, \dots, \alpha_i^{(k)},)$ is a k -vector of group indicators for each i^{th} observation and $[l_i, u_i]$ are the observed intervals and not the Turnbull intervals.

Using the permutation approach suggests that, under the null hypothesis there is no difference between groups, if the null hypothesis is true then the scores are interchangeable (see e.g., Hájek et al. (1999) and Gomez and Oller (2008)). The permutation distribution of U is determined by permuting the labels $n!$ and calculating the test statistic U for each possible rearrangement of the labels. It should be noted that this can only be done when the sample size is small. For cases when the sample size is relatively large, the central limit

theorem for exchangeable random variables applies resulting in a normal approximation with

$$E(U) = \frac{1}{n} \sum_{i=1}^n c_i = 0,$$

and variance-covariance matrix

$$\begin{aligned} V_P &= \frac{1}{n-1} \sum_{i=1}^n (U_i - \bar{U})(U_i - \bar{U})' \\ &= \frac{1}{n-1} \left(\sum_{i=1}^n c_i^2 \right) \left\{ \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})' \right\}, \end{aligned} \tag{14}$$

where $\bar{X} = n^{-1} \sum_{i=1}^n X_i$. The permutation test statistic is based on the Mahalanobis distance and is given by $G_P^2 = U' V_P^- U$ where V_P^- is the generalized inverse of matrix V_P . If the test statistic $U' V_P^- U$ is an extreme value from a χ_{k-1}^2 distribution then the null hypothesis will be rejected.

3.4 Results

In this section the survival functions based on the Kaplan-Meier estimator for both right imputation (RI) and midpoint imputation (MP) as well as the Turnbull algorithm for interval-censored data (IC) are presented for gametocyte emergence and gametocyte clearance stratified based on the covariate treatment arm.

3.4.1 Time to emergence

From Table 11 we observe the regions of mass support. Based on the Turnbull intervals there is generally more events occurring for those using SP in comparison to those using ACT. This implies that patients using SP have a higher probability of gametocytes emerging compared to the patients using ACT. It can also be observed that for both treatments events appear to be occurring in the earlier intervals. Although the last interval shows the most events, these are in fact the right censored individuals.

Table 11: Regions of mass support for gametocyte emergence per treatment group

Regions of Support	arm=SP	arm = ACT
(0,1]	0.101	0.079
(3,4]	0.116	0.059
(4,8]	0.207	0.044
(8,15]	0.093	0.024
(15,22]	0.025	0.018
(22,29]	0.006	0.013
(29,43]	0.000	0.006
(43,45]	0.452	0.758

Table 12 shows the survival rate estimates for patients for days 4,8,15,22,29 and 43 based on treatment using the three different approaches. From the table it can be observed that those using ACT have a better survival rate when it comes to gametocyte emergence regardless of the analysis approach used. Secondly, there is not much difference in estimates among the approaches. This is because most events are happening in the earlier intervals, where the “margin of error” among the alternative approaches is very small.

Table 12: Survival rate estimates for patients based on treatment

Arm	Analysis Method	Time (days)					
		4	8	15	22	29	43
SP	Right	0.788	0.579	0.480	0.456	0.450	0.450
	Midpoint	0.789	0.575	0.483	0.452	0.452	0.452
	Interval	0.784	0.576	0.483	0.458	0.452	0.452
ACT	Right	0.863	0.816	0.794	0.777	0.759	0.753
	Midpoint	0.865	0.819	0.797	0.775	0.763	0.763
	Interval	0.862	0.819	0.794	0.776	0.763	0.755

Figure 16 shows the survival curves based on the Turnbull and Kaplan-Meier methods for time to emergence that were stratified according to treatment arm. From the above figure, one observation that can be deduced is that the survivor functions of those using the alternative course of treatment ACT, is greater than the survivor distributions of those using the standard treatment SP for all three approaches. This implies that generally at time t , the estimated probability of those surviving from gametocyte emergence beyond time t is greater for individuals using ACT in comparison to those under SP. With a relatively large gap between the two strata, it may be implied that the use of treatment

is a useful predictor with regards to gametocyte emergence.

More interestingly, it can be observed that very little difference exists for the ACT survival curves estimated by the three approaches given how few of those using ACT experience gametocytemia. On the other hand, because more events are happening for those using SP the difference among the survival curves is more pronounced as the width of intervals widen. It is evident that the RI overestimates the survival curve while the MP approach underestimates or overestimates the survival curve in comparison to the NPMLE.

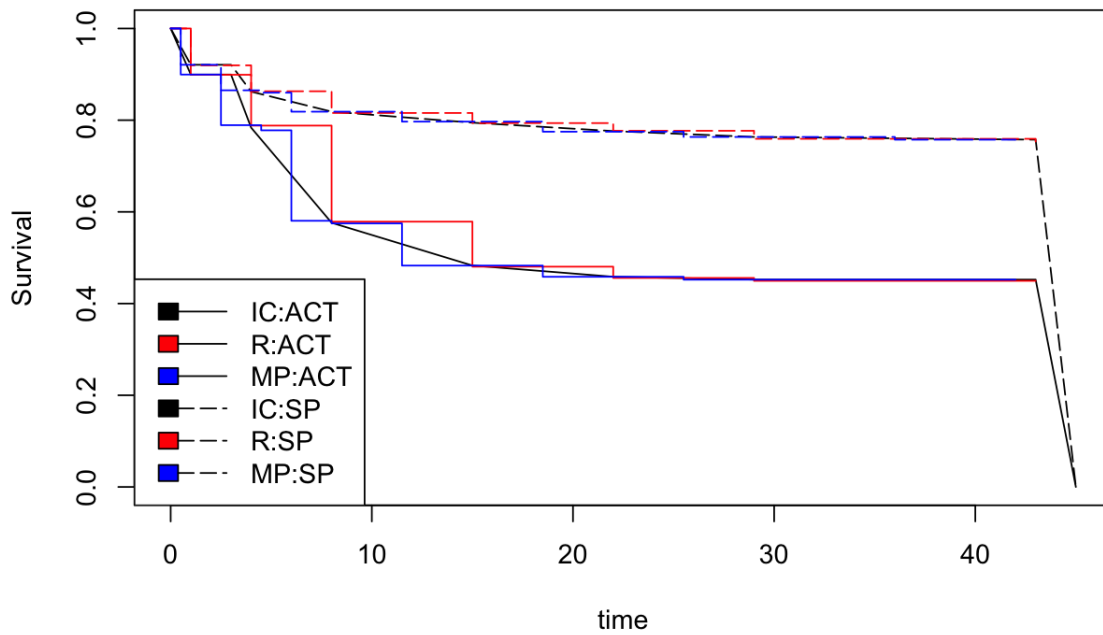


Figure 16: Comparison of gametocyte emergence survival curves based on treatment group

From Figure 16 it is very clear that there is a difference between the curves for ACT and SP. However, WLRT were performed to quantify just how different the curves from each treatment are from each other. Values for ρ and λ were varied such that the tests used simplified to the Logrank test, Wilcoxon test or emphasized late hazards for all three approaches. Table 13 confirms this as the test statistics suggest that the differences between the curves are statistically significant regardless of the weighting functions used and the analysis approach. Implying that ACT is better treatment at suppressing the emergence of gametocytes. It should be noted that, emphasizing late hazard differences resulted in the most extreme test statistics for all three approaches in comparison to the Wilcoxon and traditional logrank test. This would be expected since the gap between treatment groups regardless of the analysis approach used was largest in the later intervals. Also, it appears that the MP approach under the Wilcoxon test had the most extreme test statistic compared to the RI and IC approaches. This is because, from Figure 16 in the earlier intervals, the three survival curves for ACT are almost indistinguishable but the MP survival curve for SP is lower than that of the RI and IC curves, which could explain why the test statistic is more extreme. In addition, the Logrank test produced more significant test statistics compared to those based on the Wilcoxon test. The reason for this difference is because the Wilcoxon test emphasizes earlier differences and from Figure 16 we observe there is a narrower gap between the strata of each approach in the earlier intervals compared to the later intervals. Whereas the Logrank test assumes equal weight all throughout.

Table 13: Logrank and Wilcoxon Tests and Statistics for gametocyte emergence by treatment arm

Method	Logrank $\rho = 0, \lambda = 0$		Wilcoxon $\rho = 1, \lambda = 0$		$\rho = 0, \lambda = 1$	
	Test Statistic	P-value	Test Statistic	P-value	Test Statistic	P-value
Right	33.6	< 0.01	29.2	< 0.01	40.2	< 0.01
Midpoint	33.6	< 0.01	33.6	< 0.01	40.2	< 0.01
Interval	33.6	< 0.01	29.7	< 0.01	37.492	< 0.01

3.4.2 Time to elimination

Table 14 shows the regions of support based and their respective weights based on treatment group for time to gametocyte elimination. As before analysis is based on the 145 patients that have not been left truncated. The remainder 246 were excluded since they do not experience gametocytemia and thus cannot observe gametocyte elimination.

From the table it can be observed that patients using ACT have higher weights in the earlier intervals. Which means that patients being administered this course of treatment have a quicker recovery rate compared to those using SP.

Table 14: Regions of mass support for gametocyte elimination per treatment group

Regions of Support	arm=SP	arm = ACT
(1,4]	0.284	0.561
(4,5]	0.117	0.218
(5,8]	0.025	0.000
(8,12]	0.203	0.109
(12,15]	0.032	0.000
(15,19]	0.139	0.000
(19,22]	0.000	0.037
(22,26]	0.000	0.074
(26,29]	0.141	0.000
(29,36]	0.002	0.000
(36,40]	0.057	0.000

Table 15 shows the survival rates for patients' gametocyte elimination based on treatment group. Overall MP and IC survival estimates are much closer than RI curve, which overestimates survival rates all throughout the study period. MP in the earlier intervals appears to overestimate survival rates as with RI in comparison to the IC curve. However, in later intervals which are wider, IC over takes MP curve and the MP ends up underestimating the survival rate estimates in comparison to the IC approach. From the table we also see that ACT is better treatment in terms of clearing gametocytes given the lower survival rate estimates.

Table 15: Survival rate estimates for gametocyte elimination for patients based on treatment

Treatment	Analysis Method	Time (days)					
		4	8	15	22	29	43
SP	Right	0.934	0.587	0.391	0.239	0.148	0.037
	Midpoint	0.849	0.576	0.335	0.169	0.058	NA
	Interval	0.600	0.574	0.339	0.200	0.059	0.000
ACT	Right	0.867	0.289	0.111	0.089	0.044	0.044
	Midpoint	0.311	0.244	0.111	0.042	NA	NA
	Interval	0.220	0.220	0.111	0.091	0.074	0.000

The graphical comparison of survival distribution for time to elimination is shown in

Figure 17, it should be noted that the ACT treatment appears to speed up the elimination of gametocytes. The survival distribution function for those using ACT is much lower than those on the SP medical route. It can be inferred that, treatment, once again is a useful prognostic factor in the clearance of gametocytes.

What is interesting, is that the IC survival curves for both treatments appear to decline relatively faster in comparison to the other two curves. This may be because, under the interval-censored method and the notion of masses of regions of possible support, a lot more events/activity can be captured within an interval compared to a single observation day or time.

Alternatively, the reasoning behind this could be that, unlike the RI and MP when an event is assumed to happen at the current visit, with the IC method it is possible that the time to event has occurred earlier, that is after the previous visit but before or during the current visit.

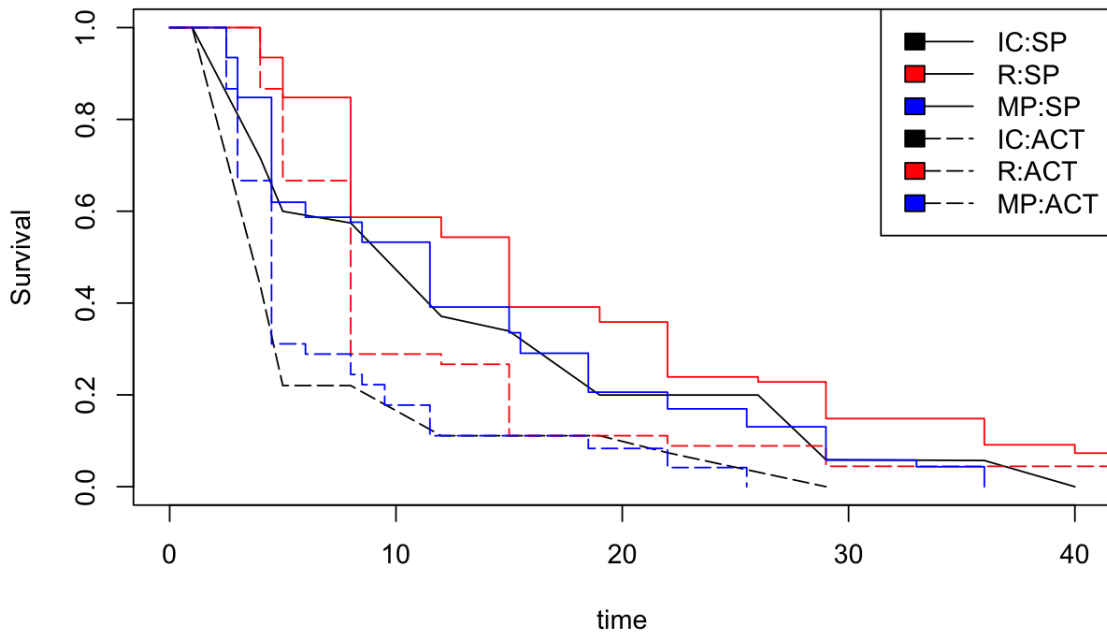


Figure 17: Comparison of gametocyte elimination survival curves based on treatment group

While a plot of the two survival functions based on the different methods is given in Figure 17 we use WLRT to assess if there truly is a difference between these two groups.

Table 16: Logrank and Wilcoxon Tests and Statistics for gametocyte elimination by treatment arm

Method	Logrank $\rho = 0, \lambda = 0$		Wilcoxon $\rho = 1, \lambda = 0$		$\rho = 0, \lambda = 1$	
	Test Statistic	P-value	Test Statistic	P-value	Test Statistic	P-value
Right	9.2	< 0.01	11.8	< 0.01	3.6	0.06
Midpoint	15.1	< 0.01	14.6	< 0.01	9.8	< 0.01
Interval	10.3	< 0.01	14.2	< 0.01	3.97	0.046

From the table above it is evident that there is indeed enough evidence to suggest a difference between the groups of patients using SP and ACT. Using both Figure 17 and the statistics in Table 16 we can go as far as saying that ACT is indeed the more effective treatment.

The results in the table above also show that emphasizing early differences results in a more significant test statistic compared to assuming equal weights or emphasizing late differences. This would be expected as looking at Figure 17 we observe that the difference between treatment groups irrespective of approach is quite large in the earlier intervals.

4 Models for survival data

4.1 Semi-parametric Proportional Hazard models

For studies interested in what factors influence subjects' survival, regression models are used to explore the relationship between survival times and various covariates. A popular regression model is the Cox proportional hazards (PH) model (Cox, 1972). It assumes a semi-parametric nature, such that the independent variables assume a parametric form, but the hazard function need not take a specific form. The hazard function, also known as the conditional failure rate, is used to approximate the risk of an event happening. It is used to estimate the instantaneous risk for example in our context, gametocyte emergence over the time interval $[t, t + \delta t)$, given the individual survived (shows no presence of gametocytes) to time t .

4.1.1 The Cox Proportional Hazards model

The Cox proportional hazards model (Cox, 1972) assumes that the hazard function at time t , given the covariate vector X is expressed as follows by Cox (1972):

$$h_i(t) = h_0(t) \exp(x_i^T \beta), \quad (15)$$

where β is a p dimensional vector of coefficients and $h_0(t)$ is the arbitrary baseline hazard. The baseline function is the hazard function for a 'standard' subject for whom the explanatory values are set to 0.

The model above can also be written as:

$$\frac{h_i(t)}{h_0(t)} = \exp(x_i^T \beta). \quad (16)$$

The model may also be expressed as follows:

$$\log \frac{h_i(t)}{h_0(t)} = x_i^T \beta. \quad (17)$$

In this form, the regression coefficient β is interpreted as the log hazard ratio for an increase of 1 in the value of a continuous covariate. For binary covariates, 1 indicates the presence of the explanatory variable and 0 as the absence, e^β is interpreted as the relative change in the hazard of gametocyte emergence for example, due to the presence of the predictor variable. Although the Cox PH model has no underlying distributional assumption, it does assume that the effect of variables on survival is constant over time. This assumption is referred to as the proportional hazards assumption.

For right-censored data fitting the PH model can be achieved using the partial likelihood estimation technique (Cox, 1972) which can be broken down into two steps. The first step of this approach is the estimation of these regression coefficients which may be

done using the maximum likelihood approach. The second step is the estimation of the baseline hazard function which can be done once the β 's have been estimated. The steps are independent of each other.

The model can equivalently be expressed in terms of the survival functions as

$$S(t|X) = S_0(t)^{\exp(X'\beta)}, \quad (18)$$

where $S_0(t)$ is the baseline survival function.

4.1.2 Interval-censored semi-parametric PH model

The right-censored Cox PH model uses a partial likelihood technique to estimate regression parameters independently from the baseline hazard function (Cox, 1972). However, for the interval-censored method, the baseline hazard and regression parameters are estimated together. In this section, the foundations of this model and the Finkelstein approach (Finkelstein, 1986) are discussed, summarizing the text from Bogaerts et al. (2017).

Let $[l_i, u_i]$ be the observed intervals for the i^{th} observation in which the survival time is observed. Also let X_i represent the vector of covariates for the i^{th} subject.

Further, let $S(t|X)$ estimate the survival function for a subject given the covariates $\mathbf{X}$.

The likelihood function expressed in terms of survival functions is proportional to:

$$\mathbf{L} = \prod_{i=1}^n \{S(l_i|X_i) - S(u_i|X_i)\}. \quad (19)$$

If it is assumed that

$$S(t) = S_0(t)^{\exp(X'\beta)},$$

then the likelihood function for equation 19 can be written as

$$L(\beta, S_0) = \prod_{i=1}^n \{S_0(l_i)^{\exp(X_i'\beta)} - S_0(u_i)^{\exp(X_i'\beta)}\}, \quad (20)$$

where β represents the p -dimensional vector of regression parameters and $S_0(t)$ represents the baseline survival function. From the above expression, it is apparent that the baseline survival function $S_0(t)$ must be defined. There are different approaches based on whether one assumes parametric, semi-parametric or non-parametric form for the baseline hazard function. Finkelstein's approach is one such semi-parametric approach (Finkelstein, 1986).

Finkelstein (1986) extended the non-parametric survival distribution conceptualized by Turnbull and used equation 20 to take into account interval-censoring.

Let $s_0 = 0 < s_1 < \dots < s_{K+1} = \infty$ be the distinct ordered survival times for the time intervals $[l_i, u_i]$.

Also, let $\alpha_{ik} = I \{s_k \in [l_i, u_i]\}$ ($k = 1, \dots, K+1$, $i = 1, \dots, n$) where I is the indicator function which indicates whether or not an observed time point falls within a specific time interval. Reparameterizing S_0 as $S_0(s_0)=1$ then $\gamma_k = \log\{-\log S_0(s_0)\} = \log\{-\log(1)\} = \infty$. Similarly, when $S_0(s_{K+1})=0$, $\gamma_{K+1} = 0$.

And so, the log-likelihood function can thus be expressed as:

$$l(\beta, \gamma) = \sum_{i=1}^n \log \left[\sum_{k=1}^{K+1} \alpha_{ik} \{e^{-\zeta_{k-1} \exp(X_i' \beta)} - e^{\zeta_k \exp(X_i' \beta)}\} \right], \quad (21)$$

where $\zeta_k = \sum_{m=0}^k \exp(\gamma_m)$.

The unknown parameters such as the regions of mass support and regression parameters in the above equation may be estimated using the Newton-Raphson iteration algorithm as proposed by Finkelstein (1986).

4.1.3 Results

Simple semi-parametric PH models were fitted to compare the time to emergence and time to gametocyte elimination for the different treatment arms based on the three different ways of dealing with interval-censored data. These approaches include: (1) event time is assumed to be an exact time equal to the right endpoint of interval (RI), (2) event time is assumed to be the midpoint of the interval in which the event occurred (MP) and (3) the data is correctly treated as interval-censored (IC). This section aims to compare the results of estimates from models built using the above event times. In this section, only single covariate models are built as the focus is not on exploring the impact of covariates on event times but rather on comparing estimates from the traditional Cox regression models, which assume exact times and the Finkelstein (1986) model, which takes into account interval-censoring. The models fitted is of the form:

$$\log \frac{h_i(t)}{h_0(t)} = \beta \times \text{arm}_i,$$

where $\text{arm}_i = 0$ if treatment SP is administered and $\text{arm}_i = 1$ if treatment is ACT

Table 17 shows the estimated regression coefficient and the standard errors for the gametocyte emergence semi-parametric models, based on the three approaches. For all three approaches, a negative parameter is estimated. This suggests that patients using the alternative treatment ACT generally have a lower “failure rate” than patients on the

Table 17: Gametocyte Emergence semi-parametric PH Models

Model	Coef	HR	Standard error	P-Value
RI Model	-1.002	0.367	0.178	< 0.001
MP Model	-1.001	0.368	0.178	< 0.001
IC Model	-1.014	0.363	0.162	< 0.001

standard treatment SP. In other words, the hazard of gametocyte emergence at any time t , is smaller for a patient using ACT compared to a patient using SP. The regression coefficient under the RI model is equal to -1.002, and the hazard ratio is $0.367 = e^{-1.002}$. This implies that receiving ACT as compared to SP treatment reduces the relative hazard of gametocyte emergence by 63.18% ($= 1 - e^{-1.002}$).

Under the MP model, the regression coefficient estimate is -1.000, and the hazard ratio is 0.368 ($= e^{-1.001}$). This implies that the relative hazard of gametocyte emergence is reduced by 63.20% for those in the ACT group compared to those in the SP group.

The regression coefficient for the IC Cox PH model is -1.014, which means the hazard ratio is 0.363 ($= e^{-1.014}$) which is similar to the result from the Cox PH model for right-censored data. This result implies that patients receiving ACT compared to those using SP reduce their risk of gametocyte emergence by 63.72%.

For all the three methods, the same consensus is reached, in that the ACT course of treatment is the better of the two treatments as it can be associated with a significantly lower hazard of gametocyte emergence than the “status quo” - SP. The differences in estimates and standard errors among all three approaches are minimal. This is because most events occurred in the earlier narrower intervals, where the assumption of when events would have taken place, has very little impact and accounting for the interval-censoring nature of the data may be unnecessary.

Table 18 summarizes the effects of treatment arm on the relative hazard of gametocyte elimination under different mechanisms.

Table 18: Gametocyte Elimination semi-parametric PH Models

Model	Arm vs SP Coef	HR	Standard error	P-Value
RI Model	0.594	1.812	0.190	0.002
MP Model	0.753	2.125	0.195	< 0.0001
IC Model	0.756	2.129	0.260	0.014

Table 18 shows the model estimates for the gametocyte elimination semi-parametric models for all three approaches. From Table 18, the estimated regression coefficient for

the RI Cox Model is 0.594; the observed hazard ratio is 1.812. This can be interpreted as patients receiving ACT have a 81.20% higher hazard of gametocyte clearance than those taking the standard course of treatment, SP.

The same conclusion is reached as well when the MP and IC Cox Models are applied. However, the IC model estimates that subjects receiving ACT increases the hazard of gametocyte elimination by 112.97%, and the MP estimates a 112.50% increase.

Under the IC and MP models, the hazard is considerably higher, which implies recovery is much quicker under the ACT treatment than was estimated when the right point of the interval was assumed as the time to the event. From this, it appears the standard Cox PH model underestimated the results, this may be because the RI approach assumes time to gametocyte clearance is at the end of an interval. Which “delays” time to gametocyte clearance as compared to assuming the event time is interval-censored or in the middle of an interval. Hence, IC and MP approaches assume that the time to event occurs “earlier on in the interval” in comparison to the RI approach.

Which might beg the question, why are the differences in estimates of the gametocyte emergence Cox PH models trivial compared to the differences in estimates among the approaches in estimating gametocyte clearance? This observation is due to the fact that the intervals in which gametocyte clearance occurs are much wider, and the much wider intervals are, the greater the difference in estimates, especially between the IC and RI approaches. Differences in estimates between the IC and MP approach are not considerably different. MP approach is therefore a superior substitute for IC compared to RI. Sun and Chen (2010) also observed that estimates from the Finkelstein non-parametric maximum likelihood estimation method (Finkelstein, 1986) produced better model estimates in comparison to mid-point and right-point imputation methods. Although they used the non-parametric Finkelstein approach Sun and Chen (2010) also confirmed that midpoint imputation approach may be a suitable alternative to Finkelstein’s approach, specifically in circumstances where intervals are narrow or there is an assessment balance.

4.2 Parametric PH models

The Cox PH model (Cox, 1972) is semi-parametric, as it does not assign a distribution function for the survival times and baseline hazard but it does assume that the explanatory variables take up a parametric form (Bogaerts et al., 2017). The Cox PH model is also highly favoured over parametric models because of the fewer assumptions. (Pourhoseingholi et al., 2007). That is, the Cox PH model does not assume a distribution for the baseline hazard. By contrast, parametric methods specify a distribution for the baseline hazard. However, there are some advantages to using a parametric model; for one, parametric models give more interpretable (easily explicable) results, which leads to better analysis (Scolas, 2016). Secondly, some authors suggest that under the right circumstances parametric models are more efficient in estimating parameters as compared to Cox models (Efron, 1977).

4.2.1 Regression models

Two popular parametric PH models are considered, the Weibull parametric model and the Exponential parametric model. These models provide alternative functions for baseline hazards defined over $0 \leq t \leq \infty$.

Concerning the Weibull model, the hazard function is dependant on parameters λ and ρ . Where λ is the scale parameter and ρ is the shape parameter, which are both strictly positive. Below are the exponential and Weibull survival functions.

Table 19: Exponential and Weibull underlying 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	$\lambda \rho t^{\rho-1}$

From the above table, it appears that the Weibull model simplifies to the Exponential model when $\rho = 1$. Now, with the exponential model, it is assumed that the baseline hazard is constant. Thus, the hazard of an event of interest is constant over time.

In contrast, the Weibull model assumes a monotone baseline hazard, which is equal to 0 at the start of the study. This steadily increases over time when $\rho > 1$ until the maximum is reached, after that it decreases steadily when $\rho < 1$ approaching 0 as t nears ∞ . In Table 19 the Weibull baseline hazard is given as:

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

Therefore, the proportional hazards Weibull model can be defined as:

$$\begin{aligned} h_i(t) &= h_0(t) \exp(x_i^T \beta) \\ &= \lambda \rho t^{\rho-1} \exp(x_i^T \beta), \end{aligned} \tag{22}$$

where λ is the scale parameter and ρ the shape parameter. The survival function may be expressed as:

$$S_i(t_i) = \exp\{-\lambda t^\rho \exp(X^T \beta)\}, \tag{23}$$

and the likelihood for the Weibull proportional hazards model is defined as:

$$\begin{aligned}
 L(\beta, \rho, \lambda) &= \prod_{i=1}^n \{h_i(t)\}^{\delta_i} \{S_i(t_i)\} \\
 &= \prod_{i=1}^n \{\lambda \rho t_i^{\rho-1} \exp(x_i^T \beta)\}^{\delta_i} \exp\{-\lambda t_i^\rho \exp(x_i^T \beta)\},
 \end{aligned} \tag{24}$$

with δ_i , as the indicator variable, where 1 represents the event occurring and 0 otherwise.

To construct the Weibull proportional hazard model likelihood function, the parameters β , ρ and λ need to be estimated. This is done using the Newton-Raphson algorithm and maximising the log likelihood function defined as:

$$\begin{aligned}
 l(\beta, \rho, \lambda) &= \sum_{i=1}^n [\delta_i \log\{h_i(t_i)\} + \log S_i(t_i)] \\
 &= \sum_{i=1}^n [\delta_i \{x_i^T \beta + \log(\lambda \rho) + (\rho - 1) \log t_i\} - \lambda \exp(x_i^T \beta) t_i^\rho].
 \end{aligned} \tag{25}$$

If the researcher is instead more interested in running an exponential model. The same steps used to derive the Weibull model may also be used by fixing ρ to the value of 1.

Unlike the PH model applied to right-censored survival times, the PH model for interval-censored data requires both the baseline hazard and regression parameters to be estimated (Bogaerts et al., 2017). Using the likelihood function given in equation 20,

$$L(\beta, S_0) = \prod_{i=1}^n \{S_0(l_i)^{\exp(X_i' \beta)} - S_0(u_i)^{\exp(X_i' \beta)}\}.$$

As shown in the above equation survival functions are included in the model. How this survival function looks is dependent on which of the error term distribution was selected for analyses (Gomez et al., 2009). Alternatively, the above function may be written in terms of density functions such that the likelihood function may be given as:

$$l(\beta, f(t)) = \sum_{i=1}^n \int_{l_i}^{u_i} w(s; \theta, Z_i) ds,$$

as inspired by Gomez et al. (2009). Where θ is the vector of unknown model parameters, Z_i is the covariate vector and $w(s; \theta, Z_i)$ is the density function. Various distributions as presented by Bogaerts et al. (2017) include the following:

exponential distribution

$$f(t; \alpha) = \frac{1}{\alpha} \exp\left(-\frac{t}{\alpha}\right), \quad (26)$$

$$h(t; \alpha) = \frac{1}{\alpha}.$$

Weibull distribution

$$f(t; \alpha; \gamma) = \frac{\gamma}{\alpha} \left(\frac{t}{\alpha}\right)^{\gamma-1} \exp\left\{-\left(\frac{t}{\alpha}\right)^\gamma\right\}, \quad (27)$$

$$h(t; \alpha; \gamma) = \frac{\gamma}{\alpha} \left(\frac{t}{\alpha}\right)^{\gamma-1}.$$

log-normal distribution

$$f(t; \alpha; \gamma) = \frac{\gamma}{t\sqrt{2\pi}} \exp\left[-\frac{\{\gamma \log(\frac{t}{\alpha})\}^2}{2}\right], \quad (28)$$

$$h(t; \alpha; \gamma) = \frac{\gamma}{t\{1 - \Phi(\gamma \log(\frac{t}{\alpha}))\}\sqrt{2\pi}} \exp\left[-\frac{\{\gamma \log(\frac{t}{\alpha})\}^2}{2}\right].$$

log-logistic distribution

$$f(t; \alpha; \gamma) = \frac{\frac{\gamma}{\alpha}(\frac{t}{\alpha})^{\gamma-1}}{\{1 + (\frac{t}{\alpha})^\gamma\}^2}, \quad (29)$$

$$h(t; \alpha; \gamma) = \frac{\frac{\gamma}{\alpha}(\frac{t}{\alpha})^{\gamma-1}}{\{1 + (\frac{t}{\alpha})^\gamma\}}.$$

4.2.2 Results

In this section, the main objective will be on comparing the parametric hazard regression of right-censored data against an interval-censored parametric hazard regression.

The second objective is to assess how the ACT treatment may influence the emergence of gametocytemia. Therefore, a parametric proportional hazard model is used. For the analysis of gametocyte elimination, an AFT model is used to investigate how the use of the ACT treatment can either speed up or slow down recovery. This will be discussed in section 4.3.2.

Table 20 found below, shows the results of the analysis based on the Weibull and exponential distributions. For comparative reasons the analysis will begin by looking at each Weibull an exponential model separately based on the three approaches. Subsequently, the estimates of the three models will be compared.

Table 20: Gametocyte Emergence Parametric PH Models

Model	<i>log_shape</i>	<i>log_scale</i>	ACT vs SP	exp(Estimate)	P-Value
Weibull					
RI Model	-1.310	6.769	-0.953	0.386	< 0.001
MP model	-0.687	5.095	-1.072	0.342	< 0.001
IC model	-0.894	5.527	-1.041	0.353	< 0.001
Exponential					
RI Model		4.314	-1.187	0.305	< 0.001
MP model		4.261	-1.220	0.295	< 0.001
IC model		4.285	-1.222	0.295	< 0.001

Seeing as $\beta < 1$ for all models regardless of approach used, it implies that the patients taking the alternative course of treatment, ACT, have a lower risk of having gametocytemia. From Table 20, the arm coefficient of the RI Weibull PH model is -0.953, and the hazard of gametocyte emergence when using ACT is reduced by 61.4% ($= 1 - \exp(-0.953)$). While the MP Weibull model estimates the coefficient to be -1.072 implying that with ACT reduces the gametocyte emergence hazard by 65.8% ($= 1 - \exp(-1.072)$). From the same table, the IC model estimated that the arm coefficient is -1.041 which implies that the receiving ACT reduces the hazard of gametocyte emergence by 64.7% ($= 1 - \exp(-1.041)$).

The estimated coefficient regressions produced by the three models are marginally different with all three models giving similar conclusions (regardless of the distribution used) that the ACT treatment is preventative or at-least reduces the risk of gametocytemia. But it is evident that the RI overestimates arm coefficient but underestimates the effect of ACT on gametocyte emergence under both distributions.

4.3 Accelerated failure time models

An alternative metric for the parametric survival models to the PH metric is that of accelerated failure times (AFT). AFT models are appropriate when the focus is on estimation of time rather than the relative hazard, as well as instances where the parametric and semi-parametric PH models do not satisfy the proportionality assumption.

4.3.1 Right censored accelerated failure time models

Suppose patients are randomly assigned to two groups. In the context of this dissertation, let one group be those who receive the standard treatment, SP and the second group, those who receive the alternative treatment, ACT.

In terms of the AFT model, the use of the alternative treatment can either speed up or slow down the progression rate along the time of passage. Similarly, the probability that an individual using ACT survives beyond time t is the probability that an individual using the SP treatment survives beyond time βt .

β is an unknown non-negative constant, which is known as the acceleration factor. For instance, when the event of interest is the emergence of gametocytes, values of β that are less than 1, then the effect of the covariate is accelerated.

If, however, $\beta > 1$, this would entail that the covariate slows down the advancement of the event of interest over time.

An individual with a covariate vector given by X the hazard function is given as

$$h_i(t) = h_0(\exp(-\beta x_i)t) \exp(-\beta x_i), \quad (30)$$

and the survival distribution:

$$S_i(t) = S_0(\exp(-\beta x_i)t). \quad (31)$$

Further for the random variable T (the time to event),

$$T = \exp(\beta x_i)T_0. \quad (32)$$

Given this relationship, the log-linear AFT model proposed by Collett (1994) is given as:

$$\log(T_i) = \alpha_0 + \beta x_i + \sigma \epsilon_i, \quad (33)$$

where α_0 is the intercept parameter is the value of $\log(T_i)$ when $x_i = 0$ and β is the unknown regression coefficient which measures the impact of $x_i = 1$ rather than $x_i = 0$, x_{ji} is the corresponding value of the j^{th} independent variable, X_j , $j = 1, 2, \dots, k$, for the i^{th} individual, $i = 1, 2, \dots, n$, ϵ is the residual variable that measures the difference between the modelled and observed times on the log scale, which is scaled by σ the scale parameter, and is assumed to have a particular probability distribution.

4.3.2 Accelerated failure time models for interval-censored data

In this section, the AFT model for interval-censored data is discussed following the text from Bogaerts et al. (2017), as well as the method that will be used in this dissertation to estimate the parameters of the AFT model.

Once again, let the observed survival data be such that the response times $T_1 \dots T_n$ are such that the survival event times are mainly known to occur within intervals $[l_i, u_i]$, with event indicator variable δ_i for $(i = 1, 2, \dots, n)$.

With covariate vector from each subject/individual i and n random variables. The general expression of the AFT model in the logarithmic form of the survival time T is again given by:

$$\log(T) = \alpha_0 + X'\beta + \sigma\epsilon.$$

While the model is linear, and most parameters from linear models are estimated using the Least Squares Method, this would not be suitable in the case of interval-censored data (Bogaerts et al., 2017).

Instead, the approach used for estimation is the general maximum likelihood theory.

From the above equation, the error term is a vital component, which throughout this section is deemed to be identical and independent, density $g(e)$. This error term, comes with an unknown location parameter $\mu \in R$ and scale $\sigma > 0$. The error term may be standardized so that the standardized error term is given as ϵ^* and the density is $g^*(\epsilon^*)$. The choice of which error term is based on the information criteria, AIC and BIC. The distribution that produces the lowest values of these statistics will be selected. Common error distributions include: the standard normal distribution which has a log-normal survival time distribution; the standard logistic error distribution which has a log-logistic survival time distribution and the extreme value error distribution which is connected to the Weibull survival time distribution.

As for the parameters, these are estimated using the maximum likelihood approach. The likelihood for the AFT interval-censored model is given by:

$$L(\Theta) = \prod_{i=1}^n L_i(\Theta) \tag{34}$$

Where $L_i(\Theta)$ integrates between the interval limits $[l_i, u_i]$ so that $L_i(\Theta)$ is given as:

$$L(\Theta) = \oint_{l_i}^{u_i} f_i(t; \theta) dt$$

and $f_i(t; \theta)$:

$$f_i(t; \theta) = \frac{1}{\sigma t} g^* \left(\frac{\log(t) - \mu - X'_i \beta}{\sigma} \right)$$

4.3.3 Results

Time to gametocyte elimination may be used to calculate the duration of gametocytemia explicitly. For this reason, AFT models were fitted as they are capable of directly estimating time. However, this section's main focus is the comparison between the standard methods, using - RI and MP approaches and the IC methods. AFT models based on different distributions were fitted with only one covariate included in the analysis - treatment arm. To calculate duration, only subjects who experienced gametocytemia may be included. As mentioned in the previous chapters, of the 391 subjects included in the initial study, only 145 experienced gametocytemia. Those excluded were an example of left truncation.

Table 21: Gametocyte Elimination AFT Models

Weibull					
Model	$\log_shape$	$\log_scale$	ACT vs SP	exp(Estimate)	P-Value
RI Model	0.354	2.825	-0.453	0.640	< 0.001
MP model	0.268	2.451	-0.627	0.534	< 0.001
IC model	0.189	2.400	-0.664	0.515	< 0.001
Log-logistic					
Model	$\log_alpha$	$\log_beta$	ACT vs SP	exp(Estimate)	P-Value
RI Model	2.448	0.858 3	-0.512	0.599	< 0.001
MP model	2.033	0.727	-0.642	0.526	< 0.001
IC model	1.955	0.658	-0.792	0.453	< 0.001
Log-normal					
Model	μ	$\log_s$	ACT vs SP	exp(Estimate)	P-Value
RI Model	2.457	-0.333	-0.461	0.631	< 0.001
MP model	2.058	-0.217	-0.572	0.565	< 0.001
IC model	1.964	-0.120	-0.719	0.487	< 0.001

The gametocyte elimination AFT model using the Weibull, log-logistic or log-normal distribution results are shown in Table 21. For comparative sake we discuss the Weibull AFT models fitted based on the different approaches. The regression arm coefficient for the RI method, is given as -0.492, and the acceleration factor for a patient taking the ACT treatment relative to those under the SP treatment is 0.611. This, implies that individuals who are taking the ACT treatment have a faster recovery rate than those taking the SP course of treatment.

The same conclusion is reached for the other approaches. However, the IC and MP Weibull

models' acceleration factors are much closer together than the acceleration factor from the RI Weibull AFT model. Although the difference in the regression coefficient is not very large, it does show that incorrectly assuming that the event time occurs at the end of an interval underestimates the effect of the ACT treatment. Furthermore, if a researcher prefers to use the traditional survival methods, then MP is the better alternative. But the question becomes how much does using RI and MP "cost" us in terms of accuracy when it comes to estimating not just parameter estimates but quantities, such as duration, from these fitted models?

Table 22: Duration estimates and standard errors based on treatment arm

Arm	RI		MP		IC	
	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
ACT	12.415	1.358	7.593	0.901	7.038	0.964
SP	19.523	1.510	14.208	1.202	13.669	1.285

To further illustrate the impact of ignoring interval-censoring Table 22 shows the estimates of duration using the different approaches and based on the treatments administered. While all three approaches agree that ACT reduces duration of gametocytemia it is apparent that the RI approach grossly overestimates duration for both treatments. The estimates from the MP and IC approaches are much lower and closer in estimation.

5 Building survival models for interval-censored data

The previous chapter focused on the differences in model estimates between models fitted using exact time points such as the right end point or the midpoint of an interval versus models that correctly accounted for interval-censoring. The analysis revealed discrepancies amongst the estimates of those models. Estimates made from the gametocyte emergence survival data resulted in minor differences. While the differences in estimates were amplified, in the case of the time to gametocyte clearance. The results suggested that where time intervals in which events occurred were wide resulted in the estimates amongst the different approaches varying relatively significantly. It was further observed that the results from the model estimates that assumed the information was right imputed were extremely biased. From the analysis of time to gametocyte clearance we could therefore see the impact of ignoring the presence of interval-censoring. Therefore, subsequent models will be built using the interval-censored methodology introduced in chapter 4 for further analysis. This chapter will extend the previous models that looked at the association between treatment arm and (1) time to gametocyte emergence and (2) time to gametocyte elimination to include potential predictors referred to in Chapter 1.

5.1 Model fitting approach

The models of interest look at investigating factors that influence gametocyte emergence and which covariates influence the time until gametocytes clear from the host's system. Given the analysis that will be done proportional hazard (PH) models will be fitted to model gametocyte emergence and AFT models will be fitted for gametocyte elimination. However, before we can start making inferences and conclusions we need to decide on an approach. The first step involves deciding on which distribution of the PH model or AFT model best suits the data. This will require comparing models of different distributions using statistics and graphical procedures to assess the goodness of fit of the model. Statistics used to assess model fit include the:

$$AIC = -2\log(\text{likelihood}) + 2(k)$$
$$BIC = -2\log(\text{likelihood}) + k\log(N),$$

where k is the number of parameters, and N is the sample size (Bogaerts et al., 2017). Lower values for either the AIC or BIC indicate better model fit.

For the gametocyte emergence PH models a further step will be conducted. In addition to using the AIC statistic to determine which distributions should be used, a decision on whether a semi-parametric model or parametric model must be made. While the semi-parametric model is the classic choice (Kosorok, 2009). A parametric model may still be chosen. To assess the model fit, the semi parametric model is made useful in examining the model fit. This is done by plotting the parametric baseline distributions under consideration against the semi-parametric fit (Anderson-Bergman, 2017). In R (R Core Team, 2019), the *icenReg* package developed by Anderson-Bergman (2017) uses the

function **diag_baseline** to visualize this diagnostic plot which uses the $\log[-\log\hat{S}(t|X)]$ transformation to assess the fit of different parametric baseline models. A parametric model that is best suited to the data will closely follow the the semi-parametric estimate in this graphical plot (Anderson-Bergman, 2017). This implies that the chosen parametric model is suitable in describing the underlying distribution.

After it has been decided on which distribution the PH and AFT models should be based on, the next step of the model building approach decides on which variables to remove from the model. To assist in deciding on whether to keep or discard a variable from the model the deviance test will be used. This test is commonly used when we want to compare nested models (models that use a subset of explanatory variables from a larger model) to assess if the reduced model is adequate or not, where a small p-value indicates inadequacy. The deviance test statistic is calculated as the difference between the log-likelihood functions of the two models being compared. So that the deviance statistic is given as:

$$D = -2\{\log l_1 - \log l_2\},$$

where $\log l_1$ and $\log l_2$ are the log likelihoods of the nested models m_1 and m_2 . D , the deviance statistic follows the χ^2 distribution with d degrees of freedom (dof), where the dof are equal to the difference in the number of parameters of the two models.

After a model has been built and refined it is necessary to verify the validity of the model. For the PH models we need to check whether the covariates satisfy the PH assumption. The **diag_covar** function in R (R Core Team, 2019) also part of the *icenReg* package developed by Anderson-Bergman (2017) is used. The function uses the fact that the difference between the strata's transformed baseline survival functions given as:

$$\log(-\log\hat{S}(t|X, \beta)) = \log(-\log(S_0(t)^{\exp(X\beta)})) = X\beta + \log(-\log(S_0(t))),$$

as stated by Anderson-Bergman (2017) which should be constant if the PH is valid. To assess if a model is correct, the function stratifies covariates across each covariate's different factor levels. To capture the difference, the overall average of the strata is subtracted off each stratum by default (Anderson-Bergman, 2017). Because each curve is being subtracted by the overall mean, it is expected that when the suitable parametric model is used, each curve is a straight line given some stochastic noise (Anderson-Bergman, 2017). Visually this means that the transformation of survival curves will be approximately parallel. If, however, the assumption does not hold, then the curves will either converge or cross each other.

It should be noted that the curves may cross by chance despite the PH assumption being satisfied (Bogaerts et al., 2017). In the event, there are gross violations of the PH assumption, possible solutions include fitting stratified models, time-varying coefficients, and fitting models that use an accelerated time metric instead of the proportional hazards metric.

It should be brought to the reader's attention that the diagnostic tools that use the R functions `diag_covar` and `diag_baseline` cannot be applied to AFT models. Anderson-Bergman (2017) states that the reason for this is that currently there is no method that estimates the AFT semi-parametric model.

This section will proceed in the following manner. Models for the time to gametocyte emergence will be fitted to assess which covariates might make a patient more susceptible to hosting the transmissible parasite. Subsequently, an analysis of time to gametocyte clearance and finally an investigation into duration of gametocytemia and which covariate patterns might lengthen duration.

5.1.1 Time to gametocyte emergence

In this section, the focus will be on assessing the impact of factors on the time to emergence. This analysis will help determine which characteristics could induce gametocyte emergence. Therefore, proportional hazards (PH) models will be used.

The model building process begins by deciding whether a semi-parametric PH model or a parametric PH model should be used. Figure 18 shows the plots of the Weibull and exponential distributions against the semi-parametric model. From these figures, the most appropriate parametric model between the Weibull and exponential is the Weibull model, since the Weibull deviates the least from the semi-parametric model.

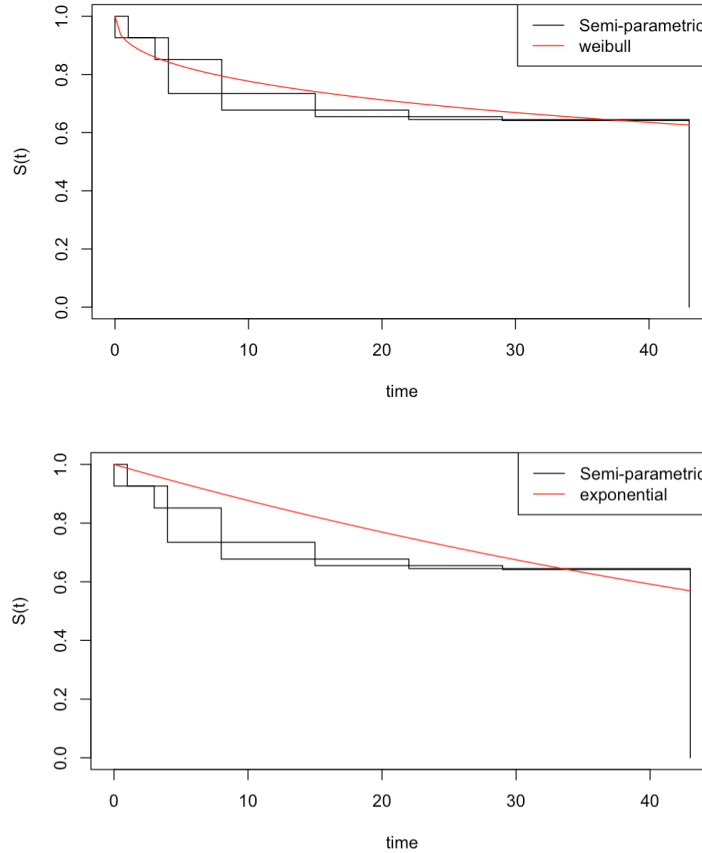


Figure 18: Examining Baseline Weibull and Exponential Distribution Fits

However, an evaluation of Table 23, the model with the lowest AIC statistic is the Cox model, indicating a better model fit. While the Weibull model was the better choice between the parametric models, the Cox model is carried forward for further analysis and refined by systematically removing the non-significant covariates.

Because we are comparing nested models, the deviance test is used to decide whether it is worth removing a covariate or not. Table 24 shows the order in which non-significant covariates were excluded from the Cox model. The selection was based on backward selection, deleting covariates with the highest p-value in the model. Note that once a covariate is deleted, it cannot re-enter the model again. In addition to the above, the deviance test was used to test the impact of the deletion of variables in the goodness of fit of a model. It appears that there was only a considerable difference between models C4 and C5 which implies deleting p_{zero} from the model will potentially reduce the model fit (p-value = 0.064). Therefore, based on the model fitting results, the best-fitted model is model C4.

Table 23: Parameter estimates and 95% C.I. for time to emergence PH models

Definition	Parameter	Cox	Weibull	Exponential
log_shape	$\log \lambda$		-0.858 (-1.028; -0.687)	
log_scale	$\log \rho$		5.549 (4.967; 6.131)	4.333 (4.148; 4.518)
logage	β_1	-0.662 (-1.135; -0.189)	-0.648 (-1.132; -0.165)	-0.804 (-1.292; -0.316)
mutat	β_2	0.250 (-0.169; 0.670)	0.232 (-0.151; 0.614)	0.364 (-0.018; 0.745)
gender	β_4	0.045 (-0.289; 0.379)	0.085 (-0.254; 0.425)	0.152 (-0.185; 0.489)
arm	β_5	-0.984 (-1.326; -0.643)	-1.029 (-1.381; -0.677)	-1.275 (-1.625; -0.925)
anemia	β_6	0.256 (-0.114; 0.626)	0.255 (-0.112; 0.622)	0.295 (-0.072; 0.662)
pzero	β_7	0.110 (-0.045; 0.265)	0.118 (-0.025; 0.260)	0.181 (0.041; 0.321)
AIC		903.767	963.054	1104.65

Table 24: Parameter estimates and 95% C.I. for nested time to emergence Cox PH models

Definition	Parameter	C1	C2	C3	C4	C5
logage	β_1	-0.662 (-1.303; -0.136)	-0.676 (-1.189; -0.262)	-0.805 (-1.298; -0.312)	-0.788 (-1.218; -0.359)	-0.829 (-1.292; -0.366)
mutat	β_2	0.250 (-0.181; 0.667)	0.253 (-0.066; 0.556)	0.251 (-0.112; 0.614)		
gender	β_4	0.045 (-0.294; 0.344)				
arm	β_5	-0.984 (-1.361; -0.604)	-0.983 (-0.140; 0.510)	-0.982 (-1.377; -0.588)	-0.992 (-1.324; -0.659)	-0.987 (-1.349; -0.625)
anemia	β_6	0.256 (-0.127; 0.488)	0.248 (-0.167; 0.523)			
pzero	β_7	0.110 (-0.030; 0.260)	0.112 (-0.035; 0.267)	0.122 (-0.010; 0.272)	0.131 (-0.010; 0.272)	
AIC		903.767	901.833	901.440	901.238	902.664
Deviance test statistic			0.066	1.607	1.798	3.426
Deviance p-value			0.797	0.205	0.180	0.064

The model C_4 can thus be interpreted as follows:

- For every one-year increase in age, the hazard of gametocyte emergence increases by 54.5% ($= 1 - \exp(-0.788)$)
- patients receiving the combinatory alternative treatment of artesunate and sulfadoxine-pyrimethamine reduce the risk of gametocyte emergence by 62.89% ($= 1 - \exp(-0.992)$) compared to patients using the sulfadoxine-pyrimethamine treatment only.
- Every ten-fold increase in the baseline asexual parasite density results in a 14.00% ($= \exp(0.131) - 1$) increase in risk of gametocyte emergence.

Next step requires model validation. The plots shown below in Figure 19 are based on model C_4 to test the validity of the PH assumption of the overall fit. A closer look at the sub-figures, it appears that the covariates *logage* and *arm* strata are largely parallel. And for the covariate *pzero* the differences in the earlier days is significant, but the differences narrow, and the strata cross but eventually become parallel and remain constant as time goes on. The crossings of these two strata suggest that the proportional hazards assumption has been violated. And from the diagram there appears to be some trend in the earlier times that then tapers off.

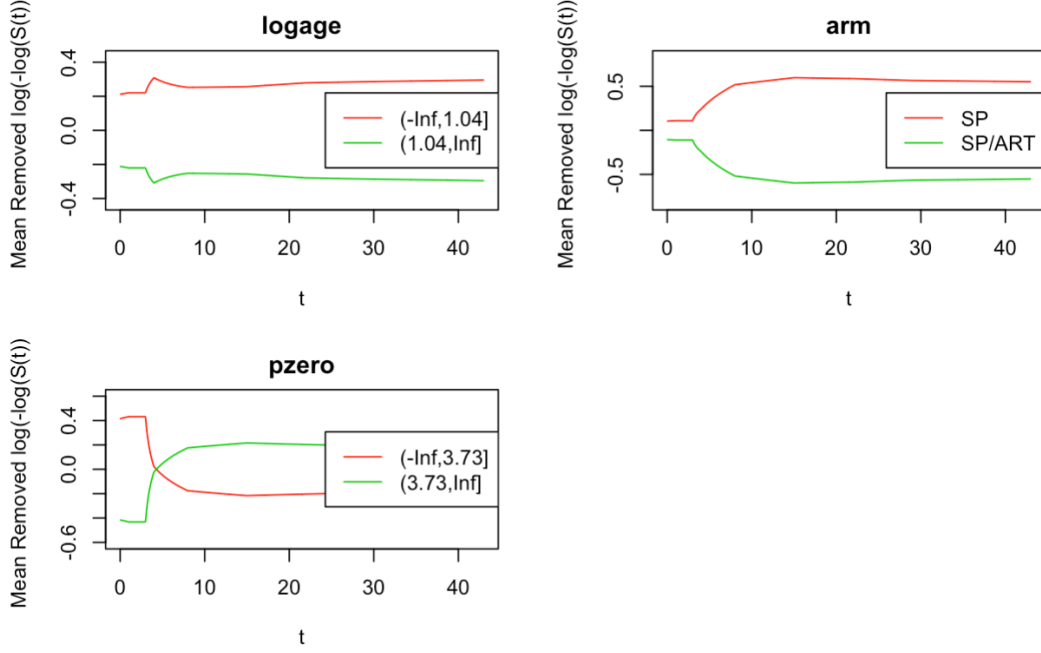


Figure 19: PH Assumption Validity

In instances where the assumption of proportional hazards is violated there are three possible solutions to fix this. One option includes fitting stratified models. These models produce different forms of the baseline hazard for each stratum such that our model is given by:

$$h_{ij}(t) = h_{0j}(t) \exp(\beta_1 x_{i1} + \dots + \beta_p x_{ip}),$$

where h_{0j} is the baseline hazard for the j^{th} stratum. From Figure 19 only the *pzero* covariate appears not to satisfy the assumption. One method that might help solve this problem, would be to fit a model that stratifies the covariate *pzero*. However, stratification does not allow us to interpret and comment (Little, 2013) on the effect of *pzero* on time to gametocyte emergence.

Alternatively, we could modify the model by using the time variable as the dependent variable instead of the hazard function by using AFT models. However, our interest is in finding which factors have an impact on proportional hazard, so this will not do.

Instead, to ensure the assumption is met, an interaction term will be introduced into the model. As previously mentioned, there appears to be some trend in the *pzero* covariate in the earlier time intervals, indicating a possible interaction between *pzero* and time. To capture this relationship *pzero* is converted to a time-varying covariate by interacting the baseline asexual density with time. The interaction term is created by multiplying *pzero* by the indicator variable which is equal to 1 when time to gametocyte emergence is greater than 4 and 0 otherwise.

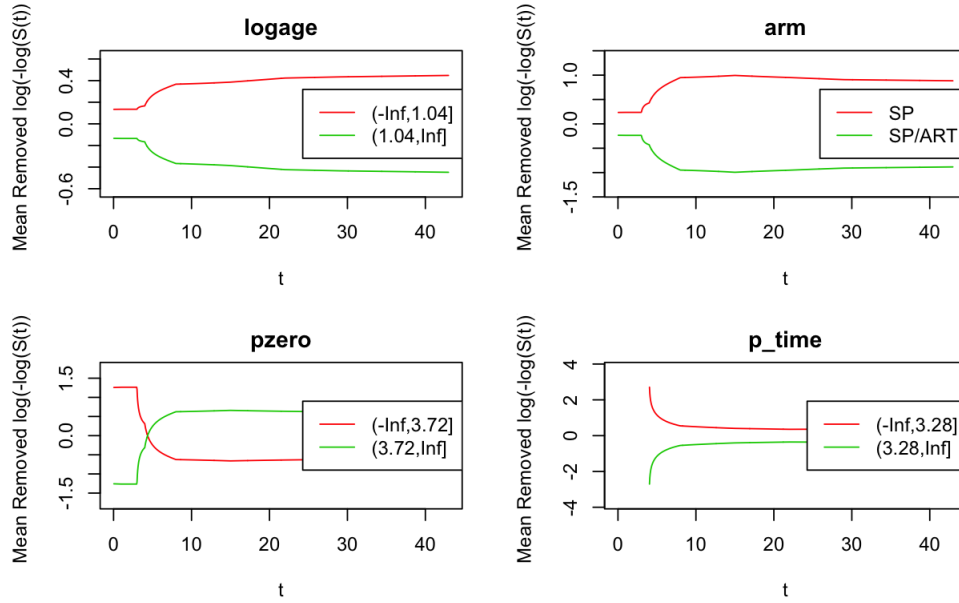


Figure 20: PH Assumption Validity with interaction

Figure 20 shows the test for proportional hazards. Table 25 shows the estimates of the final model *Modified2*. The effects of the covariates appear to be similar to the previous estimates. *Pzero* and *arm* are still significant including the interaction term given by *pzero* $\times$ indicator variable for time. However, *logage* is no longer a significant factor.

The final modified model *Modified2* can be interpreted as follows:

- Patients receiving the combinatory alternative treatment of artesunate and sulfadoxine-pyrimethamine reduce the risk of gametocyte emergence by 56.70% ($= 1 - \exp(-0.837)$) compared to patients using the sulfadoxine-pyrimethamine treatment only.
- Every ten-fold increase in the baseline asexual parasite density results in a 147% ($\exp(0.905) - 1$) increase in risk of gametocyte emergence.
- This is the modification of *pzero* effect for later time periods, it reduces the effect by 65% to $0.88 = \exp(0.905 - 1.032)$.

Table 25: Revised Parameter estimates and 95% C.I. for time to emergence PH models

Definition	Parameter	Modified1	Modified2
logage	β_1	-0.184 (-0.724, 0.357)	
arm	β_5	-0.831 (-1.205, -0.456)	-0.837 (-1.264 -0.409)
pzero	β_7	0.891 (0.652, 1.129)	0.905 (0.617 1.193)
pzero $\times$ time	β_8	-1.021 (-1.157, -0.885)	-1.032 (-1.201 -0.864)
AIC		696.009	694.560
Deviance test statistic			0.551
Deviance p-value			0.965

5.1.2 Time to gametocyte clearance

Time to gametocyte clearance is defined as the time from gametocyte emergence, the starting point (assuming it is an exact time) until time to gametocyte clearance. This time to event process is also a building block towards estimating duration which will be done in the subsequent section. In this section, we are interested in seeing how the time until gametocytes have cleared from the host's system might be affected by covariates. Since we are focusing on the time variable as a response AFT models will be used. These AFT models include the Weibull, log-normal and log-logistic models.

As previously mentioned, the analysis for time to gametocyte clearance will only include individuals that experience gametocytemia. This is because, firstly for one to experience gametocyte clearance it is on the precondition that they had experienced an emergence of gametocytes in the first place. And secondly, the inclusion of those who did not experience gametocytemia will distort the results.

Table 26 shows the estimates of the AFT models built using the interval-censored data. For AFT interval-censored models there is unfortunately no present function available in the R software to plot the different distributions to decide on the best parametric model. Therefore, we can only rely on the AIC to decide on which parametric model best describes the data. Based on the results from Table 26 the Log-normal model has the lowest AIC.

Table 26: Parameter estimates and 95% C.I. for time to clearance AFT models

Definition	Parameter	Weibull	Log-normal	Log-logistic
log_shape	$\log\lambda$	0.287 (0.141; 0.433)		
log_scale	$\log\rho$	2.372 (2.227; 2.516)		
log_alpha	$\log$			1.974 (1.826; 2.123)
log_beta	$\log\beta$			0.810 (0.638; 0.981)
mu	μ		1.979 (1.829; 2.129)	
log_s	$\log s$		-0.239 (-0.390; -0.089)	
logage	β_1	-0.255 (-0.623; -0.112)	-0.300 (-0.727 ; -0.127)	-0.334 (-0.773; -0.106)
mutat	β_2	0.256 (-0.095; 0.607)	0.296 (-0.061; 0.652)	0.308 (-0.038; 0.653)
gender	β_4	0.043 (-0.277; 0.364)	0.199 (-0.130; 0.527)	0.240 (-0.082 ; 0.562)
arm	β_5	-0.662 (-0.977; -0.347)	-0.641 (-0.970; -0.312)	-0.667 (-0.995; -0.339)
anemia	β_6	0.264 (-0.051; 0.580)	0.155 (-0.189; 0.499)	0.147 (-0.195; 0.490)
pzero	β_7	0.226 (0.085; 0.367)	0.232 (0.091; 0.373)	0.243 (0.104; 0.382)
AIC		354.283	338.029	339.512

Table 27 shown below shows the parameter estimates for the nested time to clearance AFT models as each non-significant variable is removed. The AIC in conjunction with the deviance statistics are used to decide upon the final model. In Table 27 model L4 has the lowest AIC. However, we choose to interpret model L3 that contains the mutation variables in addition to $\log_{10}(\text{age})$, treatment arm and pzero. Since the deviance p-value for L4 vs L3 ($p=0.084$) is not statistically significant at the 0.05 level, indicating that there is not enough evidence to support the removal of the mutat covariate. Model *L3* is thus interpreted as follows:

- The time to gametocyte clearance is accelerated by 36.23% ($= 1 - \exp(-0.450)$) for every ten-fold increase in age.
- Patients who do have a quintuple mutations are likely to see the deceleration of time to gametocyte clearance by 36.50% ($= \exp(0.311) - 1$) compared to those who do not have such a mutation
- Patients receiving the combinatory alternative treatment of artesunate and sulfadoxine-pyrimethamine compared to those using sulfadoxine-pyrimethamine only, accelerates the time to gametocyte clearance by 46.31% ($= 1 - \exp(-0.622)$)
- the time to gametocyte clearance decelerates by 28.40% ($= \exp(0.250) - 1$) for every ten-fold increase in the baseline asexual parasite density.

Table 27: Parameter estimates and 95% C.I. for nested time to clearance AFT models

Definition	Parameter	L1	L2	L3	L4
mu	$\log$	1.979 (1.829; 2.129)	1.978 (1.828; 2.129)	1.977 (1.826; 2.128)	1.974 (1.822; 2.127)
log_s	$\log\beta$	-0.239 (-0.390; -0.089)	-0.234 (-0.384; -0.084)	-0.230 (-0.380; -0.079)	-0.218 (-0.367; -0.068)
logage	β_1	-0.300 (-0.727; 0.127)	-0.392 (-0.770; -0.016)	-0.450 (-0.814; -0.086)	-0.447 (-0.814; -0.080)
mutat	β_2	0.296 (-0.061; 0.652)	0.276 (-0.079; 0.631)	0.311 (-0.040; 0.662)	
gender	β_4	0.199 (-0.130; 0.527)	0.177 (-0.149; 0.503)		
arm	β_5	-0.641 (-0.970; -0.312)	-0.626 (-0.954; -0.297)	-0.622 (-0.952; -0.292)	-0.625 (-0.958; -0.291)
anemia	β_6	0.155 (-0.189; 0.499)			
pzero	β_7	0.238 (0.091; 0.373)	0.236 (0.094; 0.377)	0.250 (0.110; 0.390)	0.273 (0.135; 0.412)
AIC		338.029	336.808	335.933	336.924
Deviance test statistic			(L2 vs L1) 0.778	(L3 vs L2) 1.126	(L4 vs L3) 2.991
Deviance p-value			(L2 vs L1) 0.378	(L3 vs L2) 0.289	(L4 vs L3) 0.084

5.1.3 Duration

In clinical studies, the impact of covariates on the duration of a disease is often of interest. Duration in our context refers to the length of time a patient experiences gametocytemia. This would be from the time we first observe gametocytes until the time they clear from the host's system. To estimate duration, we would need to have the survival time as the response variable instead of the hazard function. In the previous section, we used AFT models to model the time to gametocyte clearance. In this section, we will utilize the same model developed for time to gametocyte clearance, model L3, to estimate duration of gametocytemia for the subset of subjects who had gametocyte recordings during the 42-day study period.

The final model, model L3, showed that the most significant covariates include *mutat*; *pzero*; *arm* and *logage*. Using final model L3 the predicted duration estimates are shown in Table 28 we can compare patients who:

- use the combination treatment of artesunate and sulfadoxine-pyrimethamine (ACT) to patients who use sulfadoxine-pyrimethamine treatment only (SP).
- have “high” $\log_{10}$ baseline asexual parasite density to patients who have a “low” baseline asexual density
- have “high” $\log_{10}$ age in years to patients who have a “low” age in years
- have quintuple mutations to patients who do not have

For the continuous covariates $\log_{10}$ age and *pzero*, the value used as the proxy for “high” is equal to the value of the 97.5 quantiles of the covariate and “low” is based on the 2.5 quantiles. For the *logage* covariate, the values are 1.731 and 0.301 respectively. And the values used for the *pzero* variable are 5.186 for “high” and 1.348 for “low”.

Table 28: Estimated Duration of gametocytemia and Standard Errors

Pzero	Mutat	Arm	Logage	Estimated Duration	Standard Error
High	0	ACT	High	4.418	1.155
			Low	8.405	1.786
		SP	High	8.228	1.825
			Low	15.652	2.718
	1	ACT	High	6.030	1.670
			Low	11.471	2.700
		SP	High	11.229	2.684
			Low	21.362	4.260
Low	0	ACT	High	1.693	0.408
			Low	3.220	0.786
		SP	High	3.152	0.731
			Low	5.997	1.459
	1	ACT	High	2.310	0.673
			Low	4.395	1.304
		SP	High	4.302	1.217
			Low	8.184	2.416

From Table 28, the covariate patterns that result in the longest duration of gametocytemia appear to be patients:

- who are of a relatively younger age
- using the sulfadoxine-pyrimethamine treatment only
- with quintuple mutations present
- with a high baseline parasite density

5.2 Doubly interval-censored data

In the previous section, duration (which in this context is defined as the length of time a patient experiences gametocytemia) was estimated. Duration of gametocytemia was modelled using the time to gametocyte clearance AFT model $L3$. With this previous model, it was assumed that the start time was an exact time, which was equal to the day on which gametocytes were first observed or emerged. However, in this particular section, in addition to the previous method, duration is estimated using two other approaches. Duration still remains as the difference between the time to gametocyte emergence and time to gametocyte clearance. However, what will set these approaches apart from each other is how we assume the time to gametocyte emergence and time to gametocyte clearance.

The one approach will correctly apply interval-censored methods, assuming that both time to gametocyte emergence and time to gametocyte clearance are interval-censored - leading to doubly interval-censored times.

In chapters 3 and 4 the results presented revealed that between the right imputation approach and midpoint imputation approach, the midpoint approach produced estimates that were much closer to interval-censored estimates. Looking at the coefficients between the interval-censored and midpoint models the difference in estimates may not be alarming. But do these “small” differences eventually add up and result in far greater discrepancies? Hence, the last approach will assume time to gametocyte emergence and time to gametocyte clearance are exact times, equal to the midpoint of the interval in which the event has occurred. In this section we investigate how using different approaches to estimate quantities such as duration will truly impact results.

5.2.1 Methodology

In section 2.4 the concept of doubly interval-censoring was first discussed. As previously mentioned, doubly interval-censored observations occur when the time of origin and the time to event of interest are both interval-censored.

Time of origin of the survival time is the time at which the individual is considered to be at risk initially and is assumed to be known. Typically, this happens at enrollment or it could be the at the point in time when the patient enters the study in a particular state. However, as is the case of most epidemiology and clinical trials, the time of origin

can only be established at a planned visit to the doctor/clinician. The same goes with the time of event, as the event will only be recorded to have happened at the next visit. This then implies that the event is known to have occurred in-between consecutive visits. For example, in the case of the present research, the onset of gametocytemia can only be determined at a pre-planned doctor's visit. Thus, the time to gametocyte emergence is known to have occurred between visits, time to gametocyte emergence is therefore interval-censored. In addition, the time to gametocyte elimination is also said to be interval-censored as the event is known to have happened within some interval. It then follows, that the survival times are doubly interval-censored. Figure 11 in Chapter 2 shows a graphical illustration of doubly interval-censored observations.

In cases of doubly censoring, one option is to reduce doubly censored data to singly interval-censored data or right censored data (Bogaerts et al., 2017). This is done by transforming the intervals for time to onset (in this case time to gametocyte emergence) (T^O) $[l^O, u^O]$ and time to event (time to gametocyte elimination) (T^E) $[l^E, u^E]$ and combining these two to give one interval given as: $[l^E - u^O, u^E - l^O]$ (Bogaerts et al., 2017). Thereafter, applying the interval-censored methods such as the ones discussed above to the gap time $T = T^E - T^O$ using a reduced likelihood approach (Bogaerts et al., 2017). Using this method is not always suitable as mentioned by De Gruttola and Lagakos (1989). Alternatively, Dejardin and Lesaffre (2013) proposed a stochastic EM algorithm for the semi-parametric analysis of covariates on the distribution of gap time T_i for doubly-interval censored data. While the stochastic EM algorithm is the preferred approach, in this research, the reduced likelihood adhoc approach will be used and applied to the estimation of duration. In addition to the gap time, as previously mentioned, midpoint estimates for time to onset (T^O) $[l^O, u^O]$ and time to event (T^E) $[l^E, u^E]$ will be used to estimate duration. In other words, for the midpoint approach, the length of time between time to onset and time to event, will be given as the difference between the midpoint of the interval $[l^O, u^O]$ and midpoint of $[l^E, u^E]$. Standard survival analysis methods will be used to estimate duration using these midpoint estimates.

5.2.2 Results

Earlier in Chapter 5, models were built to assess risk factors that affect time to emergence; observe which covariates had an impact on time to elimination and how significant covariates either sped up or lengthened duration.

In the upcoming section AFT models are derived to assess how covariates affect the length of time that gametocytes are present in the body. In addition, the most appropriate of the AFT models is used to explicitly estimate duration of gametocytemia.

The duration results from this will be compared to the results that were estimated in section 5.1.3. The aim of the comparison is to illustrate the impact of applying different methods that take into account the interval-censoring mechanisms and how these different approaches might impact the results of the duration estimates.

The same covariates and distribution used to construct the AFT model L3 in section 5.1.3 are also used to construct the AFT models that estimate duration, when (1) the data is

correctly assumed to be doubly interval-censored and (2) data uses the midpoints of both the time to emergence and time to elimination intervals.

Table 29 shows the parameter estimates based on the three approaches. From the table it appears that the midpoint approach understates the effect of covariate. The singly interval-censored approach gives estimates that are closer to those constructed under the doubly interval-censored approach.

Table 29: Comparison of Parameter estimates of the Log-normal AFT models

Parameter	Estimate	95% CI	P-value
Doubly Interval-censored Approach			
mu	2.021	(1.837; 2.205)	0.000
log_s	-0.234	(-0.433; -0.035)	0.021
Logage	-0.441	(-0.844; -0.037)	0.032
Mutat	0.346	(-0.030; 0.723)	0.071
Arm	-0.689	(-1.076; -0.301)	< 0.001
Pzero	0.297	(0.130; 0.463)	< 0.001
Midpoint Approach			
mu	2.218	(2.089; 2.346)	0.000
log_s	-0.270	(-0.394; -0.145)	< 0.001
Logage	-0.393	(-0.706; -0.080)	0.001
Mutat	0.337	(-0.032; 0.642)	0.003
Arm	-0.432	(-0.707; 0.158)	< 0.001
Pzero	0.217	(0.101; 0.332)	< 0.001
Singly Interval Censored Approach			
mu	1.826	(2.128; 1.977)	0.000
log_s	-0.230	(-0.380; -0.079)	0.003
Logage	-0.450	(-0.814; -0.086)	0.015
Mutat	0.311	(-0.040; 0.662)	0.082
Arm	-0.622	(-0.952; -0.292)	< 0.001
Pzero	0.250	(0.110; 0.390)	< 0.001

Using the parameter estimates from the above table, duration is estimated. From Table 30 it is evident that the doubly interval-censored and singly censored approaches produce similar estimates of duration. This is to be expected given that the parameter estimates differed by only a few decimal points as shown in Table 29. Secondly, where duration is estimated to be relatively short, the midpoint approach generally overestimates and where duration is relatively long the midpoint approach underestimates the duration results compared to the doubly interval-censored approach. To explain the difference between the doubly interval-censored and midpoint duration estimates, recall in chapter 3, the survival plots that were produced for both time to gametocyte emergence and time to gametocyte elimination.

In Figure 14 of chapter 3, it was observed that in the earlier and narrower intervals, the interval-censored (IC) curve was enveloped between the right imputation (RI) and midpoint (MP) approaches.

However, in Figure 15, RI and MP curves take over the IC curve between days 1 and 4. This meant that under the MP approach, gametocyte emergence event times were underestimated in the earlier and smaller intervals, while the gametocyte elimination times were overestimated. Which implies that under the MP approach, the length of time between time to emergence and time to elimination is elongated because of this phenomena.

In other words, the event time for gametocyte emergence under the MP approach falls behind the IC. While the event time for gametocyte clearance for the MP approach falls in front of the IC. Keeping this in mind, this could explain why there is a significant difference in the estimation of the short duration estimates (which can only happen in the earlier and shorter intervals) between the midpoint approach and doubly interval-censored approach.

To explain why the midpoint underestimates duration in periods where duration is relatively long, we once again shift our focus to the Figures 14 and 15. Long duration implies that time to gametocyte clearance happened in the later and more wider intervals. Looking at Figure 15, the IC curve overtakes the MP curve from approximately day 16, and we see that the MP approach falls behind the IC curve. This then explains why the MP approach underestimates the longer duration estimates. Because in comparison to the IC curve, the MP assumes time to gametocyte clearance has occurred 'earlier'.

The difference between the doubly interval-censored approach and the singly interval-censored approach, is explained as follows. For the singly interval-censored approach, the length of time between the time to emergence and the time to clearance is generally shortened, seeing as the singly interval-censored method's starting time is equal to the upper limit of the time to emergence interval. In other words right imputation is used to define the starting time.

In contrast, the doubly interval-censored approach correctly assumes that the starting time is interval-censored. Recall that in Figure 14 the IC gametocyte emergence survival curve was shifted to the left in comparison to the RI curve. Therefore, the gametocyte emergence event times under the IC approach are earlier in comparison to the event times under the RI approach. This then inadvertently lengthens the width between the time to gametocyte emergence and time to gametocyte clearance for the doubly interval-censored approach.

Which explains why the singly interval-censored method generally underestimates the duration estimates in comparison to the doubly interval-censored method. Interestingly, for the relatively shorter duration estimates there are only minor differences between the singly interval-censored and doubly interval-censored methods. From which we can only assume that time to gametocyte emergence occurred in the regions in which the survival curves for gametocyte emergence amongst the three approaches were indistinguishable.

Table 30: Estimated Duration of gametocytemia and Standard Errors based on the Doubly interval-censored, Midpoint and Singly interval-censored approaches

Pzero	Mutat	Arm	Logage	Doubly Interval-censored			Midpoint		Singly Interval-censored	
				Estimate	S.E.		Estimate	S.E.	Estimate	S.E.
High	0	ACT	High	4.722	1.431		6.309	1.372	4.418	1.155
			Low	8.867	2.091		11.069	1.986	8.405	1.786
		SP	High	9.401	2.276		9.721	1.833	8.228	1.825
			Low	17.651	3.370		17.056	2.581	15.652	2.718
	1	ACT	High	6.677	2.120		8.838	2.045	6.030	1.670
			Low	12.538	3.184		15.506	3.160	11.471	2.700
		SP	High	13.292	3.472		13.618	2.775	11.229	2.684
			Low	24.959	5.347		23.893	4.272	21.369	4.260
Low	0	ACT	High	1.513	0.493		2.749	0.547	1.693	0.408
			Low	2.841	0.856		4.822	0.966	3.220	0.786
		SP	High	3.012	0.854		4.235	0.826	3.152	0.731
			Low	5.656	1.590		7.430	1.504	5.997	1.459
	1	ACT	High	2.140	0.792		3.850	0.929	2.310	0.673
			Low	4.018	1.401		6.755	1.679	4.394	1.304
		SP	High	4.260	1.425		5.933	1.409	4.302	1.217
			Low	7.998	2.659		10.409	2.600	8.184	2.416

From Table 30, the covariate patterns that result in the longest duration of gametocytemia regardless of the approach used appear to be patients:

- who are of a relatively younger age
- using the sulfadoxine-pyrimethamine treatment only
- with quintuple mutations present
- with a high baseline parasite density

This is the same conclusion that was found in Table 28 when singly interval-censored data was used.

These results once again reiterate the notion that in scenarios where intervals are shorter, the impact of treating an event time as an exact time or interval-censored time is negligible. However, as intervals widen, the greater the difference in results will be. Previously, it had been noted that the midpoint approach might be a suitable method for researchers who might prefer using traditional survival methods. However, the results from this analysis suggest that while the midpoint approach produces better estimates than the right imputation approach, the midpoint is still too inferior a substitute for interval-censoring. While the parameter estimates between the midpoint and interval-censoring approaches might result in minor differences estimating quantities such as duration lead to even bigger differences.

6 Discussion

This dissertation looks into the modelling of gametocytes while accommodating for different types of censoring with special focus on interval-censoring. Gametocytes are the sexual form of the *Plasmodium falciparum* parasite which are responsible for the transmission of malaria. Studies are generally designed to follow the evolution of the asexual parasite responsible for the symptoms of malaria but not necessarily the development of gametocytes. While gametocytes play a significant role in malaria transmission it is often seen that gametocytes are indirectly modelled through using asexual parasite densities.

In this dissertation, gametocytes are modelled directly using gametocyte densities. These records were taken over a 42 day study period, on days 0,3,7,14,21,28 and 42. Due to study design not only was the data left and right censored it was also interval-censored.

This dissertation was split into two parts. The first part focused on methodology comparing the results of when event times were assumed to be the right end point or midpoint of an interval with the results when data was correctly assumed to be interval-censored. Further analysis looked into treating data as doubly interval-censored data.

The second part of the dissertation focused more on answering the clinical questions. The investigation included determining which risk factors affect time to emergence? Which covariates affect the deceleration or quick advancement of gametocyte elimination? Lastly which characteristics make a patient more susceptible to carrying gametocytes for longer. Such questions were answered by the models built that take into account the interval-censoring mechanisms. This analysis provided insight which could help in the fight and curtailment of malaria transmission.

6.1 Methodological discussion

A feature of the data used in this investigation was that of censoring. Some patients entered the study with gametocytes and hence the exact day on which they first observed gametocytes is unknown as it happened prior to their enrollment in the study. Some patients are lost to follow up or exited due to treatment failure and therefore their gametocyte profiles are incomplete. All these scenarios are examples of censoring. Throughout this analysis, it is assumed that censoring is non-informative. Patients who exit the study are assumed to have the same gametocyte profiles as patients who remained in the study right up until the end. Additionally, due to the study design, patients are examined at specific follow up times.

Because of this, the actual event times are only known to have occurred within some known interval. Event times are thus interval-censored.

Interval-censored data is traditionally modeled by assuming the event time is equal to the midpoint of the interval in which the event is observed to have occurred. Alternatively, interval-censored data is reduced to an exact time equal to the time of observation of occurrence and thus completely ignoring the interval-censoring mechanisms.

Researchers and authors, then apply the standard survival methodologies, which are more well known and easier to apply. One of the purposes of this dissertation was to highlight

how using such approaches that ignore interval-censoring mechanisms compare to results that use interval-censoring techniques.

Two time-to-event processes were analyzed time to gametocyte emergence and time to gametocyte elimination. The survival curves in chapter 3 for time to emergence, reveal that using right imputation (assuming the event time is equal to the observed time) and the midpoint approach closely track the interval-censored NPMLE survival curve. In comparison, the survival curves from time to elimination paint a different story. The survival curve under the right imputation overestimated the survival times along with the mid point approach fence in the interval-censored survival curve, in the wider intervals. Which can be expected as right imputation tends to overestimate and the midpoint underestimates as seen by other authors such as Lindsey and Ryan (1998). In the earlier and shorter intervals, it is observed that the midpoint does a better job of imitating the interval-censored approach in comparison to the right imputation approach. Should authors choose to apply the more traditional methodologies in cases where data is interval-censored, it might be best to choose the midpoint approach in heavily interval-censored data. However, the interval-censoring mechanisms may be ignored only if the intervals are relatively small in width.

In Chapter 4 semi-parametric, parametric and AFT models were fitted. It was observed that the right imputation method underestimated the effect of ACT on both the time to gametocyte emergence and time to gametocyte clearance for all semi-parametric, parametric and AFT models. Once again, the midpoint approach produced similar results to the interval-censored approach compared to the right imputation. From the above applications it can be seen that analyzing interval-censored data while using incorrect methods, can lead to biased estimates and inaccurate conclusions.

Some shortcomings and limitations of this dissertation include the following. For the analysis of doubly interval-censored the reduced likelihood approach was used however a better method proposed by Dejardin and Lesaffre (2013), which is known as the stochastic EM algorithm could have been applied to the data. Using the gap time and reduced likelihood approach may not produce valid results as highlighted by De Gruttola and Lagakos (1989). But as mentioned, estimates using this approach are still appropriate given that the intervals are small (Dejardin and Lesaffre, 2013).

Secondly, in our analysis for duration, patients that were left truncated were excluded from analysis. This meant that estimates were based on the infected individuals in the cohort. Chatora (2018) argued that the results may be potentially biased, if the estimation of duration only includes patients that experienced gametocytemia. Chatora (2018) also argued that using such patients to estimate duration may lead to an overestimation of results for the whole population.

The author suggested that the estimate of duration can be made more representative by adjusting for the duration and taking into account the prevalence of gametocytemia. Chatora (2018) modified duration by including the probability of a patient developing

gametocytemia given their covariate profile with the help of a logistic model. Future researchers interested in estimating duration, may wish to consider using such an approach.

Lastly, the interval-censored analysis software is not completely developed. Some analysis was not possible because of this for example no AFT diagnostic tools are available as yet as mentioned by Anderson-Bergman (2017). Because of this, researchers might find it easier to apply standard survival analysis and ignore the interval-censoring mechanism.

6.2 Clinical discussion

Malaria has for the last few decades been a plague that has taken the lives of many millions. Despite the progress in malaria interventions, which has resulted in a reduction on a global scale, there has been an increase in the resistance of the parasite towards artemisinin. The growing resistance was being seen especially in areas such as Lao PDR, Myanmar, Cambodia, Vietnam, most malaria endemic countries in Africa and South America (WHO (2015a), WHO (2011a)). The growing resistance in treatments is worrying as it can potentially reverse the strides that were previously taken to overcome this endemic. For this reason, studies that investigate the factors that impact the prevalence of gametocytes and affect the length of time gametocytes might remain in the host's system are of grave importance especially to those living in areas heavily burdened by malaria. As looking into these matters will help come up with better strategies to fight against malaria transmission.

In this research, emphasis was focused on which covariates influence the emergence of gametocytes, modelling the time to gametocyte elimination and duration of gametocytemia. The data used were collected between 2002 and 2004 for a series of clinical trials set out in Mozambique and the north eastern province Mpumalanga, South Africa (Barnes et al. (2006), Allen et al. (2009)). The main aim of these studies was to investigate the efficacy of two treatments, sulfadoxine-pyrimethamine (SP) and the combination of artesunate and sulfadoxine-pyrimethamine (ACT) in clearing the asexual parasite. The data used were collected on days 0,3,7,14,21,28 and 42. Due to the study design and the ever increasing intervals, it is expected that interval-censoring would have an impact. Therefore, the methods used for analysis take into account interval-censoring.

Chapter 5 focused more on answering the clinical questions rather than the methodological questions. To assess the factors that influence the emergence of gametocytes, proportional hazards models were analyzed. This included the semi-parametric PH model proposed by Finkelstein (1986) and the Weibull and exponential proportional hazards. From the results it seems that the most influential factors of gametocyte emergence are the baseline asexual density, interaction term that includes the indicator variable for time and the baseline asexual density and the treatment administered. It was found that patients using a combination of artesunate and sulfadoxine-pyrimethamine reduced their relative risk of gametocytemia by 56.70%. In addition, for every ten-fold increase in the baseline asexual parasite density results in a 147% increase in the relative hazard of gametocyte emergence. Lastly, in later time periods, the effect of baseline asexual parasite density is reduced by 65%.

Surprisingly, no association was found between quintuple mutations and gametocyte emergence. This is an interesting find as it was expected that the relationship between mutation and treatment would affect prevalence of gametocytemia.

However, if we recall in section 1.4.1 Table 3, the exploratory data analysis found there to be no association between gametocyte prevalence and quintuple mutation or gender which could explain the findings. Interestingly, the same model used to derive gametocytemia prevalence excluded the covariate anemia. This finding was unexpected given that previous literature suggested a relationship between low levels of haemoglobin and gametocytemia.

Using the AFT model derived to estimate time to gametocyte clearance, model L3, it was found that age, quintuple mutation, treatment arm and baseline asexual parasite density affect the time until gametocyte are cleared from the patient's system. More specifically, for every ten-fold increase in age time to gametocyte clearance is accelerated by 36.23%. Patients who received ACT decreased the time to gametocyte clearance by 46.31% compared to those using the standard treatment. Individuals that had a quintuple mutations present saw an increase in the time to clearance by 36.50%. Finally, patient's time to gametocyte clearance decelerates by 28.40% for every ten-fold increase in the baseline asexual parasite density.

The other clinical objective was to investigate the duration of gametocytemia. This would inform of the infectivity period. Table 30 summarizes the estimated duration period based on different approaches that either assumed (1) start (time to gametocyte emergence) and end times (time to gametocyte clearance) were both interval-censored, (2) start and end times were equal to the midpoint of the interval in which the event is assumed to have occurred or (3) the start time was an exact time equal to the time gametocytes were observed to have occurred and the end time is interval-censored. All three approaches came to the same conclusion that longest duration of gametocytemia were associated with patients:

- who are of a relatively younger age
- using the sulfadoxine-pyrimethamine treatment only
- with quintuple mutations present
- with a high baseline asexual parasite density

From the results and conclusions above the following recommendations are made. Literature suggests that older patients appear to have some immunity against malaria. This may be because of their more developed immune systems. And from the results it appears that more emphasis should be turned towards preventing malaria in the young given their vulnerability. This could be in the form of educating parents on ways to reduce mosquitoes in and around the home. Or possibly investing resources into the development of anti-malarial drugs that are specific to children.

It is also recommended that sulfadoxine-pyrimethamine treatment stop being administered to patients. Presently more and more people are using artemisinin based treatments but it is essential all rural and urban areas in these malaria endemic countries adopt this

treatment. As the results highlight that artemisinin based treatments are more effective which is why they have already replaced sulfadoxine-pyrimethamine as the “standard” treatment.

Lastly, heavily burdened areas plagued by malaria should consider educating the communities on malaria and the symptoms related to malaria. This will help infected persons to identify their symptoms quicker and check themselves into clinics/hospitals sooner. That way they do not delay themselves into getting treatment. As delaying treatment could potentially give the asexual parasite more time to multiply which results in higher baseline asexual density by the time patient does get to clinic. Which means longer gametocytemia duration increasing the chances of human-to-mosquito and mosquito-to-human transmission resulting in the cycle starting again in another host.

On a deeper level, to reduce the malaria burden in these high incidence areas would also require a conversation on removing the barriers to health care. The data was collected in Mpumalanga, South Africa and southern Mozambique, these areas are relatively rural. The population in such rural areas generally have limited access to health care facilities and constrained incomes in comparison to those living in urban areas. While educating the members of society on measures of prevention and treatment is beneficial, it is imperative that we address the systemic problems associated with the health care systems such as accessibility and affordability. Therefore, to lower the malaria burden, increasing the number of health care facilities and trained staff would help detect and diagnose malaria early and treat it quickly reducing the risk of an individual carrying high level of baseline asexual parasite densities which would lead to higher level of gametocytes in their system; longer duration of gametocytemia and potential spread of infection within the community. In addition, reducing the price of anti-malarial drugs through subsidies or negotiations initiated by health departments with the respective manufacturers might allow people in these areas to afford treatments like ACT which we have seen from the analysis reduces duration of gametocytes and reduces the risk of gametocytes emerging. All in all, the end to malaria will take not just community members doing their part but also governments ensuring that their people have access to the resources, health care workers and treatments they need at an affordable price.

6.3 Concluding remarks

It is important that methods of analyzing data related to gametocytes are improved, as to better define risk factors that determine the extent of gametocyte carriage; host susceptibility and develop treatments/measures that will help to eradicate gametocytes. This dissertation illustrates the dangers of ignoring and not taking into account interval-censoring mechanisms. Further showing how results may differ amongst different methods of treating event times and could lead to biased results when time to event data is not treated correctly. While it has been found that in scenarios where data is interval-censored, researchers should apply interval-censored methods to their data. Information can be lost when censored data is being analyzed Brook (1982). To prevent this loss of information it is important that the study design is such that censoring minimized. For starters, this may

include reducing the time between visits, although, this may increase the costs of the study (through increased personnel costs or equipment needed for example) and inconvenience study participants. Leung et al. (1997) goes a steps further and states that designing a study in such a way that censoring does not occur is the best solution. For future research, finding design strategies that restrict or reduce censoring and adding this area of research to literature would be helpful as suggested by Leung et al. (1997).

References

- 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).
- Anderson-Bergman, C. (2017), ‘icenReg: Regression Models for Interval Censored Data in R’, *Journal of Statistical Software* **81**(12), 766–778.
- Barnes, K. I., Little, F., Mabuza, A., Mngomezulu, N., Govere, J., Durrheim, D., Roper, C., Watkins, B. and White, N. (2008), ‘Increased Gametocytemia after Treatment: An Early Parasitological Indicator of Emerging Sulfadoxine- Pyrimethamine Resistance in Falciparum Malaria.’, *The Journal of Infectious Diseases* **197**(11), 1605–1613.
- Barnes, K. I., Little, F., Smith, P. J., Evans, A., Watkins, W. M. and White, N. J. (2006), ‘Sulfadoxine-pyrimethamine pharmacokinetics in malaria: pediatric dosing implications’, *Clinical Pharmacology Therapeutics* **80**, 582–596.
- 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.
- Bogaerts, K., Komarek, A. and Lesaffre, E. (2017), *Survival Analysis with Interval-Censored Data: A Practical Approach with examples in R, SAS, and BUGS*, Chapman and Hall/CRC.
- Bousema, T. and Drakeley, C. (2011), ‘Epidemiology and Infectivity of Plasmodium falciparum and Plasmodium vivax Gametocytes in Relation to Malaria Control and Elimination’, *Clinical microbiology reviews* **24**(2), 377–410.
- Bousema, T., Okell, L., Shekalaghe, S., Griffin, J. T., Omar, S., Sawa, P., Sutherland, C., Sauerwein, R., Ghani, A. C. and Drakeley, C. (2010), ‘Revisiting the circulation time of Plasmodium falciparum gametocytes: molecular detection methods to estimate the duration of gametocyte carriage and the effect of gametocytocidal drugs’, *Malaria Journal* **9**(1), 136.
- Brook, R. J. (1982), ‘On the loss of information through censoring’, *Biometrika* **69**(1), 137–144.
- Campbell, M. J. and Machin, D. (1999), *Medical Statistics: A common-sense approach*, John Wiley Sons Ltd.
- Carter, J. Y., Loolpapit, M. P., Lema, O. E., Tome, J. L., Nagelkerke, N. J. D. and Watkins, W. M. (2005), ‘Reduction of the Efficacy of Antifolate Antimalarial Therapy by Folic Acid Supplementation’, *The American journal of tropical medicine and hygiene* **73**(1), 166–70.

- Chatora, T. (2018), Joint models for nonlinear longitudinal profiles in the presence of informative censoring, PhD thesis.
- Collett, D. (1994), *Modelling survival data in medical research*, London:Chapman Hall.
- Cox, D. R. (1972), ‘Regression Models and Life-Tables’, *Journal of the Royal Statistical Society* **34**(2), 187–220.
- David, H. A. and Moeschberger, M. L. (1978), *The Theory of Competing Risks*, Griffin London.
- De Gruttola, V. and Lagakos, S. W. (1989), ‘Analysis of Doubly-Censored Survival Data, with Application to AIDS’, *Biometrics* **45**(1), 1–11.
- Dejardin, D. and Lesaffre, E. (2013), ‘Stochastic EM algorithm for doubly interval-censored data’, *Biostatistics* **14**(4), 766–778.
- Distiller, G. B., Little, F. and Barnes, K. I. (2010), ‘Nonlinear mixed effects modeling of gametocyte’, *Malaria Journal* **9**(60).
- Efron, B. (1977), ‘The Efficiency of Cox’s Likelihood Function for Censored Data’, *Journal of the American Statistical Association* **72**(359), 557–565.
- Fay, M. P. (1996), ‘Rank invariant tests for interval censored data under the grouped continuous model.’, *Biometrics* **52**(3), 811–822.
- Fay, M. P. (1999), ‘Comparing several score tests for interval-censored data.’, *Statistics in Medicine* **18**(3), 273–85.
- Fay, M. P. and Shih, J. H. (1998), ‘Permutation tests using estimated distribution functions.’, *Journal of the American Statistical Association* **93**, 387–396.
- Finkelstein, D. M. (1986), ‘A Proportional Hazards Model for Interval-Censored Failure Time Data’, *Biometrics* **42**(4), 845–854.
- Finkelstein, D. M. and Wolfe, R. A. (1985), ‘A Semiparametric Model for Regression Analysis of Interval-Censored Failure Time Data’, *Biometrics* **41**(4), 933–945.
- Fleming, T. R. and Harrington, D. P. (1980), ‘A class of hypothesis tests for one and two sample censored survival data Communications in Statistics-Theory and Methods.’, (9), 763–794.
- Gentleman, R. and Vandal, A. (2019), *Icens: NPMLE for Censored and Truncated Data*. R package version 1.58.0.
- Gomez, G., Calle, M. L., Oller, R. and Langohr, K. (2009), ‘Tutorial on methods for interval-censored data and their implementation in R’, **9**(4), 259–297.
- Gomez, G. and Oller, R. (2008), ‘A New Class of Rank Tests for Interval-censored Data’.
- Groeneboom, P. and Wellner, J. (1992), *Information Bounds and Nonparametric Maximum Likelihood Estimation*, Birkhäuser.

- Hájek, J., Šidák, Z. and Sen, P. (1999), *Theory of Rank Tests*, Academic Press.
- Henderson, R., Diggle, P. and Dobson, A. (2000), ‘Joint modelling of longitudinal measurements and event time data’, *Biostatistics* **1**(4), 465–480.
- Kaplan, E. L. and Meier, P. (1958), ‘Nonparametric Estimation from Incomplete Observations’, *Journal of the American Statistical Association* **53**(282), 457–481.
- Karunajeewa, H. A. and Mueller, I. (2016), ‘How important is gametocyte clearance after malaria therapy?’, *BMC Medicine* **14**(93).
- Klein, J. P. and Moeschberger, M. L. (2003), *Survival Analysis Techniques for Censored and Truncated Data*.
- Kosorok, M. R. (2009), ‘What’s So Special About Semiparametric Methods?’, **71-A**(2), 331–353.
- Leung, K. M., Elashoff, R. M. and Afifi, A. A. (1997), ‘Censoring Issues in Survival Analysis’, *Annual Review of Public Health* **18**, 83–104.
- Lindsey, J. C. and Ryan, L. M. (1998), ‘Methods for interval-censored data’, *Statistics in Medicine* **17**(2), 219–238.
- Little, F. (2013), ‘Survival Analysis’.
- Mantel, N. (1966), ‘Evaluation of survival data and two new rank order statistics arising in its consideration’, *Cancer chemotherapy reports* **50**(3), 163–170.
- Mantel, N. and Haenszel, W. (1959), ‘Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease’, *Journal of the National Cancer Institute* **22**(4), 719–748.
- Michalakakis, Y. and Renaud, F. (2009), ‘Malaria: Evolution in vector control’, *Nature* **462**, 298–300.
- Peto, R. (1973), ‘Experimental Survival Curves for Interval-censored Data’, *Journal of the Royal Statistical Society* **22**(1), 86–91.
- Peto, R. and Peto, J. (1972), ‘Asymptotically efficient rank invariant test procedures.’, *Journal of the Royal Statistical Society, Series A* **135**(2), 185–207.
- Pourhoseingholi, M. A., Hajizadeh, E., Dehkordi, B. M., Safaee, A., Abadi, A. and Zali, M. R. (2007), ‘Comparing Cox Regression and Parametric Models for Survival of Patients with Gastric Carcinoma’, *Asian Pacific J Cancer Prev* **8**, 412–416.
- Prentice, R. L. (1978), ‘Linear rank tests with right censored data.’, *Biometrika* **65**, 167–79.
- Price, R., Nosten, F., Simpson, J. A., Luxemburger, C., Phaipun, L., Terkuile, 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.

- R Core Team (2019), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
- Rodrigues, A. S., Calsavara, V. F., Silva, F. I. B., Alves, F. A. and Vivas, A. P. M. (2018), ‘Use of interval-censored survival data as an alternative to Kaplan-Meier survival curves: studies of oral lesion occurrence in liver transplants and cancer recurrence’, *Applied Cancer Research* **38**(16).
- Roser, M. and Ritchie, H. (2019), ‘Malaria’, *Our World in Data* .
- Ross, A., Killeen, G. and Smith, T. (2006), ‘Risk factors for gametocyte carriage in uncomplicated falciparum malaria’, *The American Society of Tropical Medicine and Hygiene* **75**(2), 32–37.
- Scolas, S. (2016), Modelling interval-censored event times in the presence of a cure fraction: From building to refining, PhD thesis.
- Singh, R. S. and Totawattage, D. P. (2013), ‘The Statistical Analysis of Interval-Censored Failure Time Data with Applications’, *Open Journal of Statistics* **3**(2), 155–166.
- Sowunmi, A., Fateye, B. A., Adedeji, A. A., Fehintola, F. A. and Happi, T. C. (2004), ‘Risk factors for gametocyte carriage in uncomplicated falciparum malaria in children’, *Parasitology* **129**(3), 255–262.
- Sun, J. (2006), *The statistical analysis of interval-censored failure time data.*, New York: Springer.
- Sun, X. and Chen, C. (2010), ‘Comparison of Finkelstein’s Method With the Conventional Approach for Interval-Censored Data Analysis’, *Statistics in Biopharmaceutical Research* **2**(1), 97–108.
- Trager, W., Gill, G. S., Lawrence, C. and Nagel, R. (1999), ‘Plasmodium falciparum: enhanced gametocyte formation in vitro in reticulocyte-rich blood.’, *Exp Parasitol.* **91**, 115–8.
- Turnbull, B. (1976), ‘The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data’, *Journal of the Royal Statistical Society. Series B (Methodological)* **38**(3), 290–295.
- von Seidlein, L., Drakeley, C., Greenwood, B., Walraven, G. and Targett, G. (2001), ‘Risk factors for gametocyte carriage in Gambian children.’, *The American Society of Tropical Medicine and Hygiene* **65**(5), 523 – 527.
- Wellner, J. A. and Zhan, Y. (1997), ‘A Hybrid Algorithm for Computation of the Nonparametric Maximum Likelihood Estimator from Censored Data’, *Journal of the American Statistical Association* **92**(439), 945–959.
- WHO (2011a), *Global report on antimalarial efficacy and drug resistance: 2000–2010*, Geneva: World Health Organization.

- WHO (2011b), *Haemoglobin concentrations for the diagnosis of anaemia and assessment of severity*, Geneva, Switzerland.
- WHO (2015a), ‘Status report on artemisinin and ACT resistance’.
- WHO (2015b), *World Malaria Report 2015*, Geneva: World Health Organization.
- WHO (2019), *World Malaria Report 2019*, Geneva: World Health Organization.
- WWARN Gametocyte Study Group (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.
- Zhang, Z. and Sun, J. (2010), ‘Interval censoring’, *Statistical methods in medical research* **19**(1), 53–70.

A Appendices

A.1 Libraries and data used

```
#libraries needed
library(survival)
library(Icens)
library(icenReg)
library(icensBKL)
library(tidyverse)
library(dplyr)
library(foreign)
library(readr)
library(survminer)
library(broom)
library(haven)
library(readr)
library(rio)
library(interval)
library(smoothSurv)
library(logspline)
library(tidyverse)
library(survMisc)

outcome_table = read_dta("outcome_table.dta")
demographics_table = read_dta("demographics_table.dta")
#merge the two tables together
attach(outcome_table)
CombinedTable=outcome_table % > % right_join(demographics_table)
#save(Table, file = 'Final_tbl.RData')

games = read.dta("Game.dta")
attach(games)

#games = games[,c(1,2,20)]
games$gamedens=as.numeric(games$gamedens)
games$gamedens=log10(games$gamedens+1)
games1=games % > % pivot_wider(id_cols = subjectno, names_from = day ,values_from
= gamedens)

games1$pid=games1$subjectno
games2=games1 % > % right_join(CombinedTable,by="pid")
```

```

gamedens.table = games2[,-c(1,3,4)]

export(Table, "tbl.csv")
Covariate.tbl = read_dta("Game.dta")
attach(Covariate.tbl)
attach(gamedens.table)
Covariate.tbl = Covariate.tbl %>% group_by(subjectno) %>% slice(1) %>% un-
group() %>% mutate(muts=dhfrmax+dhpsmax) %>%
select(feverhis, pardensd0, muts, haemd0, haem, dosage, subjectno)
Covariate.tbl$pid=Covariate.tbl$subjectno
Full_Table=Covariate.tbl %>% right_join(gamedens.table) %>% select(-subjectno)
%>% select(c(pid:'42', everything()))

Full_Table=Full_Table %>% mutate(totalgmcil=rowSums(Full_Table[,2:8], na.rm =
T)) %>% select(c(pid:'42', totalgmcil, everything()))

write.csv(Full_Table, file=~ /Documents/Thesis _malaria/CHapter 1/Full_Table.csv")

#exported to "Numbers" program on Macbook for further work

```

A.2 Data imported

```

corrected.tbl = read_delim("data2021.tbl.csv", ",", escape_double = FALSE, trim_ws =
TRUE)
Table =corrected.tbl
save(Table, file = "Final.tbl.RData")
load(file = "Final.tbl.RData")

attach(Table)
x= c("MON2003_010", "MOM2004_188", "MOM2004_162", "MOM2004_129", "MOM2004_023",
"MOB2004_052", "MOB2004_047", "MOB2004_014", "MOB2004_009", "MOB2004_004",
"MOM2004_021", "MOM2004_196", "MOM2004_119", "MOB2004_074", "MOB2004_023",
"MOM2004_137", "MOM2004_158")
pid_exl= as.data.frame(x)
Table=Table %>% mutate(logage=log10(age)) %>% mutate(censor_elim=ifelse(is.na(u_elim),0,1))
%>% mutate(censor_emerg=ifelse(is.na(l_emer),0,1)) %>% mutate(pzero=log10(pardensd0))
%>% mutate(anemia = ifelse(haem<11,1,0)) %>% mutate(mutat=ifelse(muts>4,1,0))

Table= Table %>% filter(!pid %in% pid_exl$x)

```

```
Table=Table[,-1]
```

```
attach(Table)
```

A.3 Descriptive statistics

```
#right
par(mfrow=c(1,3))
plot(0, 0, col="white", las=1, xlim=c(0,42), xlab="pday", ylim=range(0:4, na.rm=T),
ylab=expression("log"[10] * "gametocyte density"), xaxt="n",cex.lab=1.4)
abline(h=seq(0,9e3,1e3), v=seq(-2,45,4), col="azure2")
axis(1, c(0,3,7,14,21,28,42), c(0,3,7,14,21,28,42))
new_pid_92 = Table$pid[13]
lines(c(0,3,7,14,21,28,42), Table[13,2:8], lwd=2, col="blue")
points(c(0,3,7,14,21,28,42), Table[13,2:8], pch=16, col="blue")
text(15.0, 4, new_pid_92, cex=1.3)
```

```
plot(0, 0, col="white", las=1, xlim=c(0,42), xlab="pday", ylim=range(0:4, na.rm=T),
ylab=expression("log"[10] * " gametocyte density"), xaxt="n",cex.lab=1.4)
abline(h=seq(0,9e3,1e3), v=seq(-2,45,4), col="azure2")
axis(1, c(0,3,7,14,21,28,42), c(0,3,7,14,21,28,42))
new_pid18 = Table$pid[8]
lines(c(0,3,7,14,21,28,42), Table[8,2:8], lwd=2, col="blue")
points(c(0,3,7,14,21,28,42), Table[8,2:8], pch=16, col="blue")
text(15.0, 4, new_pid18, cex=1.3)
```

```
plot(0, 0, col="white", las=1, xlim=c(0,42), xlab="pday", ylim=range(0:4, na.rm=T),
ylab=expression("log"[10] * " gametocyte density"), xaxt="n",cex.lab=1.4)
abline(h=seq(0,9e3,1e3), v=seq(-2,45,4), col="azure2")
axis(1, c(0,3,7,14,21,28,42), c(0,3,7,14,21,28,42))
new_pid12 = Table$pid[3]
lines(c(0,3,7,14,21,28,42), Table[3,2:8], lwd=2, col="blue")
points(c(0,3,7,14,21,28,42), Table[3,2:8], pch=16, col="blue")
text(15.0, 4, new_pid12, cex=1.3)
```

```
plot(0, 0, col="white", las=1, xlim=c(0,42), xlab="pday", ylim=range(0:4, na.rm=T),
ylab=expression("log"[10] * " gametocyte density"), xaxt="n",cex.lab=1.4)
abline(h=seq(0,9e3,1e3), v=seq(-2,45,4), col="azure2")
axis(1, c(0,3,7,14,21,28,42), c(0,3,7,14,21,28,42))
```

```

new_pid12 = Table$pid[301]
lines(c(0,3,7,14,21,28,42), Table[301,2:8], lwd=2, col="blue")
points(c(0,3,7,14,21,28,42), Table[301,2:8], pch=16, col="blue")
text(15.0, 4, new_pid12, cex=1.3)

#left
par(mfrow=c(1,3))

plot(0, 0, col="white", las=1, xlim=c(0,42), xlab="pday", ylim=range(0:4, na.rm=T),
ylab=expression("log"[10] * " gametocyte density"), xaxt="n",cex.lab=1.4)
abline(h=seq(0,9e3,1e3), v=seq(-2,45,4), col="azure2")
axis(1, c(0,3,7,14,21,28,42), c(0,3,7,14,21,28,42))
new_pid_60 = Table$pid[29]
lines(c(0,3,7,14,21,28,42), Table[29,2:8], lwd=2, col="green")
points(c(0,3,7,14,21,28,42), Table[29,2:8], pch=16, col="green")
text(15.0, 4, new_pid_60, cex=1.3)

plot(0, 0, col="white", las=1, xlim=c(0,42), xlab="pday", ylim=range(0:4, na.rm=T),
ylab=expression("log"[10] * " gametocyte density"), xaxt="n",cex.lab=1.4)
abline(h=seq(0,9e3,1e3), v=seq(-2,45,4), col="azure2")
axis(1, c(0,3,7,14,21,28,42), c(0,3,7,14,21,28,42))
new_pid129 = Table$pid[54]
lines(c(0,3,7,14,21,28,42), Table[54,2:8], lwd=2, col="green")
points(c(0,3,7,14,21,28,42), Table[54,2:8], pch=16, col="green")
text(15.0, 4, new_pid129, cex=1.3)

plot(0, 0, col="white", las=1, xlim=c(0,42), xlab="pday", ylim=range(0:4, na.rm=T),
ylab=expression("log"[10] * " gametocyte density"), xaxt="n",cex.lab=1.4)
abline(h=seq(0,9e3,1e3), v=seq(-2,45,4), col="azure2")
axis(1, c(0,3,7,14,21,28,42), c(0,3,7,14,21,28,42))
new_pid125 = Table$pid[145]
lines(c(0,3,7,14,21,28,42), Table[145,2:8], lwd=2, col="green")
points(c(0,3,7,14,21,28,42), Table[145,2:8], pch=16, col="green")
text(15.0, 4, new_pid125, cex=1.3)

#interval
par(mfrow=c(1,3))
plot(0, 0, col="white", las=1, xlim=c(0,42), xlab="pday", ylim=range(0:4, na.rm=T),
ylab=expression("log"[10] * " gametocyte density"), xaxt="n",cex.lab=1.4)
abline(h=seq(0,9e3,1e3), v=seq(-2,45,4), col="azure2")
axis(1, c(0,3,7,14,21,28,42), c(0,3,7,14,21,28,42))

```

```

new_pid_11 = Table$pid[40]
lines(c(0,3,7,14,21,28,42), Table[40,2:8], lwd=2, col="red")
points(c(0,3,7,14,21,28,42), Table[40,2:8], pch=16, col="red")
text(15.0, 4, new_pid_11, cex=1.3)

plot(0, 0, col="white", las=1, xlim=c(0,42), xlab="pday", ylim=range(0:4, na.rm=T),
ylab=expression("log"[10] * " gametocyte density"), xaxt="n",cex.lab=1.4)
abline(h=seq(0,9e3,1e3), v=seq(-2,45,4), col="azure2")
axis(1, c(0,3,7,14,21,28,42), c(0,3,7,14,21,28,42))
new_pid44 = Table$pid[60]
lines(c(0,3,7,14,21,28,42), Table[60,2:8], lwd=2, col="red")
points(c(0,3,7,14,21,28,42), Table[60,2:8], pch=16, col="red")
text(15.0, 4, new_pid44, cex=1.3)

plot(0, 0, col="white", las=1, xlim=c(0,42), xlab="pday", ylim=range(0:4, na.rm=T),
ylab=expression("log"[10] * " gametocyte density"), xaxt="n",cex.lab=1.4)
abline(h=seq(0,9e3,1e3), v=seq(-2,45,4), col="azure2")
axis(1, c(0,3,7,14,21,28,42), c(0,3,7,14,21,28,42))
new_pid19 = Table$pid[47]
lines(c(0,3,7,14,21,28,42), Table[47,2:8], lwd=2, col="red")
points(c(0,3,7,14,21,28,42), Table[47,2:8], pch=16, col="red")
text(15.0, 4, new_pid19, cex=1.3)

#CREATE SPAGHETTI PLOTS
CREATE SPAGHETTI PLOTS
Failure_tbl = Table %>% filter(PIoutcome%in% c("LCF" , "LPF" , "ETF" ,"death" ))
%>% select(c(pid:'42','Time to Emergence':'Time to Elim'))
attach(Failure_tbl)

plot(0, 0, col="white", las=1, xlim=c(0,42),xaxt="n", xlab="pday", ylim=c(0,4),
ylab=expression("log"[10] * " gametocyte density"))
axis(1, c(0,3,7,14,21,28,42), c(0,3,7,14,21,28,42))
abline(h=seq(0,9e3,1e3), v=seq(-2,45,4), col="azure2")
axis(1, 1:8, colnames(censor_frame))
clrs = rainbow(length(Failure_tbl))
for(j in 1:8){
row=3+j

lines(c(0,3,7,14,21,28,42), Failure_tbl[row,2:8], lwd=2, col=clrs[j])
points(c(0,3,7,14,21,28,42) Failure_tbl[row,2:8], pch=16, col=clrs[j])
}

```

```

plot(0, 0, col="white", las=1, xlim=c(0,42), xaxt="n", xlab="pday", ylim=c(0,4),
ylab=expression("log"[10] * "gametocyte density"))
axis(1, c(0,3,7,14,21,28,42), c(0,3,7,14,21,28,42))
abline(h=seq(0,9e3,1e3), v=seq(-2,45,4), col="azure2")
axis(1, 1:8, colnames(censor_frame))
rw1= as.matrix(c(39,38,26,21,17,7,1,5,8,3,19,35))
for(j in 1:nrow(Failure_tbl)){
row=j
lines(c(0,3,7,14,21,28,42), Failure_tbl[row,2:8], lwd=2)
points(c(0,3,7,14,21,28,42), Failure_tbl[row,2:8], pch=16)
}

```

```

Success_tbl = Table %>% filter(PIoutcome%in% c("ACPR")) %>% select(c(pid:'42','Time
to Emergence':'Time to Elim'))
attach(Success_tbl)

```

```

plot(0, 0, col="white", las=1, xlim=c(0,42), xaxt="n", xlab="pday", ylim=c(0,4),
ylab=expression("log"[10] * "gametocyte density"))
axis(1, c(0,3,7,14,21,28,42), c(0,3,7,14,21,28,42))
abline(h=seq(0,9e3,1e3), v=seq(-2,45,4), col="azure2")
axis(1, 1:8, colnames(censor_frame))
clrs = rainbow(length(Success_tbl))

```

```

for(j in 1:26){
row = 54+j
lines(c(0,3,7,14,21,28,42),Success_tbl[row,2:8], lwd=2, col=clrs[j])
points(c(0,3,7,14,21,28,42), Success_tbl[row,2:8], pch=16, col=clrs[j])
}

```

```

plot(0, 0, col="white", las=1, xlim=c(0,42), xaxt="n", xlab="pday", ylim=c(0,4),
ylab=expression("log"[10] * "gametocyte density"))
axis(1, c(0,3,7,14,21,28,42), c(0,3,7,14,21,28,42))
abline(h=seq(0,9e3,1e3), v=seq(-2,45,4), col="azure2")
axis(1, 1:8, colnames(censor_frame))
rw1= as.matrix(c(39,38,26,21,17,7,1,5,8,3,19,35))
for(j in 1:nrow(Success_tbl)){
row=j
lines(c(0,3,7,14,21,28,42), Success_tbl[row,2:8], lwd=2)
points(c(0,3,7,14,21,28,42), Success_tbl[row,2:8], pch=16)
}

```

```
LTFU_tbl = Table %>% filter(PIoutcome%in% c("LTFU")) %>% select(c(pid:'42','Time
to Emergence':'Time to Elim'))
attach(LTFU_tbl)
```

```
plot(0, 0, col="white", las=1, xlim=c(0,42), xaxt="n", xlab="pday", ylim=c(0,4),
ylab=expression("log"[10] * " gametocyte density"))
axis(1, c(0,3,7,14,21,28,42),c(0,3,7,14,21,28,42))
abline(h=seq(0,9e3,1e3), v=seq(-2,45,4), col="azure2")
axis(1, 1:8, colnames(censor_frame))
rw1= as.matrix(c(39,38,26,21,17,7,1,5,8,3,19,35))
clrs = c("yellow",rainbow(length(LTFU_tbl)))
for(j in 1:5){
d=c(8,2,3,4,6)
row=d[j]
lines(c(0,3,7,14,21,28,42), LTFU_tbl[row,2:8], lwd=2, col=clrs[j])
points(c(0,3,7,14,21,28,42), LTFU_tbl[row,2:8], pch=16, col=clrs[j])
}
```

```
plot(0, 0, col="white", las=1, xlim=c(0,42), xaxt="n", xlab="pday", ylim=c(0,4),
ylab=expression("log"[10] * " gametocyte density"))
axis(1, c(0,3,7,14,21,28,42), c(0,3,7,14,21,28,42))
abline(h=seq(0,9e3,1e3), v=seq(-2,45,4), col="azure2")
axis(1, 1:8, colnames(censor_frame))
rw1= as.matrix(c(39,38,26,21,17,7,1,5,8,3,19,35))
for(j in 1:nrow(LTFU_tbl)){
row=j
lines(c(0,3,7,14,21,28,42), LTFU_tbl[row,2:8], lwd=2)
points(c(0,3,7,14,21,28,42), LTFU_tbl[row,2:8], pch=16)
}
```

```
Table = Table %>% mutate(treatment_outcome =ifelse( PIoutcome == "LTFU"
,"LTFU", ifelse( PIoutcome == "ACPR", "Success", "Failure")))
attach(Table)
```

```
#trtoutcome
table(treatment_outcome,arm)
fisher.test(table(treatment_outcome,arm))
```

```
#gender
```

```
table(gender,arm)
fisher.test(table(gender,arm))
```

```
#mutat
table(mutat,arm)
fisher.test(table(mutat,arm))
```

```
#anemia
table(anemia,arm)
fisher.test(table(anemia,arm))
```

```
boxplot(Table$logage~ arm, xlab="Treatment",ylab= "Logage")$stats
boxplot(Table$pzero~ arm , xlab= "Treatment",ylab="Pzero")$stats
```

```
kruskal.test(arm ~Table$logage)
kruskal.test(arm~Table$pzero)
```

```
prevalence_tbl= Table% > % mutate(present = ifelse(totalgmcil>0,"Present","Absent"))% >
% select(present,logage,pzero,arm,anemia,gender,mutat)
```

```
attach(prevalence_tbl)
#arm
table(present,arm)
fisher.test(table(present,arm))
```

```
#gender
```

```
table(present,gender)
fisher.test(table(present,gender))
```

```
#mutat
```

```
table(present,mutat)
fisher.test(table(present,mutat))
```

```
#anemia
table(present,anemia)
```

```
fisher.test(table(present,anemia))
```

```
#trt_out table(treatment_outcome,present)
fisher.test(table(treatment_outcome,present))
```

```
boxplot(Table$logage~ present,ylab= "Logage")$stats
boxplot(Table$Pzero~present,ylab= "Pzero")$stats
```

```
kruskal.test(present ~Table$logage)
kruskal.test(present ~Table$Pzero)
```

```
cols = rainbow(8)
sp_tbl = Table %>% filter(arm=="SP")
df=rbind(c(19,29,57,50,34,20,5))
df= as.matrix(df)
colnames(df)= c(0,3,7,14,21,28,42)
barplot(df[1,], col = cols, main = "SP", xlab = "Day", ylab = "Frequency" ,ylim = c(0,60))
```

```
art_tbl = Table %>% filter(arm=="SP/ART")
```

```
df1=rbind(c(16,19,18,8,8,3,1))
df1= as.matrix(df1)
colnames(df1)= c(0,3,7,14,21,28,42)
barplot(df1[1,], col = cols, main = "ACT", xlab = "Day", ylab = "Frequency", ylim =
c(0,60))
```

```
par(mfrow= c(2,2))
```

```
plot(c(0,3,7,14,21,28,42),colMeans(Failure_tbl[,2:8],na.rm = T),type="b",ylim=c(0,1),col
= "black", xlab= "Day",ylab = expression("mean log"[10] * " density"), main = "Out-
come ")
lines(c(0,3,7,14,21,28,42),colMeans(LTFU_tbl[,2:8],na.rm=T), col = "red",type="b")
lines(c(0,3,7,14,21,28,42),colMeans(Success_tbl[,2:8],na.rm=T), col = "green",type="b")
legend(x="topright",legend=c("Failure", "LTFU", "Success"),fill=c("black", "red", "green"))
```

```
plot(c(0,3,7,14,21,28,42),colMeans(subset(Table[2:8], Table$gender=="F"),na.rm=T),type="b",
ylim=c(0,1),col = "black", xlab= "Day",ylab = expression("mean log"[10] * " density"),main
= "Gender")
lines(c(0,3,7,14,21,28,42),colMeans(subset(Table[2:8], Table$gender=="M"),na.rm=T),col
```

```
= "red",type = "b")
legend(x="topright",legend=c("F","M"),fill=c("black","red"))

plot(c(0,3,7,14,21,28,42),colMeans(subset(Table[2:8], Table$mutat=="0"),na.rm=T),type="b",
ylim=c(0,1),col = "black", xlab= "Day",ylab = expression("mean log"[10] * " density"),main
= "Mutat")
lines(c(0,3,7,14,21,28,42),colMeans(subset(Table[2:8], Table$mutat=="1"),na.rm=T),col =
"red",type = "b")
legend(x="topright",legend=c("≤5 mutations","5 mutations"),fill=c("black","red"))

plot(c(0,3,7,14,21,28,42),colMeans(subset(Table[2:8], Table$anemia=="0"),na.rm=T),type="b",
ylim=c(0,1),col = "black", xlab= "Day",ylab = expression("mean log"[10] * " density"),
main = "Anemia")
lines(c(0,3,7,14,21,28,42),colMeans(subset(Table[2:8], Table$anemia=="1"),na.rm=T),col
= "red",type = "b")
legend(x="topright",legend=c("No","Yes"),fill=c("black","red"))

plot(c(0,3,7,14,21,28,42),colMeans(sp_ttbl[,2:8],na.rm = T),type="b",
ylim=c(0,1),col = "black", xlab= "Day",ylab = expression("mean log"[10] * " density"),
main = "Arm")
lines(c(0,3,7,14,21,28,42),colMeans(art_ttbl[,2:8],na.rm=T), col = "red",type="b")
legend(x="topright",legend=c("SP","ACT"),fill=c("black","red"))
```

A.4 Standard survival analysis

```
#STANDARD PROCEDURE-EMERGENCE # Create an event indicator to determine
if subject is censored or not
#have to use interval to allow for both left and right censored observations
#Table= Table %>% mutate(ecensored=ifelse(emerg_censored=="3",2,2))
#cox can not handle 0s
Table = Table %>% mutate(event_emerg=ifelse(is.na(u_emer),0,ifelse(is.na(l_emer),2,1)))
%>% mutate(time2_emerg=rep(NA,391)) %>% mutate(time_emerg=ifelse('Time to
Emergence'==0,0.0001,'Time to Emergence'))
attach(Table)

# Create the Survival object:
surv.rel=Surv(time='Time to Emergence'+1,time2= time2_emerg,event =event_emerg,type="in-
terval")
print(surv.rel)
surv_fit=survfit(surv.rel~1)
```

```
summary(surv_fit)

#STANDARD PROCEDURE - ELIMINATION
#those who do not have gametocytes must be excluded - left truncated
Table = Table %>% mutate(event_elim=ifelse(is.na(u_elim),0,ifelse(is.na(l_elim),2,1)))
%>% mutate(time2_elim=rep(NA,391))
subset_tbl = Table %>% filter(totalgmCIL>0)

#no left censored just right censored
surv.rel2=Surv(subset_tbl$adjusted+1 ,subset_tbl$event_elim)
print(surv.rel2)
surv_fit2=survfit(surv.rel2~1)
summary(surv_fit2)

#survdiff(surv.rel~arm )

#COMPARISON OF ARMS
surv_arm=survfit(surv.rel~arm,conf.type="log-log")
surv_arm2=survfit(surv.rel2~subset_tbl$arm,conf.type="log-log")

summary(surv_arm)

#COX

#NB!!!!!!!!!!!!!!!!!!!!!!
#cox can not handle left censored... so turn left to right censored

Table = Table %>% mutate(event_emerg2=ifelse(event_emerg==2,1,event_emerg))
%>% mutate(ttemer= ifelse('Time to Emergence'==0,0.0001,'Time to Emergence'+1))
subset_tbl$ttemer= ifelse(subset_tbl$adjusted==0,0.0001,subset_tbl$adjusted+1)
attach(Table)

# Create the Survival object:
cox.rel=Surv(ttemer,event_emerg2,type="right")
cox.rel
cox_emer=coxph(cox.rel~arm,data=Table)

logrank.test(time = ttemer,event = event_emerg2,group = Table$arm,rho = 0, gamma=0)
```

```
logrank.test(time = ttemer,event = event_emerg2,group = Table$arm,rho = 1, gamma=0)
logrank.test(time = ttemer,event = event_emerg2,group = Table$arm,rho = 0, gamma=1)
```

```
subset_tbl = subset_tbl % > % mutate(event_elim2=ifelse(!is.na(subset_tbl$u_elim)
!(is.na(subset_tbl$l_elim)),1,0))
```

```
cox.rel2=Surv(subset_tbl$ttelim ,subset_tbl$event_elim2,type="right")
cox.rel2
attach(subset_tbl)
cox_elim=coxph(cox.rel2~arm,data=subset_tbl)
```

```
logrank.test(time = subset_tbl$ttelim ,event = subset_tbl$event_elim2,group = sub-
set_tbl$arm,rho = 0, gamma=0)
logrank.test(time = subset_tbl$ttelim ,event = subset_tbl$event_elim2,group = subset_tbl$arm,rho
= 1, gamma=0)
logrank.test(time = subset_tbl$ttelim ,event = subset_tbl$event_elim2,group = subset_tbl$arm,rho
= 0, gamma=1)
```

```
#PH MODELS
# can not have 0s
```

```
weib_elim=survreg(Surv(ttelim ,subset_tbl$event_elim)~ arm,data = subset_tbl,dist="weibull")
```

```
ic_par(Surv(time=ttemer,time2= time2_emerg,event =event_emerg2,type="interval")~
arm,data = Table,model = "ph",dist="weibull")
ic_par(Surv(time=ttemer,time2= time2_emerg,event =event_emerg2,type="interval")~ arm,data
= Table,model = "ph",dist="exponential")
```

```
ic_par(Surv(time=ttemer,time2= time2_emerg,event =event_emerg2,type="interval")~
arm,data = Table,model = "aft",dist="weibull")
ic_par(Surv(time=ttemer,time2= time2_emerg,event =event_emerg2,type="interval")~ arm,data
= Table,model = "aft",dist="loglogistic")
ic_par(Surv(time=ttemer,time2= time2_emerg,event =event_emerg2,type="interval")~ arm,data
= Table,model = "aft",dist="lnorm")
```

```
diag_baseline( ic_par(Surv(time=ttemer,time2= time2_emerg,event =event_emerg2,type="interval")~
arm,data = Table,model = "ph"))
```

```

attach(subset_tbl)

ic_par(Surv(time = subset_tbl$ttelim, time2 = subset_tbl$time2_elim,event = subset_tbl$event_elim, "interval") ~ arm, data = subset_tbl, model = "ph", dist="weibull")
ic_par(Surv(time = subset_tbl$ttelim, time2 = subset_tbl$time2_elim,event = subset_tbl$event_elim, "interval") ~ arm, data = subset_tbl, model = "ph", dist="exponential")

ri_aft=ic_par(Surv(time = subset_tbl$ttelim, time2 = subset_tbl$time2_elim,event = subset_tbl$event_elim, "interval") ~ arm, data = subset_tbl, model = "aft", dist="weibull")
ri_aft
ic_par(Surv(time = subset_tbl$ttelim, time2 = subset_tbl$time2_elim,event = subset_tbl$event_elim, "interval") ~ arm, data = subset_tbl, model = "aft", dist="loglogistic")
ic_par(Surv(time = subset_tbl$ttelim, time2 = subset_tbl$time2_elim,event = subset_tbl$event_elim, "interval") ~ arm, data = subset_tbl, model = "aft", dist="lnorm")

```

A.5 Midpoint censoring

```

Mydata = subset(Table,select=c("l_emer", "u_emer", "arm", "logage", "mutat", "gender", "anemia", "pzero", "Time to Emergence"))
attach(Mydata)
Mydata$status = ifelse( is.na(u_emer), 0, 1)
mid_low =ifelse( is.na(l_emer), 0,ifelse(l_emer==0,1,l_emer+1))
mid_upp = ifelse(is.na(u_emer),45,u_emer+1)
Mydata$midpoint = ifelse( is.na(u_emer),'Time to Emergence', (mid_low+mid_upp)/2)
Mydata$time2 = rep(NA,391)
ifelse(Table$event_emerg==2,1,Table$event_emerg)

mid_surv=Surv(Mydata$midpoint ,Mydata$status)
print(mid_surv)
mid_fit=survfit(mid_surv~1)
summary(mid_fit)

#comparison
mid_arm=survfit(mid_surv~Mydata$arm,conf.type="log-log")

#cox model
coxph(mid_surv~arm,data = Mydata)

#ph model

```

```

ic_par(Surv(time=Mydata$midpoint,time2= Mydata$time2,event =Mydata$status,type="interval")~
arm,data = Mydata,model = "ph",dist="weibull")
ic_par(Surv(time=Mydata$midpoint,time2= Mydata$time2,event =Mydata$status,type="interval")~
arm,data = Mydata,model = "ph",dist="exponential")

#aft model
ic_par(Surv(time=Mydata$midpoint,time2= Mydata$time2,event =Mydata$status,type="interval")~
Mydata$arm,data = Mydata,model = "aft",dist="weibull")
ic_par(Surv(time=Mydata$midpoint,time2= Mydata$time2,event =Mydata$status,type="interval")~
Mydata$arm,data = Mydata,model = "aft",dist="loglogistic")
ic_par(Surv(time=Mydata$midpoint,time2= Mydata$time2,event =Mydata$status,type="interval")~
Mydata$arm,data = Mydata,model = "aft",dist="lnorm")

survdif(mid_surv~Mydata$arm)

Mydata1= subset(subset_tbl,select=c("l_elim", "u_elim", "arm", "logage", "mutat", "gen-
der", "anemia", "pzero"))
attach(Mydata1)
Mydata1$status = ifelse( is.na(u_elim), 0, 1)
mid_low2 = ifelse(is.na(l_elim),0,ifelse(l_elim==0,1,l_elim+1))
mid_upp2= ifelse(is.na(u_elim),45,u_elim+1)
Mydata1$midpoint = ifelse( is.na(u_elim), l_elim, (mid_low2+mid_upp2)/2)
Mydata1$midpoint = ifelse(Mydata1$midpoint==0,0.0001,Mydata1$midpoint)
Mydata1$time2 = rep(NA,145)

mid_surv2=Surv(Mydata1$midpoint,Mydata1$status)
print(mid_surv2)
mid_fit2=survfit(mid_surv2~1)
summary(mid_fit2)

#comparison
mid_arm2=survfit(mid_surv2~Mydata1$arm,conf.type="log-log")
ggsurvplot(mid_arm2,data = Mydata1, pval = T)

survdif(mid_surv2~Mydata1$arm)
#cox model
cox_mid=coxph(mid_surv2~arm,data = Mydata1)

#ph model
ic_par(Surv(time=Mydata1$midpoint,time2= Mydata1$time2,event =Mydata1$status,type="interval")~

```

```

arm,data = Mydata1,model = "ph",dist="weibull")
ic_par(Surv(time=Mydata1$midpoint,time2= Mydata1$time2,event =Mydata1$status,type="interval")~
arm,data = Mydata1,model = "ph",dist="exponential")

#aft model
mid_aft = ic_par(Surv(time=Mydata1$midpoint,time2= Mydata1$time2,event =Mydata1$status
,type="interval")~ arm,data = Mydata1,model = "aft",dist="weibull")
ic_par(Surv(time=Mydata1$midpoint,time2= Mydata1$time2,event =Mydata1$status ,type="interval")~
arm,data = Mydata1,model = "aft",dist="loglogistic")
ic_par(Surv(time=Mydata1$midpoint,time2= Mydata1$time2,event =Mydata1$status ,type="interval")~
arm,data = Mydata1,model = "aft",dist="lnorm")
mid_aft

mid_elim = survreg(Surv(Mydata1$midpoint,Mydata1$status) arm,data = Mydata1,dist='weibull')
weib_elim = survreg(Surv(ttelim ,subset_tbl$event_elim) arm,data = subset_tbl,dist='weibull')

#plot(NPMLE2$time,c(0,1),type="l",xlab="Time",ylab="",cex.lab =1, cex.axis =
1,
xlim=c(0,50),xaxt="n")

#INTERVAL CENSORING
#INTERVAL CENSORED APPROACH

#time to emergence
t1_sub=subset(Table, select=c("l_emer","u_emer","arm"))
t1_sub$low=as.integer(unlist(t1_sub$l_emer))
t1_sub$upp=as.integer(unlist(t1_sub$u_emer))
attach(t1_sub)
#must be a dataframe and not any other type of list
t1_sub=as.data.frame(t1_sub)
t1_sub

### Replace NAs for left- and right-censored observations
t1_sub$low=ifelse(is.na(t1_sub$low),0,ifelse(t1_sub$low==0,1,t1_sub$low+1))
t1_sub$upp= ifelse(is.na(t1_sub$upp),45,t1_sub$upp+1)

NPMLE= EMICM(t1_sub[,4:5])
print(NPMLE)

```

```

NPMLE2=icsurv2cdf(NPMLE)
plot(NPMLE2$time,1-NPMLE2$cdf,type="l",xlab="Time",ylab="",cex.lab =1, cex.axis
= 1,xlim=c(0,50),xaxt="n")
title(ylab=expression(hat(S)(t)),mgp=c(2,1,0))
axis(1, c(0,3,7,14,21,28,42,45), c(0,3,7,14,21,28,42,45))
lines(surv_fit, col = "red",conf.int = F)
lines(mid_fit,col = "blue",conf.int = F)
legend(x="topright",legend=c("IC", "R", "MP"),fill=c("black", "red", "blue"))

```

A.6 Interval-censored approach

```

#INTERVAL CENSORED APPROACH
t1_sub2=subset(Table, select=c("l_elim", "u_elim", "totalgmci", "arm"))
t1_sub2=t1_sub2 %>% filter(totalgmci>0)
t1_sub2$low2=as.integer(unlist(t1_sub2$l_elim))
t1_sub2$upp2=as.integer(unlist(t1_sub2$u_elim))

#must be a dataframe and not any other type of list
t1_sub2=as.data.frame(t1_sub2)

### Replace NAs for left- and right-censored observations

t1_sub2$low2 = ifelse(is.na(t1_sub2$l_elim),0,ifelse(t1_sub2$l_elim==0,1,t1_sub2$l_elim+1))
t1_sub2$upp2= ifelse(is.na(t1_sub2$u_elim),45,t1_sub2$u_elim+1)
t1_sub2

npmle= EMICM(t1_sub2[,5:6])
print(npmle)

npmle2=icsurv2cdf(npmle)
plot(npmle2$time,1-npmle2$cdf,type="l",xlab="Time",ylab="",xlim=c(0,45),xaxt="n")
title(ylab=expression(hat(S)(t)),mgp=c(2,1,0))
axis(1, c(0,3,7,14,21,28,42,45), c(0,3,7,14,21,28,42,45))
lines(surv_fit2,col = "red",conf.int = F)
lines(mid_fit2,col = "blue",conf.int = F)
legend(x="topright",legend=c("IC", "R", "MP"),fill=c("black", "red", "blue"))
abline(h=0.5,col="green")

#COMPARISON OF ARM

```

```
f1=icfit(Surv(low,upp,type="interval2")~arm,data=t1_sub)
cols=c("blue","red")
plot(f1,shade=F,LTY=c(1,5),LEGEND = F)
lines(surv_arm, col = c("red"),lty=c(1,5))
lines(mid_arm, col =c("blue") ,lty=c(1,5))
legend(x="bottomleft",legend=c("IC:ACT","R:ACT","MP:ACT","IC:SP","R:SP","MP:SP"),
fill=c("black","red","blue"), lty = c(1,1,1,5,5,5))
```

```
f1
f1.test = ictest(Surv(low,upp,type="interval2")~arm,data=t1_sub,scores="logrank2",
method="pctl",icontrol=icfitControl(maxit = 100000))
ictest(Surv(low,upp,type="interval2")~arm,data=t1_sub,scores="logrank2",
method="pctl",icontrol=icfitControl(maxit = 100000))
ictest(Surv(low,upp,type="interval2")~arm,data=t1_sub,scores="logrank1",
method="pctl",icontrol=icfitControl(maxit = 100000))
ictest(Surv(low,upp,type="interval2")~arm,data=t1_sub,scores="wmw",
method="scoretest",icontrol=icfitControl(maxit = 100000))
```

```
#####
f2=icfit(Surv(low2,upp2,type="interval2")~arm,data=t1_sub2)
cols=c("blue","red")
plot(f2,shade=F,LTY=c(1,5),LEGEND = F)
lines(surv_arm2, col = c("red"),lty=c(1,5))
lines(mid_arm2, col =c("blue") ,lty=c(1,5))
legend(x="topright",legend=c("IC:SP","R:SP","MP:SP","IC:ACT","R:ACT","MP:ACT"),
fill=c("black","red","blue"), lty = c(1,1,1,5,5,5))
```

```
f2
ictest(Surv(low2,upp2,type="interval2")~arm,data=t1_sub2,scores="logrank2",
method="pctl",icontrol=icfitControl(maxit = 100000))
```

```
summary(f1)
summary(f2)
```

```
A = subset(t1_sub2,select=c("low2","upp2"))
B = subset(t1_sub,select=c("low","upp"))
```

```
#IC
kSampleIcens(A,t1_sub2[, "arm"],rho =0,gamma=0)
```

```
kSampleIcens(A,t1_sub2[, "arm"],rho =1,gamma=0)
kSampleIcens(A,t1_sub2[, "arm"],rho =0,gamma=1)
```

```
kSampleIcens(B,t1_sub[, "arm"],rho =0,gamma=0)
kSampleIcens(B,t1_sub[, "arm"],rho =1,gamma=0)
kSampleIcens(B,t1_sub[, "arm"],rho =0,gamma=1)
```

```
#midpoint ,
logrank.test(time =Mydata$midpoint ,event = Mydata$status,group = Mydata$arm,rho
= 0, gamma=0)
logrank.test(time =Mydata$midpoint ,event = Mydata$status,group = Mydata$arm,rho
= 1, gamma=0)
logrank.test(time =Mydata$midpoint ,event = Mydata$status,group = Mydata$arm,rho
= 0, gamma=1)
```

```
#midpoint ,
logrank.test(time =Mydata1$midpoint ,event = Mydata1$status,group = Mydata1$arm,rho
= 0, gamma=0)
logrank.test(time =Mydata1$midpoint ,event = Mydata1$status,group = Mydata1$arm,rho
= 1, gamma=0)
logrank.test(time =Mydata1$midpoint ,event = Mydata1$status,group = Mydata1$arm,rho
= 0, gamma=1)
```

```
#C = subset(Table,select=c("Time to Emergence", "time2_emerg"))
#RI
#arm2 = getElement(Table,28)
#kSampleIcens(C,arm2,rho =0,gamma=0)
#kSampleIcens(as.matrix(C),Table[, "arm"],rho =1,gamma=0)
```

```
survdif(surv.rel2~subset_tbl$arm,rho =1 )
survdif(surv.rel2~subset_tbl$arm, rho =0 )
```

```
#uncensored times cannot be equal = 0. Try replacing exact times = 0 with really
small numbers
#right censored times must be set to inf
```

```
table_sub=subset(Table,select=c("l_emer", "u_emer", "arm", "logage", "mutat", "gender", "ane-
mia", "pzero"))
table_sub2 = subset(subset_tbl,select=c("l_elim", "u_elim", "arm", "logage", "mutat", "gen-
der", "anemia", "pzero"))
```

```

table_sub$low = ifelse(is.na(table_sub$l_emer),0,ifelse(table_sub$l_emer==0,1,table_sub$l_emer+1))

table_sub$up= ifelse(is.na(table_sub$u_emer),Inf,table_sub$u_emer+1)

table_sub2$low2 = ifelse(is.na(subset_tbl$l_elim),0,ifelse(subset_tbl$l_elim==0,1,subset_tbl$l_elim+1))

table_sub2$up2 = ifelse(is.na(subset_tbl$u_elim),Inf,subset_tbl$u_elim+1)

#what do the tables look like
table_sub
table_sub2

attach(table_sub2)
#cox
model_33=ic_sp(Surv(low,up,type="interval2")~arm,data=table_sub,model="ph" , bs_samples=100)
model_34=ic_sp(Surv(low2,up2,type="interval2")~arm,data=table_sub2,model="ph" , bs_samples=100)

#diag_baseline(cbind(low2,up2) ~ arm, model = "ph",data = table_sub2, lgdLoca-
tion = "topright")

#weibull
# model_43=ic_par(Surv(low_emer,upp_emer,type="interval2")~arm,data=Table,model="ph",
dist="weibull" )
# model_44=ic_par(Surv(low_elim,upp_elim,type="interval2")~arm,data=Table,model="ph",
dist="weibull")

#aft
attach(table_sub)
ic_par(Surv(low,up,type="interval2")~arm,data=table_sub,model="ph",dist="weibull")
ic_par(Surv(low,up,type="interval2")~arm,data=table_sub,model="ph",dist="exponential")

# ic_par(Surv(low,up,type="interval2")~arm,data=table_sub,model="aft",dist="weibull")
# ic_par(Surv(low,up,type="interval2")~arm,data=table_sub,model="aft",dist="loglogis-
tic")
# ic_par(Surv(low,up,type="interval2")~arm,data=table_sub,model="aft",dist="lnorm")

```

```

attach(table_sub)

# ic_par(Surv(low2,up2,type="interval2")~arm,data=table_sub2,model="ph",dist="weibull")
# ic_par(Surv(low2,up2,type="interval2")~arm,data=table_sub2,model="ph",dist="ex-
ponential")

ic_aft=ic_par(Surv(low2,up2,type="interval2")~arm,data=table_sub2,model="aft",dist="weibull")
ic_aft
ic_par(Surv(low2,up2,type="interval2")~arm,data=table_sub2,model="aft",dist="loglogis-
tic")
ic_par(Surv(low2,up2,type="interval2")~arm,data=table_sub2,model="aft",dist="lnorm")

#PARAMETRIC diagnostics
diag_baseline(cbind(low,up) ~ arm,
model = "ph",
data = table_sub,
lgdLocation = "topright")

#duration calculations

dat1 = data.frame(arm="SP/ART")
dat2 = data.frame(arm="SP")

predict(mid_aft,dat1,type="response")
predict(mid_aft,dat2,type="response")

predict(ri_aft,dat1,type="response",interval = "confidence")
predict(ri_aft,dat2,type="response")

predict(ic_aft,dat1,type="response")
predict(ic_aft,dat2,type="response")

predict(weib_elim,newdata=dat1,type='response', se.fit=T)

predict(weib_elim,newdata=dat2,type='response', se.fit=T)

predict(mid_elim,dat1,type='response', se.fit=T)
predict(mid_elim,dat2,type='response', se.fit=T)

```

```
predict(ic_elim,dat1,type='response',se.fit = T)
predict(ic_elim,dat2,type='response', se.fit = T)
```

A.7 Building up the full model

#parametric adding covariates

```
#BUILDING UP THE FULL MODEL
```

```
#Function doesnt work where there are NAs
```

```
table_sub=na.omit(table_sub[, -c(1:2)])
```

```
attach(table_sub)
model_52=ic_sp(Surv(low,up,type="interval2")~logage+mutat+gender+arm+anemia+pzero,
data=table_sub,model="ph",bs_samples=100)
model_53=ic_par(Surv(low,up,type="interval2")~logage+mutat+gender+arm+anemia+pzero,
data=table_sub,model="ph",dist="weibull" )
model_55=ic_par(Surv(low,up,type="interval2")~logage+mutat+gender+arm+anemia+pzero,
data=table_sub,model="ph",dist="exponential" )
```

```
#unnecessary distributions
#model_56=ic_par(Surv(low_emer,upp_emer,type="interval2")~logage+mutat+
+gender+arm+anemia+pzero,data=tbl2,model="ph",dist="generalgamma" )
#model_57=ic_par(Surv(low_emer,upp_emer,type="interval2")~logage+mutat+
+gender+arm+anemia+pzero,data=tbl2,model="ph",dist="lnorm" )
#model_54=ic_par(Surv(low_emer,upp_emer,type="interval2")~logage+mutat+
+gender+arm+anemia+pzero,data=tbl2,model="ph",dist="exponential" )
```

```
#diagnostics
diag_baseline(cbind(low,up) ~ logage+mutat+gender+arm+anemia+pzero, model = "ph",
data = table_sub, lgdLocation = "topright")
```

```
diag_baseline(cbind(low,up) ~ logage+mutat+gender+arm+anemia+pzero, model =
"ph", dist="weibull", data = table_sub, lgdLocation = "topright")
```

```
diag_baseline(cbind(low,up) ~ logage+mutat+gender+arm+anemia+pzero, model =
```

```

“ph”, dist=“exponential”, data = table_sub, lgdLocation = “topright”)

diag_baseline(cbind(low,up) ~ logage+mutat+gender+arm+anemia+pzero, model =
“ph”, dist=“gamma”, data = table_sub, lgdLocation = “topright”)

diag_baseline(cbind(low,up) ~ logage+mutat+gender+arm+anemia+pzero, model =
“ph”, dist=“loglogistic”, data = table_sub, lgdLocation = “topright”)

-2*(model_52$llk) + 12
-2*(model_53$llk) + 12
-2*(model_55$llk) + 12

trt=data.frame(unlist(model_52$coefficients))
se=data.frame(unlist(sqrt(diag(model_52$var))))
z=qnorm(1.-0.05/2)

#confidence intervals
cis = data.frame(matrix(nrow = nrow(se), ncol = 2))
colnames(cis) = c(“l”, “u”)
for(i in 1:nrow(se)){
  cis[i,1]=trt[i,1]-z*se[i,1]
  cis[i,2] = trt[i,1]+z*se[i,1]

}
print(round(cis,3))

#build model by removing one by one
c1=ic_sp(Surv(low,up,type=“interval2”)~logage+mutat+gender+arm+anemia+pzero,
data=table_sub,model=“ph”,bs_samples=100)

c2=ic_sp(Surv(low,up,type=“interval2”)~logage+mutat+arm+anemia+pzero,
data=table_sub,model=“ph”,bs_samples=100)

c3=ic_sp(Surv(low,up,type=“interval2”)~logage+anemia+arm+pzero,
data=table_sub,model=“ph”,bs_samples=100)

c4=ic_sp(Surv(low,up,type=“interval2”)~logage+arm+pzero,
data=table_sub,model=“ph”,bs_samples=100)

```

```
c5=ic_sp(Surv(low,up,type="interval2")~logage+arm,data=table_sub,model="ph",
bs_samples=100)
```

```
-2*(c1$llk) + 12
-2*(c2$llk) + 10
-2*(c3$llk) + 8
-2*(c4$llk) + 6
-2*(c5$llk) + 4
#-2*(W4$llk) + 6
```

```
c1
c2
c3
c4
c5
```

```
#deviance tests
(c2$llk-c1$llk)*-2
(c3$llk-c2$llk)*-2
(c4$llk-c3$llk)*-2
(c5$llk-c4$llk)*-2
```

```
trt=data.frame(unlist(c4$coefficients))
se=data.frame(unlist(sqrt(diag(c4$var))))
z=qnorm(1.-0.05/2)
```

```
#confidence intervals
cis = data.frame(matrix(nrow = nrow(se), ncol = 2))
colnames(cis) = c("l", "u")
for(i in 1:nrow(se)){
  cis[i,1]=trt[i,1]-z*se[i,1]
  cis[i,2] = trt[i,1]+z*se[i,1]

}
print(round(cis,3))
```

```
diag_covar(cbind(low,up)~logage+arm+pzero,data=table_sub,model="ph",lgdLocation
= "right")
```

```

#create new variable
table_sub$t1 = ifelse(table_sub$up>4,1,0)
table_sub$p_time = table_sub$pzero*table_sub$t1
table_sub$p_time2 = table_sub$pzero*table_sub$t1*table_sub$t1
c4_time=ic_sp(Surv(low,up,type="interval2")~logage+arm+strata(pzero)
+p_time,data=table_sub,model="ph",bs_samples=100)

#diag_covar(cbind(low,up)~logage+arm+pzero+p_time,data=table_sub,model="ph",lgdLocation
= "right")

Int_model =ic_sp(Surv(low,up,type="interval2")~logage+arm+pzero+p_time,
data=table_sub,model="ph",bs_samples=100)
Int_model
diag_covar(cbind(low,up)~logage+arm+pzero+p_time,data=table_sub,model="ph",lgdLocation
= "right")

final_int = ic_sp(Surv(low,up,type="interval2")~arm+pzero+p_time,data=table_sub,model="ph",
bs_samples=100)
final_int

(Int_model$llk-final_int$llk)*-2

-2*(Int_model$llk) + 6
-2*(final_int$llk) + 6

trt=data.frame(unlist(Int_model$coefficients))
se=data.frame(unlist(sqrt(diag(Int_model$var))))
z=qnorm(1.-0.05/2)

#confidence intervals
cis = data.frame(matrix(nrow = nrow(se), ncol = 2))
colnames(cis) = c("l", "u")
for(i in 1:nrow(se)){
  cis[i,1]=trt[i,1]-z*se[i,1]
  cis[i,2] = trt[i,1]+z*se[i,1]

}
print(round(cis,3))

```

```
attach(table_sub2)
table_sub2=na.omit(table_sub2[, -c(1:2)])
model_44=ic_par(Surv(low2,up2,type="interval2")~logage+mutat
+gender+arm+anemia+pzero,data=table_sub2,model="aft",dist="weibull" )
```

```
model_44
```

```
model_42=ic_par(Surv(low2,up2,type="interval2")~logage+mutat
+gender+arm+anemia+pzero,data=table_sub2,model="aft",dist="lnorm" )
```

```
model_42
```

```
model_41=ic_par(Surv(low2,up2,type="interval2")~logage+mutat
+gender+arm+anemia+pzero,data=table_sub2,model="aft",dist="loglogistic" )
```

```
model_41
```

```
#Unnecessary distributions
#model_42=ic_sp(Surv(low_elim,upp_elim,type="interval2")~logage+mutat+
+gender+arm+anemia+pzero,data=tbl2,model="ph",bs_samples=10)
#model_43=ic_par(Surv(low_elim,upp_elim,type="interval2")~logage+mutat+
+gender+arm+anemia+pzero,data=tbl2,model="aft",dist="" )
#calculating AICs
-2*(model_44$llk) + 16
-2*(model_42$llk) + 16
-2*(model_41$llk) + 16
```

```
trt=data.frame(unlist(model_41$coefficients))
se=data.frame(unlist(sqrt(diag(model_41$var))))
z=qnorm(1.-0.05/2)
```

```
#confidence intervals
cis = data.frame(matrix(nrow = nrow(se), ncol = 2))
colnames(cis) = c("l", "u")
for(i in 1:nrow(se)){
  cis[i,1]=trt[i,1]-z*se[i,1]
  cis[i,2] = trt[i,1]+z*se[i,1]
```

```

    }
print(round(cis,3))

```

```

W1=ic_par(Surv(low2,up2,type="interval2")~logage+mutat
+gender+arm+anemia+pzero,data=table_sub2,model="aft",dist="lnorm" )

```

```

W1
#confidence intervals
round(confint(W1),3)

```

```

W2=ic_par(Surv(low2,up2,type="interval2")~logage+mutat +gender+arm+pzero,
data=table_sub2,model="aft",dist="lnorm" )

```

```

W2
#confidence intervals
round(confint(W2),3)

```

```

W3=ic_par(Surv(low2,up2,type="interval2")~logage+mutat+arm+pzero,
data=table_sub2,model="aft",dist="lnorm" )

```

```

W3_surv=survreg(Surv(low2,up2,type='interval2')~logage+mutat+arm+pzero,
data=table_sub2,dist='lognormal' )

```

```

W3

```

```

W4=ic_par(Surv(low2,up2,type="interval2")~logage+arm+pzero,
data=table_sub2,model="aft",dist="lnorm" )

```

```

#confidence intervals
round(confint(W3),3)

```

```

#calculating AICs
-2*(W1$llk) + 16
-2*(W2$llk) + 14
-2*(W3$llk) + 12
-2*(W4$llk) + 10

```

```

    trt=data.frame(unlist(W3$coefficients))
    se=data.frame(unlist(sqrt(diag(W3$var))))
    z=qnorm(1.-0.05/2)

    #confidence intervals
    cis = data.frame(matrix(nrow = nrow(se), ncol = 2))
    colnames(cis) = c("l", "u")
    for(i in 1:nrow(se)){
    cis[i,1]=trt[i,1]-z*se[i,1]
    cis[i,2] = trt[i,1]+z*se[i,1]

    }
    print(round(cis,3))

    #deviance tests
    (W2$llk-W1$llk)*-2
    (W3$llk-W2$llk)*-2
    (W4$llk-W3$llk)*-2

    #DURATION

    # attach(Table)
    # duration_tbl=Table %>% filter(totalgmcil>0)%>% select(c(logage,mutat
    ,gender,arm,anemia,pzero,l_elim,u_elim))
    # duration_tbl= na.omit(duration_tbl[,1:6])
    # attach(duration_tbl)
    # # model_76=ic_par(Surv(l_elim,u_elim,type="interval2")~logage+mutat
    +gender+arm+anemia+pzero,data=duration_tbl,model="aft",dist="lnorm")
    # # model_76
    # # # model_77=ic_par(Surv(l_elim,u_elim,type="interval2")~logage+mutat
    +gender+arm+anemia+pzero,data=duration_tbl,model="aft",dist="weibull")
    # # model_77
    # # # model_78=ic_par(Surv(l_elim,u_elim,type="interval2")~logage+mutat
    +gender+arm+anemia+pzero,data=duration_tbl,model="aft",dist="loglogistic")
    # # model_78
    # # # -2*(model_76$llk) + 8
    # -2*(model_77$llk) + 8
    # -2*(model_78$llk) + 8

    #quantiles to use for continous covariates
    quantile(logage, probs = c(2.5,97.5)/100)

```

```
quantile(pzero, probs = c(2.5,97.5)/100)

new_data= data.frame(pzero=5.186,mutat=0,arm="SP/ART",logage=1.731)
new_data1= data.frame(pzero=5.186,mutat=0,arm="SP/ART",logage=0.301)

new_data2= data.frame(pzero=5.186,mutat=0,arm="SP",logage=1.731)
new_data3= data.frame(pzero=5.186,mutat=0,arm="SP",logage=0.301)

new_data4= data.frame(pzero=5.186,mutat=1,arm="SP/ART",logage=1.731)
new_data5= data.frame(pzero=5.186,mutat=1,arm="SP/ART",logage=0.301)

new_data6= data.frame(pzero=5.186,mutat=1,arm="SP",logage=1.731)
new_data7= data.frame(pzero=5.186,mutat=1,arm="SP",logage=0.301)

new_data8= data.frame(pzero=1.348,mutat=0,arm="SP/ART",logage=1.731)
new_data9= data.frame(pzero=1.348,mutat=0,arm="SP/ART",logage=0.301)

new_data10= data.frame(pzero=1.348,mutat=0,arm="SP",logage=1.731)
new_data11= data.frame(pzero=1.348,mutat=0,arm="SP",logage=0.301)

new_data12= data.frame(pzero=1.348,mutat=1,arm="SP/ART",logage=1.731)
new_data13= data.frame(pzero=1.348,mutat=1,arm="SP/ART",logage=0.301)

new_data14= data.frame(pzero=1.348,mutat=1,arm="SP",logage=1.731)
new_data15= data.frame(pzero=1.348,mutat=1,arm="SP",logage=0.301)

predict(W3,new_data,type="response")
predict(W3,new_data1,type="response")
predict(W3,new_data2,type="response")
predict(W3,new_data3,type="response")
predict(W3,new_data4,type="response")
predict(W3,new_data5,type="response")
predict(W3,new_data6,type="response")
predict(W3,new_data7,type="response")
predict(W3,new_data8,type="response")
predict(W3,new_data9,type="response")
predict(W3,new_data10,type="response")
```

```
predict(W3,new_data11,type="response")
predict(W3,new_data12,type="response")
predict(W3,new_data13,type="response")
predict(W3,new_data14,type="response")
predict(W3,new_data15,type="response")
```

```
#used to extract s.e.
predict(W3_surv,new_data,type = 'response', se.fit = T)
```

A.8 Doubly censored

```
#DOUBLY CENSORED
attach(Table)
# find the upper limit and lower limit for the gap time Ti
gap_tbl=Table %>% select(totalgmcil,Du_elim,l_emer,Dl_elim,u_emer,logage,mutat ,gender,arm,anemia,pzero)
```

```
#remove those subjects without gametocytes ie the left censored and right censored patients
gap_tbl = gap_tbl %>% filter(totalgmcil>0)
```

```
#remove NA for anemia to prevent error
gap_tbl=gap_tbl %>% filter(!is.na(anemia))
```

```
attach(gap_tbl)
gap_tbl$Dl_elim = ifelse(gap_tbl$Dl_elim==0,1,gap_tbl$Dl_elim+1)
```

```
gap_tbl$Du_elim= ifelse(is.na(gap_tbl$Du_elim),Inf,gap_tbl$Du_elim+1)
```

```
gap_tbl$l_emer= ifelse(is.na(gap_tbl$l_emer ),0,ifelse(gap_tbl$l_emer ==0,1,gap_tbl$l_emer+1))
```

```
gap_tbl$u_emer= ifelse(is.na(gap_tbl$u_emer),Inf,gap_tbl$u_emer+1)
```

```
attach(gap_tbl)
gap_tbl2=gap_tbl %>% mutate(upp_gap= gap_tbl$Du_elim-gap_tbl$l_emer) %>% mutate(low_gap= gap_tbl$Dl_elim-gap_tbl$u_emer) %>% mutate(low_gap= ifelse(low_gap==0,0.0001,low_gap+0.0001)) %>%
```

```

select(low_gap,upp_gap,Du_elim,l_emer,Dl_elim,u_emer,logage,mutat ,gender,arm,anemia,pzero)

#cox model
attach(gap_tbl2)
dbl_cox = ic_sp(Surv(gap_tbl2$low_gap,gap_tbl2$upp_gap,type="interval2")~logage+arm+
pzero+gender+anemia+mutat,data=gap_tbl2, bs_samples = 100)
dbl_cox

dbl_cox2= ic_sp(Surv(gap_tbl2$low_gap,gap_tbl2$upp_gap,type="interval2")~logage+arm+
pzero+gender+mutat,data=gap_tbl2, bs_samples = 100)
dbl_cox2

dbl_cox3=ic_sp(Surv(gap_tbl2$low_gap,gap_tbl2$upp_gap,type="interval2")~logage+arm+
pzero+mutat,data=gap_tbl2, bs_samples = 100)
dbl_cox3

dbl_cox4=ic_sp(Surv(gap_tbl2$low_gap,gap_tbl2$upp_gap,type="interval2")~logage+arm+
pzero,data=gap_tbl2, bs_samples = 100)
dbl_cox4

trt=data.frame(unlist(dbl_cox3$coefficients))
se=data.frame(unlist(sqrt(diag(dbl_cox3$var))))
z=qnorm(1.-0.05/2)

#confidence intervals
cis = data.frame(matrix(nrow = nrow(se), ncol = 2))
colnames(cis) = c("l", "u")
for(i in 1:nrow(se)){
cis[i,1]=exp(trt[i,1]-z*se[i,1])
cis[i,2] = exp(trt[i,1]+z*se[i,1])

}
round(print(cis),3)

#PH model
dbl_ph= ic_par(Surv(gap_tbl2$low_gap,gap_tbl2$upp_gap,type="interval2")~logage+
arm+pzero+gender+anemia+mutat,data=gap_tbl2, model="ph", dist="weibull")
dbl_ph

```

```
dbl_ph2= ic_par(Surv(gap_tbl2$low_gap,gap_tbl2$supp_gap,type="interval2")~logage+
arm+pzero+anemia+mutat,data=gap_tbl2, model="ph", dist="weibull")
dbl_ph2
```

```
dbl_ph3 = ic_par(Surv(gap_tbl2$low_gap,gap_tbl2$supp_gap,type="interval2")~logage+
arm+pzero+mutat,data=gap_tbl2, model="ph", dist="weibull")
dbl_ph3
```

```
dbl_ph4=ic_par(Surv(gap_tbl2$low_gap,gap_tbl2$supp_gap,type="interval2")~logage+
arm+pzero,data=gap_tbl2, model="ph", dist="weibull")
dbl_ph4
```

```
dbl_ph5=ic_par(Surv(gap_tbl2$low_gap,gap_tbl2$supp_gap,type="interval2")~logage+
arm,data=gap_tbl2, model="ph", dist="weibull")
dbl_ph5
```

```
trt=data.frame(unlist(dbl_ph5$coefficients))
se=data.frame(unlist(sqrt(diag(dbl_ph5$var))))
z=qnorm(1.-0.05/2)
```

```
#confidence intervals
cis = data.frame(matrix(nrow = nrow(se), ncol = 2))
colnames(cis) = c("l", "u")
for(i in 1:nrow(se)){
cis[i,1]=exp(trt[i,1]-z*se[i,1])
cis[i,2] = exp(trt[i,1]+z*se[i,1])

}
round(print(cis),3)
```

```
(dbl_ph2$llk-dbl_ph$llk)*-2
(dbl_ph3$llk-dbl_ph2$llk)*-2
(dbl_ph4$llk-dbl_ph3$llk)*-2
(dbl_ph5$llk-dbl_ph4$llk)*-2
```

```
#AFT model
```

```
dbl_aft=ic_par(Surv(gap_tbl2$low_gap,gap_tbl2$supp_gap,type="interval2")~logage+
arm+pzero+gender+anemia+mutat,data=gap_tbl2, model="aft", dist="lnorm")
```

```
dbl_aft
```

```
#refine aft model
dbl_aft2=ic_par(Surv(gap_tbl2$low_gap,gap_tbl2$upp_gap,type="interval2")~logage+
arm+gender+pzero+mutat,data=gap_tbl2, model="aft", dist="lnorm")
dbl_aft2
```

```
dbl_aft3=ic_par(Surv(gap_tbl2$low_gap,gap_tbl2$upp_gap,type="interval2")~logage+
mutat+arm+pzero,data=gap_tbl2, model="aft", dist="lnorm")
dbl_aft3
```

```
dbl_aft_surv = survreg(Surv(gap_tbl2$low_gap, gap_tbl2$upp_gap, type = 'interval2') ~logage+mutat+
arm+pzero,data=gap_tbl2, dist = 'lognormal')
```

```
trt=data.frame(unlist(dbl_aft$coefficients))
se=data.frame(unlist(sqrt(diag(dbl_aft$var))))
z=qnorm(1.-0.05/2)
```

```
#confidence intervals
cis = data.frame(matrix(nrow = nrow(se), ncol = 2))
colnames(cis) = c("l", "u")
for(i in 1:nrow(se)){
  cis[i,1]=trt[i,1]-z*se[i,1]
  cis[i,2] = trt[i,1]+z*se[i,1]
}
round(print(cis),3)
```

```
predict(dbl_aft3,new_data,type="response")
predict(dbl_aft3,new_data1,type="response")
predict(dbl_aft3,new_data2,type="response")
predict(dbl_aft3,new_data3,type="response")
predict(dbl_aft3,new_data4,type="response")
predict(dbl_aft3,new_data5,type="response")
predict(dbl_aft3,new_data6,type="response")
predict(dbl_aft3,new_data7,type="response")
predict(dbl_aft3,new_data8,type="response")
predict(dbl_aft3,new_data9,type="response")
predict(dbl_aft3,new_data10,type="response")
predict(dbl_aft3,new_data11,type="response")
```

```

predict(dbl_aft3,new_data12,type="response")
predict(dbl_aft3,new_data13,type="response")
predict(dbl_aft3,new_data14,type="response")
predict(dbl_aft3,new_data15,type="response")

# confidence interval for treatment
trt=data.frame(unlist(mid_aft3$coefficients))
se=data.frame(unlist(sqrt(diag(mid_aft3$var))))
z=qnorm(1.-0.05/2)

#confidence intervals
cis = data.frame(matrix(nrow = nrow(se), ncol = 2))
colnames(cis) = c("l", "u")
for(i in 1:nrow(se)){
  cis[i,1]=(trt[i,1]-z*se[i,1])
  cis[i,2] = (trt[i,1]+z*se[i,1])
}
print(round(cis,3))

#data for midpoint
mid_table = subset(Table,select=c("l_emer", "u_emer", "Dl_elim", "Du_elim", "arm",
"logage", "mutat", "gender", "anemia", "pzero", "Time to Emergence", "Time to Elim", "totalgmcil"))
mid_table = mid_table %>% filter(totalgmcil>0)

attach(mid_table)
mid_table$status = ifelse( is.na(Du_elim), 0, 1)
mid_table$time2 = rep(NA,145)

mid_low_emerg =ifelse( is.na(l_emer), 0,ifelse(l_emer==0,1,l_emer+1))
mid_upp_emerg = ifelse(is.na(u_emer),45,u_emer+1)
mid_table$mp_emerg = ifelse( is.na(u_emer),'Time to Emergence', (mid_low_emerg+mid_upp_emerg)/2)

mid_low_elim =ifelse( is.na(Dl_elim), 0,ifelse(Dl_elim==0,1,Dl_elim+1))
mid_upp_elim = ifelse(is.na(Du_elim),45,Du_elim+1)
mid_table$mp_elim = ifelse( is.na(Du_elim),'Time to Elim', (mid_low_elim+mid_upp_elim)/2)

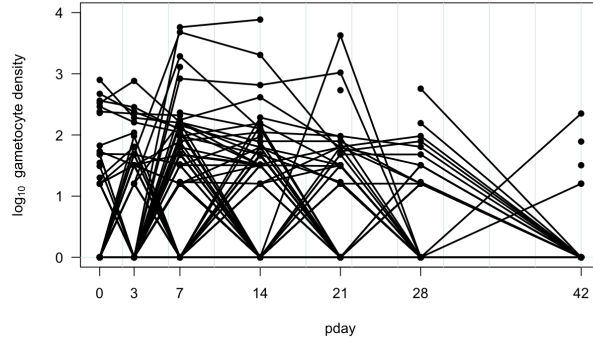
mid_table$middur = mid_table$mp_elim-mid_table$mp_emerg

```

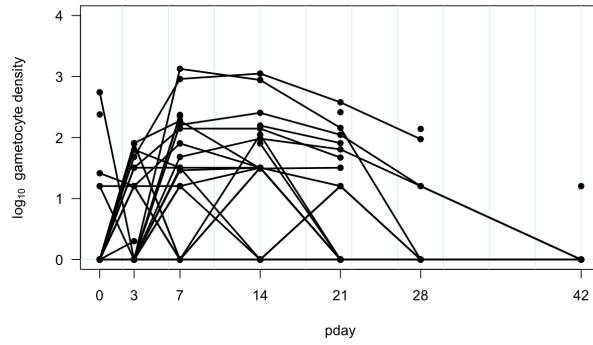
```
mid_aft3=ic_par(Surv(time=mid_table$middur,time2= mid_table$time2,event =mid_table$status
,type="interval")~
logage+mutat+arm+pzero,data = mid_table,model = "aft",dist="lnorm")
```

```
#mid_aft3=ic_par(Surv(mp_emerg,mp_elim,type="interval2")~logage+
mutat+arm+pzero,data=mid_table,model="aft", dist="lnorm")
#mid_aft3
mid_table
```

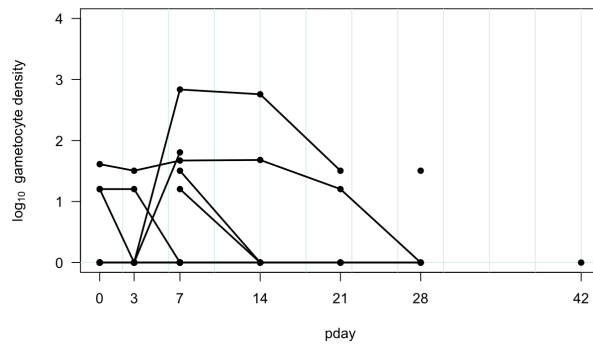
```
predict(mid_aft3,new_data,type="response")
predict(mid_aft3,new_data1,type="response")
predict(mid_aft3,new_data2,type="response")
predict(mid_aft3,new_data3,type="response")
predict(mid_aft3,new_data4,type="response")
predict(mid_aft3,new_data5,type="response")
predict(mid_aft3,new_data6,type="response")
predict(mid_aft3,new_data7,type="response")
predict(mid_aft3,new_data8,type="response")
predict(mid_aft3,new_data9,type="response")
predict(mid_aft3,new_data10,type="response")
predict(mid_aft3,new_data11,type="response")
predict(mid_aft3,new_data12,type="response")
predict(mid_aft3,new_data13,type="response")
predict(mid_aft3,new_data14,type="response")
predict(mid_aft3,new_data15,type="response")
```



(a) Success



(b) Failure



(c) LTFU

Figure 21: All Gametocyte Profile Spaghetti Plots