HOCHSCHULE
MITTWEIDA
UNIVERSITY OF
APPLIED SCIENCES

MASTER THESIS

Mr.
Deepak Singh

Statistical analysis of risk factors associated with malaria and severe malarial anemia (SMA) in children in Western Kenya

2023

Faculty of **Applied Computer and Bio Sciences**

MASTER THESIS

Statistical analysis of risk factors associated with malaria and severe malarial anemia (SMA) in children in Western Kenya

Author:
Deepak Singh

Study Programme:
Applied Mathematics for Network and Data Sciences

Seminar Group:
MA18W1-M

First Referee:
Prof. Dr. rer. nat. habil. Kristan Schneider

Second Referee:
Prof. Dr. rer. nat. habil. Thomas Kalinowski

Mittweida, December 2023

Acknowledgement

I would like to express my sincere gratitude to Prof. Dr. Kristan Schneider for all the guidance and constant help with my Master's thesis. I also want to express my gratitude to my second supervisor Prof. Dr. Thomas Kalinowski for his support. Further, I would like to thank Hochschule Mittweida for giving me the chance to study in this Master's program. Additionally, I would like to thank everyone for all the help and support through this journey.

Bibliographic Information

Singh, Deepak : Statistical analysis of risk factors associated with malaria and severe malarial anemia (SMA) in children in Western Kenya, 89 pages, 13 figures, Hochschule Mittweida, University of Applied Sciences, Faculty of Applied Computer and Bio Sciences

Master Thesis, 2023

Abstract

Malaria remains a major global health concern, particularly affecting vulnerable populations such as children in sub-Saharan Africa regions. This thesis presents a comprehensive statistical analysis of risk factors associated with mortality, malaria, and severe malarial anemia (SMA) in children residing in Western Kenya, an area with high malaria transmission rates.

The study leverages a real dataset collected from two cohorts of children, encompassing demographic, clinical, and laboratory variables. After thorough data analysis, data consists of 1654 children with repeated clinical visits. These children had multiple malaria episodes ranging between 0 to 26, 0 to 4 SMA episodes, and a total of 14 deaths. The primary objectives are to identify and quantify risk factors contributing to the incidence of mortality, malaria, and SMA.

The analytical approach involves the application of various statistical methods, including logistic regression, survival analysis, Cox models, and frailty models. We explore the influence of factors such as age, gender, hemoglobin genotypes, and SMA development.

Key findings from the analysis highlight the significant role of variables like hemoglobin levels with specific genetic and other factors such as HIV and Alphthalsimeia. Furthermore, the study investigates the temporal aspect of malaria episodes, considering recurrent events and frailty models to account for unobserved heterogeneity among study participants.

This research provides valuable insights into the epidemiology of malaria and SMA in the context of Western Kenya. These findings contribute to the development of targeted interventions and public health strategies aimed at reducing the burden of malaria-related morbidity and mortality in children. Additionally, the study underscores the importance of continued surveillance and monitoring to adapt interventions in response to evolving risk factors and challenges.

In conclusion, this thesis contributes to the broader understanding of malaria risk factors and severe malarial anemia in a specific geographical context while emphasizing the need for multifaceted approaches to combat malaria in endemic regions. It serves as a foundation for evidence-based policies and interventions, with the ultimate goal of improving the health and well-being of children in Western Kenya and similar malaria-endemic areas worldwide.

I. Contents

II. List of Figures

III. List of Tables

1 Introduction

1.1 Malaria in sub-Saharan Africa

Malaria remains a formidable global health challenge, particularly in regions where the burden of infectious diseases is most pronounced. The World Health Organization (WHO) estimates that in 2019, there were approximately 227 million reported cases of malaria across 85 endemic countries worldwide. Following the onset of the COVID-19 pandemic and associated service disruptions in 2020, the global estimate for malaria cases increased to 241 million, reflecting a surge of 14 million additional cases compared to the preceding year. Since the year 2000, there has been a consistent decline in malaria-related deaths, decreasing from $896,000$ to $562,000$ in 2015 and further to $558,000$ in 2019. However, in 2020, there was an alarming increase in malaria deaths, reaching an estimated $627,000$, marking a 12% rise from the previous year. Notably, disruptions caused by the COVID-19 pandemic contributed to 68% $(47,000)$ of the additional $69,000$ malaria deaths. The remaining $22,000$ additional deaths signify the rise in mortality between 2019 and 2020 in the absence of pandemic-related disruptions. In the most recent year, the percentage of mortality from malaria among children under the age of five climbed from 4.8% to 7.8% [1].

The report further states that the primary burden of the disease is borne by 32 nations with moderate to high malaria transmission within the WHO African Region. This region constitutes approximately 95% of all malaria cases and 96% of all malaria deaths globally, with children under the age of five contributing to 80% of all deaths [1]. From 2000 to 2019, the malaria cases per $1,000$ persons at risk declined from 368 to 222, although there was a slight increase to 233 in 2020, attributed to disruptions during the COVID-19 pandemic. Malaria mortality witnessed a 36% reduction from 2000 to 2019, decreasing from $840,000$ to $534,000$, but experienced an increase to $602,000$ in 2020. The malaria death rates declined by 63% between 2000 and 2019, dropping from 150 to 56 per $100,000$ people at risk, followed by a rise to 62 in 2020. The rise in cases and deaths between 2019 and 2020 was predominantly attributed to the WHO African Region, accounting for over 95% of the increase. Lower socioeconomic position, lack of awareness, poor infrastructure, lack of healthcare, and sparse or nonexistent intervention strategies are some of the main causes of sub-Saharan Africa's high malaria burden [1].

Predominantly, *Plasmodium falciparum (P. falciparum)* infections account for the majority of cases in the African region. Western Kenya is considered a holoendemic *P. falciparum* transmission region where *falciparum* malaria takes a prominent role as a major contributor to childhood morbidity and mortality [2]. In the holoendemic areas, severe malaria primarily presents as severe malaria anemia (SMA), defined as hemoglobin (Hgb) levels

falling below 5.0 g/dl, especially affecting children under the age of five who have not yet developed natural immunity to malaria [3]. In the past two decades, research in western Kenya has revealed that immune genes play a crucial role in the development of severe malarial anemia (SMA). Furthermore, the variability observed in these genes has a significant impact on susceptibility to *P. falciparum* infections, SMA, and mortality [4].

The high prevalence of malaria in Kenya, particularly in Western Kenya, is influenced by various factors such as favorable climatic conditions for mosquito breeding and limited access to preventive measures [5]. This heightened transmission rate in the region results in increased susceptibility to malaria-related complications among children, placing a significant strain on local healthcare systems and necessitating targeted research efforts [6]. Epidemiological studies have shown an association between malaria disease prevalence and socio-economic status, such as poor housing conditions, in the highlands of Western Kenya. Additionally, the resurgence of malaria infections and vector densities in Western Kenya has underscored the urgent need to elucidate the risk factors contributing to clinical malaria in different areas [5].

1.2 Distribution of *Plasmodium* Species

The distribution of *Plasmodium* species varies worldwide. The frequency of malaria cases and fatalities varies according to the season. Data on prevalence must be connected to each country's endemicity and season [7]. The distribution of *Plasmodium* species varies worldwide. All malarial infections in humans are caused by five species of the genus Plasmodium. The predominant culprits are usually *Plasmodium falciparum* or *Plasmodium vivax*. However, infections can also be attributed to *Plasmodium ovale*, *Plasmodium malaria*, and, in specific areas of Southeast Asia, the monkey malaria *Plasmodium knowlesi* [8].

***Plasmodium falciparum*:** *P. falciparum* is widespread in almost all malaria-endemic countries, and it is the predominant malaria species in the African Region, Southeast Asia, and Western Pacific regions [1, 9]. In Asia, both *P. falciparum* and *P. vivax* are prevalent, with *P. falciparum* being more common in certain regions [7].

***Plasmodium vivax*:** Due to the widespread presence of the Duffy negative phenotype, which hinders the entry of *P. vivax* merozoites into red blood cells, *P. vivax* infections are rare in central and western Africa. Only 5% of malaria cases in eastern and southern Africa are attributed to *P. vivax*. In Asia, *P. vivax* and *P. falciparum* are the two predominant species [7, 10].

***Plasmodium ovale*:** Due to the difficulties in diagnosing *P. ovale* malaria, accurately estimating the actual impact of the disease poses a challenge. *P. ovale* is prevalent in both Asia and sub-Saharan Africa [7]. In Africa, both *Plasmodium ovale curtisi* and *Plas-*

modium ovale wallikeri, two species of *P. ovale*, are common and have been identified in regions such as Congo-Brazzaville, Uganda, and Equatorial Guinea [10].

***Plasmodium malaria*:** *P. malaria* is widespread in various regions, including the Amazon Basin in South America, sub-Saharan Africa, Southeast Asia, several islands in the western Pacific, and Indonesia [10]. Positive cases have been identified in pregnant women in Nigeria, and during the rainy season, some instances were reported in four villages in Mulanda and Uganda [7].

***Plasmodium knowlesi*:** *P. knowlesi* infection, impacting both humans and monkeys, is exclusive to the Southeast Asia region, the species' original discovery habitat. Initially identified in a human in the Pahang rainforest in Peninsular Malaysia in 1965, *Plasmodium knowlesi* is predominantly found in forested areas [7].

1.3 Etiology and Symptom Complexity

Malaria is caused by a parasite of the genus *Plasmodium* and transmitted to humans through the bite of an infected female Anopheles mosquito [11]. Malaria is caused by a diminutive protozoan that falls within the Plasmodium species group, comprising various subspecies. The Plasmodium genus represents an amoeboid intracellular parasite with the ability to accumulate malaria pigment, an insoluble byproduct of hemoglobin metabolism [11]. There are five major *Plasmodium* parasites that cause malaria in humans: *P. falciparum*, *P. vivax*, *P. malaria*, *P. ovale*, and *P. knowlesi* [8].

Malaria's initial symptoms are vague and variable, including fever, headache, weakness, myalgia, chills, dizziness, abdominal pain, diarrhea, nausea, vomiting, anorexia, and pruritus [12]. Shivering, fever, and sweating can be considered the three main early-stage symptoms along with some other supporting symptoms [13]. In malaria caused by *P. vivax* and *P. ovale*, this concept appears in three stages: tremor (cold period), fever (hot period), and sweating (sweating period) [13]. In the endemic areas, malaria patients may show other symptoms such as headache, nausea, vomiting, diarrhea, limpness, and muscle pain. Malaria has symptoms that are similar to dengue fever, typhoid fever, the common cold, respiratory tract infection, dyspepsia, and pneumonia [14]. This symptom overlap with other tropical diseases reduces diagnostic specificity, which can encourage the indiscriminate use of antimalarials and compromise the standard of care for patients with non-malarial fevers in the endemic areas [12].

In General, malaria is classified as asymptomatic, uncomplicated, or severe. All *Plasmodium* species can produce asymptomatic malaria, in which the patient has circulating parasites but no symptoms [15]. *Plasmodium* species can cause uncomplicated malaria and typically, symptoms start 7 to 10 days after the original mosquito bite [13]. There are no clear-cut clinical or laboratory signs of severe organ dysfunction, but symptoms can

include fever, moderate to severe shaking chills, profuse perspiration, headache, nausea, vomiting, diarrhea, and anemia. Infection with *P. falciparum* is typically what causes severe malaria, while it can also rarely be brought on by *P. vivax* or *P. knowlesi* [14]. The risks of cerebral malaria include severe anemia, end-organ damage, coma, respiratory problems (including edema and hyperpneic syndrome), hypoglycemia, and acute kidney disease. Hyperparasitaemia and higher mortality are frequently linked to severe malaria [15, 16].

1.4 Life cycle of Malaria Parasite

The life cycle of the malaria parasite in humans is very complex and distinguished by an exogenous sexual phase (sporogony), in which replication occurs in many Anopheles mosquito species, and an endogenous asexual phase (schizogony), in which replication occurs in vertebrate hosts [17]. The sexual phase of the parasite's life cycle occurs in mosquitoes and the asexual phase of the life cycle occurs in humans [11, 18]. In Anopheles mosquitos, Plasmodium reproduces sexually (by merging the parasite's sex cells). The parasite reproduces asexually (via cell division) in humans, first in liver cells and then repeatedly in red blood cells (RBCs). The *Plasmodium* life cycle takes approximately 8 to 35 days within the mosquito, after which the *Plasmodium* is infectious. When an infected Anopheles mosquito bites a human, it takes blood and at the same time, it injects saliva into the human's bloodstream which contains the infectious form of the parasite, the sporozoite [17, 18].

Each sporozoite generates tens of thousands of merozoites inside the liver cells (hepatocyte), and upon release from the liver, each merozoite is capable of invading an RBC. Depending on the infecting species, the tissue phase can be completed in 8 to 25 days for *P. falciparum*, 8 to 27 days for *P. vivax*, 9 to 17 days for *P. ovale*, and 15 to 30 days for *P. malaria* [17, 18]. This period is known as the prepatent period. The erythrocyte is where asexual division begins, and it is where the parasites go through their various stages of development [10, 17]. Because of its distinctive appearance, the early trophozoite is frequently referred to as the "ring form". Multiple rounds of nuclear division without cytokinesis mark the end of this trophic stage, resulting in the formation of schizonts (as depicted in figure 1.1 by Renu [18]). Each mature schizont contains approximately 20 merozoites, which are released after RBC lysis to infect additional uninfected RBCs [18]. This repeated intraerythrocytic cycle of invasion-multiplication-release-invasion lasts approximately 48 hours in *P. falciparum*, *P. ovale*, and *P. vivax* infections, and 72 hours in *P. malaria* infections. A small percentage of merozoites in RBCs eventually differentiate into micro and macro-gametocytes (male and female, respectively), which have no further activity within the human host [10, 17, 18].

For the virus to spread to new victims through female Anopheles mosquitoes, these gametocytes are crucial. Before any gametocytes are generated, asexual erythrocytic

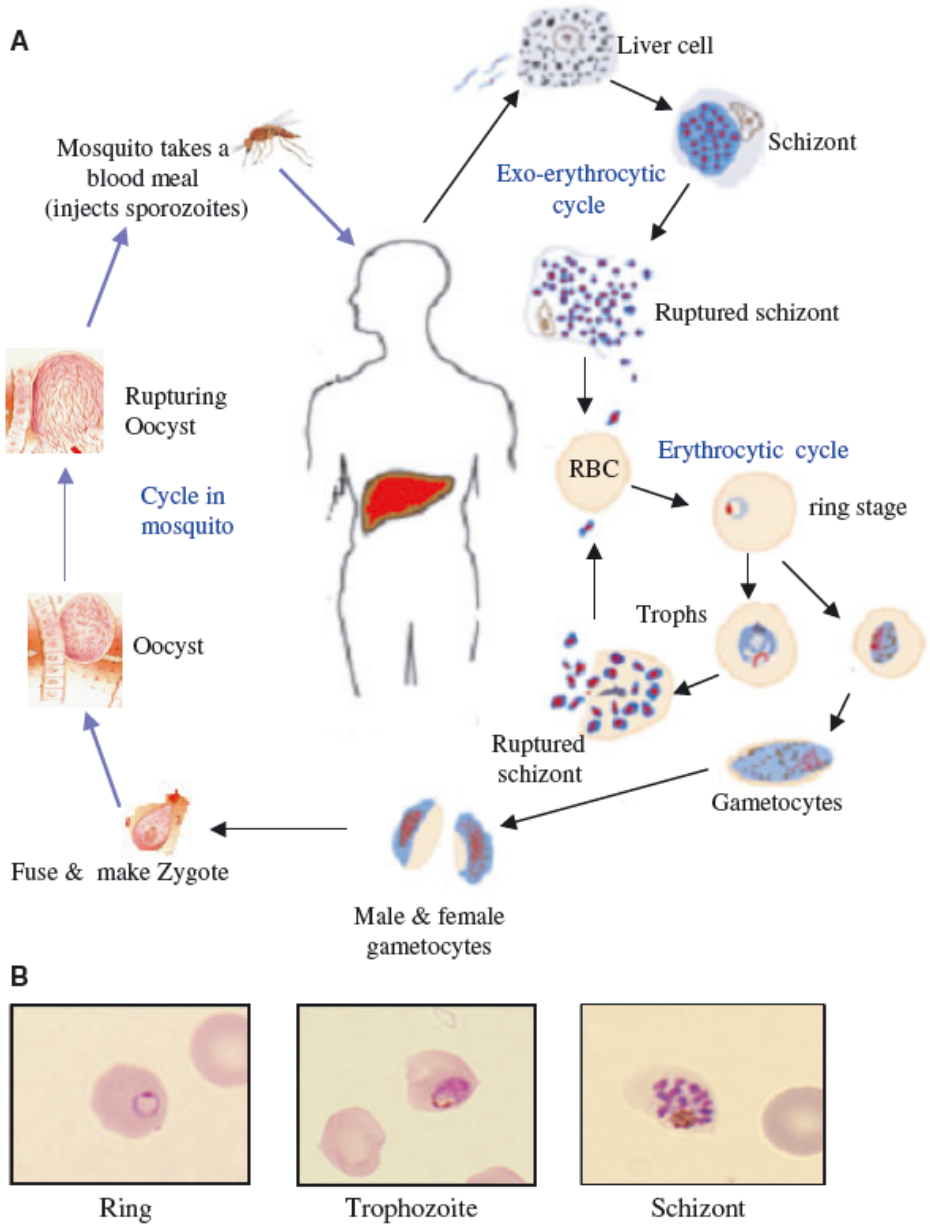

Figure (1.1) (A): Life cycle of malaria parasite *P. falciparum*. (B): Different intraerythrocytic stages of development of *P. falciparum*

schizogony often lasts for several cycles [19]. Erythrocytic schizogony in *P. falciparum* lasts 48 hours, and gametocytogenesis lasts 10 to 12 days. Infections with *P. vivax* and *P. ovale*, and gametocytes begin to develop on the fifth day of the original assault and thereafter grow in number. Infections with P. malaria and gametocytes begin to appear anywhere between 5 and 23 days after the primary attack [8, 10, 17, 18].

During the sexual phase, when a mosquito feeds on blood from an infected person, it may ingest these gametocytes into its midgut, where exflagellation of microgameto-

cytes results in microgametes while macrogametocytes create macrogametes. A zygote is created after the fusion of these gametes and fertilization. This develops into an ookinete, which then becomes an oocyst by penetrating the midgut cell wall [8, 17]. The single extracellular phase of the Plasmodium life cycle, the oocyst formation, is the longest developmental stage (takes 3 to 30 days). Before releasing the contagious sporozoites, the falciparum oocysts mature for 11 to 16 days. When the oocyst ruptures, the sporozoites migrate to the salivary glands, where they can be transmitted to another host. After 10 to 18 days, this form of the parasite is found in the salivary glands, and the mosquito is infected for another one to two months. The Plasmodium life cycle is restarted when an infected mosquito bites a susceptible host [17, 18].

1.5 Severe Malaria and Severe Malarial Anemia

In sub-Saharan Africa, severe malarial anemia (SMA) is one of the main causes of pediatric morbidity, hospitalization, and mortality [20]. Morbidity and mortality are highest in young children in areas with high stable transmission because acquired protective immunity is insufficient in protecting against severe disease [21]. One of the age groups most affected by malaria is children under the age of five. Children are more likely to experience severe malaria symptoms than adults, such as severe anemia, hypoglycemia, and brain malaria [22].

The epidemiology of helminth infections, inherited red cell abnormalities, and nutritional deficiencies all coexist with the epidemiology of malaria, making anemia frequently multifactorial and the distribution of hemoglobin concentrations in healthy individuals lower and broader than in temperate regions [23]. Anemia is defined differently in malaria. The most common classification is based on hemoglobin concentration, which is used in higher malaria transmission areas. In patients with acute malaria, hemoglobin (Hb) concentrations between 8 g/dL (grams per deciliter) and 11 g/dL are considered mild anemia, Hb concentrations between 5 g/dL and 8 g/dL are considered moderate, and Hb concentrations less than 5 g/dL are considered severe anemia [23, 24]. Since anemia quickly sets in with acute malaria, most symptomatic individuals have already lost at least 1 g of hemoglobin per decilitre (100 mL) of blood before seeking medical help. Rapid growth occurs in the liver and spleen. The anemia is hemolytic, therefore haptoglobin and haemopexin concentrations are decreased, unconjugated bilirubin may be elevated, and red cell indices are often normal. Leukocyte counts are typically in the low-normal range, and platelet counts are almost always low. [25]. The complicated etiological basis of the multifactorial disease known as malarial anemia (MA) is only partially understood. One of the primary clinical manifestations of severe malaria induced by P. falciparum is severe MA [8, 26].

Severe malaria is a multisystem, multi-organ disease, children frequently exhibit a mix of the classic clinical phenotypes: cerebral malaria (CM), severe malarial anemia (SMA),

respiratory distress, and hypoglycemia. Out of which, CM and SMA are most common in children [27]. Severe malarial anemia (SMA) is one of the main causes of pediatric morbidity, hospitalization, and mortality in sub-Saharan Africa. SMA is described by the World Health Organization as hemoglobin (Hb) < 5.0 g/dL when malaria parasites are present [20]. However, this definition is only partially useful for the epidemiological analysis of this illness as well as for the therapeutic management of patients. P. falciparum is the most common cause of SMA, but P. knowlesi and P. vivax can also cause severe disease [23]. Vital organ malfunction and even mortality have been linked to malaria infections. Clinical or biochemical evidence of vital organ failure characterizes severe malaria. P. falciparum infections cause the majority of fatal cases of severe malaria [28].

1.6 Addressing Research Questions

This study endeavors to address four pivotal research inquiries. The primary focus is identifying factors influencing mortality, and three distinct scenarios are examined. Initially, a Cox proportional hazard regression model was employed on survival data to scrutinize the association between survival time and specified predictor variables concerning all-cause mortality based on the first clinical visit. This model enables us to draw inferences about the impact of predictor variables on the hazard of experiencing the event. In the second scenario, logistic regression is applied to discern risk factors for mortality based on the initial clinical visit. Lastly, in the third scenario, a frailty model was performed to uncover risk factors associated with mortality while accommodating unobserved heterogeneity among children.

The second primary inquiry revolves around identifying the covariates influencing the frequency of malaria and SMA episodes. This question is tackled through two distinct approaches with the same covariates. Initially, a Poisson regression model is applied, followed by a multiple-event per subject Cox model (Andersen-Gill) in the second approach. These models aid in revealing risk factors associated with the frequency of malaria and SMA episodes throughout the follow-up period. The third major question mirrors the second major question but is primarily focused on the second occurrences of Malaria and SMA events. A multiple-event per subject Cox model (Frailty) is employed to delineate the relationship between the risk factors and malaria episodes. The final research question is based on analyzing the impact of various risk factors on malaria and SMA episodes, given the first episode. A multiple-event per subject Cox model (Frailty) is performed and the frailty term was used to account for any unobserved heterogeneity among different children.

The significance level is set at 5%, and a model search is conducted based on minimizing the AIC, utilizing forward-backward selection to identify the most optimal model among all the candidates. R studio (R version $4.3.0$) was used throughout the thesis.

2 Regression Models and Model Selection

Regression models are extensively employed in epidemiological studies and various other fields. These models, a class of statistical models, play a crucial role in analyzing the statistical relationship between a response variable and covariates or explanatory variables. Named for their formation from a linear combination of predictors and coefficients, these models are referred to as linear models [29]. Regression analysis is a prominent method for establishing and understanding this relationship. The primary goal is to discern how independent variables influence the dependent variable and make predictions based on these relationships. Through regression analysis, we can assess the impact of multiple variables, including covariates and factors, within the same model [30, 31]. Regression models are extensively employed in epidemiological studies and various other fields. Selecting the appropriate model depends on the characteristics of the data. Visualizing the data through scatter plots aids in assessing its distribution, and trying different models is recommended when the distribution is unclear [32].

This chapter explores various regression models, starting with Linear Regression and Multilinear Regression for multiple predictors. Logistic Regression is discussed for binary outcomes, and Poisson Regression is introduced for count data and event occurrences. The Generalized Linear Model (GLM) unifies these models, providing a flexible approach. Finally, the chapter concludes with insights into Model Selection, emphasizing the importance of choosing the most appropriate model for a given dataset. Together, these sections provide a comprehensive understanding of regression modeling, catering to various statistical scenarios.

2.1 Simple Linear Regression Model

A simple linear regression model is employed to establish a linear relationship between the two variables, say x and y. It is well-suited for scenarios involving continuous data and when the distribution approximates normality [31, 32]. The most common equation for this model is expressed as

$$y = \beta_0 + \beta_1 x + \varepsilon. \tag{2.1}$$

Here, x is also referred to as the covariate, predictor, explanatory, or independent variable, while y is termed the outcome, predicted, response, or dependent variable. β_0 is the intercept or constant, β_1 represents corresponding regression coefficients, and ε is known as error or disturbance. The model aims to identify the best-fit line by determining the regression coefficient that minimizes the total error ε. It is assumed commonly that ε follows a normal distribution with zero mean and some variance, which is constant

across the independent variables. The standard model for simple linear regression is given as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, 2, ..., n. \tag{2.2}$$

Here, $(X_i, Y_i), i = 1, ..., n$ are dependent and independent variables, respectively, and $\varepsilon_1, ..., \varepsilon_n$ represents error which is independent and identically distributed (IID) with,

$$E(\varepsilon) = 0, Var(\varepsilon) = \sigma^2$$

Here, the response variable (independent variable) follows a (conditional) normal distribution with the mean and variance expressed as

$$E(y_i) = \beta_0 + \beta_1 x_i, Var(y_i) = \sigma^2.$$

The y_i are (conditionally) independent given the covariates x_i. We can estimate the unknown parameters β_0 and β_1 using the least squares (LS) method [31–33]. These estimated values are determined as the minimizer of the sum of the squared deviations.

$$LS(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2. \tag{2.3}$$

2.2 Multi-linear Regression Model

A multi-linear regression model, also known as the multiple regression model, extends the principles of the simple linear regression model by accommodating multiple predictors or covariates. These predictors can include binary, continuous, or multi-categorical variables, which may be derived through transformations of the original covariates [29]. The foundational assumption related to the error term aligns with that of the simple linear regression model. For a continuous variable y and continuous or appropriately coded categorical regressors $x_1, ..., x_k$, denoted as $(y_i, x_{i1}, ..., x_{ik}), i = 1, ..., n$, the model is written as

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \varepsilon_i, i = 1, 2, ..., n. \tag{2.4}$$

Here, β_0 represents the intercept, and $\beta_1, ..., \beta_k$ are the unknown parameters or regression coefficients. The estimation of these unknown parameters can be achieved through a similar least squares method [31, 32].

2.3 Logistic Regression Model

A logistic regression model is well-suited for scenarios where the dependent variable is dichotomous or binary, making it distinct from the linear model. This distinction lies in the choice of the parametric model and underlying assumptions [29]. The logistic regression model is chosen for its flexibility, ease of use, and ability to offer clinically meaningful interpretations. In case of only one predictor variable, it is termed a simple logistic regression model, and when multiple predictors, including both categorical and continuous variables, are involved, it is referred to as a multiple logistic regression model [34, 35].

The logistic regression model's popularity stems from its foundation on the logistic function, characterized by an S-shaped curve that ranges between 0 and 1 [36]. Figure 2.1 by Kleinbaum et al. [36] provides a representation of an S-shaped curve. The logistic function is represented as:

$$f(x) = \frac{1}{1 + e^{-x}}$$

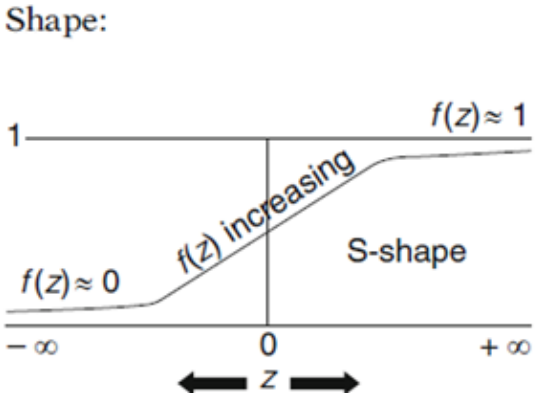

Figure (2.1) Logistic function representation

The logistic regression model consistently characterizes probabilities within the range of 0 to 1. This ensures that risk estimates fall between 0 and 1, a feature not always guaranteed in alternative models. As illustrated in figure 2.1, the logistic function initiates at $z = -\infty$ with a value of 1. It gradually decreases as z approaches ∞, ultimately converging to 0 [36].

Epidemiologists find the S-shape of the logistic function compelling when interpreting the risk associated with various values of the variable z, which serves as an index aggregating contributions from multiple risk factors [29, 36]. The sigmoid curve indicates that the impact of z on an individual's risk is negligible for low z values until a certain

threshold is reached. Beyond this threshold, the risk escalates rapidly for a range of intermediate z values, maintaining an extremely high level once z becomes sufficiently large [36].

The logistic model in terms of probability is denoted as $E(Y|X)$, indicating the occurrence of an event given the value of the predictor. This probability is represented as $P(X)$ and can be expressed as

$$P(x) = E[y|x] = \beta_0 + \beta_1 x. \tag{2.5}$$

Similar to a simple linear regression model, the logistic model utilizes the logistic function in the equation. The transformed simple logistic regression model is given as

$$P(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}, \tag{2.6}$$

where β_0 and β_1 are unknown coefficients. The model can also be expressed in terms of odds of the outcome with the predictor x as

$$\frac{P(x)}{1 - P(x)} = \exp(\beta_0 + \beta_1 x).$$

In the context of a binary predictor variable x taking values of 0 or 1, the odds ratio for these values can be expressed using the above equation

$$\frac{P(1)/[1 - P(1)]}{P(0)/[1 - P(0)]} = \exp(\beta_1).$$

The above model, with some transformation, becomes linear in the predictor, and it can be written as

$$\log\left[\frac{P(x)}{1 - P(x)}\right] = \beta_0 + \beta_1 x.$$

The corresponding multiple logistic regression model for multiple predictors is a direct generalization of the version for a single predictor. For a binary outcome Y and predictors $p(x_1, x_2, ..., x_p)$, the model can be expressed as

$$\log\left[\frac{P(x_1, x_2, ..., x_p)}{1 - P(x_1, x_2, ..., x_p)}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p, \tag{2.7}$$

which can be re-expressed in terms of outcome probabilities as

$$P(x_1, x_2, ..., x_p) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)}, \tag{2.8}$$

where x_p's are independent variables, and the terms $\beta_0, ..., \beta_p$ are unknown parameters that need to be estimated based on the obtained x's. The interpretation of the regression

coefficients is as follows

- For a given predictor x_j, the coefficient β_j gives the change in log odds of the outcome associated with a unit increase in x_j, for arbitrary fixed values for the predictors $x_1, ..., x_{j-1}; x_{j+1}, ..., x_p$
- The exponentiated regression coefficients $\exp(\beta_j)$ represents the odd ratio associated with a one unit change in x_j

The unknown parameters of the model are estimated by maximizing the likelihood ratio. In-depth discussions about these models exceed the scope of this overview. For a more comprehensive understanding, readers are recommended to explore works by Kleimbaum [36], [29], [34], [32], and other pertinent sources referenced in this thesis.

2.4 Poisson Regression Model

The Poisson distribution belongs to the exponential family of distributions which shares a similar form and properties with other members of this family [34]. An important characteristic of the Poisson distribution is that its mean and variance are expected to be equal. However, this assumption may not always align with real-world scenarios. To address this, a common practice is to consider a situation where the variance is proportional to the mean. The estimation of this proportionality factor is known as the scale parameter. In practical terms, when the scale parameter exceeds 1, indicating that the variance is larger than what is assumed by a specific distribution, the data is referred to as overdispersed [32, 36].

Count data analysis is a common scenario, especially when examining the frequency of specific events within a fixed time period at a constant average rate [32]. Linear models are not suitable for count data due to the inherent characteristics of count data—non-negativity and a mean always greater than zero. In such cases, the Poisson model proves to be invaluable for analyzing the rate of events when individuals are subject to different follow-up times. This stands in contrast to logistic regression, which focuses solely on the occurrence or non-occurrence of events and is primarily used for calculating odds ratios [31, 32].

The fundamental assumption is that for a single observation i, the outcome variable (Y_i) follows a Poisson distribution with mean Y_i, where Y_i is related to the predictor variables through the log link function [31]. The model can be expressed as

$$Y_i = \log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}, \tag{2.9}$$

where $x_{i1}, x_{i2}, ..., x_{ik}$ are the predictor variables, $(\beta_0, \beta_1, ..., \beta_k)$ are regression coefficients, and $\log(\mu_i)$ is the natural logarithm of the mean of the Poisson distribution

(μ_i).

The effect of the covariates on the rate Y_i is similar to the effect on the odds in the logistic model. The exponential of a particular coefficient is the estimated relative rate associated with the relevant variable. The maximum likelihood method is used to estimate the coefficients. The likelihood for a set of observations $Y_1, Y_2, ..., Y_n$ with corresponding means $\mu_1, \mu_2, ..., \mu_n$ is given by

$$L(\beta_0, \beta_1, ..., \beta_k) = \prod_{i=1}^{n} \frac{e^{-\mu_i} \mu_i^{Y_i}}{Y_i!}. \tag{2.10}$$

The log-likelihood function, which is often used for numerical optimization, is provided as

$$l(\beta_0, \beta_1, ..., \beta_k) = \sum_{i=1}^{n} (-\mu_i + Y_i \log(\mu_i) - \log(Y_i!)). \tag{2.11}$$

The maximum likelihood estimation (MLE) involves finding the values of $\beta_0, \beta_1, ..., \beta_k$ that maximize this log-likelihood function [31, 32].

2.5 Generalised Linear Models

Generalized Linear Models (GLM) serve as a natural extension of classical linear models, encompassing a wide range of regression approaches [36]. Their ability to adapt to various data types and distributions makes them invaluable in situations where conventional models may encounter limitations. The essence of GLMs lies in their capacity to align the analysis with the inherent characteristics of the data, offering a tailored and effective approach to regression modeling. For more in-depth understanding and practical applications, readers are encouraged to explore relevant literature on Generalized Linear Models [31, 32].

GLM follows similar assumptions as logistic and linear models, where the effect of covariates is modeled through a linear predictor [31]. We represent the mean response $E(Y)$ with μ, and a function of the mean is represented by $g(\mu)$. For p independent variables, the generalized linear model can be given as

$$g(\mu) = \beta_0 + \sum_{h=1}^{p} \beta_h X_h. \tag{2.12}$$

GLM is comprised of three key components: a random component, a systematic component, and the link function. It is required for the random component that the outcome variable Y follows a distribution from the exponential family. In the systematic component, the predictors are combined in the model as a linear function of the parameters. The link function establishes a linear relationship between the mean response and the

fixed linear set of parameters. It "links" the mean response function with the fixed linear combination of predictors and parameters. GLM offers large degrees of flexibility to choose each of the features of the model to fulfill the requirements [32,36]. From above equation, the link function g "links" $E(Y)$ with $\beta_0 + \sum b_h X_h$.

GLM can also be expressed in terms of the inverse of the link function, which is the mean μ.

$$\mu = g^{-1}(g(\mu)).$$

We notice that this inverse function is modeled in terms of the predictor (X) and their coefficients β, as $g^{-1}(X,\beta)$. The maximum likelihood method is used to estimate the model parameters. If the response variables (Y_i) are independent, then the likelihood is the product of each observation's contribution (L_i) to the likelihood, i.e.,

$$L = \prod_{i=1^K} L_i. \tag{2.13}$$

If the response variables are not independent, then the likelihood can be complicated or intractable. Such situations are handled differently to model the problems as mentioned in [31] and [36].

Having discussed several models, the critical question arises: Which model is the most suitable for our data, effectively capturing the underlying relationships? Model selection is a fundamental challenge in statistics, demanding careful consideration. In the upcoming section, we will delve into diverse model selection mechanics and criteria to address this crucial aspect.

2.6 Model Selection

Model selection plays a pivotal role in statistical modeling, involving the choice of the most suitable statistical model to effectively depict the relationship between predictor and response variables [37]. The significance of selecting the right model cannot be overstated, as it profoundly influences the accuracy and validity of statistical inferences drawn from the model. Opting for an inappropriate model may result in flawed predictions, imprecise estimations, and incorrect conclusions. Conversely, choosing the correct model enhances the likelihood of accurate predictions, precise estimations, and well-founded conclusions [38]. In essence, for a given dataset, model selection entails picking the most fitting model from a class of models. This process is extensively applied across various disciplines such as epidemiology, economics, biology, and beyond [39].

In statistics, several methods are employed for model selection, among which informa-

tion criteria and cross-validation stand out. These approaches are designed to pinpoint the model that strikes the optimal balance between complexity and goodness of fit to the data [37, 40].

2.6.1 Model selection based on information criteria

Model selection based on information criteria, such as the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), is a widely employed strategy. These criteria assign a score to each model based on the data, producing a ranked list of candidate models from the best to the worst [37]. AIC and BIC assess the balance between a model's goodness of fit and the number of parameters it employs. A lower AIC or BIC indicates a better model fit, with a preference for simpler models featuring fewer parameters [39]. Let us delve into AIC and BIC approaches briefly.

Akaike Information Criterion (AIC): The Akaike Information Criterion (AIC) is a widely used model selection method designed to choose the best statistical model from a set of candidate models. Developed by Hirotugu Akaike in 1973, AIC operates on the principle of striking a balance between a model's goodness of fit and the number of parameters it employs. For a more detailed theoretical and mathematical exploration, refer to [40]. The AIC algorithm selects the model $\mathscr{M}$ that minimizes the loss, typically measured logarithmically [37]. The AIC is defined as follows:

$$AIC = -2L + 2k. \tag{2.14}$$

Here, L represents the log-likelihood of the model $\mathscr{M}$, and k represents the number of parameters used in the model $\mathscr{M}$. The AIC penalizes models with more parameters, discouraging overfitting, and aims to select models that achieve a good fit with a relatively small number of parameters. Lower AIC values indicate a better trade-off between model complexity and fit to the data [39].

In the case of autoregressive order selection, the following is also commonly used for the k-order model,

$$AIC_k = n \log e_k + 2k, \tag{2.15}$$

where e_k is the average sample prediction error based on the quadratic loss. This can be derived assuming that autoregressive noises are Gaussian with autoregressive of different orders [39].

In practical applications, AIC values for each candidate model are computed, and the model with the lowest AIC value is chosen as the optimal model. It is common to compare the differences in AIC values between candidate models rather than the absolute values [39]. The model with the lowest AIC or the smallest AIC difference is deemed

the most favorable. The AIC method finds extensive utility in linear regression analysis and various parametric modeling approaches. Its application extends to nonlinear regression analysis, survival analysis, time series analysis, and other statistical domains [37].

The AIC method stands out for its computational simplicity, requiring no additional assumptions. Its versatility lies in the ability to compare models with varying numbers of parameters, facilitating the selection of a parsimonious model that guards against overfitting. A noteworthy aspect is its applicability to both nested and non-nested models [37, 39].

However, certain limitations accompany the AIC method. It assumes that the compared models are equally plausible, a condition that may not always hold. Furthermore, the AIC method lacks a provision for quantifying the uncertainty associated with model selection, a factor of significance in certain applications [37].

Bayesian Information Criterion (BIC): Similar to the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), also known as the Schwarz Criterion (SC), is a model selection strategy [41]. The BIC measures a model's goodness of fit and includes penalties for models with a large number of parameters [42]. The BIC's primary goal is to find the model, designated as $\mathscr{M}$, that minimizes the following expression:

$$BIC = -2L + k \log n. \tag{2.16}$$

The logarithm of the sample size (n) replaces the constant 2 in the penalty term, distinguishing it from AIC. Model selection relies on the BIC value, and the model with the smallest BIC value is considered the most suitable model [37].

The BIC method presents various advantages compared to other model selection techniques. It is straightforward to calculate, does not necessitate additional assumptions, and is applicable to both nested and non-nested models. The approach tends to prefer simpler models, aiding in the selection of a parsimonious model with reduced susceptibility to overfitting [37].

2.6.2 Model selection based on cross-validation

The cross-validation method is implemented as follows: Initially, the data is randomly divided into the training set and the validation set [39]. For a dataset with n observations, the training set comprises n_t data points, where $1 \leq n_t \leq n-1$, and the validation set includes $n_v = n - n_t$ data points. The model undergoes training on the training set and is subsequently validated on the remaining data to determine the average validation loss. This entire process is independently repeated multiple times, with different training and validation sets each time. The model's performance is assessed based on the smallest

average validation loss. The model with the smallest average validation loss is chosen and retrained on all data for subsequent predictions [37, 39]. A specific type of cross-validation known as k-fold cross-validation is commonly employed. In this method, data is randomly partitioned into k subsets of about similar size, where k is a positive integer. It is widely utilized because of its low computing complexity [39].

AIC is both asymptotically efficient and minimax optimal in the nonparametric framework. This implies that the predictive performance of the selected model asymptotically equals the best among the candidate models, although it is sensitive to the sample size [38, 43]. On the other hand, in the parametric framework, BIC is consistent and asymptotically efficient [41, 42]. Despite these favorable properties, both AIC and BIC have their limitations. AIC is not asymptotically efficient in a parametric setting and requires at least two correct candidate models. However, caution is advised in their use, and interpretation should be context-specific to the given problem [39]. For further exploration of model selection methods and criteria, the reader can refer to [43], [37], and [41].

3 Survival Analysis Fundamentals

This chapter of the study serves as an overview of survival analysis, a statistical methodology crucial for studying time-to-event data. The initial section covers the intricacies of the data generation process, exploring aspects such as censoring and truncation, while establishing essential terminology and notation. The subsequent section provides an understanding of basic survival and hazard functions, fundamental components for understanding event occurrences over time. Finally, the chapter concludes by exploring the intricate relationship between survival and hazard functions, shedding light on their interconnected roles in the analysis of time-dependent outcomes.

3.1 Survival Analysis

The problem of analyzing the expected time until one or more events occur exists in several fields, such as epidemiology, engineering, economics, public health, demography, medicine, and biology [44]. This process of time-to-event analysis or duration analysis is called survival analysis, and this time is called survival time or duration time, or transition time. In other words, survival analysis is a set of statistical methods for data analysis in which the outcome variable of interest is time until an event happens. Survival analysis is a statistical discipline predominantly observed in the fields of biology and medicine. [45, 46].

In this context, "time" can refer to days, weeks, months, or years from the initiation of a study until the occurrence of an event of interest. An event can encompass various outcomes such as death, organ failure, disease incidence, patient recovery, and more, depending on the specific research requirements [45]. To illustrate the concept of survival analysis, consider two examples. The first example is a study of individuals over several years to see who develops heart disease. Here, the event would be “developing heart disease,” and the outcome would be “the time in years until a person develops heart disease.” For the second example, take a malaria study on children for a period of time to identify a malaria-positive case. Here, the event is a “malaria-positive case,” and the outcome is measured as “the time until the positive case is seen” [45].

3.1.1 Data generation processes

In survival analysis, how data is generated or collected holds significant importance. Understanding the data generation process is crucial, as it carries substantial implications for subsequent analysis [46]. There are four primary approaches to sampling processes that yield survival time data :

Stock sample: In stock sampling, a random sample of individuals currently experiencing the state of interest is selected. Subsequently, these individuals, at a later time, are typically (though not always) interviewed to determine when they entered this state, denoted as the spell start date. For instance, consider modeling the duration of spells for unemployment benefits. In this case, a sample is drawn from individuals currently unemployed, and their reported unemployment start dates are recorded during the data collection process [46].

Inflow sample: In the case of an inflow sample, data is collected by randomly sampling individuals as they enter the state of interest. Subsequently, follow-up observations are conducted either until a predetermined date (which may be uniform for all individuals) or until the conclusion of the spell. For example, one might take a sample of every person initiating the receipt of unemployment insurance (UI) when studying the duration of UI spells [46].

Outflow sample: In an outflow sample, data is gathered through a random sampling of individuals who exit the state of interest. From the previous example of UI spells, the sample would comprise individuals who conclude their UI receipt [46].

Population sample: In a population sample, data is obtained through a general population survey, where the sampling process is unrelated to the specific process of interest. Respondents in this sample are typically asked about their current and/or past instances of the phenomenon under consideration, including information about the start and end dates of relevant spells [46].

Various combinations of sample types can be utilized to generate survival time data. For instance, a sample of spells can be created by considering all events that occurred between two specified dates, like June 1 and December 31 of a particular year. This approach accommodates spells that are already in progress when the observation window starts (similar to the stock sample case) and those that commence during the window (akin to the inflow sample case) [46].

The longitudinal data in these four sample types can be acquired from three distinct survey types or databases:

Administrative records: Administrative records, such as the government's benefits system database, can provide information about events like unemployment benefit spells. These records might serve as the sole source of information about individuals or be supplemented with a social survey that seeks additional details from the relevant population [46].

Cross-section sample survey, with retrospective questions: In a cross-section sample survey with retrospective questions, respondents are asked to recall and provide information about their spells during the survey. For instance, to determine the duration of

marriages, questions may cover the current marriage status (married/divorced), whether the individual has ever been married, and specific dates of marriage and divorce. Similar data collection approaches can be employed by employers for employment-related information [46].

Panel and cohort survey, with prospective data collection: Panel and cohort surveys, involve prospective data collection, where repeated interviews or observations are conducted at different times on the sample of interest to gather longitudinal information [46].

In data generation, a combination of survey methods can be employed. For instance, retrospective questions may be incorporated to gather information about experiences in a panel survey before the survey initiation. This approach recognizes that the nature of information about spells can vary, and such variations have significant implications for the appropriate data analysis [46].

3.1.2 Censoring and Truncation

Censoring and truncation are the two main aspects of incomplete survival time data. Censoring occurs when we have partial information about individual survival time, but the exact survival time is unknown [44]. To illustrate censoring, consider a study on cancer patients monitored until they exit the state of remission. If the study concludes that a patient is still in remission (without the event occurring), their survival time is considered censored. Censoring typically happens for three reasons: an individual does not experience the event before the study concludes, irregular follow-up, or withdrawal from the study [45, 47]. Let us focus on two primary types of censoring:

Right censoring: Right censoring is the most common type encountered and data of this nature is known as right-censored data. In right censoring, an individual is followed up until the end of the study, but the event of interest does not occur during this time. Thus, we know the duration (censored time) until the event did not happen, but the exact time remains unknown [48, 49]. For instance, in a cancer diagnosis study where the event of interest is death, a person diagnosed with cancer is observed until the end of the study and if the person does not die during the study, the event of interest did not occur. In this scenario, we have information about the person's survival time, but the precise survival time remains undisclosed [50].

The definition of right censoring is now presented mathematically, and its illustration is comprehended through an example. Given a lifetime X and a fixed censoring time C_r (C_r for "right" censoring time) for a specific individual. Assuming that $X's$ are independent and identically distributed (IID) with probability density function (PDF) $f(x)$ and survival function $S(x)$. The exact lifetime X of a person is known if and only if X is less than or equal to C_r. The individual is a survivor if X is greater than C_r, and event time is

censored at C_r. We can represent the data from an experiment by pairs of random variables (T, δ), where $\delta = 1$ if the lifetime corresponds to an event, and $\delta = 0$ if it is censored, and T is equal to X if the lifetime is observed, it is equal to C_r if it is censored, i.e., $T = \min(X, C_r)$ [44, 48].

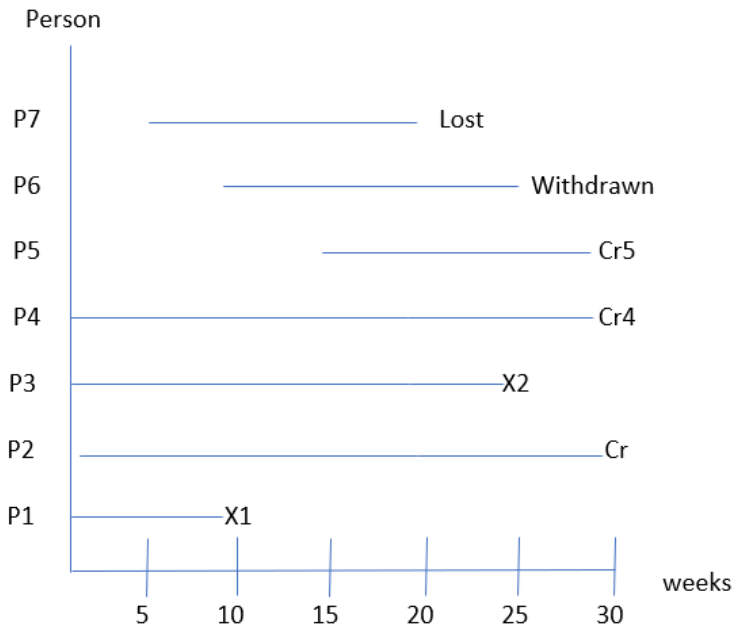

Figure (3.1) Right Censoring

The concept can be explored with graphical representation from Figure 3.1. Here, X denotes a person who experienced the event and C_r denotes censored time. Let us assume that five individuals participated in a study of 30 weeks. In the below figure, first-person (P_1) experienced the event at week 10 and P_3 at week 25, so survival time for (P_1) is 10 weeks and 25 weeks for (P_3), and survival time is not censored. On the other hand, (P_2) and (P_4) were part of the study till the end but they did not experience the event. Here, survival time is censored as can only assert that the survival time is at least 30 weeks. (P_5) joined the study at week 15 and did not experience the event till the end of the study and the censored time for this person is 15 weeks. (P_6) joined the study in week 10 but withdrew from the study at week 25, so the survival time is censored after 10 weeks. (P_7) enters the study in week five and is followed until week 20, when the person is lost to follow up so censored time is 15 weeks [44, 45].

As stated by David, this concept can be extended to situations where subjects enter the study at various time points and are observed until a predefined time at which the study ends. In such instances, a practical approach is to normalize the starting times for each subject to 0 [45]. This form of right censoring is commonly referred to as generalized Type I censoring, and figure 3.2 illustrates a concise example. The dataset used in this study is also characterized by right-censored data

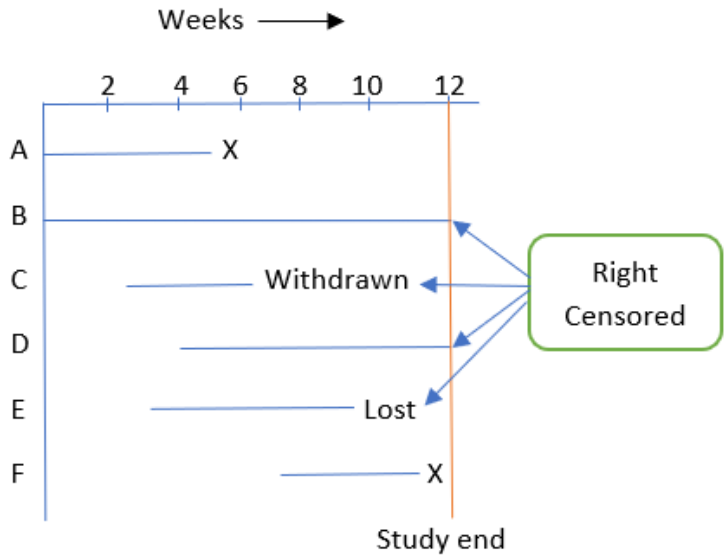

Figure (3.2) Type I censoring

Another form of censoring involves continuing the study until the failure of the initial r individuals (where r is predetermined and $r < n$, with n being the total number of individuals). This type of right censoring is referred to as Type II censoring and is commonly employed in assessing the lifespan of equipment. In this scenario, all items undergo testing simultaneously, and the test concludes when r out of the n items experience failure [48]. It is important to emphasize that both r, denoting the number of failures, and $n - r$, representing the number of censored observations, are fixed integers. This censoring is also termed progressive type II censoring and an example is shown in figure 3.1 [44]. Out of six people, (A and F) get the event, and (B, C, D, and E) are censored.

Person	Survival Time	Status (1=Failed, 0=Censored)
A	5.0	1
B	12.0	0
C	3.8	0
D	8.0	0
E	6.0	0
F	3.5	1

Table (3.1) Type II censoring

Left censoring: This type of censoring occurs when an individual's actual survival time is less than or equal to the observed survival time. In such instances, the event of interest has already occurred for an individual before the start of the study. Consequently, the precise duration of the event is not known [46, 49]. To illustrate, consider a study aiming to estimate the age at menarche (the age at first menstruation). An individual within a certain age range is enrolled in the study for this purpose. If an individual in the

study has already experienced her first menstruation, we know that the age at menarche is less than her current age (age at enrollment), but the exact age is not known [51].

Similarly, as above, let X denote the lifetime (spell time) of a specific individual and C_l denote censoring time (left censoring time). The precise lifetime is known if and only if X is greater than or equal to C_l. In the context of the left censoring scheme, the data can be represented by a pair of random variables (T, E), where E denotes the exact lifetime observed. If the lifetime is observed, T equals X, and E is set to one; otherwise, if the exact lifetime X is not observed, E is set to zero [44].

Interval censoring: Interval censoring arises when individuals are part of periodic follow-up studies, and the event time falls within an interval. For instance, consider a malaria study on children where observations are made periodically [44]. The intervals of time when a child tests positive for malaria will be noted. In this scenario, one knows the interval during which a child tested positive for malaria. Interval censoring is a special case that includes both right-censoring and left-censoring [44]. To illustrate, if the first malaria test is negative at time t_1 and the second is positive at time t_2, left censoring occurs when t_1 is 0, and t_2 is the known upper bound on the true survival time. Conversely, right censoring occurs when t_2 is infinity, and t_1 is the known lower bound on the true survival time. In figure 3.3, the first person $(P1)$ indicates interval censoring, and the second person $(P2)$ indicates left censoring.

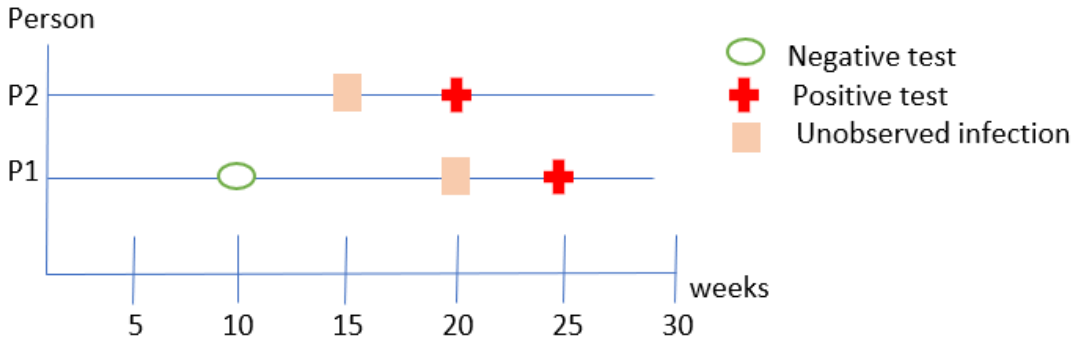

Figure (3.3) Left and interval Censoring

Truncation is also a significant aspect of survival time data. In the case of truncation, there is a systematic exclusion of survival times from the sample, and the sample selection effect is dependent on survival time itself [46]. Survival data is truncated when only individuals whose event time falls within a specific observational window are observed. An individual whose event time falls outside of this range is not observed, and the investigator has no information on this subject. In simple terms, truncation occurs when an observation is not recorded due to that observation being below or above a certain threshold, and the data is incomplete [44]. There are two types of truncation:

Left truncation: Left truncation occurs when only those who have survived for more

than a certain time are included in the observation sample, meaning those with shorter survival times are not observed. Delayed entry and stock sampling with follow-up is also known as left truncation [44, 46]. To illustrate, consider a survival study of retired people at a retirement center. The age at death and the age at which an individual entered the retirement center (the truncation event) are recorded. An individual must survive to a certain age to enter the retirement center, and those who died earlier are not part of the data because they had no chance to enter the study. These individuals are considered left truncated [44].

Right truncation: Right truncation occurs when only those who have experienced the exit event by a specific date are included in the sample, leading to the systematic exclusion of relatively long survival times [44, 46]. To illustrate, consider a study of sampling patients with transfusion-induced AIDS. The time between infection at the moment of transfusion and the onset of clinical AIDS was calculated retrospectively using transfusion times. The database only includes those who had AIDS by a certain date, based on their waiting period from transfusion to that date. Patients who received blood transfusions before that particular date but developed AIDS after it are not tracked and are right-truncated [52].

3.1.3 Terminology and Notation

After gaining a foundational understanding of survival analysis, let us introduce basic mathematical terminology and notation. Firstly, we represent an individual's survival time with the capital letter T, a random variable. As T signifies time, it encompasses all possible non-negative numbers, indicating any duration equal to or greater than zero. Subsequently, small t denotes a specific time value of interest for T. For instance, if we are assessing whether a person survives more than five years after cancer therapy, t would be five, and we seek to determine if T surpasses this threshold. Lastly, the letter $d(0,1)$ defines a random variable representing censorship status. Therefore, $d = 1$ indicates that the event occurred during the study period, implying the survival time is uncensored. Conversely, $d = 0$ denotes that the event did not occur during the study period, indicating censored survival time. Additionally, $d = 0$ is assigned if an individual fails to follow up or withdraws from the study within the study period [53]. Consequently, we have

$$\begin{aligned} T &= \text{survival time } (T \geq 0), \\ t &= \text{specific realisation of} T, \end{aligned}$$

and D is a random variable indicating whether the survival time is censored, i.e.,

$$D = \begin{cases} 1 & \text{if not censored (i.e., the event occurred during the study)}, \\ 0 & \text{if censored (i.e., the event did not occur, lost to follow-up or withdraws)}. \end{cases}$$

Realizations of D will be denoted by d.

In the following section, two crucial quantitative terms in survival analysis are introduced: the survivor function and the hazard function.

3.2 Basic Survival and Hazard functions

These concepts are used widely in any survival analysis scenario. The survivor function is denoted by $S(t)$ and the hazard function is denoted by $H(t)$. $S(t)$ stands for the probability that an individual survives longer than some specific time t, indicating that it is the probability of the random variable T exceeding the specific time t [53]. For a continuous random variable T with a cumulative distribution function (CDF) $F(t)$, and probability density function (PDF) $f(t)$, $F(t)$ is also referred to as the failure function in survival analysis [46]. The failure function can be written as:

$$[T \leq t] = F(t), \tag{3.1}$$

and $S(t)$ ca be written as:

$$[T > t] = 1 - F(t) = S(t). \tag{3.2}$$

The $f(t)$ denotes the frequency of events per unit time and is given by:

$$\begin{aligned} f(t) &= \lim_{\delta_t \to \infty} \frac{[t \leq T \leq t + \delta_t]}{\delta_t} \\ &= \frac{\partial F(t)}{\partial t} \\ &= -\frac{\partial S(t)}{\partial t} \\ &= -S'(t) \end{aligned} \tag{3.3}$$

where δ_t is an infinitesimally small interval of time. The unconditional probability of having a spell of length exactly t, i.e., leaving the state in a tiny interval of time $[t, t + \delta_t]$, is represented by $f(t)\delta_t$ [46]. Theoretically, the survival curve can be drawn as in figure 3.4.

Here, t ranges from 0 to infinity, and $S(t)$ is plotted as a smooth curve. All survivor functions have the following theoretical properties:

- They are non-increasing
- At time $t = 0$, $S(t) = 1$. In other words, the probability of surviving past time t is 1

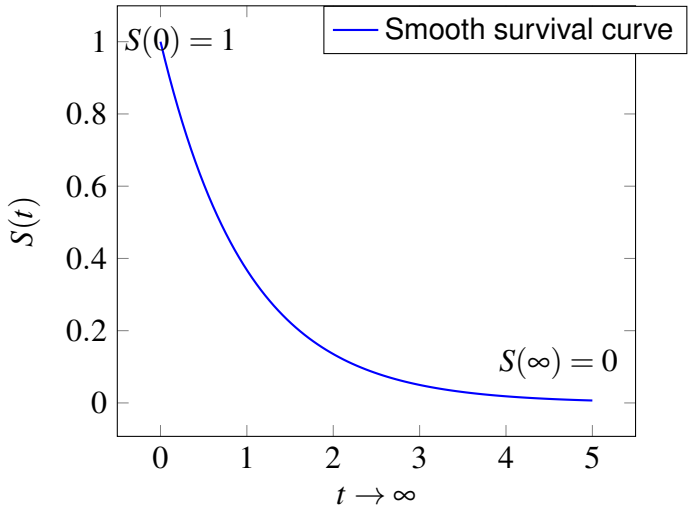

Figure (3.4) Theoretical survival curve

- At time $t = \infty$, $S(t) = 0$. As the time tends to infinity, eventually nobody will survive so the $S(t)$ will fall to 0

The graph represents a survivor function that typically starts at 1 and goes down to 0. In the real-world data, the graphs obtained are usually step functions, as shown in figure 3.5.

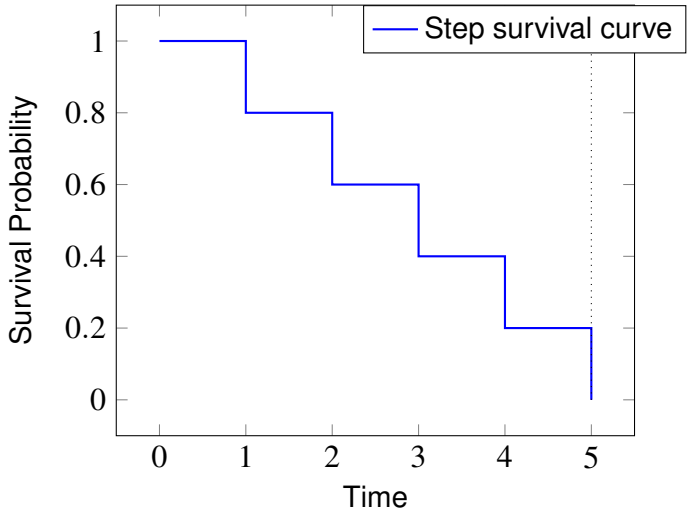

Figure (3.5) Practical survival curve

The hazard function $h(t)$ is the instantaneous potential time for the event to occur, given no previous events. It is equal to the limit of δ_t (tends to 0), of a survival probability statement divided by δ_t, where δ_t is a small interval of time. It can be written as:

$$h(t) = \lim_{\delta_t \to \infty} \frac{[t \leq T \leq t + \delta_t \big| T \geq t]}{\delta_t}. \tag{3.4}$$

Integrating $h(t)$ over t gives us the cumulative hazard function as:

$$H(t) = \int h(t)dt. \tag{3.5}$$

In this context, the conditional probability determines the likelihood of an individual surviving in the time interval between t and $t + \delta_t$, given that the survival time is greater than or equal to t. It is crucial to note that the hazard function value serves as a risk indicator rather than a probability of the event occurring. The risk of an event increases with the hazard function's value. Consequently, the result of dividing the numerator (conditional probability) by the denominator (small time interval) is not a probability but a rate. The scale for this ratio does not range from 0 to 1, as it would for a probability, but rather from 0 to infinity, depending on the unit of time measurement (e.g., days, weeks, months, or years) [44, 53].

The hazard function $h(t)$, similar to the survivor function, can be graphed as t varies in value [44]. The graph of $h(t)$ can start anywhere and move up or down in any direction over time, unlike the survivor function. The hazard function has the following properties for a particular value of t:

- It is always non-negative, i.e., is always greater than or equal to zero
- It does not have an upper bound

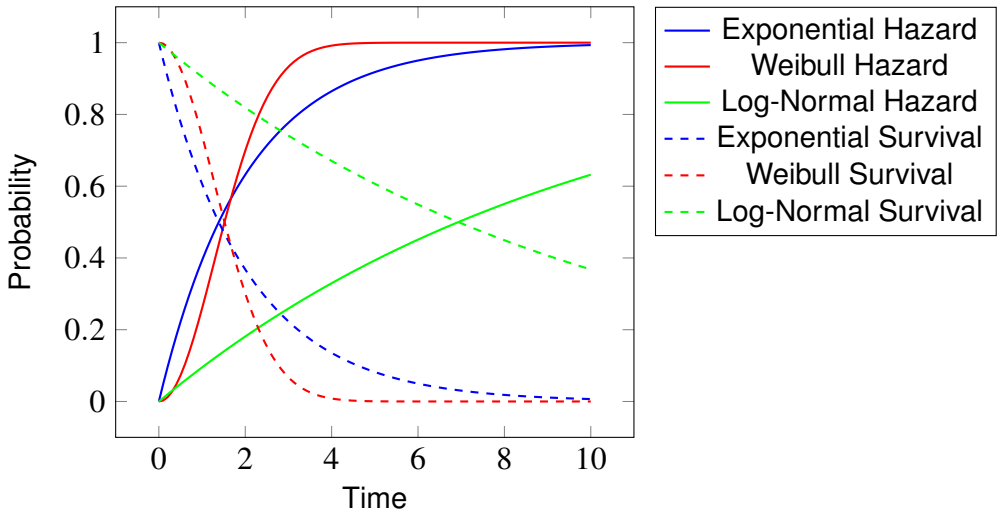

Figure (3.6) Hazard and Survival functions

Figure 3.6 depicts three distinct types of survival and hazard functions. In contrast to the survivor function, the hazard function focuses more on the likelihood of failure than on success or the occurrence of the event. Due to its conditional nature, the hazard function is occasionally referred to as the conditional failure rate. In a way, the hazard function provides complementary information to the survivor function [33, 53].

The survival function is more intuitive for interpreting survival data as it directly characterizes the survival experience in a study. However, the hazard function also holds significance for the following reasons [44, 53]:

- It measures an instantaneous potential
- It can be used to identify a specific model that fits the data, such as exponential, Weibull, or lognormal
- It is the way of performing mathematical modeling of survival data. That means the hazard function interprets the survival model

In the following section, we will look at the mathematical aspects of the relationship between survivor and hazard functions.

3.3 Relation between survivor and hazard function

The relation between $S(t)$ and $h(t)$ can always be defined, regardless of the function (distribution) used. They are related in such a way if one is known the other one can be derived [53]. Let us look at the mathematical perspective of the relation between $S(t)$ and $h(t)$ [46, 53]. Taking the following information from the previous section.

$$\begin{aligned} S(t) &= [T > t], \\ F(t) &= [T \leq t], \\ S(t) &= 1 - F(t). \end{aligned}$$

Furthermore, the hazard function is given as

$$h(t) = \lim_{\delta_t \to \infty} \frac{[t \leq T \leq t + \delta_t \big| T > t]}{\delta_t}.$$

Let us consider the numerator from the hazard function without the limit

$$[t \leq T \leq t + \delta_t \big| T > t].$$

This is a form of conditional probability

$$[A \big| B] = \frac{[A \cap B]}{\Pr[B]}, \textit{if } [B] \textit{ is not equal to } 0.$$

We can use the above transformation and $h(t)$ can be written as

$$h(t) = \lim_{\delta_t \to \infty} \left(\frac{[t \leq T \leq t + \delta_t] \cap [T > t]}{[T > t].\delta_t} \right).$$

The above formula is derived using two basic facts, the probability of $[t < T \leq \delta_t]$ and $[T > t]$ is $[t < T \leq t + \delta_t]$ and we know $[T > t]$ is $S(t)$. Now re-arranging the above equation gives the following:

$$\begin{aligned} h(t) &= \lim_{\delta_t \to \infty} \left(\frac{[T \leq t + \delta_t] - [T \leq t]}{S(t)\delta} \right), \\ &= \lim_{\delta_t \to \infty} \left(\frac{F(t + \delta_t) - F(t)}{\delta} \right) . \frac{1}{S(t)}, \\ &= \frac{\partial F(t)}{\partial t} . \frac{1}{S(t)}, \end{aligned}$$

which can be written in the following form

$$h(t) = \frac{f(t)}{S(t)}. \tag{3.6}$$

Equation (3.6) represents the relationship between $h(t)$ and $S(t)$. It can further be deduced in terms of $H(t)$. Using the chain rule over the log function, it can be worked as

$$h(t) = \frac{d}{dt}(1 - S(t)) . \frac{1}{S(t)},$$

$$= -\frac{d}{dt} S(t) . \frac{1}{S(t)},$$

$$h(t) = -\frac{d}{dt} \ln S(t). \tag{3.7}$$

We can interpret the hazard function from equation (3.7). It is the negative natural logarithm of survival rate differentiated over time t. This can further be represented in the form of $H(t)$ as

$$H(t) = -\ln S(t). \tag{3.8}$$

The equation (3.8) above signifies the relationship between the cumulative hazard function and the survival function. It can be understood as the cumulative hazard function at time t being the negative logarithm of the survival function at the same time point [46]. The following section provides the introduction of commonly used distributions in survival analysis along with their corresponding survival and hazard functions.

3.4 Common Survival and Hazard functions

In this section, a quick recap on the survival and hazard functions derived from the most common distributions in survival analysis. Additionally, our discussion will encompass Cox proportional hazard models, including aspects like estimation, interpretation, and relevant assumptions. Furthermore, we will explore an extension of the Cox proportional hazard model known as the Andersen-Gill model.

Exponential distribution: The exponential distribution is employed to model the waiting time until a specific event occurs, commonly applied in the analysis of survival data [54]. If a random variable T follows an exponential distribution with parameters (λ), then the probability density function (PDF) of the exponential distribution is expressed as:

$$f(t) = \lambda \exp(-\lambda t), \lambda > 0, t > 0,$$

and the mean (or expected lifetime) and its variance are given as:

$$E(t) = \frac{1}{\lambda},$$

$$Var[T] = \frac{1}{\lambda^2}.$$

For the exponential model, the survival function ($S(t)$) and the hazard function ($h(t)$) are given as follows

$$S(t) = \exp(-\lambda t), \tag{3.9}$$

$$h(t) = \lambda, \tag{3.10}$$

where $\lambda > 0$ is called the rate parameter and $t \geq 0$.

Weibull distribution: Weibull distribution was first proposed by Rosin and Rammler in 1933 and later it was proposed by Werody Weibull for the life span of materials [54]. It can be found in many studies where analysis of lifetime data is involved [55]. A random variable T with parameters λ and p. The PDF is given by

$$f(t) = \lambda p(\lambda p)^{p-1} \exp(-(\lambda t)^p). \tag{3.11}$$

The corresponding mean (or expected lifetime) and its variance are expressed as

$$E[T] = \frac{1}{\lambda}\Gamma\left(1 + \frac{1}{p}\right), \tag{3.12}$$

$$Var[T] = \frac{1}{\lambda^2}\Gamma\left(1+\frac{2}{p}\right) - \frac{1}{\lambda^2}\Gamma\left(1+\frac{2}{p}\right)^2. \tag{3.13}$$

Here, Γ is the Gamma function defined by

$$\Gamma(x) = \int_0^\infty t^{x-1}\exp(-t)dt. \tag{3.14}$$

For the Weibull distribution, the survival and hazard functions are given as

$$S(t) = \exp(-(\lambda t)^p), \tag{3.15}$$

$$h(t) = \lambda p(\lambda p)^{p-1}. \tag{3.16}$$

Here, $t \geq 0$, $\lambda > 0$ is a rate parameter, and $p > 0$ is the shape that allows control of the behavior of the hazard function. The following observation can be drawn from the hazard function as

- $h(t)$ decreases when $p < 1$,
- $h(t)$ increases when $p > 1$,
- $h(t)$ is constant when $p = 1$ and reduces to exponential distribution with constant hazard function

Using equations (3.15) and (3.16), the following relation can be deduced as

$$f(t) = h(t)S(t). \tag{3.17}$$

Log-logistic distribution: A random variable T follows the log-logistic distribution if $\ln(T)$ follows a logistic distribution with parameters p and λ, the PDF of log-logistic distribution is given as

$$f(t) = \lambda p(\lambda t)^{p-1}(1+(\lambda t)^p)^{-2}. \tag{3.18}$$

The mean (or expected lifetime) and its variance can be written as

$$E[T] = \exp\left(\mu + \frac{\sigma^2}{2}\right), \tag{3.19}$$

$$Var[T] = \left(\exp(\sigma^2) - 1\right)\exp(2\mu + \sigma^2). \tag{3.20}$$

For this distribution, the survival and hazard functions are provided as:

$$S(t) = \frac{1}{1+(\lambda t)^p}, \tag{3.21}$$

$$h(t) = \frac{\lambda p(\lambda t)^{p-1}}{1+(\lambda t)^p}. \tag{3.22}$$

It depends on the rate parameter $\lambda > 0$, shape parameter $p > 0$, and $t > 0$ [54]. As it depends on p, we can observe the following behaviors of the hazard function:

- $h(t)$ decreases from infinity when $p < 1$
- $h(t)$ decreases from lambda when $p = 1$
- $h(t)$ first increases and then decreases when $p > 1$. So, $h(t = 0) = 0$ and it is maximum at $t = (p-1)^{1/p}$

Log-normal distribution: This distribution is commonly used and it has a good ability to fit the data for reliability analysis. It relates to the normal distribution and can fit the data well with a particular relationship with the normal distribution. Let $N(\mu, \sigma^2)$ denote a normal distribution with mean (μ) and variance (σ) square [54,55]. A random variable T follows the log-normal distribution if $\ln(T)$ follows a normal distribution. The PDF of the log-normal distribution is given as

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{\ln t - \mu}{\sigma}\right)^2\right\} = \frac{1}{t\sigma}\phi\left(\frac{\ln t - \mu}{\sigma}\right), t > 0,$$

where $\phi(z) = (1/\sqrt{2\pi})\exp -z^2/2$ is the PDF of $N(0,1)$. The mean and variance of the log-normal distribution are presented as

$$E(T) = \exp\left(\mu + \frac{\sigma^2}{2}\right),$$

$$Var(T) = \{\exp(\sigma^2) - 1\}\exp(2\mu + \sigma^2).$$

The corresponding survival and hazard functions for the log-normal distribution are given as

$$S(t) = 1 - \phi\left(\frac{\ln t - \mu}{\sigma}\right),$$

$$h(t) = \frac{1}{t\sigma}\frac{\phi\left(\frac{\ln t-\mu}{\sigma}\right)}{\left\{1-\phi\left(\frac{\ln t-\mu}{\sigma}\right)\right\}}.$$

Here $t > 0$ and the hazard function have similar behavior to the Log-Logistic model for $t > 1$. It has two parameters as normal distribution has two parameters, μ belongs to the Real Number and the standard deviation $\sigma > 0$. These parameters can be obtained from the following transformation, $\mu = -\ln(\lambda)$ and $\sigma = t^{-1}$.

Gamma distribution: It is one of the popular distributions in statistics. It is sometimes used with and without log-normal and Weibull distributions [55]. If a random variable T follows a gamma distribution, then its PDF is defined as

$$f(t) = \frac{\lambda^{\beta}}{\Gamma(\beta)} x^{\beta-1} \exp(-\lambda t), t > 0.$$

Where $\lambda > 0$ is the scale parameter and $\beta \geq 0$ is the shape parameter. Gamma distribution changes to exponential distribution at $\beta = 1$.

The mean and variance for the gamma distribution are provided as

$$E(T) = \frac{\beta}{\lambda},$$

$$Var(T) = \frac{\beta}{\lambda^2}.$$

The survival function is represented as

$$S(t) = \frac{1}{\Gamma(\beta)} \int_t^{\infty} \lambda (\lambda t)^{\beta-1} \exp(-\lambda t) dt.$$

The corresponding hazard function is increasing for $\beta > 1$ and is provided as

$$h(t) = \lim_{t \to +\infty} = 0, \lim_{t \to \infty} = \lambda.$$

It is decreasing when $\beta < 1$ and is written as

$$h(t) = \lim_{t \to +\infty} = \infty, \lim_{t \to \infty} = \lambda.$$

For $\beta = 1$, the hazard function is constant.

In the forthcoming chapter, an overview of the Cox Proportional Hazard model and the Frailty model will be presented.

4 Cox Proportional Hazard Model and Frailty Model

This chapter provides an overview of advanced survival analysis models, with a primary focus on the Cox Proportional Hazard (PH) Model. The subsections within this section provide insights into the estimation of Cox PH model parameters using Maximum Likelihood (ML) and offer guidance on interpreting hazard ratios. A critical aspect of survival analysis is the proportional hazard assumption, and the chapter explores two approaches for its assessment: a graphical approach and the Goodness of Fit (GOF) approach. Moving beyond the Cox PH model, the Andersen-Gill (AG) Model is introduced as an extension suitable for recurrent events in time-to-event data.

The chapter also ventures into Frailty Models, including discussions on univariate frailty models, the relationship between marginal and conditional hazards, shared frailty models, and gamma frailty models. These models play a pivotal role in addressing unobserved heterogeneity and individual variability in survival analysis, providing a robust framework for nuanced modeling. Overall, this chapter equips readers with a comprehensive understanding of advanced survival analysis techniques and their practical applications.

4.1 Cox Proportional Hazard Model

The Cox-proportional hazard model, introduced by Cox in 1972, stands out as a widely embraced semi-parametric model, particularly in clinical and biomedical studies for the analysis of survival time data [56]. Unlike fully parametric models, it does not assume a specific distribution for the outcome or dictate the baseline hazard function, making it a versatile tool [46]. This model finds extensive use in clinical trials, enabling the exploration of survival differences attributed to treatment and prognostic factors [57]. Its semi-parametric nature allows for the analysis of censored survival data without making assumptions about the baseline hazard function [46]. In essence, the Cox proportional hazard model serves as a regression-based approach to unravel the relationship between events and a set of covariates [58]. The hazard, representing the instantaneous event probability at a given time, provides insights into the likelihood of an individual experiencing the event around that specific point in time [57, 58].

A mathematical representation of the con-proportional hazard model is given as

$$h(t,X) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p). \tag{4.1}$$

Here, $h(t,X)$ is the hazard at time t, $X = (x_1, x_2, ..., x_p)$ are a set of p covariates (pre-

dictors) and $(\beta_1, \beta_2, ..., \beta_p)$ are model parameters describing the effect of predictors on the overall hazard and $h_0(t)$ is the baseline hazard function. The values of the baseline hazard function act as the hazard function when all covariates are zero. The fundamental assumption is that covariates exert a multiplicative effect on the hazard function, and this effect remains constant over time. This implies that the ratio of the hazard function to the baseline hazard remains constant over time [49, 58]. The interpretation of the Cox model is facilitated through the hazard ratio, where a hazard ratio greater than 1 suggests a higher likelihood of the event occurring, while a hazard ratio less than 1 indicates a lower likelihood of the event occurring [58]. Viewing the Cox model's predicted quantity as a relative risk, rather than an absolute risk, underscores the importance of exponentiating the linear predictors to ensure non-negativity [49]. The insight is further enhanced by taking the logarithm of both sides of the model equation.

$$\log(h(t,X)) = \log(h_0(t)) + (\beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p). \tag{4.2}$$

It is evident from the above equation that the logarithm of the baseline hazard function serves as a time-varying intercept. This formulation aligns with a log-linear model, indicating that the hazard exhibits a linear change concerning continuous predictors [32]. The Cox proportional hazards model allows for obtaining hazard ratio estimates that accommodate various model factors, such as age, sex, and race, owing to the model's regression framework [58].

At times, the interest lies in obtaining numerical estimates for the regression coefficients, denoted as β_i, as opposed to focusing on the hazard's shape [54]. To provide a concise summary of the overall outcome, let us reframe the aforementioned equation as

$$\log\left(\frac{h(t,X)}{h_0(t)}\right) = \sum_i \beta_i X_i. \tag{4.3}$$

The above formulation is linear in X_i and β_i, highlighting the link to the regression model. In a simple summary

$$\log(HR_0) = \log\left(\frac{\textit{group hazard}}{\textit{baseline hazard}}\right) = \sum_i \beta_i X_i. \tag{4.4}$$

Here, the group hazard encompasses all the effects of the covariates X_i, excluding the baseline hazard from these effects.

For example, take one covariate ($p = 1$) as gender, and take values $X_1 = 1$ for female and $X_2 = 0$ for male. Then, from the above equation, one can derive,

$$\log\left(\frac{\textit{hazard female}}{\textit{baseline hazard}}\right) = \beta_1, \tag{4.5}$$

$$\log\left(\frac{\textit{hazard male}}{\textit{baseline hazard}}\right) = 0. \tag{4.6}$$

Taking the difference between the above equations,

$$\log\left(\frac{\text{hazard female}}{\textit{baseline hazard}}\right) - \log\left(\frac{\textit{hazard male}}{\text{baseline hazard}}\right) = \log\left(\frac{\textit{hazard female}}{\textit{hazard male}}\right) = \beta_1. \tag{4.7}$$

The log hazard ratio of the hazard for females and males is β_1, which is a direct interpretation of the regression coefficient β_1. Transforming the above equation, the following is given as the hazard ratio

$$\frac{\text{hazard female}}{\text{hazard male}} = \exp(\beta_1). \tag{4.8}$$

The hazard ratio provides a means of interpreting the effects of non-binary covariates in a similar fashion. An advantageous aspect of this model lies in its capability to estimate parameters without the need for baseline hazard function estimation, eliminating the necessity for parametric assumptions [54]. The Cox model's robustness makes it appealing for various reasons. In scenarios involving survival time with censoring, the Cox model is preferred over the logistic model. The estimation of β's (a part of the exponential) is possible even without specifying the baseline hazard part [53]. Subsequently, let us look into the method of estimating the Cox model using Maximum Likelihood (ML).

4.1.1 Estimation of Cox PH Model using Maximum Likelihood (ML)

The estimation of Cox model parameters involves maximizing the likelihood function, denoted as L, to find the optimal value for the parameter β [49]. In the Cox model, the absence of an explicitly specified baseline hazard function results in an undefined full likelihood for the model. Therefore, Cox introduced a partial likelihood, denoted as $PL(\beta)$, which focuses on the partial probabilities for individuals who experience an event, excluding the explicit consideration of probabilities for censored individuals [49, 53, 54].

The partial likelihood is expressed as the product of individual likelihoods for each failure time. These probabilities signify the likelihood of an individual experiencing an event at a specific time, considering that the event occurred to this individual out of all individuals at risk at that time [49]. It is assumed that at each event time, only one event occurs, indicating the absence of ties between event times [49, 54]. The formula for the partial likelihood is given as:

$$PL(\beta) = \prod_{t_i:\textit{event at } t_i} \frac{h_0(t)e^{\beta x}(t_i)}{\sum_{j\in R_{(t_i)}} h_0(t)e^{\beta x_j}}. \tag{4.9}$$

$R(t_j)$ denotes the risk set at time t_i, referring to the individuals who are still part of the sample just before time t_i. Here, x_i represents the value of x for each individual who experienced an event at time t_i. The product is taken over ordered event times. Baseline hazards are eliminated from the equation, resulting in the simplified partial likelihood:

$$PL(\beta) = \prod_{t_i:\text{event at } t_i} \frac{e^{\beta x}(t_i)}{\sum_{j:t_j \geq t_i} e^{\beta x_j}}. \tag{4.10}$$

The coefficients are obtained by taking the partial derivative of the maximum likelihood function with respect to each coefficient [59]. An attractive feature of the Cox model is that, despite leaving the baseline hazard unspecified, it allows for the estimation of the coefficients in the exponential part of the model [44].

As noted earlier, the assumption of no ties between event times may not always hold and ties could occur in practice [54]. To address this issue, the two most widely used extensions are exact methods, the Breslow approximation [60] and Efron approximation [61]. For further details on these methods, readers are encouraged to consult the respective literature.

4.1.2 Interpreting Hazard Ratios

Let us consider the hazard ratio between two groups instead of two individuals. These groups are the treatment group and control group, which will have a more intuitive meaning in the medical context [54]. Using the above equations, their corresponding hazards can be denoted as

$$h(t, X^{\textit{treatment}}) = h_0(t) \left(\sum_i^p \beta_i X_i^{\textit{treatment}} \right),$$

$$h(t, X^{\text{control}}) = h_0(t) \left(\sum_i^p \beta_i X_i^{\textit{control}} \right).$$

Assuming that the PH holds regardless of the complexity of the individual hazards, the hazard ratio is constant over time. One gets,

$$HR(\textit{T vs. C}) = \frac{h(t, X^{\text{treatment}})}{h(t, X^{\text{control}})}.$$

Here, HR denotes the Hazard Ratio between the treatment and control groups at any given time t and it is always a positive real-valued number [54].

Comparing survival curves between treatment and control groups, as seen in a log-rank test, survival curves offer a binary distinction. In contrast, HR provides insights into

both, the magnitude and direction of this difference [54]. The interpretation of HR can be summarized as:

- $HR(T \textit{ vs. } C) > 1$: Treatment group has higher hazard than control group,
- $HR(T \textit{ vs. } C) = 1$: There is no difference between both groups,
- $HR(T \textit{ vs. } C) < 1$: The control group has higher hazard than treatment group.

For better understanding, $HR(T \textit{ vs. } C) = 1.4$ means that the hazard for the treatment group is increased by 40% in comparison to the control group, and $HR(T \textit{ vs. } C) = 0.6$ means that the hazard for the treatment group is decreased by 40% in comparison to the control group [54].

4.2 Proportional Hazard Assumption Check

As previously stated, in the Cox proportional hazard model, the hazard ratio remains constant over time for any two specifications of predictors. Simply put, the hazard for one individual is proportional to that of another, where the proportionality constant does not depend on time. However, this assumption may be violated if hazard graphs are not approximately parallel or cross for two or more categories of a predictor [54]. It is important to note that the assumption might not be met even if hazard functions do not cross. Therefore, it is advisable to employ approaches such as graphical analysis, goodness-of-fit tests, and consideration of time-dependent variables to assess and validate this assumption [53].

4.2.1 Graphical Approach

David and Mitchel described two graphical approaches that will be discussed for assessing the PH assumption. The first method the most common involves comparing estimated $-\ln(-\ln)$ survivor curves for different combinations of categories of variables of interest. For example, figure 4.1 compares males with females demonstrating that the PH assumption is satisfied for the variable "Sex". The second approach is to compare observed survivor curves with predicted ones [53].

In figure 4.2, the variable "Sex", the observed curve is derived without incorporating it into the PH model, while the predicted curve is obtained by including this variable in the PH model. If the observed and predicted curves closely align, one can conclude that the PH assumption is met. Figure 4.2 indicates that the PH assumption is reasonably satisfied [53]. Let us provide a brief overview of these two approaches.

In the $\log - \log$ survival curve method, a double natural log transformation of the estimated survival curve is performed, expressed as $-ln(-\ln)$. Given that the logarithm of

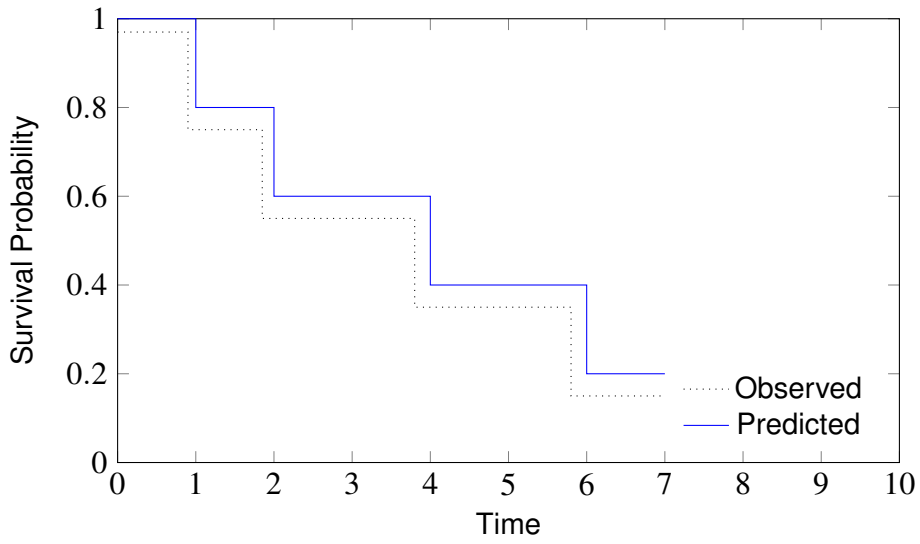

Figure (4.1) $-\ln(-\ln)$ comparison approach for the sex variable

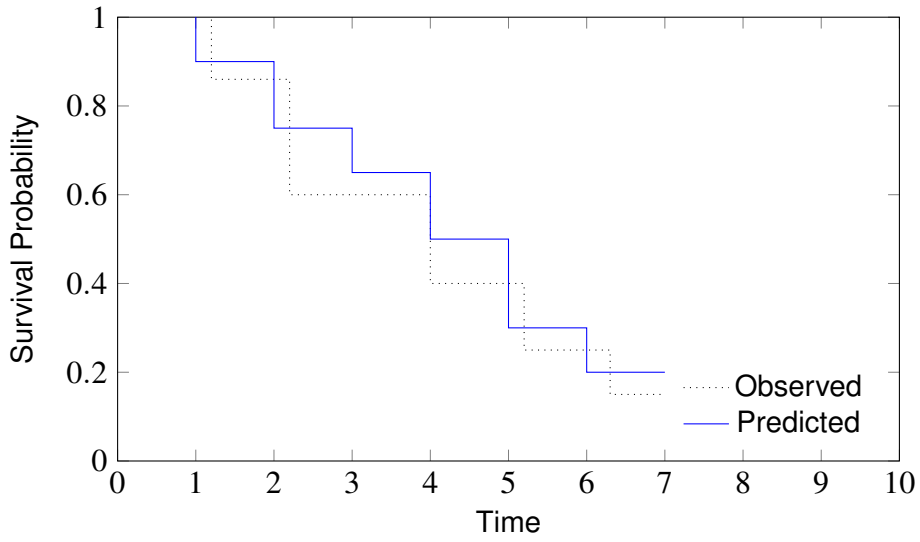

Figure (4.2) Observed vs. predicted approach for the sex variable

a probability is inherently a negative number and the logarithm of a negative number is undefined, the negative operation is applied to the first logarithm before the second log transformation [53]. This process is essential for handling the logarithmic transformation appropriately. The mathematical representation of the log transformation is elucidated by starting with the formula for the survival curve corresponding to the hazard function in the Cox PH model.

$$h(t,X) = h_0(t)\exp(\sum_{i=1}^{p}\beta_i X_i). \tag{4.11}$$

Using the relationship between any hazard function and its corresponding survival func-

tion, the corresponding survival function is given as

$$S(t,X) = [S_0(t)]^{\exp(\sum_{i=1}^{p} \beta_i X_i)}. \tag{4.12}$$

Here, $S_0(t)$ denotes the baseline survival function. Taking the first $\log$ of the above equation, it gives,

$$\ln S(t,X) = \exp(\sum_{i=1}^{p} \beta_i X_i) \ln(S_0(t)), 0 \leq S(t,X) \leq 1.$$

This produces a negative value because $\ln(S_0(t))$ is negative. Here, taking the negative of the above equation, then applying the second log,

$$\begin{aligned}
\ln[-\ln S(t,X)] &= \ln\left[-\exp(\sum_{i=1}^{p} \beta_i X_i) \ln(S_0(t))\right], \\
&= \ln\left[-\exp(\sum_{i=1}^{p} \beta_i X_i)\right] + \ln[-\ln(S_0(t))], \\
&= \sum_{i=1}^{p} \beta_i X_i + \ln[-\ln(S_0(t))],
\end{aligned}$$

which can further be written as

$$-\ln[-\ln S(t,X)] = (\sum_{i=1}^{p} \beta_i X_i) - \ln[-\ln(S_0(t))].$$

Consider two individuals by the specific covariates,

$$X_1 = (X_{11}, X_{12}, ..., X_{1p}),$$

$$X_2 = (X_{21}, X_{22}, ..., X_{2p}).$$

By applying the above transformation,

$$\ln[-\ln S(t,X_1)] - \ln[-\ln S(t,X_2)] = \sum_{i=1}^{p} (X_{1i} - X_{2i}), \textit{which does not depend on t}$$

Alternatively, through algebraic manipulation, the aforementioned equation can be given by expressing the $\log - \log$ survival curve for individual X_1 as the $\log - \log$ curve for individual X_2 plus a linear sum term that is independent of time (t) [53].

$$\ln[-\ln S(t,X_1)] = \ln[-\ln S(t,X_2)] + \sum_{i=1}^{p} (X_{1i} - X_{2i}) \tag{4.13}$$

Equation 4.13 indicates that if the estimated $\log - \log$ survival curves for individuals are plotted on the same graph and the Cox PH model is used, these plots would be

approximately parallel. The parallelism of log-log survival plots for the Cox PH model provides a graphical way to evaluate the PH assumption. In other words, one should anticipate that empirical plots of log-log survival curves for several people will be roughly parallel if a PH model is acceptable for a particular collection of variables [54].

In the context of assessing the PH assumption using predicted survivor curves, two strategies can be employed. These strategies align with the ones previously discussed, with the key distinction lying in the comparison between observed and predicted survival curves [54].

1. Verifying PH assumption for variables one-at-a-time, or
2. Verifying PH assumption after adjusting other variables

As stated by David and Mitchel in the one-at-a-time strategy, data is stratified by categories of the predictor, and Kaplan-Meier (KM) curves are generated for each category, referred to as observed survival curves. For predicted curves, the Cox proportional hazards (PH) model is fitted for the predictors under examination. Substituting the predictor values for each category into the estimated survival curve formula yields separate estimated survival curves. The closeness between observed and predicted curves determines the satisfaction or violation of the PH assumption: close alignment indicates fulfillment, while substantial discrepancies signal a violation. The second strategy mirrors this approach, with the predicted curve adjusted for other variables, akin to the earlier method [53].

4.2.2 The Goodness of Fit (GOF) Approach

This approach is more statistical than the aforementioned graphical approaches. Many statistical tests are available to verify the PH assumption, but the most popular one is called the Schoenfeld residuals. This method is a variation of the original method proposed by Schoenfeld in 1982 [62]. This variation was given by Harrel and Lee in 1986 and is based on the residuals defined by Schoenfeld. The general idea to accept or reject the null hypothesis is based on the p-value [53]. If it is less than 0.05 then the null hypothesis is not accepted or not true, and if it is greater than 0.05, then the null hypothesis is accepted or true. The threshold for the p-value should be predefined before running the test for accepting/rejecting the null or alternative hypothesis [53, 63].

The Schoenfeld residuals are defined for each individual that experiences an event and for each predictor in the model [54]. The Schoenfeld residual is calculated as the observed value of the predictor minus the weighted average of the predictor at time t, and these weights are hazards for each individual. The PH model can be represented as

$$h(t,X) = h_0(t) * \exp(\beta_1 X_1 + \beta_2 X_2 + ... + \beta_i X_i)$$

where X_i's are predictors for each individual. The idea of this is, that if the PH assumption is true for a particular covariate, then the Schoenfeld residuals for that covariate will not be related to survival time [53]. The following three process steps are used for the implementation of the test:

1. Schoenfeld residuals are obtained for each predictor by running the Cox PH model
2. Order the failures using rank (the first one to get event gets rank 1, and so on)
3. Examine the correlation between the variables generated in the initial and subsequent stages. The null hypothesis posits that the correlation between Schoenfeld residuals and ranked failure time is zero

If the null hypothesis is rejected, it indicates a violation of the proportional hazards (PH) assumption [53, 54]. The Schoenfeld residual for the i-th individual with k covariates experiencing an event at time t_i is given by

$$r_{ik} = X_{ik} - \overline{X}_k(\beta, t_i), \tag{4.14}$$

where X_{ik} is the individual value for subject i and $\overline{X}_k(\beta, t_i)$ the weighted average of the covariate values for the individuals at risk at t_i, indicated by $R(t_i)$, and defined as

$$\overline{X}_k(\beta, t_i) = \sum_{j \in \mathscr{R}(t_i)} X_{jk} w_j(\beta, t_i).$$

The function for weight for all individuals at risk is given as:

$$\textit{[subject j fails at } t_i\textit{]} = w_j(\beta, t_i) = \frac{\exp(\beta^T X_j)}{\sum_{l \in \mathscr{R}(t_i)} \exp(\beta^T X_l)}.$$

In equation 4.14, the Schoenfeld residual is evaluated for a β by fitting a Cox PH model. A vector $(r_k = (r_{1k}, r_{1k}, ..., r_{nk}))$ is generated for each covariate k which is used to compare with the vector of rank values by a correlation test [44, 54].

4.3 Andersen-Gill (AG) Model

The Andersen-Gill (AG) model, based on a counting process framework, extends the Cox model, which is articulated in relation to increments in the count of events over time. This model focuses on the time elapsed since randomization for a specific treatment or exposure until an event transpires [64]. Referred to as the total time scale or time since study entry, it employs a shared baseline hazard function for all events and computes a universal parameter for the factors under consideration. The AG model is a widely used statistical model for analyzing time-to-event data with recurrent events and time-varying covariates [65]. It is a generalization of the Cox proportional hazards model, which assumes that the correlation between event times for a person can be explained by past events [66]. This implies that the time intervals between events are conditionally uncor-

related, given the covariates. This model is particularly appropriate when correlations among events for each individual are influenced by observed covariates [64, 67].

The AG model is well-suited for situations where correlations among events for each individual are influenced by measured covariates [67]. The data input in the counting process style is characteristic of the AG model, where each subject is represented as a series of observations with recurrence times given as $(t_0, t_1], (t_1, t_2], ..., (t_m, \text{last follow-up time}]$ [65]. In this model, each recurrent event for the i^{th} subject is assumed to follow a proportional hazard model, i.e.,

$$h_i(t) = h_0(t) \exp \beta_k X_i(t). \tag{4.15}$$

Here, $h_0(t)$ is the baseline hazard function, β is a vector of regression coefficients and X_i is a vector of observable covariates. In this model, the risk of a recurrent event for a subject follows the standard Cox proportional hazards assumption, but the number of recurrences is not taken into account. Each subject's risk intervals contribute to the risk set for every event, regardless of the number of events for each individual [65, 67].

Andersen-Gill model is specifically designed to handle correlated recurrent events, making it well-suited for longitudinal studies and medical research for repeated event occurrence. The Poisson regression, on the other hand, is a more general model for count data, it may not be as well-suited for recurrent event analysis in survival studies [**?**, 65].

4.4 Frailty Model

The frailty concept proves to be a valuable tool for integrating random effects, correlation, and latent variations into survival models [59]. Fundamentally, frailty introduces an imperceptible, variable proportionality factor that alters the hazard function of an individual or a cohort of related individuals [68]. The notion of frailty dates back to 1920 when Greenwood and Yule delved into the idea of "accident proneness." The term "frailty" found its initial application in the realm of univariate survival models through the work of Vaupel et al. in 1979 [59]. While not explicitly coined, Clayton's groundbreaking 1978 paper, examining the incidence of chronic diseases in families, extensively employed the concept of frailty [69].

In the realm of medical statistics, it is a widely accepted notion that every individual exhibits distinct characteristics or dissimilarities. Consequently, the nature of diseases or responses to treatment varies among individuals, implying that certain individuals may be at a higher risk than others [59]. This variability in individuals suggests the existence of different frailties, where more frail individuals are prone to experiencing events earlier than their counterparts. Such heterogeneity is commonly attributed to biological variation, recognized as one of the most influential sources of diversity in medicine and biology. Regrettably, this variation is often treated as a mere nuisance

rather than being critically considered [65]. Conventional clinical trials typically draw conclusions based on the average treatment effect, overlooking the intricate biological variations among participants [70].

Recent research has shown interest in additional random effects in survival models. Frailty, in this context, can be examined at both the individual and group level [65]. It is characterized as an imperceptible random effect shared among individuals or within a group. In the medical context, it is evident that frailty tends to escalate with age, compounded by other unobservable factors [71].

The frailty serves as a stochastic element introduced to account for variability not explained by other predictors in a model, stemming from unobserved factors [72]. To illustrate, consider a model with age variable and dichotomous smoking status variables as the sole predictors. If we examine two individuals at the age of 35, the survival function can be conceptualized in two ways: individual survival curves and an average over a theoretically large population. Without frailty being part of the model, the survival functions would be the same, but event times may vary. Now, introducing frailty into the model, not only will event times vary, but different individual survival functions may emerge due to unobserved factors. These unobserved factors may cause a high variability in survival times compared to what would be expected in a model without the frailty component [53].

4.4.1 Univariate Frailty Model

Frailty models are commonly viewed as extensions of proportional hazard models, with univariate frailty focusing on a single outcome variable. In such models, the frailty term introduces an unobserved random effect that alters the hazard function for each individual in the sample [44]. The objective is to accommodate unobserved heterogeneity among individuals, influencing their risk of experiencing the outcome of interest. In the context of independent individuals' lifetimes, the frailty captures this heterogeneity, signifying the impact of unobserved risk factors in a proportional hazards model [73]. Univariate frailty models prove particularly beneficial when dealing with clustered survival data, where individuals are grouped in clusters like families, hospitals, or geographical regions. Here, the frailty term represents the cluster-specific effect on the hazard function, enabling the estimation of between-cluster variation in the hazard function [72].

A frailty model incorporates a multiplicative random effect, the frailty, explicitly into the hazard function to address unobserved variability [68]. Let us understand the mathematical representation of this model. The hazard function for an individual, conditioned on the frailty Z, is expressed as:

$$h(t \mid Z) = Zh(t) \tag{4.16}$$

Here, the frailty $Z > 0$ is assumed to follow some non-negative distribution with mean 1 and variance is some parameter that is typically estimated from the data. $S(t)$ is the survival function and may include parameters and covariates. $h(t) \equiv h(t \mid Z = 1)$ is the conditional hazard for an individual with $Z = 1$ or simply conditional hazard. Individuals with $Z > 1$ are more frail with increased hazard and decreased probability of survival. Individuals with $Z < 1$ are less frail with decreased hazard and an increased probability of survival. Individuals with $Z = 1$ have average frailty. It is a very simple model that assumes that the frailty is constant over time [48, 68]. For simplicity, covariates are not expressed in the above equation. The corresponding conditional survival function for an individual is given by:

$$S(t \mid Z) = exp(-ZH(t)). \tag{4.17}$$

$H(t) = \int_0^t h_0(t)dt$ is the cumulative hazard function and survival function $S(t \mid Z = 1) = exp(-H(t))$ corresponds to the survival of an individual with average frailty value 1. The survival of a population of individuals with varying frailty values is called the marginal survival curve associated with $H(t)$ [68]. It is the expectation of $S(t \mid Z)$ with respect to Z and denoted by $\overline{S}$.

$$\bar{S} = E[exp(-ZH(t))] \tag{4.18}$$

In contrast to $S(t), \overline{S}(t)$ has a population-averaged interpretation. In case of no covariates, $\overline{S}(t)$ may be a weighted average of individual survival curves and this weighing depends on the distribution of Z. The marginal hazard from the survival function $S(t) = d/dt[-logS(t)]$ is written as:

$$\overline{h}(t) = \frac{E[Zexp(-ZH(t))]}{E[exp(-ZH(t))]}h(t) = E[Z \mid T \geq t]h(t) \tag{4.19}$$

In this context, a similar interpretation involves a weighted average, specifically the weighted average of individual hazards for individuals alive at time t. The weighting is contingent on the distribution of Z among individuals alive at time t. When all frailties at time t are equal to 1, the conditional and marginal hazards are equal [68]. However, if the frailties at time t differ from 1, the survivors' frailty distribution at that time behaves analogously to the survivors' distribution of an observable covariate. In the presence of observed covariates, it is assumed that the proportional hazard assumption holds conditional on the frailty [48, 73].

$$h(t \mid Z) = Zh_o(t)exp(\beta^T x) \tag{4.20}$$

The interpretation for the average population holds conditional on x for marginal survival and marginal hazard. To put it another way, for a hypothetical population of individuals with given covariate values x. This is the same as the interpretation of the marginal hazard. Individuals with higher risks die earlier, regardless of whether the differences are due to observed covariates x or frailty. As a result, the survivor population is more

homogeneous and at a lower risk for events than the general population at time zero [48, 68].

The choice of the frailty distribution is a crucial aspect of frailty models. Among the commonly employed frailty distributions are the Gamma distribution, the Positive Stable distribution, the Power Variance Function (PVF), the Compound Poisson distribution, and the Log-normal distribution [59]. Univariate frailty models are frequently utilized in this context. For instance, Aalen and Tretli (1999) applied the compound Poisson distribution to testicular cancer data. Their model posited that a subset of men is predisposed to testicular cancer, leading to a selection effect over time [65].

4.4.2 Relation between Marginal and Conditional Hazard

Before deriving the relation, it is crucial to comprehend the Laplace transformation as it proves highly valuable in explicating the connection between conditional hazards, marginal hazards, and survival functions [68]. The Laplace transformation provides a unique specification for the distribution of Z, setting the stage for a more nuanced exploration of their interrelation.

$$\mathscr{L}(c) = E[exp(-cZ)].$$

Note that $\mathscr{L}(0) = 0$. Expectation is derived by taking the negative derivative of $\mathscr{L}$ evaluated at 0, i.e., $E(Z) = -\mathscr{L}^{'}(0)$. Further, the k-th derivative is given as $\mathscr{L}^{k}(0) = E(Z^{k})$. The variation for the square coefficient is defined as:

$$CV^{2}(Z) = \frac{\mathscr{L}^{''}(0)}{(\mathscr{L}^{'}(0))^{2}} - 1.$$

Back to the frailty model, the Laplace transformation of the marginal survival function can be re-written as:

$$\overline{S}(t) = \mathscr{L}(H(t)),$$

and the marginal hazard can be given as:

$$\overline{h}(t) = \frac{d}{dt}[-logS(t)] = -\frac{\mathscr{L}^{'}H(t)}{\mathscr{L}(H(t))}(h(t)).$$

The above formula describes the relation between marginal and conditional hazards. The Laplace transformation of the frailty distribution of survivors can further be expressed in terms of the Laplace transform of Z [68].

4.4.3 Shared Frailty Model

The shared frailty model emerges as a valuable approach in scenarios characterized by the clustering of individuals, such as families, communities, or hospital settings [48]. This model is particularly applicable when the frailty effect is believed to be shared within these clusters, introducing a shared vulnerability that influences the overall survival dynamics within the group [69]. This type of model becomes especially relevant when examining data involving the lifespans of related individuals or recurrent events in the same person, where the assumption of independence among clustered survival times is untenable [65]. The shared frailty model, aptly named, assumes that clusters of individuals within a group share a common frailty, making it a widely employed and effective tool in frailty modeling [72].

The concept of shared frailty was initially introduced by David Clayton in 1978 in the context of a bivariate case, albeit without explicitly referring to frailty. Subsequently, this concept underwent comprehensive exploration by Hougaard in 2000 [69]. The shared frailty model closely resembles the univariate model, with the key distinction lying in the interpretation of frailty as a measure of relative risk that is shared among individuals within a specific group or cluster [72]. The mathematical formulation of this model will be elucidated further.

Consider n clusters, where cluster i encompasses n_i subjects associated with frailty Z_i $(1 \leq i \leq n)$. Denoting X_{ij} $(1 \leq i \leq n, 1 \leq j \leq n_i)$ as the vector of covariates at time T_{ij} for the jth individual in the i^{th} cluster. Then, the hazard function is given as:

$$h(t \mid X_{ij}, Z_i) = Z_i h_0(t) \exp(\beta X_{ij}). \tag{4.21}$$

In the above equation, $h_0(t)$ represents the baseline hazard, and β is a parameter to be estimated. The frailties Z_i are treated as independent and identically distributed random variables with a density function denoted as $f(z)$. These Z_i values are assumed to be independent samples from a distribution with a mean of 1 and an unknown variance. The model assumes that lifetimes are conditionally independent given the shared frailty within the clusters, establishing a dependence structure between lifetimes within the same cluster [48, 69].

Again, using the relation between the survival and hazard functions, the joint conditional survival functions can be derived. Hence, for the individuals in the ith cluster, it can be given as:

$$\begin{aligned} S(t_{i1}, ..., t_{ni} \mid X_{ij}, Z_i) &= S(t_{i1} \mid X_{i1}, Z_i) ....S(t_{in_i} \mid X_{in_i}, Z_i), \\ &= \exp\left(-Z_i \sum_{j=1}^{n_i} H_0(t_{ij}) \exp(\beta X_{ij}) \right). \end{aligned}$$

where $H_0(t) = \int_0^t h_0(s)ds$ is the cumulative baseline hazard function and $X_i = (X_{i1}, ..., X_{in_i})$ represents the covariates matrix of the individuals in the ith cluster.

Furthermore, the marginal survival function with respect to frailty Z_i is the expectation of the above survival function and it can be stated as:

$$\begin{aligned} S(t_{i1}, ..., t_{ni} \mid X_i) &= E[S(t_{i1}, ..., t_{in_i} \mid X_{ij}, Z_i)], \\ &= E\left[\exp\left(-Z_i \sum_{j=1}^{n_i} H_0(t_{ij}) \exp(\beta X_{ij})\right)\right], \\ &= \mathscr{L}\left(\sum_{j=1}^{n_i} H_0(t_{ij}) \exp(\beta X_{ij})\right). \end{aligned}$$

Here, $\mathscr{L}$ represents the Laplace transform of the frailty variable. Therefore, the multivariate survival function is expressed in the form of the Laplace transformation of the frailty distribution, which is evaluated at the cumulative baseline hazard [69].

This equation assumes independence between the frailties, which might not be realistic. The shared frailty model considers the possibility of dependence within clusters and is given as:

$$S(t_{11}, ..., t_{nn_n} \mid X_1, ..., X_n) = \prod_{i=1}^{n} \mathscr{L}\left(\sum_{j=1}^{n_i} H_0(t_{ij}) \exp(\beta X_{ij})\right). \tag{4.22}$$

The frailty in both shared and unshared frailty models is essentially the same; it is a random effect to cover a source of variation brought on by latent or unobservable causes [69]. The shared and unshared frailties, however, are applied to separate sets of data, which affects how they are interpreted and how they are estimated [48].

In unshared frailty models, survival is assumed to be independent between subjects in the study. In shared frailty, the frailty shared by the subjects accounts for dependence among subjects [69]. Shared frailty is a method for accounting for data correlation caused by unobservable characteristics shared within subject groups [48].

5 Data Analysis

This chapter encompasses various data analysis tasks, including data collection, cleaning, exploration, and preprocessing. In data collection, we gain insights into different aspects of data collection concerning this study. Subsequently, in the data cleaning and exploration phase, issues such as missing values, outliers, and inconsistencies are addressed to ensure the quality and reliability of the data. Furthermore, the data will be explored to comprehend its structure, distribution, and patterns, which may involve employing summary statistics, data visualization, and obtaining preliminary insights. Finally, data will be prepared for analysis by transforming and encoding variables, handling categorical data, and scaling or normalizing data, if necessary.

5.1 Data Collection Method

5.1.1 Site of the study

The study took place in Siaya County, situated in western Kenya, and spanned two cohorts. The initial cohort extended from March 2004 to October 2005, while the second cohort covered the period from February 2009 to December 2012 [4]. Siaya County is a rural area predominantly inhabited by the Luo ethnic group, constituting more than 96% of the population [3]. This region is recognized as a holoendemic *P. falciparum* transmission area, with *P. falciparum* malaria is a significant contributor to childhood mortality and morbidity. Severe malarial anemia (SMA) contributes significantly to the occupancy rates of hospital beds at Siaya County Referral Hospital (SCRH), leading to substantial morbidity and mortality within the hospital setting. [2].

5.1.2 Study design and participants

Children under five years of age exhibiting symptoms of malaria, irrespective of gender, were eligible for inclusion in the study [2]. Upon enrollment, participants underwent comprehensive examinations to gather clinical and demographic information with details about the illness history. Children, who tested negative for *P. falciparum*, had a recent history of hospitalization for any reason, received blood transfusions, reported antimalarial therapy use in the past two weeks, or had cerebral malaria were excluded from the study [4]. Given that malaria is the primary cause of pediatric hospitalization in the region and prior antimalarial use might influence the immunologic and outcome variables under investigation, this selection strategy was deemed appropriate [74].

Hemoglobin levels served as the basis for categorizing study participants with *P. fal-*

ciparum malaria (any density) into uncomplicated malaria (UM) ($Hb \geq 5.0g/dL$) and severe malarial anemia (SMA) ($Hb < 5.0g/dL$) [2]. A study explored the impact of HIV-1/2 and bacteremia status on the severity of malarial anemia, with these infections being characterized in all participants. Treatment adhered to the guidelines outlined by the Ministry of Health (MOH), Kenya. Written informed consent from the parent/legal guardian of each study participant was obtained for enrollment [3, 4].

5.1.3 Longitudinal follow-up

Throughout the follow-up period, parents/guardians were requested to visit their children once every three months (i.e., a quarterly visit). If parents or guardians did not bring their children back for the regular quarterly visit, study officials visited the homes to inquire about the children's health (including mortality) [3]. Furthermore, parents/-guardians were asked to bring their child to the hospital during any fever episode(s) or other diseases (i.e., acute visit) for adequate clinical management and longitudinal documentation of childhood illnesses [2]. Participants in the study had a thorough physical and laboratory workup for proper clinical management at each acute and quarterly visit (e.g., complete blood count, malaria parasitemia determination, examination of viral and bacterial infections, etc.). Data on mortality were gathered throughout the follow-up period either from hospital records or verbal autopsy in cases when the death occurred outside the hospital [3, 4].

5.1.4 Laboratory measures

Blood samples were obtained for laboratory diagnostics and an HIV exposure test was performed. Bacterial comorbidity was determined using culture and sensitivity tests, and the determination of Glucose six phosphate dehydrogenase (G6PD) deficiency (levels: normal, intermediate, deficient) was done [4]. To better understand chronic anemia caused by genetic factors, investigations were conducted on sickle cell traits and α-thalassemia deletions [2].

5.2 Data Cleaning, Exploration and Preprocessing

The raw data comprises a combination of two clinical cohorts with multiple clinical visits of each patient. There are $19,894$ observations, encompassing 41 distinct categorical and numerical variables. The next steps involve cleaning the data and preparing for our models. After the data is cleaned, a small explanation of the final variables will be provided for better understanding.

Ensuring the data's completeness is a crucial step in any analysis. R-Studio serves as the coding platform for all tasks in the study. In figure 5.1, the y-axis represents

the variable names, while the x-axis indicates the percentage of missing rows for each variable.

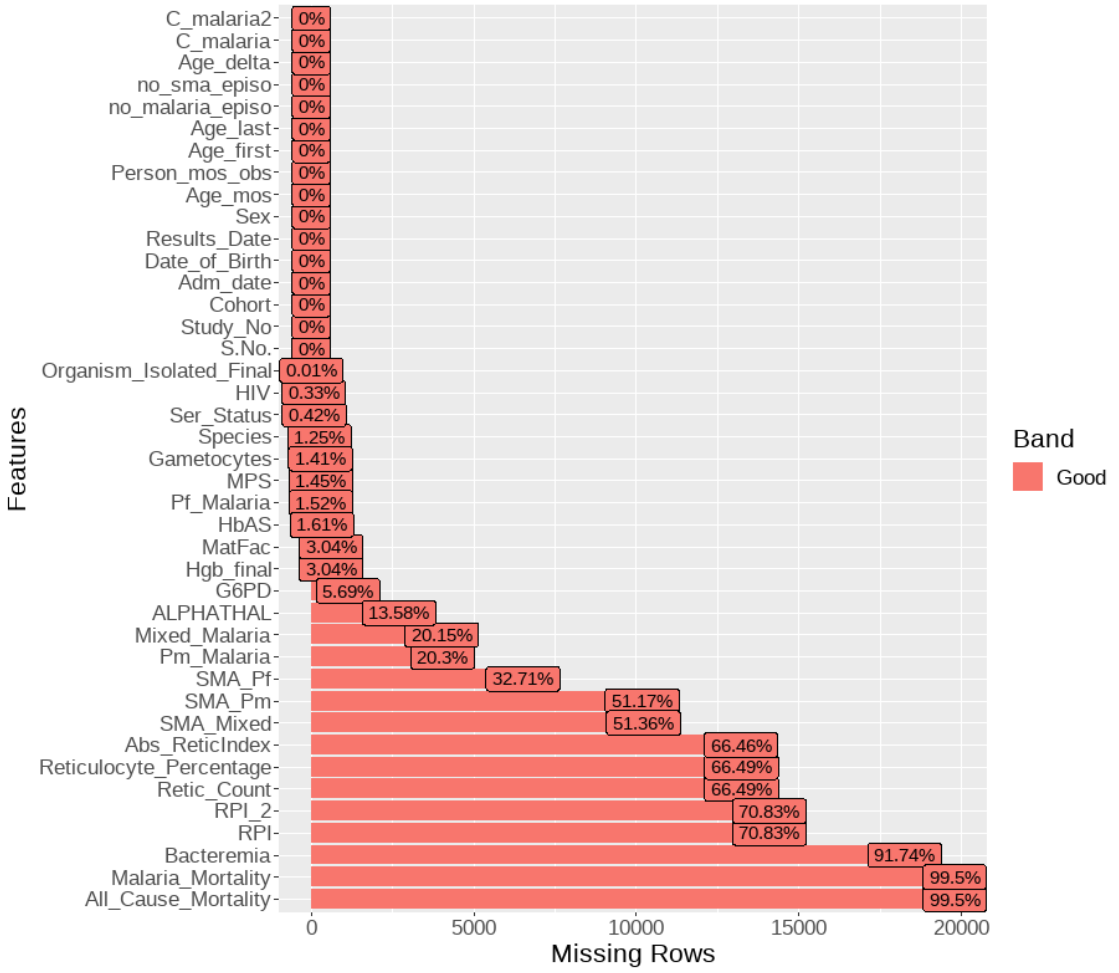

Figure (5.1) Missing data info.

Notably, from the top (C_malaria2 to S.No.), there are no missing rows. Following that, from (Organism_Isolated_Final to G6PD), less than 10% of rows exhibit missing data. However, beyond this point, the percentage of missing rows increased significantly. Specifically, from (Malaria_Mortality to All_Cause_Mortality), over 99% of rows have missing data. This observation prompts a deeper exploration of the dataset to identify factors that could explain the high prevalence of missing data.

After exploring the dataset, it becomes evident that several variables contain substantial missing values. For instance, "Retic_Count" and "Reticulocyte_Percentage" exhibit $13,228$ missing values out of $19,894$ observations. "Abs_ReticIndex" is not far behind with $13,236$ missing values, while 'RPI' and "RPI_2" show $14,090$ missing values. "SMA_mixed" is particularly affected, with $10,218$ missing values, while "SMA_Pm" and "SMA_Pf" have $12,097$ missing values, respectively. Furthermore, "Organism_Isolated_Final" and "Bacteremia" both record $18,244$ and $18,250$ missing values, respectively. "All_Cause_Mortality"

and "Malaria_Mortality" rank the highest in terms of missing data, with $19,794$ missing data.

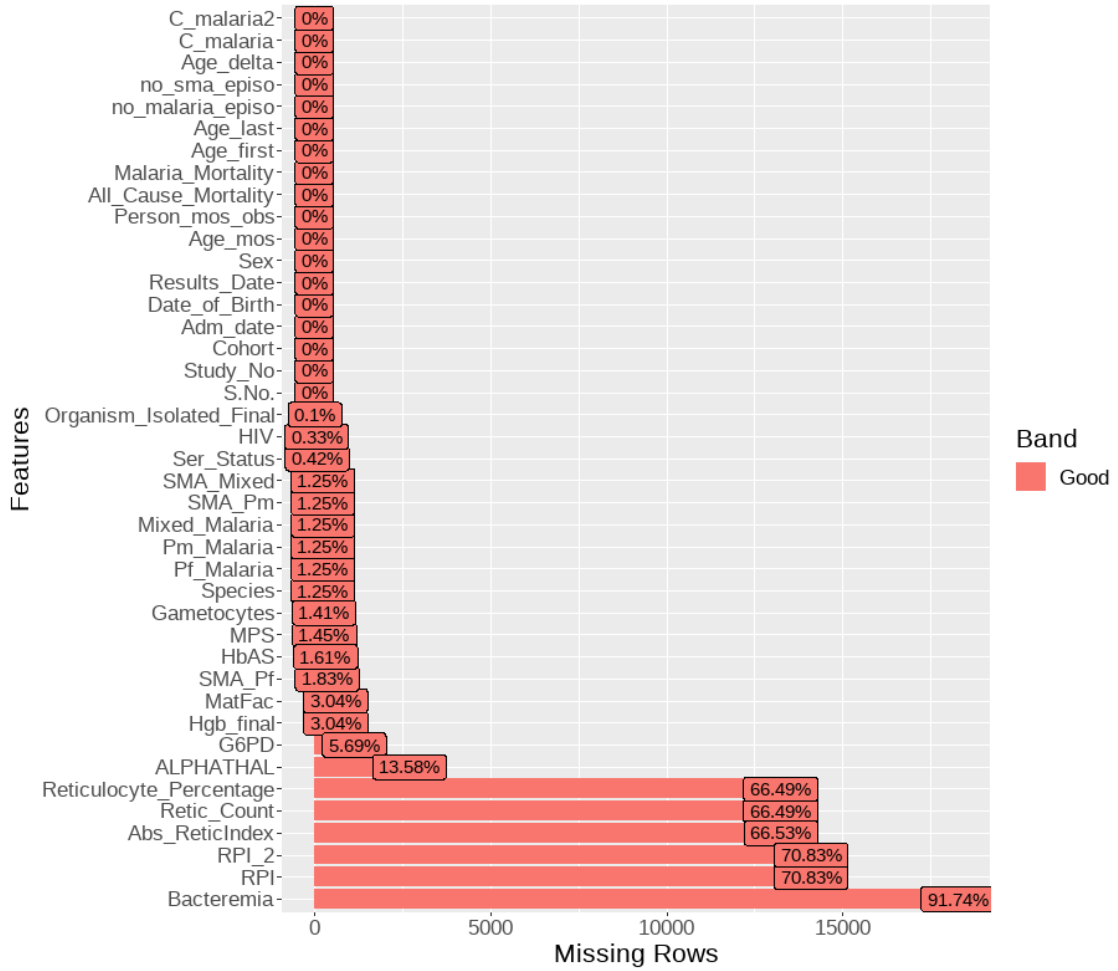

Figure (5.2) Missing data information

During the data exploration, some missing values were attributable to data entry errors that could be rectified through data modification. For instance, consider the "Species" variable, a categorical variable with various categories. Filtering this variable for 'Pf' (P. falciparum), the possible values for "Pf_Malaria" become binary (0 or 1). Consequently, we can replace the missing values (NA's) in "Pf_Malaria" with 0, effectively eliminating NA's and simplifying the variable into two categories.

A similar approach can be applied to "Pm_Malaria" and "Pm_Malaria," further reducing the number of missing values in these variables. Furthermore, rectifying the data for the "All_Cause_Mortality" variable by examining the last clinical visit status for each patient can aid in minimizing missing values for this variable. These data modifications have proven effective in significantly reducing the amount of missing data in several columns. As depicted in figure 5.2, our data modification efforts have yielded a substantial reduction in missing data for several key variables, including "SMA_Mixed," "SMA_Pm,"

"Mixed_Malaria," "Pm_Malaria," "Pf_Malaria," and "SMA_Pf."

After conducting further analysis and discussions, the decision was made to eliminate variables with more than 50% missing data. Additionally, data collection errors were addressed, leading to the removal of additional variables. Consequently, the final dataset includes data from $1,654$ unique individuals across two cohorts, resulting in a total of $16,535$ observations with 22 variables. In this study, the significance of malaria status applies regardless of the specific malaria species, and the same applies to severe malarial anemia (SMA). Figure 5.3 provides a visual representation of the missing data information in the cleaned final dataset, which is devoid of any missing values. This dataset will be utilized for subsequent statistical analyses and modeling.

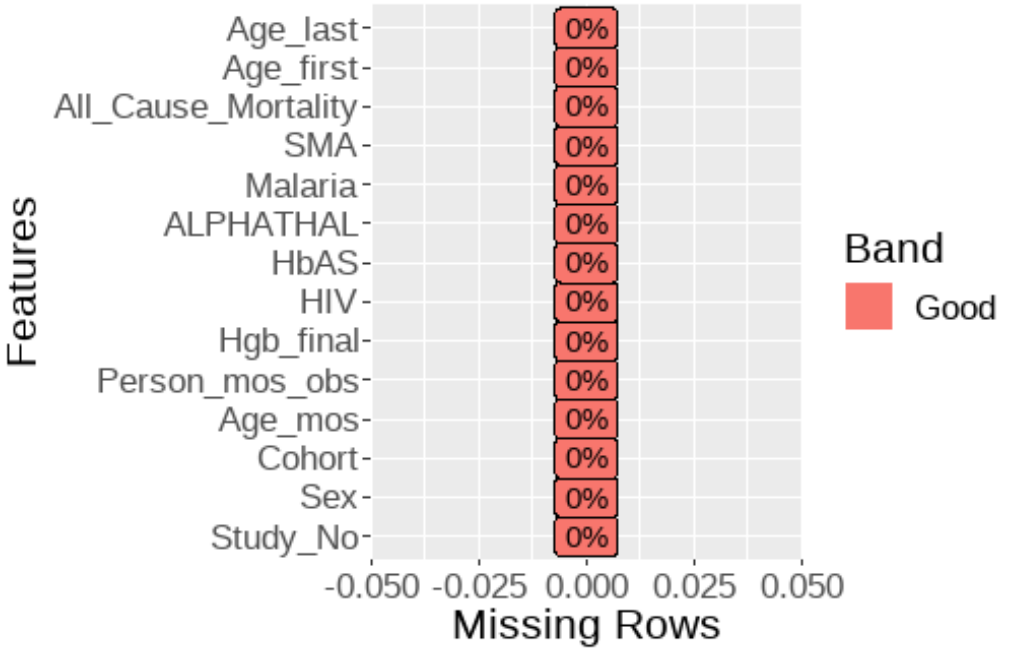

Figure (5.3) Missing data information in final dataset

The distribution of metric variables was evaluated using a combination of box plots, histograms, and Q-Q plots to assess normality. Descriptive statistics, including the median with the interquartile range and the mean with the standard deviation, were calculated for each metric variable. For those variables displaying a normal distribution, one-way ANOVAs, and two-sample t-tests were employed for between-group comparisons. In instances where normality assumptions were not met, Kruskal-Wallis tests and Mann-Whitney U tests were utilized. Furthermore, the distribution of categorical variables across two groups, Malaria and Severe Malarial Anemia (SMA), was examined using a Chi-square test for homogeneity.

Moreover, it was observed that certain variable names were not easily comprehensible. Therefore, a table 5.1 has been generated to provide concise descriptions of the variables.

Variable Name	**Short Description**
Study_No	Refers to patient number
Sex	Gender (Male/Female)
Cohort	Cohort number (1 or 2)
Age_mos	Age in months
person_mos_obs	Observation time in months since enrollment
Hgb_final	Hemoglobin level (g/dl)
HIV	HIV status (0 = Negative, 1= Positive)
HbAS	Sickle cell status (AA, AS, SS)
ALPHATHAL	Alpha-thalassemia (α- thalassemia status) ($\alpha\alpha/\alpha\alpha, \alpha/\alpha\alpha, \alpha/\alpha$)
Malaria	Malaria status (0 = Negative, 1 = Positive)
SMA	Severe Malaria Anemia status (0 = Negative, 1 = Positive)
All_Cause_Mortality	Dead or alive due to any cause (0 = Alive, 1 = Dead)
Age_first	Age in months at enrollment
Age_last	Age in months at the last clinical visit
Age_delta	Difference of Age_last and Age_first

Table (5.1) Variable Descrition

6 Results

The chapter unveils crucial insights in four sections. First, "Characteristics upon Enrollment" details participants' demographic and clinical data. Second, in "Mortality Models", sophisticated techniques analyze factors affecting mortality. Third, "Malaria and SMA Frequency Models" explores the prevalence and patterns of malaria and SMA cases. Lastly, the "Frailty model for recurrent events" is to assess the impact of various covariates on the survival time, considering frailty in the study subjects.

6.1 Characteristics at Enrollment

In this study, the data, consisting of $1,391$ observations, was classified into two categories: aparasitemic ($n = 253$) and parasitemic ($n = 1138$). The parasitemic group was further stratified into two subgroups: malaria (hemoglobin; Hb ≥ 5.0 g/dL, $n = 872$) and severe malarial anemia (SMA, Hb < 5.0 g/dL, $n = 266$). Table 6.1 displays the demographic, clinical, and laboratory characteristics of the study participants. Although the distribution of the two cohorts was evenly balanced, Cohort 1 (49.4%) and Cohort 2 (50.6%), exhibited significant differences between the aparasitemic and parasitemic groups. Gender distribution was equal, with 50% females and 50% males, and these proportions were comparable ($p = 0.975$). Children in the SMA group were significantly younger ($p < 0.001$) compared to the aparasitemic and malaria groups.

Clinical measurements indicated that in comparison to the aparasitemic and malaria groups, children with SMA exhibited a significant decrease in hemoglobin levels ($p < 0.001$). Co-infections were infrequent, as the majority of participants tested negative for HIV (96.1%), while a small minority tested positive (3.9%). The presence of HIV differed significantly among the groups ($p < 0.001$). Further, the distribution of sickle cell genotypes differed across the groups ($p = 0.022$), with a higher frequency of Hb_AS carriage in children with malaria and the lowest frequencies among children with SMA. The distribution of α-thalassemia deletion variants also varied across the groups ($p < 0.001$), with the highest frequencies of single and double deletions observed in children with SMA, demonstrating a significant difference.

After understanding the diverse patient outcomes at the time of enrollment, our next step involves employing a range of statistical models. These models will be instrumental in the identification of risk factors that are significantly associated with mortality, malaria infection, and SMA within our medical research study.

Table (6.1) Demographic, clinical, and laboratory characteristics of dataset

Characteristic	Total	Aparasitemic	Malaria (Hb $\geq$ 5.0 g/dL)	SMA (Hb<5.0 g/dL)	p-value[1]
No. of Participants	**1,391**	**253**	**872**	**266**	
Demographic parameters					
Study Cohort, n (%)					< 0.001
One (1)	687 (49.4%)	166 (65.6%)	373 (42.8%)	148 (55.6%)	
Two (2)	704 (50.6%)	87 (34.4%)	499 (57.2%)	118 (44.4%)	
Sex, n (%)					0.975
Female	696 (50.0%)	128 (50.6%)	436 (50.0%)	132 (49.6%)	
Male	695 (50.0%)	125 (49.4%)	436 (50.0%)	134 (50.4%)	
Age, Months					< 0.001
Mean (SD)	13.590 (8.092)	13.554 (8.811)	13.974 (7.806)	12.367 (8.207)	
Clinical measurements					
Hemoglobin, g/dL					< 0.001
Mean (SD)	7.463 (2.534)	9.577 (2.567)	7.860 (1.837)	4.149 (0.710)	
Co-infections					
HIV, n (%)					< 0.001
Negative (0)	1337 (96.1%)	243 (96.0%)	849 (97.4%)	245 (92.1%)	
Positive (1)	54 (3.9%)	10 (4.0%)	23 (2.6%)	21 (7.9%)	
Genetic variants					
Sickle Cell status, n (%)					< 0.001
Hb_AA	1168 (84.0%)	200 (79.1%)	724 (83.0%)	244 (91.7%)	
Hb_AS	205 (14.7%)	44 (17.4%)	144 (16.5%)	17 (6.4%)	
Hb_SS	18 (1.3%)	9 (3.6%)	4 (0.5%)	5 (1.9%)	
α-Thalassemia, n (%)					0.022
$\alpha\alpha/\alpha\alpha$	590 (42.4%)	112 (44.3%)	373 (42.8%)	105 (39.5%)	
$\alpha/\alpha\alpha$	520 (37.4%)	75 (29.6%)	336 (38.5%)	109 (41.0%)	
α/α	281 (20.2%)	66 (26.1%)	163 (18.7%)	52 (19.5%)	

6.2 Mortality Models

This section addresses the first research question to identify risk factors associated with mortality using three different models.

First, a Cox proportional hazard model was performed with patient-level data ($n = 1,391$) with covariates to risk factors associated with all-cause mortality. In table 6.2, the covariates that emerged from the model are presented with results. The observed HR (0.567) along with ($95\%CI : 0.431 - 0.732, p < 0.001$) indicate that for every unit increase in hemoglobin, there is a 43.84% lower risk of mortality. The negative coefficient (-0.577) suggests that lower hemoglobin levels are associated with a higher risk of mortality among these children. Additionally, childrens with the $(\alpha/\alpha\alpha)$ genotype have a lower risk of mortality as compared to $(\alpha\alpha/\alpha\alpha)$, as indicated by a hazard ratio of 0.144 ($95\%CI : 0.018 - 1.174, p = 0.070$). However, for those with the (α/α) genotype, the difference in mortality risk is not statistically significant ($HR = 0.940, 95\%CI : 0.295 - 2.993, p = 0.917$).

Further, HIV-positive children have substantially higher mortality risk, with a hazard ratio of 3.922 ($95\%CI : 1.285 - 11.974, p = 0.016$), as compared to HIV-negative children. Age is associated with a lower risk of mortality. For each one-year increase in age, there was a 6.49% reduction in the risk of mortality, as indicated by a hazard ratio of

Table (6.2) Cox model analysis for mortality

	Estimate	HR	SE	p	(95% CI)
Hgb_final	-0.577	0.562	0.135	<0.001	(0.431, 0.732)
ALPHATHAL $(\alpha/\alpha\alpha)$	-1.940	0.144	1.072	0.070	(0.018, 1.174)
ALPHATHAL (α/α)	-0.062	0.940	0.591	0.917	(0.295, 2.993)
HIV_1 (+)	1.367	3.923	0.569	0.016	(1.285, 11.974)
Age_last	-0.067	0.935	0.030	0.026	(0.882, 0.992)
Test Results					
Test	**Test Stat**	**df**			
Observations	1,391				
No. of events	14				
AIC	120.55				
Log Likelihood	-55.273				
Concordance	0.976				
Wald Test	44.700	5			

0.935 $(95\%CI : 0.882 - 0.992, p = 0.026)$. As age increases, the body develops a better immune system or better disease resistance.

Next, a predictive model is performed for mortality during the initial clinical encounter, employing a logistic regression framework. To this end, a dataset is compiled comprising information from $1,391$ patients, specifically focusing on their first clinical visit characteristics. Additionally, we a new variable was introduced, denoted as "Malaria_status_last_visit," designed to discern the malaria status of patients during their most recent clinical visit. In this context, a value of 0 signifies a negative malaria diagnosis, while a value of 1 indicates a positive diagnosis of malaria. The introduction of this variable serves the purpose of investigating potential associations between patient mortality and those who tested positive for malaria during their last clinical visit.

Table (6.3) Logistic model for mortality

	Estimate	SE	p	(95% CI)
(Intercept)	-7.675	1.176	<0.001	(-10.777, -5.862)
Malaria_last_visit (+)	1.992	1.176	0.090	(-0.104, 5.024)
HIV_1 (+)	3.487	1.030	<0.001	(1.322, 5.659)
Test Results				
Test	**Test Stat**	**df**		
Observations	1,391			
AIC	49.663			
Null deviance	54.800	1,390		
Residual deviance	43.663	1,388		

The model examines the influence of several covariates and the results drawn up by the model are presented in table 6.3. The model emerges HIV as the only significant

covariate. The coefficient for Malaria_status_last_visit is estimated to be 1.992 ($p = 0.090$). The coefficient for HIV-positive estimates to be 3.487 ($p < 0.001$). This indicates that children with HIV have significantly higher odds of mortality compared to those without HIV. Additionally, the model suggests that children with malaria-positive status at their last visit have increased odds of mortality compared to those without malaria-negative status, although the effect is not statistically significant.

Finally, a Cox Proportional Hazards model is employed with the addition of the frailty term to account for unobserved heterogeneity or variability that may exist among different children. It helps to address the potential presence of unmeasured factors that can influence the risk of the event of interest (in this case, all-cause mortality). The model runs the complete dataset ($n = 16,535$) with covariates from the final dataset.

As per table 6.4, age exhibits a mild effect on mortality, with an estimated hazard ratio (HR) of 0.950 $(95\%CI : 0.887 - 1.018)$, suggesting that for each additional month of age, the risk of mortality decreases slightly, though this trend did not reach statistical significance ($p = 0.150$). Malaria status (malaria (+)) emerged as a factor influencing mortality. Children with a positive malaria status show a higher hazard of mortality, with an estimated HR of 0.276 and $p = 0.034$, as compared to children without malaria. However, this effect was moderately significant.

Table (6.4) Cox model for mortality with frailty

	Estimate	**HR**	**SE**	**Chi_sq**	**p**	**(95% CI)**
Age (months)	-0.051	0.950	0.035	2.11	0.150	(0.887, 1.018)
Malaria (+)	-1.287	0.276	0.608	4.48	0.034	(0.084, 0.909)
Hgb_final	-0.690	0.502	0.132	27.15	<0.001	(0.387, 0.650)
HIV (+)	2.320	10.171	0.569	16.59	<0.001	(3.332, 31.051)
Frailty (Study_No)	-	-	-	0.00	0.950	-
Test Results						
Test	**Test Stat**	**df**				
Observations	16,535					
No. of events	14					
AIC	192.85					
Concordance	0.96					
Variance of random effect	<0.001					
Likelihood ratio test	58.98	4				

Hemoglobin levels (Hgb_final) have a substantial impact on mortality, demonstrating an estimated HR of 0.502 $(95\%CI : 0.387 - 0.650)$. This implies that higher hemoglobin levels are strongly associated with a reduced risk of mortality ($p < 0.001$). Lastly, HIV status (HIV (+)) emerges as a significant risk factor for mortality. Children with HIV infection exhibited a markedly elevated hazard of mortality, as reflected by the estimated HR of 10.171 $(95\%CI : 3.332 - 31.051)$, with a p-value indicating a highly significant association ($p < 0.001$). Notably, the frailty term associated with Study_No does not significantly impact mortality ($p = 0.95$). Suggesting that the unmeasured factors that can influence the risk of mortality, are not significant.

From a clinical perspective, these findings emphasize the relevance of age, malaria status, hemoglobin levels, and HIV status as critical determinants of All-Cause Mortality. Increasing age and HIV infection are identified as prominent risk factors for mortality, while higher hemoglobin levels are associated with a reduced mortality risk.

6.3 Malaria and SMA frequency Models

This section focuses on the second aspect of our research question. In malaria-endemic regions like western Kenya, children frequently endure multiple malaria episodes, which can sometimes culminate in life-threatening complications such as SMA [4]. Throughout the study duration, there are a total of $6,728$ malaria and 384 SMA episodes. The initial segment of the investigation focuses on finding the association between the covariates and the frequency of malaria and SMA episodes. To achieve this, a generalized linear model (Poisson regression) is used to discern the factors linked to the incidences of malaria and SMA. This inquiry was further divided into two parts.

In the first part, the data is selected assuming that the children are observed since birth with covariates such as Sex, Age_at_first_visit, Cohort, HIV, Alphathal, and HbAS. The dataset consists of $1,391$ children and the observation time is taken from birth to the last visit. A Poisson rate regression, with the logarithm of Age_at_last_visit (at enrollment) being the offset variable (rate regression), is applied. A variable specifically incorporated into the model to account for differences in exposure duration among children, thus ensuring our analysis accurately reflects the underlying associations. The results unearthed noteworthy associations, revealing the multiplicative change in incident risk, as represented by the Incident Risk Ratio (IRR).

As per the model results from table 6.5, male children exhibit higher malaria episodes ($Estimate = 0.068,\ p = 0.006$), as reflected in the IRR of 1.070, indicating a 6.78% higher risk of malaria compared to female children. Age at enrollment is strongly associated with malaria episodes, with a significant negative coefficient ($Estimate = -0.033$, $p < 0.001$). This IRR of 0.966 indicates that for every one-year increase in age, the risk of malaria decreases by 3.32%. Older children, therefore, experienced lower malaria episodes.

Additionally, children with HIV ($Estimate = -0.538,\ p < 0.001$) show 42.63% reduced risk in comparison to children without HIV. Although the Alpha-Thalassemia variable is selected by the model, it does not appear to have a significant association with malaria count. On the other hand, protection against multiple bouts of malaria over time is also associated with co-inheritance of heterozygosity. For both the sickle cell disease HbAS ($Estimate = -0.205,\ p < 0.001$) and HbAS (SS) ($Estimate = -0.739,\ p < 0.001$) do not indicate a significant effect of malaria counts.

Table (6.5) Poisson model for malaria count (from birth)

	Estimate	IRR	SE	p	(95% CI)
(Intercept)	-1.500	0.224	0.031	<0.001	(-1.558, -1.436)
Sex (Male)	0.068	1.070	0.024	0.006	(0.020, 0.115)
Age_first_visit	-0.032	0.968	0.002	<0.001	(-0.035, -0.029)
HIV (+)	-0.538	0.584	0.093	<0.001	(-0.727, -0.361)
ALPHATHAL $(\alpha/\alpha\alpha)$	0.011	1.011	0.027	0.676	(-0.042, 0.065)
ALPHATHAL (α/α)	-0.074	0.929	0.034	0.027	(-0.140,-0.009)
HbAS (AS)	-0.205	0.815	0.037	<0.001	(-0.279, -0.133)
HbAS (SS)	-0.739	0.478	0.159	<0.001	(-1.067, -0.442)
Test Results					
Test	**Test Stat**	**df**			
Observations	1,391				
AIC	6,937.38				
Null deviance	3,314.5	1,390			
Residual deviance	2744.6	1,383			

In the case of SMA episodes with the same covariates, table 6.6 presents that Age at enrollment, HbAS, and HIV variables show the same effect as in the case of the number of malaria episodes. In addition, cohort variable is selected by the model indicating that cohort 2 ($Estimate = -0.177,\ p = 0.101$) children show a non-significant decrease in SMA counts compared to Cohort 1. However, the association is not statistically significant.

Table (6.6) Poisson model for SMA count (from birth)

	Estimate	IRR	SE	p	(95% CI)
(Intercept)	-4.014	0.018	0.105	<0.001	(-4.222, -3.810)
Age_first_visit	-0.053	0.948	0.008	<0.001	(-0.068, -0.039)
Cohort (2)	-0.178	0.837	0.108	0.101	(-0.391, 0.033)
HIV (+)	0.873	2.395	0.205	<0.001	(0.448, 1.254)
HbAS (AS)	-0.758	0.469	0.197	<0.001	(-1.165, -0.392)
HbAS (SS)	0.820	2.269	0.322	0.011	(0.120, 1.394)
Test Results					
Test	**Test Stat**	**df**			
Observations	1,391				
AIC	1,906.33				
Null deviance	1,315.5	1,390			
Residual deviance	1,204.1	1,385			

In the second part, the data is compiled from enrollment until the last visit, this data has ($n = 1,295$) observations and the same covariates. Again a Poisson rate regression model, with the (logarithm of) Person_mos_obs being the offset variable (rate regression).

Table (6.7) Poisson model for malaria count (from enrollment)

	Estimate	**IRR**	**SE**	**p**	**(95% CI)**
(Intercept)	-1.440	0.237	0.032	<0.001	(-1.502, -1.378)
Sex (Male)	0.065	1.067	0.025	0.008	(0.017, 0.113)
Age_first_visit	-0.005	0.995	0.002	0.005	(-0.008, -0.001)
Cohort (2)	0.066	1.068	0.026	0.012	(0.015, 0.117)
HIV (+)	-0.388	0.678	0.095	<0.001	(-0.580, -0.207)
ALPHATHAL $(\alpha/\alpha\alpha)$	-0.004	0.996	0.028	0.898	(-0.058, 0.051)
ALPHATHAL (α/α)	-0.073	0.930	0.034	0.031	(-0.140, -0.007)
HbAS (AS)	-0.203	0.816	0.037	<0.001	(-0.277, -0.130)
HbAS (SS)	-0.706	0.494	0.161	<0.0001	(-1.038, -0.406)
	Model fit				
Test	**Test Stat**	**df**			
Observations	1,315				
AIC	7,196.3				
Null deviance	3,209.1	1,314			
Residual deviance	3,109.6	1,306			

Table 6.7 shows that male children exhibit a significantly higher malaria count compared to female children ($Estimate = 0.065,\ p = 0.008$). The estimate suggests a 6.49% increase in the risk of malaria counts for males compared to females. Age at the first clinical visit shows a significant negative association with Malaria count ($Estimate = -0.005,\ p = 0.005$). For every additional year in age at the first visit, there is a 0.46% reduction in the risk of malaria. Cohort 2 participants exhibited a significantly higher malaria count compared to cohort 1 ($Estimate = 0.066,\ p = 0.0117$).

Additionally, the impact of Alphathal genotypes on malaria does not show a significant association with malaria count, the $(\alpha/\alpha\alpha)$ genotype ($Estimate = -0.004,\ p = 0.898$) and the (α/α) ($Estimate = -0.073,\ p = 0.031$). Further, hemoglobin variants have been associated with malaria resistance and susceptibility. Both the HbAS (AS) ($Estimate = -0.203,\ p < 0.001$) and the HbAS (SS) ($Estimate = -0.706,\ p < 0.001$) exhibit significant negative association with malaria count. The estimates correspond to Incident Risk Ratios (IRRs) of 0.816 and 0.494, suggesting an 18.32% lower and 50.56% lower risk, respectively.

For SMA episodes with the same covariates, Table 6.8 shows that age at the first clinical visit exhibits a robust negative association with SMA count, signifying a 2.66% ($IRR = 0.974,\ p < 0.001$) decrease in the risk of SMA for every additional year in age at the first visit. Notably, HIV-positive children display a substantially 168.85% ($IRR = 2.687,$

$p < 0.001$) higher SMA count compared to their HIV-negative children. Furthermore, hemoglobin types emerge as a significant factor, with both HbAS (AS) and HbAS (SS) demonstrating distinct associations with SMA count. The Incident Risk Ratios (IRRs) for HbASAS and HbASSS are 0.4746 ($IRR = 0.475,\ p < 0.001$) and 2.4484 ($IRR = 0.2.447,\ p = 0.005$), indicating a 52.54% lower and 144.84% higher risk of SMA counts, respectively.

Table (6.8) Poisson model for SMA count (from enrollment)

	Estimate	**IRR**	**SE**	**p**	**(95% CI)**
(Intercept)	-4.014	0.018	0.104	<0.001	(-4.220, -3.813)
Age_first_visit	-0.027	0.974	0.007	<0.001	(-0.041, -0.013)
HIV (+)	0.989	2.687	0.216	<0.001	(0.538, 1.387)
HbAS (AS)	-0.745	0.475	0.197	<0.001	(-1.154, -0.380)
HbAS (SS)	0.895	2.447	0.322	0.005	(0.196, 1.470)
	Model fit				
Test	**Test Stat**	**df**			
Observations	1,315				
AIC	2,169.5				
Null deviance	1.542.2	1,314			
Residual deviance	1,489.3	1,310			

After assessing the influence of covariates on the occurrence of malaria and SMA episodes, the analysis extended to examining the impact of these variables on the time intervals between recurrent events. This analysis was conducted using a multiple-event per subject Cox proportional hazard model (Andersen-Gill). Two distinct datasets were created with covariates Sex, Age_last, Cohort, HIV, and HbAS. One dataset, consisting of $7,578$ observations, is designed for modeling malaria recurrent events, while the other, with $1,654$ observations, is intended for modeling SMA recurrent events.

The covariates that emerge from the model are shown in table 6.9. Age at the last visit is a crucial predictor, with a negative coefficient ($IRR = 0.984, p < 0.001$), indicating that older children have a lower risk of malaria incidence. The coefficient for sex indicates a slight positive, however, this difference is not statistically significant. Hemoglobin Genotypes HbAS (AS) and HbAS (SS) play a significant role in the risk of malaria incidents. Inheritance of either the HbAS (SS) mutant genotype ($P < 0.001$) or the HbAS (AS) variant ($P < 0.001$) decreases the longitudinal hazard for malaria infections. The negative coefficient for HIV (+) suggests that children with HIV have a lower risk of malaria incidence ($IRR = 0.619,\ p = 0.001$) than children without HIV. This result is intriguing and should be interpreted cautiously. It might be due to the influence of HIV treatment or the complex relationship between HIV and other health conditions. Children in cohort 2 have a slightly lower risk of malaria incidence than those in Cohort 1 ($IRR = 0.9003$, $p = 0.008$).

Table (6.9) AG model for malaria incidence

	Estimate	IRR	SE	p	(95% CI)
Sex (Male)	0.051	1.053	0.024	0.190	(0.975, 1.137)
Age_last	-0.017	0.984	0.001	<0.001	(0.981, 0.986)
Cohort (2)	-0.105	0.900	0.024	0.008	(0.834, 0.972)
HIV (+)	-0.480	0.620	0.094	<0.001	(0.463-0.828)
HbAS (AS)	-0.231	0.794	0.037	<0.001	(0.710, 0.888)
HbAS (SS)	-0.838	0.432	0.159	<0.001	(0.314, 0.595)
	Model fit				
Test	**Test Stat**	**df**			
Observations	7,578				
Events	6,728				
AIC	90,854.49				
Wald Test	189.310***	6			
LR Test	331.883	6			
Score (Logrank) Test	317.448	6			

In the case of SMA, the same model is applied with the same covariates. The covariates and the results derived from the model are presented in table 6.10.

Table (6.10) AG model for SMA incidence

	Estimate	IRR	SE	p	(95% CI)
Age_first	-0.043	0.958	0.008	<0.001	(0.943, 0.973)
Cohort (2)	-0.215	0.807	0.108	0.050	(0.651, 1.000)
HIV (+)	0.906	2.474	0.205	<0.001	(1.633-3.749)
HbAS (AS)	-0.737	0.478	0.197	0.001	(0.305, 0.749)
HbAS (SS)	0.872	2.393	0.322	0.027	(1.103, 5.190)
	Model fit				
Test	**Test Stat**	**df**			
Observations	1,654				
Events	384				
AIC	5,316.72				
Wald Test	72.680	5			
LR Test	89.221	5			
Score (Logrank) Test	88.419	5			

The model suggests that older children (at enrollment) have a lower risk of recurrent incidence of SMA ($P < 0.001$). The hazard decreases by a factor of 0.958 for each additional month of age. Cohort 2 is associated with a slightly reduced risk of SMA compared to Cohort 1 ($IRR = 0.807$, $p = 0.050$). The difference is statistically significant but relatively small p-value. The results indicate that children with the HbAS (AS) genotype have a significantly lower risk incidence of SMA ($IRR = 0.478$, $p = 0.001$), while those

with the HbAS (SS) genotype have a higher risk of getting SMA incidence ($IRR = 2.393$, $p = 0.027$). This is in line with known clinical evidence that certain hemoglobin genotypes are associated with susceptibility or resistance to diseases like SMA. Co-infection with malaria and HIV is a strong predictor of increased susceptibility to SMA incidence ($P < 0.001$). The positive coefficient for HIV (+) ($IRR = 2.474, P < 0.001$) suggests that children with HIV have a higher susceptibility to SMA incidence.

In the last part, the goal is to analyze the impact of various covariates on survival time with a specific focus on the occurrence of malaria and SMA. Multiple-event per subject Cox model with frailty on Study_no is applied to account for unobserved or unmeasured variables that may affect survival. As in the above case, two different models are presented, one for malaria and the other for the SMA. The dataset (n=16535) was used in both the models with covariates as Sex, Age_first, Cohort, HIV, and HbAS.

Table (6.11) Frailty model for malaria

	Estimate	**HR**	**SE**	**Chisq**	**p**	**(95% CI)**
Sex (Male)	0.048	1.050	0.037	1.70	0.190	(0.976, 1.128)
Age_first	-0.047	0.955	0.002	367.88	<0.001	(0.950, 0.960)
Cohort (2)	0.062	1.064	0.038	2.63	0.100	(0.987, 1.147)
HIV (+)	-0.337	0.714	0.126	7.17	0.007	(0.558, 0.914)
HbAS (AS)	-0.203	0.816	0.054	14.24	<0.001	(0.735, 0.907)
HbAS (SS)	-0.713	0.490	0.202	12.50	<0.001	(0.330, 0.728)
frailty (Study_No))	-	-	-	1,349.08	<0.001	-
	Model fit					
Test	**Test Stat**	**df**				
Observations	16,535					
Events	6,728					
AIC	89,823.65					
Variance of random effect	0.210					
Likelihood ratio test	2694	633				

In the malaria model, the focus was on examining the influence of various covariates on the time between the events, with the event of interest being malaria. Table 6.11 shows the covariates that are accounted for by the model. The analysis reveals that the sex variable lacks statistical significance ($p = 0.19$), implying that gender does not independently impact the risk of malaria. Age at the first visit emerges as highly significant in predicting malaria risk ($p < 0.001$), suggesting that older children exhibit a lower risk of malaria. While the cohort variable is retained in the model, it does not achieve statistical significance as a predictor of malaria risk ($p = 0.10$). The unexpected negative coefficient (-0.337, $HR = 0.714$) associated with children with HIV ($p = 0.007$) implies a lower risk of contracting malaria compared to HIV-negative children, warranting further investigation. Hemoglobin Genotypes (HbAS (AS) and HbAS (SS)) exhibit a strong association with malaria risk ($p < 0.001$ for both), indicating a reduced risk compared to those with the genotype HbAS (AA). This aligns with the well-established protective effect of the sickle cell trait against malaria. The sizable chi-square statistic (1349.08)

linked to the frailty term suggests the presence of significant unobserved factors (frailties) contributing to the variation in malaria risk among children.

In the other model to investigate the factors influencing the risk of SMA, table 6.12 shows the emerged variables and results. The coefficient (-0.762) is negative with an HR of 0.466746 and $p < 0.001$ for HbAS (AS), which suggests that children with this genotype have a 46.67% lower risk of experiencing SMA compared to those without HbAS (AA) genotype. Conversely, a positive coefficient (0.0787) is observed for the HbAS (SS) genotype ($p = 0.0145$), indicating a 119.61% higher risk of experiencing the event compared to those without the HbAS (AA) genotype. Cohort is statistically significant ($p = 0.034$), suggesting that children in Cohort 2 have a 20.43% lower risk of experiencing SMA compared to those in Cohort 1. Age continues to exhibit a similar effect as in the previous model, with older children having a reduced risk of SMA ($p < 0.001$). The negative coefficient for HIV (+) (-0.043089) with a hazard ratio of (2.302) and $p < 0.001$ implies that children with HIV face a 130.20% higher risk of experiencing the event (likely severe malaria) compared to those without HIV.

Table (6.12) Frailty model for SMA

	Estimate	**HR**	**SE**	**p**	**(95% CI)**
Age_first	-0.043	0.958	0.008	<0.001	(0.944, 0.972)
Cohort (2)	-0.229	0.796	0.108	0.034	(0.644, 0.983)
HIV (+)	0.834	2.302	0.205	<0.001	(1.540, 3.442)
HbAS (AS)	-0.762	0.467	0.197	<0.001	(0.318, 0.686)
HbAS (SS)	0.787	2.196	0.322	0.015	(1.169, 4.127)
	Model fit				
Test	**Test Stat**	**df**			
Observations	16,535				
Events	384				
AIC	5,343.97				
Concordance	0.665				
Likelihood ratio test	87.81	5			
Wald test	81.77	5			

6.4 Modeling the Hazard based on first Malaria and SMA events

This section addresses the final research question, examining how various risk factors impact the time until the occurrence of malaria or SMA events, given a malaria or SMA episode. A Cox Proportional Hazard (PH) model is employed with a gamma-distributed frailty term to consider unobserved heterogeneity among children. In the malaria model, a factor variable named First_mal is introduced, where each children's initial malaria

episode is denoted by 1, and subsequent events are marked as 0. A parallel strategy is applied to the SMA model, creating the variable first_sma. These variables serve as covariates in their respective models. The dataset encompasses $16,535$ observations, with $6,728$ events and covariates as Sex, Age_mos, First_mal, Hgb_final, HIV, ALPHATHAL, and HbAS.

The results of the first model are shown in table 6.13. Age is associated significantly with survival time, with a negative coefficient of -0.024 ($p < 0.001$), indicating that higher age is linked to a decreased hazard of malaria. Children with the first malaria episode have a substantially lower hazard ($HR = 0.204, p < 0.001$). Lower hemoglobin levels (Hgb_final) are associated with a 14.3% higher hazard ($HR = 0.857, p < 0.001$) of malaria. Children with HIV indicate a 51.7% lower hazard of malaria compared to those without HIV ($HR = 0.593, p < 0.001$). Alpha-thalassemia does not significantly impact the hazard of malaria events. Both genetic factors HbAS (AS) and HbAS (SS) also play a role and are associated with a significantly lower hazard. Children with HbAS (AS) had a 20% lower hazard ($HR = 0.800, p < 0.001$), while HbAS (SS) showed a 70.8% lower hazard ($HR = 0.292, p < 0.001$) as compared to HbAS (AA). The frailty term, representing unobserved study-specific effects, significantly contributed to the model ($p < 0.001$), emphasizing the importance of accounting for study-specific variations.

Table (6.13) Analysis of subsequent malaria event using Frailty model

	Estimate	**HR**	**SE**	**p**	**(95% CI)**
Sex (Male)	0.047	1.049	0.040	0.240	(0.967, 1.135)
Age_mos	-0.024	0.977	0.001	<0.001	(0.974, 0.980)
First_mal	-1.587	0.204	0.045	<0.001	(0.187, 0.223)
Hgb_final	-0.155	0.857	0.007	<0.001	(0.845, 0.868)
HIV (+)	-0.522	0.593	0.130	<0.001	(0.460, 0.766)
ALPHATHAL $(\alpha/\alpha\alpha)$	0.001	1.001	0.045	0.980	(0.916, 1.0894)
ALPHATHAL (α/α)	-0.123	0.884	0.055	0.025	(0.794, 0.985)
HbAS (AS)	-0.223	0.800	0.058	<0.001	(0.713, 0.897)
HbAS (SS)	-1.231	0.292	0.215	<0.001	(0.192, 0.445)
frailty (Study_No))	-	-	-	<0.001	-
	Model fit				
Test	**Test Stat**	**df**			
Observations	16,535				
Events	6728				
AIC	1,11,535.1				
Concordance	0.749				
Likelihood ratio test	4126	770			
Variance of random effect	0.306				

The second model is to investigate the factors influencing the survival time of children with SMA. The analysis, based on $16,535$ children with 384 events, yielded significant associations as indicated by table 6.14. Age, Observation time, and Hemoglobin level

Table (6.14) Analysis of subsequent SMA event using Frailty model

	Estimate	**HR**	**SE**	**p**	**(95% CI)**
Age_mos	-0.136	0.872	0.012	<0.001	(0.852, 0.893)
First_sma	1.604	4.973	0.187	<0.001	(3.444, 7.181)
Hgb_final	-0.610	0.543	0.048	<0.001	(0.494, 0.597)
	Model fit				
Test	**Test Stat**	**df**			
Observations	16,535				
Events	384				
AIC	3,997.81				
Concordance	0.958				
Wald test	776.3	3			
Score (Logrank) Test	2614	3			

exhibit similar results as in the case of malaria. Higher age ($HR = 0.872, p < 0.001$) and lower hemoglobin levels ($HR = 0.543, p < 0.001$) associate with reduced hazard. However, children with the first SMA episode (First_sma) significantly increased the hazard ($HR = 4.973, p < 0.001$). This suggest that children with a previous SMA episode have a higher risk of subsequent SMA episode.

7 Conclusion

In unraveling the intricate determinants of all-cause mortality, this comprehensive study employed a diverse set of statistical models. At the individual level, the findings highlight the importance of hemoglobin levels, genetic factors, HIV status, and age in influencing the risk of all-cause mortality in the study population. On the multiple-visit data, lower hemoglobin levels are associated with an increased hazard of mortality. The hazard of mortality for children with malaria at first visit is 0.276 times that of the reference group, indicating a lower risk. Conversely, HIV-positive children have a hazard of mortality approximately 10 times higher than those without HIV.

Factors associated with the frequency of malaria and SMA episodes were investigated on the count data. In the first part, assuming that the children were observed since birth as they were quite young at the first clinical visit. Sex, age at the first clinical visit, HIV-positive, and the HbAS genotype are associated significantly with the frequency of malaria episodes. Lower age at the first clinical visit and male children were the factors influencing the frequency of malaria episodes. On the other hand, HIV-positive and, both the HbAS genotypes HbAS (AS) and HbAS (SS) are associated with a lower count of malaria episodes. In the case of SMA episodes, age at the first clinical visit and HbAS genotypes show a similar effect as in the case of malaria episodes. However, HIV-positive children show a significantly higher count of SMA episodes. In the second part, similar models were used on the study observation level data since enrollment. Sex, age at the first clinical visit, and HIV provide similar results as in the case of the first part of malaria episodes. In addition, HbAs genotypes and Alphathalasemia are associated with a lower count of malaria episodes. For the frequency of SMA episodes, Lower age at the first clinical visit and HIV-positive indicate an increased risk of SMA episodes. The HbAS (AS) genotype is less likely to experience SMA episodes, while those with the HbAS (SS) genotype are more likely to do so.

A similar investigation as above was conducted on the multiple events per subject data to explore the survival time until the occurrence of a subsequent malaria episode. The results indicate that age at the last visit, cohort, HIV status, and HbAS genotypes as significant factors associated with the hazard of experiencing a subsequent malaria episode. Higher age at the last visit, HIV-positive and HbAS genotypes play a protective role on the hazard of malaria episodes, and children in cohort 2 have a lower hazard compared to the Cohort 1. Age, HIV, cohort, and HbAS genotypes show a similar effect in the case of SMA episodes. The following two models are similar to the above models in addition to the random effect term. The results are somewhat similar to the previous models for malaria and SMA episodes. The significant presence of unobserved heterogeneity among different children only in the case of malaria episodes.

The last analysis scrutinized the impact of various risk factors on the time until the occurrence of a second malaria and SMA episode given a first episode. Older age demonstrates a protective effect, lowering the hazard. A prior malaria event and lower hemoglobin level significantly increase the hazard, while HIV-positive status lowers it. Certain genetic factors, such as HbAS (SS), are associated with a significantly reduced hazard. Similar results are found for factors influencing the hazard of subsequent SMA episodes. Older age is associated with reduced hazard, while longer observation periods increase it. Having experienced a prior SMA episode significantly elevates the hazard, as does lower hemoglobin levels.

This extensive study on children in western Kenya illuminates key determinants of all-cause mortality, malaria, and SMA events. Individual factors, including lower hemoglobin levels, HIV status, and age, significantly influence mortality risk. Older age emerges as a protective factor against malaria, indicative of acquired immunity. Notably, the sickle cell trait exhibits a protective role against both malaria and SMA. The analysis of the frequency and survival time of malaria episodes underscores the complex interplay of age, hemoglobin, HIV status, and genetic factors. Unraveling these intricate dynamics enhances our understanding of health outcomes in the region, guiding potential interventions for improved public health.

Appendix: Implementation: R Code

Data analysis

```
# Get the required libraries
library(readr)
library(dplyr)
library(survival)
library(survminer)
library(MASS)
library(ggplot2)
library(coxme)
library(tidyverse)
library(tidyquant)
library(qwraps2)
library(gtsummary)
library(skimr)
library(xtable)
library(stargazer)
library(utile.tables)

# Load the data and fix the date format for variables
   containing date data
data_orig <-
   read_csv('C:/Users/Dell/Desktop/ThesisData.csv',
                col_types = cols (
                Adm_date = col_date(
                format = "%d/%m/%Y"),
                Results_Date =
                col_date(format = "%d/%m/%Y"),
                Date_of_Birth =
                col_date(format = "%d/%m/%Y")))

# Set the right type of variables
data <- data_orig %>% mutate(across(c(Study_No,
                          Cohort,
                          Sex,
                          Species,
                          Organism_Isolated_Final,
                          HbAS,
                          G6PD,
```

```
                              ALPHATHAL,
                              Pf_Malaria,
                              Pm_Malaria,
                              Mixed_Malaria,
                              SMA_Pf,
                              SMA_Pm,
                              SMA_Mixed,
                              Ser_Status,
                              HIV,
                              Organism_Isolated_Final,
                              Bacteremia,
                              All_Cause_Mortality,
                              Malaria_Mortality,
                              RPI_2), factor))

# Plot the missing data information
plot_missing(data_orig, group=c("Good"=1.0),
    theme_config=list(text = element_text(size = 16)))

# Selection of final data set after data analysis
main_data <- subset(data,
                    select = c(2:10,27,30:34,36:41,43))
final_data <- na.omit(main_data)

# Looking at different statistics of the data
dim(final_data)
names(final_data)
str(final_data)
glimpse(final_data)
summary(final_data)

# Creating a copy of @final_data@
no_na_data <- final_data
```

Listing (1) Data analysis: Loading libraries and data analysis

Mortality models

Cox regression model

```
# Getting only last visit data of each patient
sel <- unlist(lapply(split(no_na_data,
```

```
        no_na_data$Study_No),
        function(x) {
        a <- rep(FALSE,
        length(x$Person_mos_obs))
        a[which.max(x$Person_mos_obs)]
        <- TRUE a }))

# Creating a variable called diff.time
no_na_data$diff.time <- unlist(lapply(split(no_na_data,
                                no_na_data$Study_No),
                                function(x)
                                {a <- x$Age_mos
                                a - c(0,
                                a[-length(a)])}))

data_1 <- no_na_data[sel,] # Creating a new dataset with
   the above changes

# Fit the Cox regression model
mod1 <- coxph(Surv(diff.time,
            as.numeric(All_Cause_Mortality))
            ~ Sex + Hgb_final + HbAS + ALPHATHAL
            + HIV + Cohort + Age_last,
            data = data_1)

# Perform model selection using stepAIC
fit_mod1 <- stepAIC(mod1,
                direction = "both",
                trace= TRUE )

# Get the summary of the model
summary(fit_mod1)
```

Listing (2) Cox regression model: Data preparation and model fitting

Logistic regression model

```
# Getting first clinical visit data
first_obs_data <- no_na_data %>%
  group_by(Study_No) %>%
  slice(1)

# Create a variable for the last visit malaria status
```

```r
first_obs_data$Malaria_status_last_visit <-
   data_1$Malaria

# Fit the logistic regression model
logit_model <- glm(All_Cause_Mortality ~ Sex + Age_first
                + Malaria_status_last_visit + Hgb_final +
                HbAS + ALPHATHAL + HIV,
                data = na.omit(first_obs_data),
                family = "binomial")

# Perform model selection using stepAIC
logit_mod1 <- stepAIC(logit_model,
                direction = "both",
                trace= TRUE )

# Get the summary of the model
summary(logit_mod1)
```

Listing (3) Logistic regression model: Data preparation and model fitting

Frailty models

```r
# Fit the model
frail_mod2 <- coxph(Surv(diff.time,
                as.numeric(All_Cause_Mortality)) ~
                Sex + Age_mos + Person_mos_obs
                + Malaria + Hgb_final + HIV + HbAS
                + frailty(Study_No,
                distribution = "gamma"),
                data = na.omit(no_na_data))

# Perform model selection using stepAIC
frail_mod1_step <- stepAIC(frail_mod2,
                        direction = "both",
                        trace= TRUE )

# Get the summary of the model
summary(frail_mod1_step)
```

Listing (4) Frailty models: Fit the frailty model

Models for count data

Data preparation

```r
# Prepare a dataset since birth
bir_data <- final_data %>%
  group_by(Study_No) %>%
  summarize(
    SMA_count = sum(as.numeric(SMA == 1)),
    Malaria_count = sum(as.numeric(Malaria == 1)),
    Sex = first(Sex),
    Age_at_first_visit = first(Age_first),
    Age_at_last_visit = first(Age_last),
    Cohort = first(Cohort),
    HbAS = first(HbAS),
    Alphathal = first(ALPHATHAL),
    HIV = first(HIV),
    Person_mos_obs = first(Age_last)
  )

# Prepare a dataset since enrollment
enrl_data <- final_data %>%
  group_by(Study_No) %>%
  summarize(
    SMA_count = sum(as.numeric(SMA == 1)),
    Malaria_count = sum(as.numeric(Malaria == 1)),
    Sex = first(Sex),
    Age_at_first_visit = first(Age_first),
    Age_at_last_visit = first(Age_last),
    Cohort = first(Cohort),
    HbAS = first(HbAS),
    Alphathal = first(ALPHATHAL),
    HIV = first(HIV),
    Person_mos_obs = first(Age_delta)
  )
```

Listing (5) Data preparation: Data preparation for count data

Poisson regression models (since birth)

```r
# Model for malaria
poi_mal_mod <- glm(Malaria_count ~ Sex
```

```
                    + Age_at_first_visit + Cohort +
                    HIV + Alphathal + HbAS
                    + offset(log(Age_at_last_visit)),
                    data = na.omit(bir_data),
                    family = poisson)

# Get the summary of the model with model selection
   using stepAIC
poi_mal_mod_step <- stepAIC(poi_mal_mod,
                        direction = "both",
                        trace= TRUE )

# Get the summary of the model
summary(poi_mal_mod_step)

# Model for SMA
poi_sma_mod <- glm(SMA_count ~ Sex + Age_at_first_visit
                + Cohort + HIV + Alphathal + HbAS
                + offset(log(Age_at_last_visit)),
                data = na.omit(bir_data),
                family = poisson)

# Perform model selection using stepAIC
poi_sma_mod_step <- stepAIC(poi_sma_mod,
                        direction = "both",
                        trace= TRUE )

# Get the summary of the model
summary(poi_sma_mod_step)
```

Listing (6) Poisson regression model: Models for malaria and SMA

Poisson regression models (since enrollment)

```
# Model for malaria
poi_mal_mod2 <- glm(Malaria_count ~ Sex
                    + Age_at_first_visit
                    + Cohort + HIV + Alphathal + HbAS
                    + offset(log(Person_mos_obs)),
                    data = na.omit(Enrlm_data),
                    family = poisson)

# Perform model selection using stepAIC
```

```
poi_mal_mod2_step <- stepAIC(poi_mal_mod2,
                             direction = "both",
                             trace= TRUE )

# Get the summary of the model
summary(poi_mal_mod2_step)

# Model for SMA
poi_sma_mod2 <- glm(SMA_count ~ Sex + Age_at_first_visit
                    + Cohort + HIV + Alphathal + HbAS
                    + offset(log(Person_mos_obs)),
                    data = na.omit(Enrlm_data),
                    family = poisson)

# Perform model selection using stepAIC
poi_sma_mod2_step <- stepAIC(poi_sma_mod2,
                             direction = "both",
                             trace= TRUE )

# Get the summary of the model
summary(poi_sma_mod2_step)
```

Listing (7) Poisson regression model: Models for malaria and SMA

The Andersen-Gill models

```
# Data preparation for malaria model
sel2 <- data_event$Age_mos==data_event$Age_last |
   data_event$Malaria=="1"

ag_data_mal <- data_event[sel2,]
ag_data_mal$Prev_Event_Age <- ave(ag_data_mal$Age_mos,
   ag_data_mal$Study_No, FUN = function(x) c(0,
   x[-length(x)]))

# Data preparation for the SMA model
sel3 <- data_event$Age_mos==data_event$Age_last |
   data_event$SMA=="1"
ag_data_sma <- data_event[sel3,]
ag_data_sma$Prev_Event_Age <- ave(ag_data_sma$Age_mos,
   ag_data_sma$Study_No, FUN = function(x) c(0,
   x[-length(x)]))
```

```r
## Model for malaria
ag_mod_mal <- coxph(Surv(Prev_Event_Age, Age_mos,
                  as.numeric(Malaria))~ Sex
                  + Age_last + Cohort + HIV
                  + HbAS, id = Study_No,
                  data = na.omit(ag_data_mal))

# Perform model selection using stepAIC
ag_mod_mal_step <- stepAIC(ag_mod_mal,
                      direction = "both",
                      trace= TRUE )

# Get the summary of the model
summary(ag_mod_mal_step)

## Model for SMA
ag_mod_sma <- coxph(Surv(Prev_Event_Age, Age_mos,
                  as.numeric(SMA)) ~ Sex
                  + Age_first + Cohort + HIV
                  + HbAS, id = Study_No,
                  data = na.omit(ag_data_sma))

# Perform model selection using stepAIC
ag_mod_sma_step <- stepAIC(ag_mod_sma,
                      direction = "both",
                      trace= TRUE )

# Get the summary of the model
summary(ag_mod_sma_step)
```

Listing (8) The Andersen-Gill models: Data preparation and model fitting

The Frailty models

```r
# Data preparation for the model
data_frail <- final_data
data_frail$Study_No <- droplevels(data_frail$Study_No)
data_frail$Prev_Event_Age <- unlist(lapply(split
                                  (data_frail$Age_mos,
                                  data_frail$Study_No),
                                  function(x)
                                  {(c(0,
```

```
                                        x))[1:length(x)]}))

# Model for malaria
frail_mal_mod <- coxph(Surv(Prev_Event_Age, Age_mos,
                     as.numeric(Malaria)) ~ Sex
                     + Age_first + Cohort + HIV
                     + HbAS + frailty(Study_No,
                     distribution = "gamma"),
                     data = na.omit(data_frail))

# Perform model selection using stepAIC
frail_mal_mod_step <- stepAIC(frail_mal_mod,
                          direction = "both",
                          trace= TRUE )

# Get the summary of the model
summary(frail_mal_mod_step)

# Model for SMA
frail_sma_mod <- coxph(Surv(Prev_Event_Age, Age_mos,
                     as.numeric(SMA)) ~ Sex
                     + Age_first + Cohort + HIV
                     + HbAS + frailty(Study_No,
                     distribution = "gamma"),
                     data = na.omit(data_frail))

# Perform model selection using stepAIC
frail_sma_mod_step <- summary(stepAIC(frail_sma_mod,
                          direction = "both",
                          trace= TRUE ))

# Get the summary of the model
summary(frail_sma_mod_step)
```

Listing (9) The Frailty models: Data preparation and model fitting

The Frailty models

```
# Data preparation
Frlty_data <- no_na_data
                %>% group_by(Study_No) %>%
                        mutate(First_mal =
```

```
                          ifelse(sum(Malaria == 1) > 0
                          & row_number() ==
                          which(Malaria == 1)[1], 1, 0),
                          First_sma =
                          ifelse(sum(SMA == 1) > 0
                          & row_number() ==
                          which(SMA == 1)[1], 1, 0))

# Model for malaria
frail_mal1 <- coxph(Surv(diff.time, as.numeric(Malaria))
                  ~  Sex + Age_mos + Person_mos_obs
                  + First_mal + Hgb_final + HIV
                  + ALPHATHAL + HbAS + frailty(Study_No,
                  distribution = "gamma"),
                  data = na.omit(Frlty_data))

# Perform model selection using stepAIC
frail_mal1_step <- summary(stepAIC(frail_mal1,
                      direction = "both",
                      trace= TRUE ))

# Get the summary of the model
summary(frail_mal1_step)

# Model for SMA
frail_sma1 <- coxph(Surv(diff.time, as.numeric(SMA))
                  ~  Sex + Age_mos + Person_mos_obs
                  + First_sma + Hgb_final + HIV
                  + ALPHATHAL + HbAS + frailty(Study_No,
                  distribution = "gamma"),
                  data = na.omit(Frlty_data))

# Perform model selection using stepAIC
frail_sma1_step <- summary(stepAIC(frail_sma1,
                      direction = "both",
                      trace= TRUE ))

# Get the summary of the model
summary(frail_sma1_step)
```

Listing (10) The Frailty models: Data preparation and model fitting

Appendix: Mathematical Symbols

IID	Independent and identically distributed.
PDF	Probability density function.
CDF	Cumulative distribution function.
E	Expectation (Mean).
Var	Variance.
LS	Least square.
L	Maximum likelihood.
l	Log-likelihood.
C_r	Right censoring.
C_l	Left censoring.
$S(t)$	Survivor function.
$h(t)$	Hazard function.
$H(t)$	Cumulative Hazard function.
$F(t)$	Failure function.
HR	Hazard Ratio.
IRR	Incident risk ratio interval
PL	Partial likelihood.
$h_0(t)$	Baseline hazard.
Z	Frailty.
$\exp$	Exponential.
$\mathscr{L}$	Laplace function.
CV	Variation coefficient.
$\overline{S}(t)$	Marginal survival function.
$\overline{h}(t)$	Marginal hazard function.

Bibliography

[1] World Health Organization et al. *World malaria report 2022*. World Health Organization, 2022.

[2] Lily E Kisia, Qiuying Cheng, Evans Raballah, Elly O Munde, Benjamin H McMahon, Nick W Hengartner, John M Ong'echa, Kiprotich Chelimo, Christophe G Lambert, Collins Ouma, et al. Genetic variation in csf2 (5q31. 1) is associated with longitudinal susceptibility to pediatric malaria, severe malarial anemia, and all-cause mortality in a high-burden malaria and hiv region of kenya. *Tropical Medicine and Health*, 50(1):1–15, 2022.

[3] Evans Raballah, Samuel B Anyona, Qiuying Cheng, Elly O Munde, Ivy-Foo Hurwitz, Clinton Onyango, Caroline Ndege, Nicolas W Hengartner, Maria Andreína Pacheco, Ananias A Escalante, et al. Complement component 3 mutations alter the longitudinal risk of pediatric malaria and severe malarial anemia. *Experimental Biology and Medicine*, 247(8):672–682, 2022.

[4] Evans Raballah, Kristen Wilding, Samuel B Anyona, Elly O Munde, Ivy Hurwitz, Clinton O Onyango, Cyrus Ayieko, Christophe G Lambert, Kristan A Schneider, Philip D Seidenberg, et al. Nonsynonymous amino acid changes in the α-chain of complement component 5 influence longitudinal susceptibility to plasmodium falciparum infections and severe malarial anemia in kenyan children. *Frontiers in Genetics*, page 2567, 2022.

[5] Walters M Essendi, Anne M Vardo-Zalik, Eugenia Lo, Maxwell G Machani, Guofa Zhou, Andrew K Githeko, Guiyun Yan, and Yaw A Afrane. Epidemiological risk factors for clinical malaria infection in the highlands of western kenya. *Malaria journal*, 18(1):1–7, 2019.

[6] John M. Marshall Prathiba M. De Silva. Factors contributing to urban malaria transmission in sub-saharan africa: A systematic review. *Journal of Tropical Medicine*, 2012(4):10, 2012.

[7] Beatrice Autino, Alice Noris, Rosario Russo, and Francesco Castelli. Epidemiology of malaria in endemic areas. *Mediterranean journal of hematology and infectious diseases*, 4:e2012060, 01 2012.

[8] Nicholas J White, Sasithon Pukrittayakamee, Tran Tinh Hien, M Abul Faiz, Olugbenga A Mokuolu, and Arjen M Dondorp. Malaria. *The Lancet*, 383(9918):723–735, 2 2014.

[9] Carlos A Guerra, Priscilla W Gikandi, Andrew J Tatem, Abdisalan M Noor, Dave L Smith, Simon I Hay, and Robert W Snow. The limits and intensity of plasmodium falciparum transmission: implications for malaria control and elimination worldwide. *PLoS medicine*, 5(2):e38, 2008.

[10] Elizabeth A Ashley, Aung Pyae Phyo, and Charles J Woodrow. Malaria. *The Lancet*, 391(10130):1608–1621, 2018.

[11] Jasminka Talapko, Ivana Škrlec, Tamara Alebić, Melita Jukić, and Aleksandar Včev. Malaria: the past and the present. *Microorganisms*, 7(6):179, 2019.

[12] Noppadon Tangpukdee, Chatnapa Duangdee, Polrat Wilairatana, and Srivicha Krudsood. Malaria diagnosis: a brief review. *The Korean journal of parasitology*, 47(2):93, 2009.

[13] A. Bartoloni and L Zammarchi. Clinical aspects of uncomplicated and severe malaria. *Mediterranean journal of hematology and infectious diseases*, 4(1):e2012026, 01 2012.

[14] Yulianti Paula Bria, Chung-Hsing Yeh, and Susan Bedingfield. Significant symptoms and nonsymptom-related factors for malaria diagnosis in endemic regions of indonesia. *International Journal of Infectious Diseases*, 103:194–200, 2021.

[15] Roll Back Malaria et al. World malaria report 2005. *World Health Organization and UNICEF*, 2005.

[16] WHO. World malaria report 2020: 20 years of global progress and challenges. *World malaria report 2020: 20 years of global progress and challenges*, page 299, 2020.

[17] Dejen Nureye. History, life cycle, diagnosis and prevention of malaria: Introductory concepts and new advances. *parasite*, page 14, 1897.

[18] Renu Tuteja. Malaria: an overview. *The FEBS Journal*, 274(18):4670–4679, 2007.

[19] Alan F Cowman, Julie Healer, Danushka Marapana, and Kevin Marsh. Malaria: biology and disease. *Cell*, 167(3):610–624, 2016.

[20] Charles O Obonyo, John Vulule, Willis S Akhwale, and Diederick E Grobbee. In-hospital morbidity and mortality due to severe malarial anemia in western kenya. In *Defining and Defeating the Intolerable Burden of Malaria III: Progress and Perspectives: Supplement to Volume 77 (6) of American Journal of Tropical Medicine and Hygiene*. American Society of Tropical Medicine and Hygiene, 2007.

[21] Jane Crawley, Cindy Chu, George Mtove, and François Nosten. Malaria in children. *The Lancet*, 375(9724):1468–1481, 4 2010.

[22] Rehema H Simbauranga, Erasmus Kamugisha, Adolfine Hokororo, Benson R Kidenya, and Julie Makani. Prevalence and factors associated with severe anaemia amongst under-five children hospitalized at bugando medical centre, mwanza, tanzania. *BMC hematology*, 15(1):1–9, 2015.

[23] Nicholas J. White. Anaemia and malaria. *Malaria Journal*, 17(1), oct 19 2018.

[24] Nicholas J. White. Severe malaria. *Malaria Journal*, 21(1), oct 6 2022.

[25] World Health Organization. *Guidelines for the treatment of malaria*. World Health Organization, 3rd ed edition, 2015.

[26] Irene Ule Ngole Sumbele, Sharon Odmia Sama, Helen Kuokuo Kimbi, and Germain Sotoing Taiwe. Malaria, moderate to severe anaemia, and malarial anaemia in children at presentation to hospital in the mount cameroon area: a cross-sectional study. *Anemia*, 2016, 2016.

[27] Richard-Fabian Schumacher and Elena Spinelli. Malaria IN CHILDREN. *Mediterranean Journal of Hematology and Infectious Diseases*, 4(1):e2012073, nov 7 2012.

[28] WHO A Practical Handbook. Management of severe malaria. *WHO*, 2012.

[29] Walter T Ambrosius. *Topics in biostatistics*. Springer, 2007.

[30] Evangelos C Alexopoulos. Introduction to multivariate regression analysis. *Hippokratia*, 14(Suppl 1):23, 2010.

[31] Ludwig Fahrmeir, Thomas Kneib, Stefan Lang, and Brian D Marx. Regression: Models, methods and applications. In *Regression: Models, methods and applications*, pages 23–84. Springer, 2022.

[32] Eric Vittinghoff, David V Glidden, Stephen Shiboski, and Charles E McCulloch. Regression methods in biostatistics: linear, logistic, survival, and repeated measures models. In *Regression methods in biostatistics: linear, logistic, survival, and repeated measures models*, pages 509–509. 2012.

[33] Brandon George, Samantha Seals, and Inmaculada Aban. Survival analysis and regression models. *Journal of nuclear cardiology*, 21(4):686–694, 2014.

[34] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[35] Todd G Nick and Kathleen M Campbell. Logistic regression. *Topics in biostatistics*, pages 273–301, 2007.

[36] David G Kleinbaum, K Dietz, M Gail, Mitchel Klein, and Mitchell Klein. *Logistic regression*. Springer, 2002.

[37] Gerda Claeskens, Nils Lid Hjort, et al. Model selection and model averaging. *Cambridge Books*, 2008.

[38] Andrew T. Tredennick, Giles Hooker, Stephen P. Ellner, and Peter B. Adler. A practical guide to selecting models for exploration, inference, and prediction in ecology. *Ecology*, 102(6), May 2021.

[39] Jie Ding, Vahid Tarokh, and Yuhong Yang. Model selection techniques: An overview. *IEEE Signal Processing Magazine*, 35(6):16–34, 2018.

[40] Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

[41] Jerald B Johnson and Kristian S Omland. Model selection in ecology and evolution. *Trends in ecology & evolution*, 19(2):101–108, 2004.

[42] Walter Zucchini. An introduction to model selection. *Journal of mathematical psychology*, 44(1):41–61, 2000.

[43] Petre Stoica and Yngve Selen. Model-order selection: a review of information criterion rules. *IEEE Signal Processing Magazine*, 21(4):36–47, 2004.

[44] John P Klein and Melvin L Moeschberger. *Survival analysis: techniques for censored and truncated data*, volume 1230. Springer.

[45] David G. Kleinbaum. *Survival Analysis*. Springer New York, 1996.

[46] Stephen P Jenkins. *Survival analysis*, volume 42. Citeseer, 2005.

[47] Taane G Clark, Michael J Bradburn, Sharon B Love, and Douglas G Altman. Survival analysis part i: basic concepts and first analyses. *British journal of cancer*, 89(2):232–238, 2003.

[48] David G. Kleinbaum and Mitchel Klein. Survival analysis. 2005.

[49] Christiana Kartsonaki. Survival analysis. *Diagnostic Histopathology*, 22(7):263–270, 2016.

[50] Margaret Wahutu, Sara K Vesely, Janis Campbell, Anne Pate, Alicia L Salvatore, and Amanda E Janitz. Pancreatic cancer: a survival analysis study in oklahoma. *The Journal of the Oklahoma State Medical Association*, 109(7-8):391, 2016.

[51] Christine Wohlfahrt-Veje, Annette Mouritsen, Casper P Hagen, Jeanette Tinggaard, Mikkel Grunnet Mieritz, Malene Boas, Jørgen Holm Petersen, Niels E Skakkebæk, and Katharina M Main. Pubertal onset in boys and girls is influenced by pubertal timing of both parents. *The Journal of Clinical Endocrinology & Metabolism*, 101(7):2667–2674, 2016.

[52] JD Kalbfleisch and Jerald Franklin Lawless. Estimating the incubation time distribution and expected number of cases of transfusion-associated acquired immune deficiency syndrome. *Transfusion*, 29(8):672–676, 1989.

[53] David G. Kleinbaum and Mitchel Klein. *Survival Analysis*. Springer New York, 2012.

[54] Frank Emmert-Streib and Matthias Dehmer. Introduction to survival analysis in practice. *Machine Learning and Knowledge Extraction*, 1(3):1013–1038, 2019.

[55] Nanami Taketomi, Kazuki Yamamoto, Christophe Chesneau, and Takeshi Emura. Parametric distributions for survival and reliability analyses, a review and historical sketch. *Mathematics*, 10(20):3907, 2022.

[56] William R Lane, Stephen W Looney, and James W Wansley. An application of the cox proportional hazards model to bank failure. *Journal of Banking & Finance*, 10(4):511–531, 1986.

[57] Ritesh Singh and Keshab Mukhopadhyay. Survival analysis in clinical trials: Basics and must know areas. *Perspectives in clinical research*, 2(4):145, 2011.

[58] Mike J Bradburn, Taane G Clark, Sharon B Love, and Douglas Graham Altman. Survival analysis part ii: multivariate data analysis–an introduction to concepts and methods. *British journal of cancer*, 89(3):431–436, 2003.

[59] David D Hanagal. *Modeling survival data using frailty models*. Springer, 2011.

[60] Norman Breslow. Covariance analysis of censored survival data. *Biometrics*, pages 89–99, 1974.

[61] Bradley Efron. The efficiency of cox's likelihood function for censored data. *Journal of the American statistical Association*, 72(359):557–565, 1977.

[62] David Schoenfeld. Partial residuals for the proportional hazards regression model. *Biometrika*, 69(1):239–241, 1982.

[63] Frank E Harrell et al. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*, volume 608. Springer, 2001.

[64] Leila DAF Amorim and Jianwen Cai. Modelling recurrent events: a tutorial for analysis in epidemiology. *International journal of epidemiology*, 44(1):324–333, 2015.

[65] Terry M. Therneau and Patricia M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer New York, 2000.

[66] Frank E Harrell. Regression modeling strategies. *Bios*, 330(2018):14, 2017.

[67] Mani Thenmozhi, Visalakshi Jeyaseelan, Lakshmanan Jeyaseelan, Rita Isaac, and Rupa Vedantam. Survival analysis in longitudinal studies for recurrent events: applications and challenges. *Clinical Epidemiology and Global Health*, 7(2):253–260, 2019.

[68] Theodor A Balan and Hein Putter. A tutorial on frailty models. *Statistical methods in medical research*, 29(11):3424–3454, 2020.

[69] Andreas Wienke. *Frailty models in survival analysis*. CRC press, 2010.

[70] Odd O Aalen. Heterogeneity in survival analysis. *Statistics in medicine*, 7(11):1121–1137, 1988.

[71] Luc Duchateau and Paul Janssen. *The frailty model*. Springer, 2008.

[72] Andreas Wienke et al. Frailty models. *Rostock, Germany Max Planck institute for demographic research*, 2003.

[73] Philip Hougaard. Frailty models for survival data. *Lifetime data analysis*, 1:255–273, 1995.

[74] JOHN MICHAEL ONG'ECHA, Christopher C Keller, Tom Were, Collins Ouma, Richard O Otieno, Zachary Landis-Lewis, Daniel Ochiel, Jamie L Slingluff, Stephen Mogere, George A Ogonji, et al. Parasitemia, anemia, and malarial anemia in infants and young children in a rural holoendemic plasmodium falciparum transmission area. 2006.

Erklärung

Hiermit erkläre ich, dass ich meine Arbeit selbstständig verfasst, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und die Arbeit noch nicht anderweitig für Prüfungszwecke vorgelegt habe.

Stellen, die wörtlich oder sinngemäß aus Quellen entnommen wurden, sind als solche kenntlich gemacht.

Mittweida, 01.12.2023