HOCHSCHULE MITTWEIDA
University of Applied Sciences

MASTER THESIS

Mr.
Arjit Basu, M.Sc.

Exploration of immune responses and reticulocyte production in children with malaria

Mittweida, 30/11/2023

Faculty of **Applied Computer Sciences and Biosciences**

MASTER THESIS

Exploration of immune responses and reticulocyte production in children with malaria

Author:
Arjit Basu

Course of Study:
Applied Mathematics in Network and data science

Seminar Group:
MA18w1-M

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

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

Submission:
Mittweida, 00.00.30/11/2023

Defense/Evaluation:
Mittweida, 1/2024

Bibliographic Description:

Basu, Arjit:
Exploration of immune responses and reticulocyte production in children with malaria. – 30/11/2023. – 44 S.
Mittweida, Hochschule Mittweida – University of Applied Sciences, Faculty of Applied Computer Sciences and Biosciences, Master Thesis, 1/2024.

Abstract:

This thesis comprehensively explores factors contributing to malaria-induced anemia and severe malarial anemia (SMA). The study utilizes a comprehensive dataset to investigate immunological interactions, genetic variations, and temporal dynamics. Findings highlight the complex interplay between immune markers, genetic traits, and cohort-specific influences. Notably, age, HIV status, and genetic variations emerge as crucial factors influencing anemia risk. The incorporation of Poisson regression models sheds light on the genetic underpinnings of SMA, emphasizing the need for personalized interventions. Overall, this research provides valuable insights into the multifaceted nature of malaria-induced complications, paving the way for further molecular investigations and targeted interventions.

Contents

List of Figures and Tables

List of Figures

List of Tables

Acknowledgment

I would like to express my sincere gratitude to my thesis advisors, **Prof. Dr. rer. nat. habil. Kristan Schneider** and **Prof. Dr. rer. nat. habil. Thomas Kalinowski**, for their invaluable guidance, support, and unwavering commitment throughout the research process. Their expertise and constructive feedback significantly contributed to developing and refining this thesis.
My appreciation extends to my family and friends for their understanding, encouragement, and patience during the demanding phases of this academic journey.
Lastly, I want to express my gratitude to all the participants and sources that contributed to the data collection process.
This thesis would not have been possible without the collective support and encouragement of these individuals and organizations. Thank you for being an integral part of this academic endeavor.

1 Introduction

1.1 Global Prevalence of Malaria

Malaria, a parasitic disease transmitted through the bite of infected female Anopheles mosquitoes, prevails in tropical and subtropical regions. According to the latest World Health Organization (WHO) malaria report, 2021 saw 247 million cases in 84 countries affected by malaria. Tragically, there were 619,000 malaria-related deaths in 2021, a slight decrease from 625,000 in 2020 [1]. The vast majority of malaria cases (95%) and deaths (96%) occur in sub-Saharan Africa.

Malaria has been particularly devastating for African children, with a relentless toll since the year 2000. In the WHO African region, children under the age of 5 account for a staggering 78.9% of all malaria-related deaths. It's worth noting that some deaths might have gone unrecorded [1].

From 2000 to 2019, the incidence of malaria cases decreased from 372.6 to 225.5 cases per 1,000 population at risk. However, due to disruptions in healthcare services caused by the COVID-19 pandemic, malaria case incidence increased to 233.6 per 1,000 population at risk in 2020 but subsequently declined to 229.4 in 2021 [1].

In the WHO South-East Asia Region, there were 5.4 million malaria cases, contributing 2% of the global malaria burden. India alone accounted for about 79% of all malaria cases in the region [1].

Out of the 120 Plasmodium species that infect reptiles, mammals, and humans, only six are known to frequently infect humans. Among these, *Plasmodium falciparum* is the most deadly, causing high levels of blood-stage parasites that can affect critical organs across all age groups. This is a leading cause of severe anemia in sub-Saharan African children, where the majority of malaria deaths occur [2].

Plasmodium vivax, while generally causing milder malaria, can still lead to severe and recurrent episodes with significant associated morbidity. *Plasmodium ovale curtisi*, *Plasmodium ovale wallikeri*, and *Plasmodium malariae* are less studied, but share similarities with *Plasmodium vivax* in terms of illness severity. *Plasmodium knowlesi*, primarily found in Southeast Asia, is zoonotic and can result in severe malaria [2].

All these Plasmodium species can be distinguished through microscopic examination of stained blood smears. The incidence of malaria infection is influenced by environmental factors such as altitude, climate, vegetation, and the implementation of control measures. Poverty, natural disasters, and conflicts are often linked to higher transmission rates. Less common modes of transmission include from mother to child or via blood transfusions, although the latter is rare in settings with proper blood donor screening protocols [2].

Climate change poses a significant risk, as it can expand the range of malaria in tropical highland areas. Consequently, various Plasmodium species are found in different regions of the world [2].

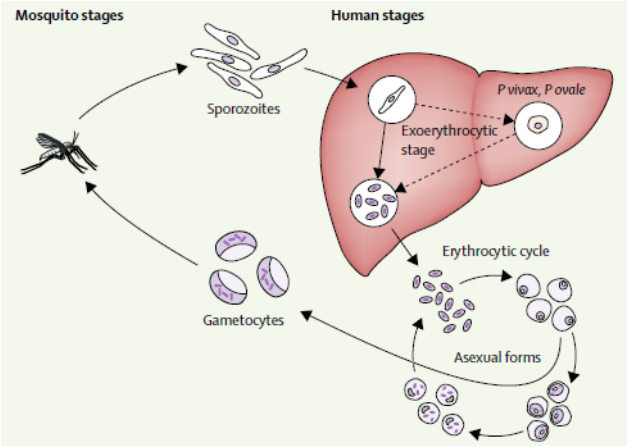

Figure 1.1: Human stages of the malaria lifecycle.

The human phase of the malaria life cycle is depicted in Figure 1.1. Sporozoites are introduced into the human host through the bite of an infected female anopheles mosquito. These parasites then undergo an initial stage within the liver, which typically lasts for 1 to 2 weeks before transitioning into the blood stage. During this blood stage, they undergo multiple rounds of asexual replication, leading to an increase in parasite numbers and the onset of malaria symptoms in the human host [2].

Within this population of blood-stage parasites, some undergo a switch to sexual development, giving rise to both female and male gametocytes. These unique transitional stages are responsible for transmitting malaria back to the mosquito during a blood meal.

In the mosquito's mid-gut, male gametocytes undergo exflagellation, and the resulting male and female gametes fuse to form a zygote. This zygote then transforms into a mobile ookinete and passes through the mosquito's gut wall. Subsequently, the oocyst releases sporozoites, which migrate to the mosquito's salivary glands, completing the full lifecycle of the malaria parasite. In *vivax* and *ovale* infections, a proportion of sporozoites become dormant hypnozoites, leading to relapses months or even years after the initial infection [2].

1.2 Inflammatory response to malaria

Once the erythrocytic cycle produces a parasitemia above a certain threshold, roughly 100 parasites per μL, malaria symptoms start to manifest. For *Plasmodium falciparum*, an incubation period of 10–14 days is typical, while other infecting species may exhibit varying incubation periods [2].

A classical malaria attack unfolds in three distinct stages: the cold stage, the hot stage, and the sweating stage. The clinical presentation of malaria encompasses a wide range of non-specific signs and symptoms, including fever, chills, headache, nausea, vomiting, muscle aches, joint pain, and jaundice. In malaria, the presence of parasitic antigens, in conjunction with various host cellular factors, triggers the release of cytokines from inflammatory cells such as macrophages, neutrophils, and endothelial cells. Elevated levels of cytokines are associated with conditions like anemia, liver dysfunction, and fever, while also contributing to the control of the parasite. As a result, cytokines play a crucial role in the pathogenesis of malaria. Innate immunity serves as the first line of defense against malaria. Inflammatory cells recognize *Plasmodium* PAMPs (pathogen-associated molecular patterns) such as glycosylphosphatidylinositol (GPI), hemozoin, and DNA via pattern recognition

receptors (PRRs). GPI, for instance, is the first molecular compound identified as a PAMP for the malaria parasite. It stimulates the release of tumor necrosis factor-alpha (TNF-alpha) and interleukin-1 (IL-1) while increasing the expression of nitric oxide synthase [3].

Following the introduction of *Plasmodium* sporozoites into the human body through mosquito bites, these sporozoites interact with three primary cell populations: CD11c+ antigen-presenting cells, hepatocytes, and Kupffer cells. Within hepatocytes, the parasite undergoes alterations in its antigenic structure. However, it remains unclear which antigens are responsible for activating cytotoxic T cells (CD8+ T cells) during this stage of infection. Protective cytotoxic T cells appear to become activated not only against *Plasmodium* antigens but also against those that are typically expressed only after hepatocyte infection. These cells trigger the production of pro-inflammatory cytokines like interferon-gamma (IFN-gamma) and TNF-alpha, which are critical for generating nitric oxide, ultimately leading to the parasite's demise within the hepatocyte. Cytotoxic T cells play a central role in the immune response against the pre-erythrocytic stage of the infection [3]. Further investigation of the inflammatory response will be conducted in the future sections of this thesis.

1.3 *Plasmodium falciparum* Malaria-Induced Anemia

Plasmodium falciparum, the most perilous among its counterparts, can induce a wide array of pathophysiological disturbances, resulting in multiple organ involvement and systemic disorders in African children. The spectrum of infections ranges from asymptomatic cases to more general malaria symptoms, such as fever and chills, which can escalate into severe, life-threatening complications. These complications include hyperparasitemia, hypoglycemia, hyperlactatemia, kidney failure, metabolic acidosis, cerebral malaria, severe anemia, and even respiratory distress [4].

The age of the host plays a pivotal role in the clinical course of the infection. For instance, children typically exhibit signs of severe anemia and hypoglycemia, while non-resident, malaria-naïve adults are susceptible to jaundice and can progress to renal failure and respiratory distress due to pulmonary edema [4]. Children under 5 years of age are one of the most vulnerable groups affected by malaria, and their morbidity increases with repeated malaria infections. A concerning statistic is that 2% of children who recover from cerebral malaria develop learning impairments and disabilities due to brain damage caused by the infection. The lack of immunity in children makes them susceptible to developing severe malaria, particularly cerebral malaria. Malaria during pregnancy can lead to low birth weight and an increased risk of mortality in the first month after birth. Repeated infections can also lead to complications such as severe anemia.

According to the World Health Organization (WHO), anemia is a condition in which Red Blood Cells (RBCs or erythrocytes) are unable to provide sufficient oxygen to the body tissues. WHO also defines anemia pathologically in children under 5 years of age as having hemoglobin (Hgb) levels less than $11g/dL$ [5]. Anemia is not a standalone diagnosis but a presentation of an underlying condition, in this case, malaria. In areas with holoendemic *Plasmodium falciparum* transmission, malarial anemia is the leading cause of morbidity and mortality. Severe anemia induced by malaria (SMA) can be defined as a hemoglobin (Hgb) concentration less than $5g/dL$ along with a parasite count greater than $10,000/\mu L$, distinguishing it from other diseases with similar presentations [6].

Multiple factors underlie the development of SMA, including both direct and indirect destruction of parasitized and non-parasitized RBCs. The *Plasmodium falciparum* parasite proliferates by infecting RBCs and progressing through various stages of its cycle. When the parasite population reaches an unsustainable level, it ruptures the cell membrane to transmit and infect new RBCs. This process leads to the destruction of both newly infected RBCs and non-parasitized cells. Both types of cells display parasite antigens on their surface, which deform and alter the cell membranes, causing premature phagocytosis and destruction by the reticuloendothelial system. When RBCs burst, they release parasite and hemoglobin waste products. Activated leukocytes take up the hemoglobin waste, known as hemozoin, which is a polymerized heme. This stimulates the innate immune system, leading to the synthesis and secretion of inflammatory cytokines, chemokines, growth factors, and mediators.

The destruction of RBCs in malaria, along with an imbalance between pro- and anti-inflammatory events, leads to the modification of erythroid cell proliferation, resulting in SMA and other malaria-related pathophysiological changes [6]. It's worth noting that defining malaria-attributable anemia can be challenging in some cases, as numerous prevalent comorbidities may also lead to dyserythropoiesis and inflammation-induced functional iron deficiency. The bone marrow's ability to respond to decreasing Hb levels can, however, be assessed by analyzing reticulocyte counts [7].

1.4 Introduction to RPI index

Reticulocytes are immature red blood cells (RBCs) that originate in the bone marrow, specifically from orthochromatic normoblasts through a process known as nuclear exclusion. These immature cells are subsequently released into the peripheral blood after undergoing a maturation period within the bone marrow. As they circulate in the bloodstream, reticulocytes continue to differentiate until they become fully mature RBCs [8].

The measurement of reticulocytes in the peripheral blood, often referred to as "reticulocyte counting", is a commonly performed test and serves as a valuable indicator of the functional status of the bone marrow. Reticulocyte counting provides insights into several aspects, including the activity of erythropoiesis (the production of RBCs) within the bone marrow, the rate at which reticulocytes are released from the bone marrow into the peripheral blood, and the rate at which these reticulocytes mature into fully functional RBCs [8].

In cases of anemia, characterized by a reduced number of RBCs, the reticulocyte count can be informative. Anemic patients with properly functioning bone marrow typically exhibit reticulocytosis, meaning an increased number of reticulocytes in the peripheral blood. Conversely, anemic patients with bone marrow dysfunction produce fewer reticulocytes, resulting in a decreased number of reticulocytes in their peripheral blood, a condition known as reticulocytopenia [8].

The reticulocyte count is usually reported as a percentage, with the normal mean percentage reticulocyte count by NMB light microscopy falling within the range of 1.0% to 1.5%, and 3% being the upper limit of normal. The relative reticulocyte count may be misleading when the RBC count is abnormal and/or erythropoietic stimulation to the bone marrow is occurring, such as in cases of severe anemia. To address this, a Packed cell volume (PCV) correction, known as the reticulocyte index, is applied to specimens to compensate for the decrease in mature RBCs [8].

$$\text{Reticulocyte Index} = \text{Reticulocyte count(\%)} \times \frac{\text{patient pcv}}{0.45}$$

A shift correction is also made which gives us the corrected count called Reticulocyte production (maturation) index (RPI) with the following formula.

$$\text{Reticulocyte Production Index} = \frac{\text{Reticulocyte index}}{\text{maturation time in peripheral blood}}$$

It's essential to note that only patients with an intact hematopoietic system will provide accurate results for the relationship between hematocrit and the reticulocyte maturation time. Impaired erythropoietin production or compromised bone marrow can lead to a low reticulocyte count, resulting in an incorrect RPI [8].

1.5 Structure and scope of the thesis

Our thesis focuses on a comprehensive investigation into the dynamics of immune responses and reticulocyte production in children affected by malaria. The overarching goal is to uncover intricate connections between immune markers, anemia, and malaria, utilizing a diverse dataset that encompasses blood count, antibody levels, and genetic information. The primary objective is to understand the nuanced relationships between anemia and immune markers, considering both their levels and genetic variations. Our analytical approach spans a range of statistical techniques, with a strong emphasis on R programming for implementation.

Our specific aims include conducting descriptive analyses to elucidate the relationships between immune markers, anemia, and malaria through correlation assessments, visualizations, and cross-tabulations. Logistic regression is employed to identify relevant factors contributing to anemia events, with a focus on model selection based on AIC (Akaike information criterion). Additionally, linear regression is applied to assess immune marker levels in relation to various identified covariates. Longitudinal analysis is conducted using Poisson rate regression to understand the temporal patterns of anemia events.

To ensure the robustness of our findings, statistical tests incorporate Welch correction where necessary, and model selection is performed meticulously. By addressing these specific aims, our thesis aims to contribute valuable insights into the complex interplay between immune responses, anemia, and malaria in pediatric populations. This research has the potential to advance our understanding of disease mechanisms and may inform public health interventions.

Chapter 2 provides an in-depth introduction to our methodology, delving into the intricacies of regression models and various aspects of survival analysis. This chapter serves as the foundation for understanding the analytical frameworks employed throughout our research.

In **Chapter 3**, we present the interpretation of both regression and survival models, offering insightful discussions on the outcomes derived from our cross-sectional and longitudinal analyses. This section provides a comprehensive understanding of the relationships uncovered between immune responses, anemia, and malaria in children.

The culmination of our research findings and insights is encapsulated in **Chapter 4**. Here, we draw conclusions from our analyses, highlighting key patterns and significant observations. Additionally, we outline avenues for future research, identifying areas that warrant further exploration to advance our understanding and contribute to the broader scientific discourse on this subject.

These chapters collectively form a cohesive narrative that navigates through our methodology, results interpretation, and the implications of our findings. The structure guides the reader seamlessly from the foundational aspects of our approach to the conclusive insights drawn from our comprehensive analyses.

2 Introduction to Methodology

Multi-variable analysis employs several variables to predict a single outcome. It explores relationships between two or more independent variables and a single dependent variable, finding applications in diverse fields like medicine, epidemiology, and pharmaceutical research, primarily serving prognosis and diagnosis purposes [9].

Various multi-variable methods, such as linear regression, logistic regression, and discriminant analysis, play crucial roles in health science. Linear regression, for instance, assesses the relationship between independent and dependent variables, making it valuable for predicting outcomes based on variable interactions. Logistic regression, on the other hand, is employed when the outcome is binary, such as disease presence or absence [9, 10].

Survival analysis, a statistical method investigating the time until a specific event occurs, focuses on events like death or disease onset. Whether it's a child succumbing to illness within a month or after 18 months, understanding contributing factors is vital for prediction and prevention [11]

The core objective of survival analysis is to assess how explanatory variables, covariates, or independent variables, influence "survival time". These covariates, such as treatment received or tumor size, help examine factors impacting the duration of survival. Another perspective frames survival analysis in terms of risk, determining the likelihood of an event, like cancer relapse, within a specific time unit for an individual in remission [11, 12]

In parallel, regression analysis, including linear and logistic regression, widens the analytical scope. Linear regression facilitates the understanding of relationships between variables, predicting outcomes based on variable interactions. Logistic regression, suited for binary outcomes, provides valuable insights when dealing with dichotomous events, such as disease presence or absence. The integration of regression analysis with survival analysis enriches the analytical toolkit, offering a comprehensive approach to exploring relationships and predicting outcomes in complex datasets [10, 13, 12].

2.1 Linear Regression

In this section, we delve into a broader category of regression models. These models are expressed as a weighted sum of independent or predictor variables, making them more versatile. The key idea here is to linearize the model concerning the predictor variables, which allows us to interpret the parameters in these regression models more easily.

Let's establish a notation for these regression models:

- Y represents the response variable, often referred to as the dependent variable.
- $\mathbf{X} = X_1, X_2, \ldots, X_p$ represents a collection or vector of predictor variables. These are also known as co-variables, independent variables, or descriptor variables. It's important to note that these predictor variables are assumed to remain constant for a given individual or subject within the population of interest.

- $\beta = \beta_0, \beta_1, \ldots, \beta_p$ denotes the set of regression coefficients or parameters. β_0 is an optional intercept parameter, and $\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients or weights corresponding to$X_1, X_2, \ldots, X_p$.

To express a weighted sum of the X using matrix or vector notation, we can write it as follows [10]:

$$\beta\mathbf{X} = \beta_0 + \beta_1 X_1 + ...\beta_p X_p. \tag{2.1}$$

A regression model expresses a relationship between predictor $\mathbf{X}$ and response Y. For instance, $E(Y|\mathbf{X})$ denotes the expected value of Y given $\mathbf{X}$. This could be written as

$$E(Y|\mathbf{X}) = \beta\mathbf{X}. \tag{2.2}$$

It is possible that the $\beta\mathbf{X}$ and $E(Y|\mathbf{X})$ are not linearly related. In such cases, a transformation function can be applied to $E(Y|\mathbf{X})$ to account for any non-linear relationships. The linear regression model essentially adds random error to the weighted sum of predictors in the equation [11, 14, 15].

Equation 2.3 shows the mathematical equation

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon. \tag{2.3}$$

In a regression model, ϵ represents the random error associated with the model. It is assumed that the mean of these error components is zero and follows an approximate normal distribution.

Many regression models are available to predict a wide variety of outcome variables. The choice of model typically depends on the analysis goal and the type of response variable [15].

A linear regression model is used to model the association between a single continuous response variable, such as Y, and explanatory variables, such as X_p. This model has the same form as shown in 2.3. The parameter β is unknown, and we try to estimate it. The most popular estimation method is the least squares method, in which we try to estimate the coefficients β to minimize the residual sum of squares [14, 15].

$$\sum \epsilon_i^2 = \epsilon^T \epsilon = (y - X\beta)^T (y - \beta\mathbf{X}). \tag{2.4}$$

Differentiating with respect to beta and setting it to zero gives us an expression for the least square estimate of the beta parameter, denoted by $\hat{\beta}$.

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y. \tag{2.5}$$

We should note that equation 2.4 does not assume the validity of model 2.5; it simply determines the best fit for the data.

2.2 Logistic Regression

Linear regression is not always the best choice. Logistic regression is more suitable for binary events like mortality, with one or many independent variables. Examining multiple variables is crucial to understanding how each contributes to the model. For instance, when evaluating 30-day mortality rates for septic patients admitted to an Emergency Department, considering patient characteristics, provider practices, and hospital variables is crucial. Examining multiple independent variables is essential since sepsis involves many factors that an isolated evaluation may not capture [13].

In order to understand which independent variables contribute to which binary outcome, it is important to refer to Equation 2.6 [13].

$$E(Y|\boldsymbol{X}) = \frac{\exp\left(\beta_0 + \beta_1 X_1 + \beta X_2 + \cdots + \beta_p X_p\right)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_p X_p)} = \frac{\exp(\boldsymbol{\beta X})}{1 + \exp(\boldsymbol{\beta X})}, \tag{2.6}$$

or equivalently

$$\log \frac{\exp(E(Y|\boldsymbol{X}))}{1 - \exp(E(Y|\boldsymbol{X})} = \boldsymbol{\beta X}. \tag{2.7}$$

The equation for logistic regression has a similar structure to that of linear regression, consisting of p independent variables $X_1, \ldots, X_p)$ and beta coefficients (β). However, it is specifically designed for analyzing binary outcomes. Instead of predicting a continuous outcome like linear regression, it predicts the probability of one binary outcome.

To ensure the predicted values fall between 0 and 1, a log transformation is applied to the linear regression equation. Logistic regression identifies the strongest linear combination of independent variables that contribute to the likelihood of the outcome. This process is called maximum likelihood estimation.

Logistic regression can handle different variables, such as continuous, ordinal, and categorical. However, the selection of independent variables must be justifiable. One way to determine the relevance of each variable is through a p-value test. Alternatively, all relevant independent variables can be included, although this may decrease the model's generalizability beyond the current study sample [13].

2.3 Poisson Rate Regression

Let's delve into the Poisson Regression model, a crucial statistical analysis tool. This model relies on the probability distribution of the dependent variable, which, in this case, is the Poisson distribution. This distribution is widely used to represent counts of rare events that take place within a specific time frame. As a discrete distribution, it can only assume non-negative integer values. It is defined by a single parameter, denoted as λ, which serves as both the distribution's mean and variance. This parameter is akin to the mean of the normal distribution, since it describes the average number of occurrences of discrete events [16]. The Poisson distribution is represented as:

$$P(Y = y|\lambda) = \frac{e^{-\lambda}\lambda^{y}}{y!}. \tag{2.8}$$

The parameter λ can be estimated by a set of predictors $\mathbf{X} = (X_1, \ldots, X_k)$ and the expression can be written as:

$$\lambda(\beta\mathbf{X}) = \exp(\beta\mathbf{X}) = \exp(\beta_1 X_1 + \beta_2 X_2 + \ldots \beta_k X_k). \tag{2.9}$$

The aim of Poisson regression is to formulate a regression equation that can precisely forecast the anticipated value of the dependent variable Y (representing the number of events in a subgroup) as a function of, λ while fitting the data observed. The equation employs N to represent the overall follow-up time or population size of the subgroup. Poisson's regression is essentially a generalized linear model, in which the systematic effects are multiplicative, the error distribution conforms to the Poisson distribution, and the link function is the natural log [17].

$$E(Y) = N\lambda(\beta\mathbf{X}). \tag{2.10}$$

Assuming the sample size is N, the likelihood function for Poisson regression is given by:

$$L(Y;\beta) = \prod \frac{exp(-N\lambda)[N\lambda]^{Y}}{y!}. \tag{2.11}$$

One can estimate the β parameter by maximizing the Likelihood function. Once this estimation is obtained, various goodness-of-fit measures can be applied and residuals can be calculated [17, 18, 16].

2.4 Concepts of Censoring and truncation

Survival analysis often deals with cases where the response variable (survival time) is incompletely determined for some subjects, and this is known as "censoring". A common example of censoring is when a patient participating in a five-year follow-up study of survival after a fatal disease diagnosis is still alive at the end of the study. In this situation, the patient's survival time is considered "censored" after five years, indicating that their actual survival time exceeds five years [12, 19].

Censoring can take various forms, with right-censoring being the most common type. In right-censoring, the true survival time of individuals becomes incomplete on the right side of the follow-up period. This occurs when the study concludes, the individual is lost to follow-up, or they are withdrawn from the study. Such data is often referred to as "right-censored" or "censored type-1" [12].

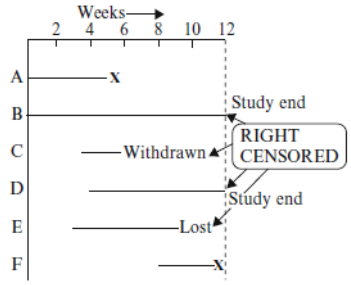

Figure 2.1: Right-censored: true survival time is equal to or greater than the observed survival time

Left-censored data occurs when an individual's actual survival time is equal to or less than the observed survival time for that individual. For example, in a scenario where we monitor individuals until they test positive for HIV, the failure event is recorded when a subject first tests positive for the virus. However, we might not have precise information regarding the exact moment of the initial exposure to the virus, which means we lack precise knowledge of when the failure event occurred.

In this situation, the survival time is considered left-censored because the genuine survival time, which terminates at exposure, is shorter than the follow-up duration, which concludes when the subject tests positive for the virus. In other words, if a person is left-censored at time t, we know they had an event between time 0 and t, but we do not know the exact time of the event [12].

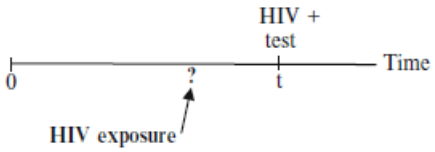

Figure 2.2: Left-censored: true survival time is less than or equal to the observed survival time

Interval censoring is used when an event is believed to have occurred within a certain time range. This type of censoring is applied when subjects are questioned or tested at fixed time points during a specified follow-up period. For example, in a longitudinal panel study, data might be collected from subjects once every 2 years. In clinical research, patients may be required to visit a clinic once every month for a period of several years [11, 19]

In an interval-censoring scenario, a patient might test negative at month 9 and positive at month 10, leaving us with the knowledge that the event occurred between the 9th and 10th clinic visits. Still, we would have no information about exactly when the event occurred [11].

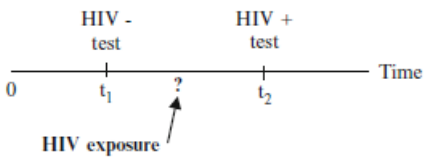

Figure 2.3: Interval-censored: true survival time is within a known time interval

There can be some confusion regarding whether observations are censored or truncated. Strictly speaking, truncation refers to cases in which subjects do not appear in the data because they are not observed. In contrast, censoring refers to cases when subjects are known to fail within a particular

episode, but the exact failure time is unknown. In other words, truncation involves missing data because subjects are not part of the study, while censoring involves subjects who are part of the study but have unknown failure times [12].

2.5 Hazard, and Survival functions.

In survival analysis, the response variable is denoted as T, and it typically represents the time until an event occurs, which is a continuous variable [10]. Instead of defining the statistical model for the response variable T in terms of the expected failure time, it is advantageous to define it in terms of the survival function, $S(t)$. The survival function is defined as:

$$S(t) = \text{Prob}\{T > t\} = 1 - F(t). \tag{2.12}$$

Here in 2.12,$F(t)$ represents the cumulative distribution function for T. When the event of interest is death, $S(t)$ signifies the probability that death occurs after time t, or in other words, the probability that the subject will survive at least until time t. It's important to note that the survival function must be non-increasing as time t increases.

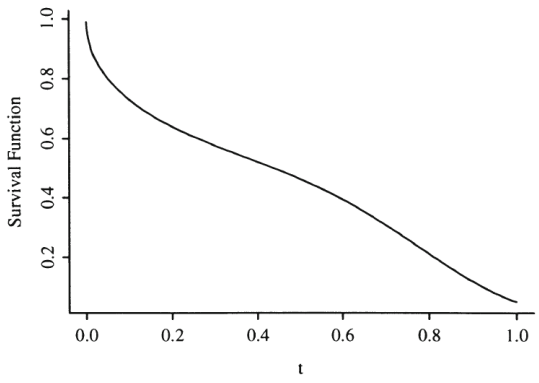

Figure 2.4: The curve for the survival function with respect to time.

In a specific example, subjects may face varying risks of experiencing the event over time. Early on, the risk is significantly elevated, leading to a sharp decline in the survival function $S(t)$. During certain time intervals, such as between 0.1 and 0.6, the risk remains relatively low, resulting in a relatively flat survival function $S(t)$. However, beyond $t = 0.6$, the risk begins to rise once more, causing the survival function $S(t)$ to decrease more rapidly again [10].

Figure 2.5 illustrates the cumulative hazard function, denoted as $\Lambda(t)$. This function is associated with the survival function presented in Figure 2.4. The cumulative hazard function, $\Lambda(t)$, conveys the accumulated risk or hazard up to a given time point t. As you will see later, it is the negative of the natural logarithm of the survival function [10].

The cumulative hazard function, $\Lambda(t)$, is a non-decreasing function as time, t, increases. This behavior signifies that the accumulated risk of the event either increases or remains constant over time. In Figure 2.5, this trend is visually evident.

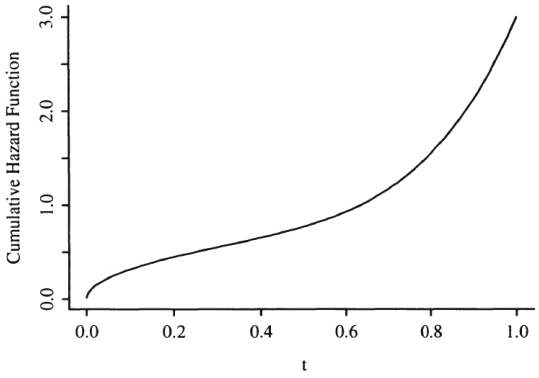

Figure 2.5: The curve for the cumulative hazard function.

$$\Lambda(t) \; = \; -\log S(t). \tag{2.13}$$

The hazard function, denoted as $\lambda(t)$, is indeed a crucial concept in survival analysis. It represents the instantaneous event (death, failure) rate at time t, given that the event hasn't occurred before time t. In essence, it helps us understand the probability of an event happening in a small interval around a specific time, conditional on the event not having occurred prior to that time.

Studying the hazard function provides insights into the underlying mechanisms and forces of risk over time, making it a valuable tool in survival analysis.

You can formally define the hazard function as follows:

$$\lambda(t) = \lim_{u \to 0} \frac{\text{Prob}\{t < T \leq t + u | T > t\}}{u}. \tag{2.14}$$

Equation 2.14 represents the hazard function, denoted as $\lambda(t)$, and it's derived using the law of conditional probability. The equation is indeed modified to express the ratio of certain functions, leading to a fundamental understanding of the hazard function.

Here's how the equation progresses:

$$\lambda(t) = \lim_{u \to 0} \frac{\text{Prob}\{t < T \leq t + u\}/\text{Prob}\{T > t\}}{u}$$
$$,$$
$$= \lim_{u \to 0} \frac{[F(t+u) - F(t)]/u}{S(t)}$$
$$,$$
$$= \frac{\delta F(t)/\delta t}{S(t)}$$
$$,$$
$$= \frac{f(t)}{S(t)} \quad . \tag{2.15}$$

The curve for the hazard function can be seen in Figure 2.6.

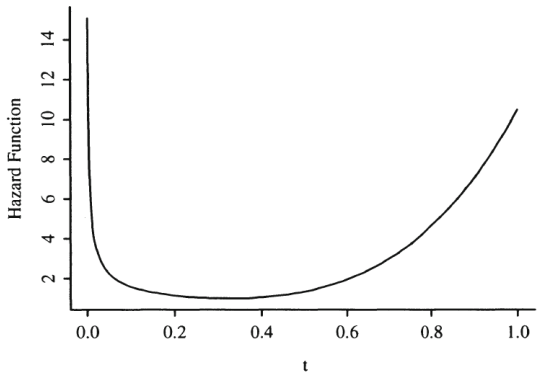

Figure 2.6: The curve for the Hazard function

In Equation 2.15, $f(t)$ represents the probability density function (PDF) of the event time T evaluated at time t. It's essentially the derivative or slope of the cumulative distribution function $1 - S(t)$. The PDF gives us the rate at which the event is occurring at a specific time point t. Understanding the hazard function $\lambda(t)$ is crucial for assessing risk and modeling event occurrence over time in survival analysis [10].

$$\frac{\delta \log S(t)}{\delta t} = \frac{\delta S(t)/\delta t}{S(t)} = -\frac{f(t)}{S(t)}. \tag{2.16}$$

Equating from the equations 2.14 and 2.15 the hazard function can also be written as:

$$\lambda(t) = -\frac{\delta \log S(t)}{\delta t}. \tag{2.17}$$

Working backward from equation 2.16, we can obtain the integral of $\lambda(t)$.

$$\int_0^t \lambda(u)\delta u = \Lambda(t) = -\log S(t). \tag{2.18}$$

In equation 2.13, $\Lambda(t)$ was defined as the cumulative distribution function of $\lambda(t)$. Thus, the area under $\lambda(t)$ can be expressed as $\Lambda(t)$. This means that we can obtain two functions out of three by either moving forward or backward, given that we have one side of the equation. Essentially, there are three different ways to describe the same distribution [10].

2.6 Cox proportional hazards model

This section presents a way to generalize the survival model to a survival regression model. By adding predictor variables $\mathbf{X} = X_1, X_2, \ldots, X_k$ to the sample, we make it heterogeneous.

The widely used survival regression specification is to multiply the hazard function $\lambda(t)$ with $exp(\beta\mathbf{X})$. Thus, the survival model has a hazard function for the failure time given the predictors $\mathbf{X}$:

$$\lambda(t|X) = \lambda(t)\exp(\beta\mathbf{X}). \tag{2.19}$$

The proportional hazards (PH) model is a type of hazard modeling that uses regression analysis. In this model, the term $\lambda(t)$ in $\lambda(t|\mathbf{X})$ represents the underlying hazard function. In Cox's semi-parametric PH model, $\lambda(t)$ can be left unspecified while still enabling estimation of β. Whether β_0 is included in $\beta\mathbf{X}$ depends on whether the underlying hazard function has a constant scale parameter. The $exp(\beta\mathbf{X})$ term is known as the relative hazard function and is often of interest because it provides information on the relative impacts of the predictors [10].

The PH model can be linearized concerning $\beta\mathbf{X}$ using the following identities:

$$\log\lambda(t|\mathbf{X}) = \log\lambda(t) + \beta\mathbf{X}. \tag{2.20}$$

the following assumptions must be considered in the PH model [10]:

- The true form of the underlying functions (λ) should be specified correctly.
- The relationship between the predictors and log hazard should be linear in its simplest form. In the absence of interaction terms, the predictors should also operate additively.
- How the predictors affect the distribution of the response should be by multiplying the hazard $\exp(\beta\mathbf{X})$ or equivalently by adding $\beta\mathbf{X}$ to the log hazard at each t. The effect of the predictors is assumed to be the same at all values of t since $\log\lambda(t)$ can be separated from $\beta\mathbf{X}$.

The Cox proportional hazards model is a widely used survival model due to its semi-parametric nature. It assumes that predictors have an impact on the hazard function, without requiring knowledge of the nature of the function itself. Initially, the focus is not on estimating the hazard function, as Cox argued that when the PH model holds, estimating beta coefficients is more important than understanding the hazard function [10].

The Cox PH model assumes that predictors act multiplicative on the hazard function, but does not assume that the hazard function is constant. Regressors are linearly related to the log of hazard. This model is particularly useful when the true hazard function is complex or unknown. It primarily focuses on the effects of predictors rather than the shape of the hazard function. The Cox PH model uses only the rank ordering of the failure and censoring times, which makes it less affected by outliers in the failure times than fully parametric methods. Additionally, the Cox model includes the log-rank test as a special case for comparing survival between two groups [10].

The Cox PH model is as efficient as parametric models such as the Weibull model with PH for estimating and testing regression coefficients, even when all assumptions of the parametric model are satisfied. When the assumptions of a parametric model are not true, the Cox analysis outperforms the parametric analysis in terms of efficiency. Cox showed that by conditioning the log-likelihood function in a specific way, it's possible to derive a valid estimate of beta that doesn't require estimating the hazard function, as it drops out of the new likelihood function [10].

Cox's method focuses on utilizing data related to the relative hazard function $\exp(\beta\mathbf{X})$ to estimate β. To explain Cox's estimator, let us assume that $t_1 < t_2 < t_3 < ... < t_i$, represents unique ordered failure times in a sample of n subjects. For now, let us assume that there are no tied failure times (although tied censoring times are allowed), meaning that $k = n$.

Consider a set of individuals who are at risk of failing just before the failure time t_i. This set is called the risk set at time t_i, denoted by R_i. R_i includes subjects with failure/censoring time $Y_j \geq t_i$. In other words, R_i is the set of subjects j who has not failed or been censored by time t_i [10].

The conditional probability that an individual i failed at t_i, given that the subjects in the set R_i are at risk of failing and given that exactly one failure occurs at t_i, is the estimator for β.

$$\text{Prob\{Subject } i \text{ fails at } t_i | R_i \text{ and one faliure at } t_i\} = \frac{\text{Prob\{subject } i \text{ fails at } t_i | R_i\}}{\text{Prob\{One failure at } t_i | R_i\}} \tag{2.21}$$

The conditional probability can be calculated by applying the rules of conditional probability [10, 20]. This probability is equal to:

$$\frac{\lambda(t_i)\exp(\beta X_i)}{\sum_{j \in R_i} \lambda(t_i)\exp(\beta X_j)} = \frac{\exp(\beta X_i)}{\sum_{Y_j \geq t_i} \exp(\beta X_j)}. \tag{2.22}$$

It is important to note that the likelihood for β that is referred to as Partial Likelihood by Cox is not dependent on $\lambda(t)$. To calculate the Partial Likelihood, we multiply the individual likelihoods (which are the conditional probabilities mentioned above) of overall failure times. This is made possible by the fact that these conditional probabilities are themselves conditionally independent across the different failure times [10, 20].

$$L(\beta) = \prod_{Y_i \, uncensored} \frac{\exp(\beta X_i)}{\sum_{Y_j \geq Y_i} \exp(\beta X_j)}. \tag{2.23}$$

The Log partial likelihood is:

$$\log L(\beta) = \sum_{Y_i \, uncensored} \{\beta X_i - \log[\sum_{Y_j \geq Y_i} \exp(\beta X_j)]\}. \tag{2.24}$$

By using the partial log-likelihood of an ordinary log-likelihood, we can obtain valid (partial) maximum likelihood estimates of β. It's worth noting that adding a constant to the predictors **X** doesn't alter this log-likelihood formulation. The Cox model is designed for the relative hazard and cannot directly estimate the underlying hazard $\lambda(t)$. This is also why an intercept term is not needed and therefore cannot be estimated.

Once we have derived the log-likelihood, we can use it to obtain point and interval estimates of hazard ratios, an estimate of β, estimated standard errors of β, confidence limits, and statistical tests. Unlike the parametric survival model, the Cox model does not require a choice of the underlying survival function to estimate the survival function. As a result, fitting a Cox model does not provide a direct estimate of $S(t|X)$ [10, 21].

2.7 Frailty model

In the realm of PH models, heterogeneity may take on one of two forms: observed or unobserved. Unobserved heterogeneity encompasses unmeasurable factors such as unknown health conditions, occupation, and lifestyle, which are typically disregarded during analysis in favor of measurable variables. It is worth noting that any population analysis represents an average person rather than an individual, and the average risk of death at any given age is that of the part of the population that is alive at that age. As individuals at high risk tend to pass away earlier than those at low risk, the composition of the population changes over time, a phenomenon commonly known as frailty or unobserved heterogeneity [22].

Frailty is a term used to describe the phenomenon where individuals who share common characteristics, such as age, gender, and weight, have varying levels of risk for mortality. This represents an unobservable random effect that is shared among subjects with similar unmeasured risks in a mortality rate study. To analyze this, a frailty model separates the hazard function into three multiplicative components: frailty, the baseline hazard function, and the linear predictor. For example, in data categorized by genotype, the hazard function given frailty Z can be expressed as the product of Z, $h_0(t)$ (the baseline hazard function), and $\exp(\beta \mathbf{X})$ (the exponential of the linear predictor). Frailty in this context refers to the shared frailty of a specific group of individuals and is always non-negative.

A group with the same genotype can have a shared frailty. The extent of variability among the groups is determined by the variability of Z, and its distribution is described by a probability density function. Therefore, it is crucial to determine whether the population being studied has any unmeasured covariates or shows signs of heterogeneity for the frailty model to be applicable [22, 23].

When examining a data set, it's essential to take into account the mixed hazards or variation in the population. Univariate data sets can reveal this variation through the distinct survival patterns of each person in the population. One method of observing this is by plotting hazard functions over time and noting that the hazard rate isn't consistent but represents a blend of hazards. When this occurs, a frailty model can be employed to fit the data.

In the instance mentioned above, where data is categorized by genotype, a multivariate setting is present. In these scenarios, a multivariate frailty model would be fitting. This is because it would be pertinent to model heterogeneity among groups of people or within groups of individuals, assuming underlying unobserved heterogeneity exists [23].

A shared frailty model considers the impact of a common, but unobserved and random covariate (frailty) on the hazard rate of each group. To model this frailty, a one-parameter gamma distribution is used, with its density function denoted as $g(z)$. The frailty's mean value is set at 1, while the variance (θ) measures the level of association. Higher values of θ indicate greater heterogeneity among subgroups and a stronger association among subjects within each subgroup.

A shared frailty model incorporates an unobserved, random covariate known as "frailty" to affect the hazard rate of different groups. This frailty is typically modeled using a one-parameter gamma distribution, denoted by its density function $g(z)$. The frailty's mean value is set at 1, while the parameter θ quantifies the level of association. Higher θ values indicate greater heterogeneity among subgroups and a stronger association among subjects within each subgroup.

$$g(z) = \frac{z^{(1/\theta-1)} \exp(-z/\theta)}{\Gamma(1/\theta)\theta^{1/\theta}}. \tag{2.25}$$

In empirical applications, the parameters of the frailty distribution are estimated, and individual frailties are predicted using observed survival data [23, 24, 25]. This frailty definition assumes that individuals are born with a specific level of relative frailty that remains constant throughout their lives. Formally, the hazard function for the frailty model can be expressed as:

$$\lambda(t|Z) = Z\lambda(t)\exp(\beta\mathbf{X}). \tag{2.26}$$

For a successful estimation procedure, it is crucial that the frailty model, characterized by the distribution of Z and the underlying baseline hazard $h_0(t)$, is identifiable [23, 24, 25].

2.8 Model selection criterion

Regression analysis can accommodate a variety of variable types, including continuous (e.g., age), ordinal (e.g., visual analog scale), and categorical (e.g., race). However, the selection of these variables should be well-justified. For instance, a p-value test with a significance level of $p <= 0.05$ can be employed to determine the relevance of variables in the model. Alternatively, all relevant independent variables can be included, particularly if they hold clinical significance, regardless of their statistical performance. Nevertheless, this approach comes with the risk of the model becoming mathematically unstable and may hinder its generalizability beyond the current study sample.

In addition to variable selection, choosing the appropriate type of logistic regression model for the study is crucial. The model-building strategy is intricately linked to the selection of independent variables, and both aspects must be considered simultaneously. There are three general approaches to structuring a regression model: direct, sequential, and step-wise. It's important to note that these strategies can yield different results from the same data, and they are not interchangeable [13].

- The direct (simultaneous)method utilizes all the variables simultaneously and makes no assumption about the order or the relative worth of these variables.
- The sequential (hierarchical) model on the other hand sequentially adds the variables, thus giving us a better idea of how the model is improving.
- The stepwise regression identifies the relevant independent variables for keeping based on a predefined statistical criterion influenced by the unique characteristics of the sample being analyzed.

When using stepwise regression, we consider statistical criteria to score the best model fit with more significance. it is important to determine which model should be given more significance. This is where the Akaike information criterion (AIC) comes in, as it can be used to compare and select the most appropriate model.

For example, we run a frailty model with forward and backward stepwise regression based on AIC to identify a subset of predictor variables that best explain variability in survival data while accommodating random effects or shared frailty. The process begins with an initial model that includes all potential predictors and the random frailty component. Through iterative steps, forward selection adds predictors and assesses their contribution to the model, guided by the AIC score. Conversely, backward selection removes predictors that do not significantly impact model fit. This combined approach aims to balance the inclusion of relevant predictors and model parsimony. The final model obtained through these steps represents the most pertinent set of predictors for the frailty model, which can aid in a more accurate understanding of survival data.

AIC is based on the K-L information loss principle, which requires that the chosen model be as close to reality as possible, i.e., with the least K-L information loss. The K-L information loss between the models can be expressed as the difference between two statistical expectations [26].

First, such expectation cannot be computed or estimated but is constant across models and can be removed. The relevant term is the second expectation, $E[\log(g(x|\theta))]$, where E is the expectation operator, log is the natural logarithm, x represents the response variable to be predicted by the

model (x represents hypothetical data), and θ represents a vector of unknown parameters. This second term also cannot be computed or estimated. Akaike found that if a second expectation was taken over an estimated, θ then that quantity could be estimated, and this result provided the link between K-L information and the maximized log-likelihood. Akaike's key finding focused on the double expectation,

$$E\{E[\log(g(x|\hat{\theta}(y)))]\}. \tag{2.27}$$

where y represents data and θ is the vector of parameter estimates based on these data. Akaike found that for large sample sizes (n) this double expectation can be estimated very simply as $log(L) - K$, where K is a correction for asymptotic bias and is merely the total number of estimable parameters in the model [26, 27].

$$E\{E[\log(g(x|\hat{\theta}(y)))]\} = \log(L(\hat{\theta}|y)) - K. \tag{2.28}$$

Multiplying both sides by -2 to get the expression for AIC [27, 28]:

$$\text{AIC} = -2\log(L(\hat{\theta}|y)) + 2K. \tag{2.29}$$

Statisticians use the term "deviance" to refer to the calculation of $-2\log(L)$, which can be computed easily. At the application level, it is necessary to calculate the AIC for each model in the set, and choose the model with the smallest AIC value, indicating that the model has the least K-L information loss.

With a comprehensive grasp of the core survival analysis models and selection criteria, the groundwork is now laid for the subsequent chapter. In the following section, our attention will shift towards an in-depth exploration of the dataset under study. We will systematically apply the previously mentioned methodologies to analyze the data, offering valuable insights into our research.

3 Data Description, Analysis, and Interpretations

Two distinct datasets will be examined in this study. The first dataset is cross-sectional, encompassing the comprehensive results obtained on the subjects' enrollment day. This dataset is rich in covariates, presenting a multitude of factors to explore for their significance in the study. Nevertheless, certain covariates will be omitted from the analysis due to their lack of relevance or the presence of missing data, leading to the presence of more NA values in the dataset.

The second dataset is the longitudinal data, involving multiple follow-up visits with relatively fewer test results. Although this dataset contains fewer covariates, it compensates with a substantial amount of data within those covariates. Survival models will be constructed based on the longitudinal data, providing a deeper understanding of the dynamic aspects of the study over time.

3.1 Data origin and description

The information was obtained from a study carried out at Siaya County Referral Hospital (SCRH), located in Western Kenya, an area with high transmission of *P.falciparum*. The inhabitants of the study area are mainly from the Luo ethnic group, representing more than 96% of the population.

The study recruited children aged between 2 and 70 months who presented with suspected malaria infections or reported for routine vaccinations at SCRH. After conducting malaria parasite screening, children with varying degrees of malarial anemia $(n = 1319)$ and controls without parasites $(n = 335)$ were enrolled. Children were excluded from the study if they had non-*falciparum* parasite strains, confirmed cerebral malaria, had been previously hospitalized for any reason, or had used antimalarial therapy in the two weeks before the study.

The current study included two cohorts recruited and followed with the same parameters over time: cohort 1 $(2003 - 2005; n = 777)$ and cohort 2 $(2007 - 2012; n = 877)$.

Following enrollment (Day 0), a total of 1654 children were scheduled for follow-up visits on Day 14 (if they had a fever at the time of enrollment) and then every three months for 36 months. If any parent or guardian failed to bring their child for the scheduled quarterly follow-up visit, the study team would trace them at their residence to check the child's health status, including any mortality. The location of each residence was identified by a GIS/GPS surveillance system. Furthermore, parents and guardians were asked to bring their child to the hospital whenever they experienced a fever (acute febrile episode(s)). For comprehensive clinical management of the patients, physical evaluations and laboratory tests, including complete blood counts (CBC), malaria parasitemia measures, and evaluation of bacteremia where indicated, were performed at enrollment, day 14, and each acute and quarterly visit. All acute episodes and scheduled visits were managed by the guidelines of the Ministry of Health-Kenya [29].

During the first visit, multiple laboratory procedures were conducted on the individuals and recorded. Blood samples were collected from the heel and/or finger-pricks, which were less than 100 μL in volume, and also through venipuncture which involved the collection of $1 - 3\ mL$ of blood. These

samples were used to determine the density of parasites causing malaria. Thin blood smears were prepared using Giemsa staining and examined to determine the density of asexual malaria parasites. Since coinfection affects the severity of malarial anemia in Siaya, all children were tested for bacterial infections and HIV-1. Parents or legal guardians of participating children were provided with pre- and post-test HIV and AIDS counseling. None of the children were undergoing antiretroviral therapy for HIV-1 at the time of enrollment. Genetic factors, sickle cell traits, and alpha-thalassemia deletions were investigated to further understand chronic anemia. Sickle cell status was determined using alkaline cellulose acetate electrophoresis on Titan III plates (Helena BioSciences, Sunderland, UK) [29].

The data analysis was conducted using R version $4.2.1$. The data from both cohorts, namely cohort $1(2003-2005)$ with $n = 777$ and cohort $2(2007-2012)$ with $n = 877$, were combined into a single dataset. However, the cohort was considered as a categorical covariate to adjust for the potential changes in malaria incidence over time. The analysis was performed on both cross-sectional and longitudinal levels [29].

3.2 Cross-sectional Analysis

Two new response variables have been defined to indicate new events in the data. They are **RPI_anemic** and **RPI_SMA**. If Hgb (hemoglobin) is less than 11 along with RPI being less than 2, **RPI_anemic** is marked as 1. Similarly, if Hgb is less than 5 and RPI is less than 2, **RPI_SMA** is marked as 1.

When the RPI threshold is set as 2, it suggests that all the subjects have some predisposition leading to low RPI levels. This could mean that the subjects have certain conditions that do not allow the bone marrow to function properly. To explore this further, we need to see if the distribution of RPI levels is similar in malaria-positive and malaria-negative cases, which will be investigated in the following sections. The Table 3.1 shows how Malaria and Anemia cases are distributed among various categories and the associated percentages. Keep in mind, our definition of anemia here includes RPI less than 2 and Hgb less than 11.

	mal-neg	**mal-pos**	**Total**
Anemia-neg	138	217	355
Row Percent	38.9%	61.1%	21.5%
Column Percent	44.2%	16.2%	
Total Percent	8.4%	13.2%	
Anemia-pos	174	1120	1294
Row Percent	13.4%	86.6%	78.5%
Column Percent	55.8%	83.8%	
Total Percent	10.6%	67.9%	
Total	312	1337	1649
Row Percent	18.9%	81.1%	100%
Pearson's Chi-squared test	$\text{Chi}^2 = 117.4012$, d.f. = 1, $p < 2e-16$		

Table 3.1: Malaria and anemia cases cross-table with marginals.

Table 3.1 shows us the following distribution Among the total cases (1649in this dataset).

- Approximately 81.1% of the cases in the study have Malaria (Malaria-positive), while 18.9% do not have Malaria (Malaria-negative).
- Also, Of the total cases, 78.5% have Anemia (Anemia-positive), and 21.5% do not have Anemia (Anemia-negative).
- Among those with Anemia (Anemia-positive), a significant majority 86.6% also have Malaria, while a smaller proportion 13.4% have Anemia without Malaria.
- Among those without Anemia (Anemia-negative), a majority 61.1% have Malaria, while 38.9% do not have Malaria.

The performed statistical tests, specifically Pearson's Chi-squared tests, indicate a strong association between Malaria and Anemia. The p-values, being less than 2e-16 (extremely low), suggest that the relationship between these two conditions is statistically significant. The odds ratio for Malaria-positive (mal-pos) is approximately 4.09. This means that individuals with Malaria are about four times more likely to have Anemia compared to those without Malaria. This means that the presence of one condition is associated with the presence of the other.

A similar association is also seen in Table 3.2, which shows the distribution of SMA and malaria.

	mal-neg	**mal-pos**	**Total**
SMA-neg	289	1092	1381
Row Percent	20.9%	79.1%	83.7%
Column Percent	92.6%	81.7%	
Total Percent	17.5%	66.2%	
SMA-pos	23	245	268
Row Percent	8.6%	91.4%	16.3%
Column Percent	7.4%	18.3%	
Total Percent	1.4%	14.9%	
Total	312	1337	1649
Row Percent	18.9%	81.1%	100%

Pearson's Chi-squared test	$\text{Chi}^2 = 22.29613$, d.f. = 1, p =$2.34e-06$

Table 3.2: Malaria and SMA cross-table with marginals.

- The majority (79.1%) of individuals in the SMA-neg group also have Malaria, while only 20.9% of those with SMA are Malaria-negative
- In the SMA-positive group, 91.4% have Malaria, and only 8.6% do not.

This close association is likely because Malaria can lead to the destruction of red blood cells, resulting in anemia. SMA is a severe form of anemia often associated with Malaria. The Chi-squared test indicates that the relationship is statistically significant with a p-value less than 0.05, emphasizing the importance of considering both conditions when diagnosing and treating patients in regions where Malaria is prevalent.

The strong association between Malaria and Anemia/SMA is likely due to several interconnected factors. Malaria is caused by a parasite infecting red blood cells, and destroying them. This causes a decrease in hemoglobin levels, which can result in anemia and SMA. SMA is characterized by a low concentration of red blood cells or a low hemoglobin content, often caused by various factors, including the chronic loss of red blood cells due to Malaria. Additionally, the symptoms of Malaria,

such as fever and fatigue, may lead to reduced dietary intake and nutritional deficiencies, further contributing to anemia and then SMA. The cases in front of us also had RPI less than 2 suggesting a predisposition causing insufficient erythropoiesis which can also cause anemia or SMA among individuals with the malaria parasite.

There appears to be a correlation between malaria and significant metric variables, including RPI levels, reticulocyte count, IFN-g (interferon gamma) levels, and Hgb. The Table 3.3 compare assorted hematological and immunological parameters between a stratified sample of malaria-positive and malaria-negative cases.

Characteristic	Whole Sample	mal-neg	mal-pos	p-value
RPI	$0.62(0.94)[0.92]0.87$	$0.66(0.85)[0.93]0.93$	$0.62(0.95)[0.92]0.86$	$\mathbf{0.5}$
ReticPer	$1.40(3.31)[2.67]2.30$	$1.00(2.34)[1.71]1.50$	$1.60(3.46)[2.89]2.60$	$\mathbf{< 0.001}$
RBCx1012L	$3.63(1.20)[3.54]1.97$	$4.65(1.17)[4.35]1.15$	$3.39(1.13)[3.35]1.80$	$\mathbf{< 0.001}$
IFNg	$22(36)[27]22$	$26(30)[32]18$	$20(37)[27]23$	$\mathbf{< 0.001}$
IFNa	$31(62)[51]43$	$40(68)[68]66$	$29(60)[47]38$	$\mathbf{< 0.001}$
HgbgdL	$7.30(2.51)[7.47]4.10$	$10.30(2.58)[9.55]2.65$	$6.80(2.24)[6.99]3.50$	$\mathbf{< 0.001}$
Hct	$24(8)[24]13$	$33(8)[30]8$	$23(7)[23]11$	$\mathbf{< 0.001}$
1 Median (SD) [Mean] IQR 2 Wilcoxon rank sum test				

Table 3.3: Distribution of whole sample and malaria groups. All the significant p-values are in bold.

After analyzing the data, it was found that there is no significant difference in RPI values between individuals with and without malaria in the studied population. However, the percentage of reticulocytes in the blood is significantly higher in those who are malaria-positive ($p < 0.001$), which is expected as reticulocytes are new red blood cells that increase during malaria infection. It was also observed that hemoglobin levels are significantly lower in the malaria-positive group ($p < 0.001$), as malaria infection leads to a decrease in hemoglobin concentration. Furthermore, the levels of IFNg and IFNa (interferon-gamma and alpha, respectively) are significantly lower in those with malaria ($p < 0.001$), indicating a weaker immune response compared to those without malaria. Lastly, hematocrit values are significantly lower in the malaria-positive group ($p < 0.001$), which is consistent with what is observed in individuals with malaria.

These measurements are relevant and play a very big part in the individual's survival in anemia and SMA. The correlation of these variables among each other was investigated, along with their p-values for statistical significance.

Hemoglobin levels are a pivotal marker of health, especially when it comes to malaria. Hemoglobin levels also vary when erythropoiesis is not functional. Hemoglobin is negatively correlated with reticulocyte count. This suggests that as hemoglobin levels increase, reticulocyte count tends to decrease. This finding may be associated with the response to anemia. As hemoglobin levels rise, the body may reduce reticulocyte production, reflecting improved oxygen-carrying capacity. However, the strength of this negative correlation varies among different subgroups based on Malaria status and erythrocyte sufficiency.

Variables	Malaria_all	Malaria_neg	Malaria_pos
hgb vs Retic count	$-0.34654\ (8.85 \times 10^{-47})$	$-0.27912\ (7.94 \times 10^{-07})$	$-0.33441\ (1.21 \times 10^{-35})$
hgb vs IFNg	$0.15004\ (6.42 \times 10^{-05})$	$0.09892\ (0.2555)$	$0.14842\ (0.0004)$
hgb vs RPI	$0.16494\ (2.02 \times 10^{-10})$	$0.2012\ (0.0037)$	$0.16558\ (3.23 \times 10^{-09})$
Retic count vs IFNg	$-0.058\ (0.1277)$	$-0.13352\ (0.1299)$	$-0.04145\ (0.3271)$
Retic count vs RPI	$0.71498\ (2.76 \times 10^{-230})$	$0.65154\ (2.85 \times 10^{-26})$	$0.72896\ (8.68 \times 10^{-210})$
IFNg vs RPI	$-0.00728\ (0.8584)$	$-0.12977\ (0.2738)$	$-0.00166\ (0.9696)$

Variables	sufficient erythropoesis
hgb vs Retic count	$-0.44788\ (0.044)$
hgb vs IFNg	$0.02084\ (0.6444)$
hgb vs RPI	$0.20887\ (3.19 \times 10^{-14})$
Retic count vs IFNg	$0.02513\ (0.5778)$
Retic count vs RPI	$0.6413\ (8.04 \times 10^{-151})$
IFNg vs RPI	$0.03376\ (0.4545)$
Variables	**insufficient erythropoesis**
hgb vs Retic count	$-0.83501\ (9.16 \times 10^{-85})$
hgb vs IFNg	$0.18938\ (0.0058)$
hgb vs RPI	$-0.1213\ (0.1098)$
Retic count vs IFNg	$-0.18976\ (0.0074)$
Retic count vs RPI	$0.44579\ (6.32 \times 10^{-10})$
IFNg vs RPI	$-0.19312\ (0.044)$
coefficient (p-value)	

Table 3.4: Correlation of Malaria-related Variables

There is a positive correlation between hemoglobin and IFNg, primarily significant in the context of Malaria. Higher hemoglobin levels appear to be associated with increased IFNg levels, indicating a potential relationship between hemoglobin status and immune response. This association may suggest that improved hemoglobin levels are related to enhanced immune function, particularly in the presence of Malaria.

Hemoglobin is positively correlated with the Reticulocyte Production Index. In most subgroups, an increase in hemoglobin is associated with a rise in RPI. This correlation implies that better hemoglobin levels are linked to increased reticulocyte production, a process that signifies the body's response to anemia. The significance of this correlation varies, with the strongest associations observed in the Malaria-positive subgroups.

The rest of the correlations are mostly weak and not statistically significant across all subgroups. Further investigation with the help of mathematical modeling has been done. We use regression models to see what kind of effect the significant variables have on our response variables and vice versa.

3.2.1 Logistic Regression

As our response variables are binary, we will use logistic regression to see which independent variables influence the likelihood of the event happening. we will investigate three response variables signifying Anemia with an RPI less than 2, SMA with an RPI less than 2, and Malaria. Table 3.5 will explain the influence of the independent variables on the binary response variable RPI_anemic.

Variable	Estimate	Std. Error	z value	p-value
(Intercept)	1.821×10^{1}	3.781×10^{0}	4.815	$\mathbf{1.47 \times 10^{-6}}$
ReticPer	-7.039×10^{-1}	2.313×10^{-1}	-3.043	$\mathbf{0.002346}$
RBCx1012L	-2.390×10^{0}	6.894×10^{-1}	-3.467	$\mathbf{0.000526}$
RPI	-3.838×10^{0}	6.107×10^{-1}	-6.285	$\mathbf{3.27 \times 10^{-10}}$
IFNg	-1.767×10^{-2}	5.357×10^{-3}	-3.299	$\mathbf{0.000972}$
IL4	-3.512×10^{-2}	2.045×10^{-2}	-1.718	0.085804
IL15	7.823×10^{-3}	4.302×10^{-3}	1.818	0.069016
IL17	2.677×10^{-2}	1.552×10^{-2}	1.724	0.084680
TNFa	1.757×10^{-2}	9.036×10^{-3}	1.944	0.051848
IP10	1.116×10^{-3}	4.916×10^{-4}	2.270	$\mathbf{0.023216}$
MIG	-2.252×10^{-3}	6.705×10^{-4}	-3.358	$\mathbf{0.000784}$
MCP1	-2.645×10^{-3}	8.713×10^{-4}	-3.036	$\mathbf{0.002398}$
Malaria	1.789×10^{0}	6.902×10^{-1}	2.592	$\mathbf{0.009544}$
Null Deviance: 561.84 on 597 degrees of freedom **Residual Deviance**: 109.64 on 584 degrees of freedom **AIC**:137.64				

Table 3.5: Logistic regression model for anemia. All the significant p-values are in bold.

Reticulocyte percentage in the blood can be an indicator of the body's response to anemia. A negative coefficient suggests that a higher Reticulocyte Percentage is associated in individuals with anemia, indicating that an increased presence of reticulocytes may be linked to improved red blood cell production.

The negative coefficient implies that higher red blood cell counts are associated with a reduced likelihood of the individuals having anemia. as higher red blood cell counts generally lead to better oxygen-carrying capacity.

RPI is a measure used to determine the body's reticulocyte production in response to anemia. A negative coefficient indicates that a higher RPI is linked to a lower likelihood of being classified as "anemic". This means that a stronger reticulocyte response to anemia is associated with a decreased probability of developing anemia.

Research suggests that lower levels of interferon-gamma (IFNg), a cytokine responsible for immune response, may be linked to immune system dysregulation or a different immune response in individuals diagnosed with anemia.

IP10 is a chemokine produced in response to interferon, and elevated levels may indicate greater immune activation and inflammation in individuals with anemia. This suggests a potential link between immune activation and anemia.

MIG is a chemokine, which is a type of signaling protein involved in immune responses. It is produced in response to interferon-gamma (IFNg) and is involved in immune response, so lower levels may indicate an immune response towards anemic individuals.

MCP1 is a chemokine involved in the recruitment of monocytes to sites of inflammation, and lower levels may suggest differences in inflammatory responses in individuals with anemia.

Having Malaria significantly increases the likelihood of anemia. The coefficient suggests that Malaria-positive individuals are at a higher risk of being anemic compared to Malaria-negative individuals.

Further, we investigate the relationship between another response variable for the event of SMA and RPI less than 2 shown in 3.6. Reticulocyte percentage having a positive coefficient may be associated with a greater likelihood of SMA. This indicates individuals with SMA show a higher percentage of reticulocytes.

A negative coefficient for RBC counts implies that individuals who experience SMA will have a lower number of red blood cells. This behavior is expected, as SMA is caused by the destruction of red blood cells by the malaria parasite. A high Negative coefficient indicates that a more robust reticulocyte response to anemia is linked to a lower risk of developing severe anemia.

IL-13 is often considered an anti-inflammatory cytokine. It can suppress the production of pro-inflammatory cytokines, such as tumor necrosis factor-alpha (TNFa). The negative coefficient for IL-13 suggests that lower levels of IL-13 are associated with a reduced likelihood of severe anemia. This could indicate that individuals with lower IL-13 levels may have variations in immune and inflammatory responses that are related to a lower risk of severe anemia.

TNF-α is a pro-inflammatory cytokine that plays a central role in the body's immune responses and inflammation. It promotes the recruitment of immune cells to sites of infection, injury, or inflammation. the positive coefficient for TNF-α suggests that higher TNF-α levels are associated with an increased likelihood of severe anemia. This implies that elevated TNF-α may be indicative of an enhanced inflammatory response, which could contribute to the risk of severe anemia.

Variable	Estimate	Std. Error	z value	p-value
(Intercept)	-1.508×10^{2}	1.625×10^{3}	-0.093	0.926070
ReticPer	1.011×10^{1}	2.046×10^{0}	4.942	$\mathbf{7.73 \times 10^{-7}}$
RBCx1012L	-1.353×10^{0}	5.105×10^{-1}	-2.650	$\mathbf{0.008043}$
RPI	-5.576×10^{1}	1.151×10^{1}	-4.845	$\mathbf{1.27 \times 10^{-6}}$
IFNa	-1.885×10^{-2}	1.210×10^{-2}	-1.558	0.119270
IL5	-4.380×10^{-1}	2.189×10^{-1}	-2.001	$\mathbf{0.045375}$
IL13	-3.711×10^{-2}	1.110×10^{-2}	-3.342	$\mathbf{0.000831}$
IL17	-5.216×10^{-2}	2.366×10^{-2}	-2.204	$\mathbf{0.027520}$
TNFa	3.723×10^{-2}	1.136×10^{-2}	3.277	$\mathbf{0.001048}$
IP10	-9.881×10^{-4}	5.289×10^{-4}	-1.868	0.061730
Eotaxin	2.888×10^{-2}	1.774×10^{-2}	1.628	0.103599
Null Deviance: 570.881 on 597 degrees of freedom **Residual Deviance**: 71.321 on 586 degrees of freedom **AIC**:95.321				

Table 3.6: Logistic regression model for SMA. All the significant p-values are in bold.

Lastly, we look into malaria as a response variable to investigate the immune response. Table 3.7 below shows the Logistic Regression model for malaria. An increase in RBC is associated with a 0.4020 decrease in the log odds of the presence of Malaria while controlling for other covariates. This finding underscores the role of RBC count in the manifestation of malaria. suggesting that malaria infection may lead to decreased red blood cell counts due to hemolysis and erythrocyte destruction.

An increment in IFNg levels leads to a 0.0262 increase in the log odds of a malaria presence, considering other factors. This suggests that during malaria infection, the body's immune response involves the production of IFNg to combat the parasite.

IL1Ra demonstrates high significance in the model. rise in IL1Ra corresponds to a 0.000335 reduction in the log odds of the presence of Malaria while accounting for other variables. This emphasizes the regulatory role of IL1Ra in the context of malaria-related immune responses. Suggesting an immune response where the body releases less IL1Ra, potentially to counteract the pro-inflammatory effects of IL-1 in malaria.

The higher IL2R levels will result in an elevation in the log odds of malaria, implying that IL2R levels are associated with individuals with malaria. as IL2R is involved in T-cell activation and proliferation, a key component of the adaptive immune response to malaria.

A decrease in IL4 suggests a shift in immune response towards a Th1 response in malaria, as IL4 is primarily associated with a Th2 response. Th1 responses are more effective in combating intracellular pathogens like malaria parasites.Th1 and Th2 responses are two distinct types of immune responses generated by T-helper cells (Th cells) in the adaptive immune system. They play essential roles in shaping the body's defense mechanisms against different types of pathogens and foreign substances.

Lower IL6 levels are strongly correlated with an elevated likelihood of "Malaria All", emphasizing the importance of IL6 in malaria-related immune dynamics. Indicating that the body is modulating its inflammatory response. IL6 is involved in fever response and acute-phase reactions.

Elevated IL10 levels are observed during Malaria, indicating a regulatory response to control excessive inflammation and tissue damage caused by the infection.

Lower IL17 levels are strongly correlated with an increased likelihood of Malaria. This highlights the role of IL17 in modulating immune responses in malaria, which may be a mechanism to limit tissue damage associated with malaria infection.

Lower GMCSF levels are associated with a heightened likelihood of Malaria, emphasizing the significance of this factor in malaria-related immunity. This may be reflecting the down regulation of granulocyte and macrophage production in response to malaria, potentially as a means to control inflammation and tissue damage.

Reduced MIP1a levels during Malaria suggest a modulation of the chemotactic response for macrophages, which are key players in the immune response to malaria.

MIP1B is highly significant in the model. Higher MIP1B levels are associated with an increased likelihood of Malaria, highlighting its importance in malaria-related immune responses. This also indicates an enhanced chemotactic response for macrophages during malaria, potentially recruiting more immune cells to combat the infection. Lower IP10 levels may be associated with a decrease in IFNg-mediated chemotaxis of immune cells, which could impact the immune response to malaria.

Variable	Estimate	Std. Error	z value	p-value
(Intercept)	5.132×10^{-1}	1.159×10^{0}	0.443	0.657882
ReticPer	1.555×10^{-1}	8.934×10^{-2}	1.740	0.081820
RBCx1012L	-4.020×10^{-1}	1.894×10^{-1}	-2.123	$\mathbf{0.033783}$
IFNg	2.619×10^{-2}	1.016×10^{-2}	2.579	$\mathbf{0.009922}$
IL1Ra	-3.351×10^{-4}	9.275×10^{-5}	-3.613	$\mathbf{0.000303}$
IL2R	8.753×10^{-4}	2.236×10^{-4}	3.915	$\mathbf{9.06 \times 10^{-5}}$
IL4	-6.894×10^{-2}	1.523×10^{-2}	-4.528	$\mathbf{5.95 \times 10^{-6}}$
IL6	-2.438×10^{-3}	5.784×10^{-4}	-4.215	$\mathbf{2.50 \times 10^{-5}}$
IL10	9.369×10^{-3}	1.885×10^{-3}	4.970	$\mathbf{6.71 \times 10^{-7}}$
IL17	-1.987×10^{-2}	8.128×10^{-3}	-2.445	$\mathbf{0.014487}$
GMCSF	-1.743×10^{-3}	6.637×10^{-4}	-2.626	$\mathbf{0.008628}$
MIP1a	-1.385×10^{-3}	6.822×10^{-4}	-2.030	$\mathbf{0.042357}$
MIP1B	8.189×10^{-3}	1.367×10^{-3}	5.989	$\mathbf{2.11 \times 10^{-9}}$
IP10	-9.949×10^{-4}	2.770×10^{-4}	-3.592	$\mathbf{0.000328}$
Eotaxin	-2.022×10^{-2}	9.807×10^{-3}	-2.062	$\mathbf{0.039236}$
RPI_anaemic	9.890×10^{-1}	5.088×10^{-1}	1.944	0.051913
Null Deviance: 439.80 on 597 degrees of freedom **Residual Deviance**: 226.61 on 582 degrees of freedom **AIC**: 258.61				

Table 3.7: Logistic regression model for malaria. All the significant p-values are in bold.

The significant findings shed light on the body's intricate response to this parasitic disease. For instance, lower red blood cell counts (RBCx1012L) are associated with an increased likelihood of malaria infection, reflecting the well-known hemolytic effects of the parasite. Elevated levels of IFNg indicate an intensified immune response, likely triggered to combat the malaria parasite. The reduction in IL4 suggests a shift towards a Th1 immune response, more effective against intracellular pathogens like Plasmodium. Notably, a decrease in IL6 levels signifies the modulation of the inflammatory response. Higher levels of anti-inflammatory cytokine IL10 suggest a regulatory mechanism to control inflammation and tissue damage. The changes in other cytokines (IL2R, IL17, GMCSF, MIP1a, MIP1B) and chemokines (IP10) illustrate a complex immune response involving T-cell activation, chemotaxis modulation, and macrophage recruitment, all aimed at optimizing the host's defense against malaria. These findings emphasize the dynamic and multifaceted nature of the immune response to malaria, demonstrating how the body orchestrates a balanced yet effective defense against the parasite while striving to limit excessive inflammation and immunopathology.

Further investigation is needed when it comes to these metric variables showing the immune response in individuals with malaria, anemia, and SMA.

3.2.2 Linear Regression

A linear regression model will show us how some metric variables will behave when other variables are varying. We are interested to know how and what kind of pathological or immune response our study group shows. We will be looking into Interferon-gamma and how other measurements react to the changes in interferon-gamma. Table 3.8 shows how the changes in the measurement of interferon-gamma will change the other immune response variables in our study group.

Variable	Estimate	Std. Error	t value	p-value
(Intercept)	1.026×10^{2}	2.436×10^{1}	4.210	$\mathbf{3.04e-05}$
MO_A	-2.510×10^{0}	1.622×10^{0}	-1.548	0.122387
RBCx1012L	6.613×10^{0}	2.981×10^{0}	2.219	$\mathbf{0.026969}$
Hct	-1.407×10^{0}	5.123×10^{-1}	-2.746	$\mathbf{0.006258}$
MCHC	-1.676×10^{0}	5.479×10^{-1}	-3.058	$\mathbf{0.002351}$
RDW	-5.647×10^{-1}	3.364×10^{-1}	-1.679	0.093790
IL2R	-2.200×10^{-3}	8.753×10^{-4}	-2.513	$\mathbf{0.012289}$
IL4	2.410×10^{-1}	8.893×10^{-2}	2.709	$\mathbf{0.006977}$
IL5	2.234×10^{0}	5.153×10^{-1}	4.336	$\mathbf{1.76e-05}$
IL6	-1.684×10^{-2}	4.575×10^{-3}	-3.681	$\mathbf{0.000258}$
IL7	1.573×10^{-1}	5.537×10^{-2}	2.841	$\mathbf{0.004690}$
IL17	8.130×10^{-2}	5.017×10^{-2}	1.621	0.105757
IFNa	1.695×10^{-1}	2.567×10^{-2}	6.600	$\mathbf{1.08e-10}$
MIP1B	-9.392×10^{-3}	3.202×10^{-3}	-2.933	$\mathbf{0.003520}$
IP10	4.192×10^{-3}	1.732×10^{-3}	2.420	$\mathbf{0.015880}$
MIG	1.826×10^{-2}	4.306×10^{-3}	4.242	$\mathbf{2.65e-05}$
Eotaxin	-1.637×10^{-1}	5.667×10^{-2}	-2.889	$\mathbf{0.004035}$
RANTES	-1.123×10^{-5}	5.262×10^{-6}	-2.134	$\mathbf{0.033334}$
MCP1	1.233×10^{-2}	4.110×10^{-3}	3.001	$\mathbf{0.002832}$
RPI_anaemic	-1.246×10^{1}	3.324×10^{0}	-3.748	$\mathbf{0.000200}$
Residual standard error: 25.47 on 486 degrees of freedom **Multiple R-squared**: 0.364 **Adjusted R-squared**:0.3391 **F-statistic**: 14.64 on 19 and 486 DF, p-value: $< 2.2e-16$				

Table 3.8: Linear regression model for interferon-gamma. All the significant p-values are in bold.

An increase in RBC count was associated with a rise in IFNg levels. This may be attributed to the role of erythrocytes in immune responses, particularly during malaria infection. Higher RBC counts could indicate a more robust immune response.

Hct levels were negatively associated with IFNg levels. A decrease in hematocrit suggests a lower proportion of red blood cells, which might reflect the severity of anemia and potentially hinder the immune response.

Increased MCHC was negatively linked to IFNg levels. Higher MCHC may imply more concentrated hemoglobin within red blood cells, possibly affecting immune cell function or the severity of anemia.

The negative association between IL2R and IFNg suggests that higher levels of IL2R are associated with reduced IFNg production. This could be explained by the role of IL2R as part of the IL2 receptor complex, which is involved in T-cell activation and proliferation. A higher expression of IL2R may lead to increased IL2 signaling, which can promote the expansion of regulatory T-cells and suppress the Th1 immune response, where IFNg plays a critical role. Therefore, elevated IL2R levels may contribute to the down regulation of IFNg, potentially favoring immune tolerance or shifting toward Th2-type responses.

The positive association between IL4 and IFNg suggests that higher IL4 levels are associated with increased IFNg production. IL4 is known for its ability to promote Th2-type immune responses, characterized by the production of cytokines like IL4 and IL5. The positive correlation with IFNg

might indicate that in some contexts, IL4 could contribute to a balanced Th1/Th2 response. This could be particularly relevant in the context of malaria, where the immune response needs to be finely tuned to combat the parasite effectively.

The positive correlation between IL5 and IFNg implies that elevated IL5 levels are associated with increased IFNg production. IL5 is primarily associated with eosinophil activation and Th2-type responses. Its positive association with IFNg suggests a complex interplay between Th1 and Th2 responses during malaria infection. IL5 may contribute to the overall immune response by enhancing IFNg production.

The negative correlation between IL6 and IFNg indicates that higher levels of IL6 are associated with reduced IFNg production. IL6 has pleiotropic effects and can influence different immune responses. In this case, its negative correlation with IFNg might imply a regulatory role. IL6 could be dampening the production of IFNg or, conversely, a reduced IFNg response may lead to elevated IL6 levels as part of the immune regulatory feedback loop.

The positive correlation between IL7 and IFNg suggests that higher levels of IL7 are associated with increased IFNg production. IL7 is a cytokine known for its role in T-cell development and homeostasis. The positive association might indicate that IL7 contributes to the maintenance and expansion of IFNg-producing T-cells, enhancing the Th1-type immune response, which is crucial in combating infections like malaria.

The strong positive correlation between IFNa and IFNg is noteworthy. IFNa, part of the type I interferon family, is typically associated with antiviral responses. In the context of malaria, the positive association might imply that the host's immune system is activating multiple defense mechanisms, including both IFNa and IFNg, to combat the infection. This suggests a complex and multifaceted immune response.

The negative correlation between MIP1B and IFNg suggests that higher MIP1B levels are associated with reduced IFNg production. MIP1B is involved in the recruitment and activation of monocytes and macrophages. A negative association with IFNg might indicate that a strong monocyte/macrophage response, potentially driven by MIP1B, could down regulate the Th1 response.

The positive correlation between IP10 and IFNg indicates that higher IP10 levels are associated with increased IFNg production. IP10 is known for its role in recruiting T-cells and other immune cells to sites of infection. This positive association suggests that IP10 contributes to the recruitment and activation of IFNg-producing T-cells, reinforcing the Th1-type immune response.

The strong positive correlation between MIG and IFNg implies that higher MIG levels are associated with increased IFNg production. MIG is induced by IFNg and plays a role in recruiting immune cells, especially T-cells. The strong positive association underscores the feedback loop that can occur during a Th1 response, where IFNg induces MIG production, which in turn recruits more IFNg-producing T-cells.

The negative correlation between Eotaxin and IFNg suggests that higher Eotaxin levels are associated with reduced IFNg production. Eotaxin is involved in eosinophil recruitment and is typically associated with Th2-type responses. The negative correlation with IFNg could indicate that in some cases, Eotaxin might dampen the Th1 response, potentially favoring a shift towards a Th2 response during malaria infection.

The negative correlation between RANTES and IFNg implies that higher RANTES levels are associated with reduced IFNg production. RANTES plays a role in recruiting immune cells, including T-cells and monocytes. The negative association could suggest that during malaria infection, RANTES might contribute to a regulated immune response, potentially down regulating the Th1 response when necessary.

The positive correlation between MCP1 and IFNg indicates that higher MCP1 levels are associated with increased IFNg production. MCP1 is involved in the recruitment of monocytes, which can play a role in the immune response against malaria. The positive correlation suggests that MCP1 contributes to the recruitment of immune cells involved in Th1 responses.

Being a binary variable, the positive coefficient suggests that anemia is positively associated with IFNg production. Anemia can be a common complication of malaria. The positive association might indicate that, in this context, the presence of anemia triggers an immune response characterized by increased IFNg production, potentially to counter the effects of anemia and the underlying infection.

Other questions we would like to investigate are the rate of anemia events and the survivability of the subjects. For that, we move to the longitudinal analysis.

3.3 Longitudinal analysis

The Longitudinal data is not as extensive as the crossectional data. Thus, we make some variables indicating anemia, SMA, malaria, and anemia, as well as malaria and SMA. Anemia is defined as Hgb less than 11 and SMA is defined as Hgb less than 5.

Table 3.9 shows the other relevant variables for our analysis, along with the distribution of the data. 3.9 summarizes key characteristics within a sample population, distinguishing between the 2 cohorts. The median age of the subjects seems to be higher in the cohort, 2 suggesting that the people in the cohort 2, on average, are older than those in the other cohort. Hemoglobin levels also exhibit a significant difference $(p < 0.001)$ between the 2 cohorts, with slightly higher median values in cohort 2. Cohort-based differences are evident in HIV status $(p < 0.001)$, with a higher proportion of subjects from cohort 1 being HIV-negative and more subjects from cohort 1 testing HIV-positive. Sickle cell variants demonstrate significant cohort-based differences $(p < 0.001)$.

Furthermore, the presence of anemia is significantly different between both cohorts $(p < 0.001)$, with a higher prevalence in cohort one. Additionally, genetic variations show significant cohort-based differences, impacting the distribution of genotypes. These findings provide insights into how cohorts may show different health outcomes and genetic characteristics within this diverse sample population.

Characteristic	Whole Sample	cohort 1	cohort 2	p-value
N	$10,3311$	$7,2321$	$3,099$	
Age (mos)	$25(15,36)$	$22(13,34)$	$30(21,39)$	< 0.001
Hgbfinal	$9.80(8.40,11.01)$	$10.00(8.30,11.30)$	$9.55(8.60,10.48)$	< 0.001
HIV				< 0.001
0	$10,088(98\%)$	$7,000(97\%)$	$3,088(100\%)$	
1	$243(2.4\%)$	$232(3.2\%)$	$11(0.4\%)$	
HbAS				< 0.001
AA	$8,824(85\%)$	$6,238(86\%)$	$2,586(83\%)$	
AS	$1,406(14\%)$	$893(12\%)$	$513(17\%)$	
SS	$101(1.0\%)$	$101(1.4\%)$	$0(0\%)$	
RPI_anaemic	$7,439(72\%)$	$4,846(67\%)$	$2,593(84\%)$	< 0.001
RPI_SMA	$304(2.9\%)$	$269(3.7\%)$	$35(1.1\%)$	< 0.001
Malaria_all				< 0.001
0	$6,011(58\%)$	$4,357(60\%)$	$1,654(53\%)$	
1	$4,320(42\%)$	$2,875(40\%)$	$1,445(47\%)$	
mal_SMA	$196(1.9\%)$	$174(2.4\%)$	$22(0.7\%)$	< 0.001
mal_anaemic	$3,618(35\%)$	$2,325(32\%)$	$1,293(42\%)$	< 0.001
IFNG_A_1616G				< 0.001
AA	$2,809(27\%)$	$2,616(36\%)$	$193(6.2\%)$	
AG	$5,161(50\%)$	$3,366(47\%)$	$1,795(58\%)$	
GG	$2,361(23\%)$	$1,250(17\%)$	$1,111(36\%)$	
IFNG_G_183T				< 0.001
GG	$9,647(93\%)$	$6,762(94\%)$	$2,885(93\%)$	
GT	$638(6.2\%)$	$424(5.9\%)$	$214(6.9\%)$	
TT	$46(0.4\%)$	$46(0.6\%)$	$0(0\%)$	

Table 3.9: Characteristics of the whole sample and cohort stratification.

3.3.1 Poisson Regression

The initial analysis we do will be Poisson rate regression. The primary focus of the analysis is to estimate the rate of SMA events, anemia events, and malaria events, which can be thought of as the number of events occurring over a specific period in a given population or group of individuals. The use of the offset as the logarithm of age in months accounts for varying observation periods among individuals. The observation periods are from the birth of the child to the last visit. In some cases, it could be that the first visit is the only visit. This is crucial as the rate of SMA events is expected to increase with longer observation times. The offset allows us to focus on the event rate per unit of observation time, providing a more accurate assessment of the impact of predictor variables as the individuals grow older. The first response variable we investigate will be the number of SMA events. The results of the Poisson regression are shown in Table 3.10.

Cohort 2 demonstrates a significantly lower risk of SMA compared to cohort 1, indicating that individuals in the second cohort may experience a protective effect against severe malaria. This finding aligns with the previous model suggesting potential advancements in preventative measures or changes in environmental conditions between cohorts.

Examining the influence of hemoglobin genotypes, individuals with “AS” sickle cell trait exhibit a decreased rate of SMA, underlining the protective role of the sickle cell trait. On the contrary, those with “SS” sickle cell disease show an elevated rate of SMA events, even though statistically sickle cell disease is not significant the model emphasizes the importance of understanding the diverse impact of different hemoglobin genotypes on malaria susceptibility.

The presence of HIV is associated with an increased risk of SMA, reinforcing the well-established relationship between immunosuppression and malaria severity. Additionally, the genetic factor “GT” for interferon-gamma exhibits a marginally significant negative association with SMA events, suggesting a potential protective effect. The observed statistical significance is marginally above the predetermined threshold, although not definitive. Nevertheless, we have decided to consider it significant, given the context of the analysis and the available data. However, the association of “TT” genetic factor is not statistically significant, indicating a need for further exploration.

Variable	Estimate	Std. Error	z value	p-value
(Intercept)	-4.19435	0.06701	-62.591	$\mathbf{< 2 \times 10^{-16}}$
cohort-2	-1.52364	0.18070	-8.432	$\mathbf{< 2 \times 10^{-16}}$
HbAS-AS	-0.75279	0.22184	-3.393	$\mathbf{0.00069}$
HbAS-SS	0.56994	0.35981	1.584	0.11320
HIV	0.75095	0.24427	3.074	$\mathbf{0.00211}$
ifngg-GT	-0.58006	0.30765	-1.885	0.05937
ifngg-TT	0.44214	0.71027	0.622	0.53362
Null deviance: 1097.81 on 836 degrees of freedom **Residual deviance**: 953.71 on 830 degrees of freedom **AIC**: 1420.1 **Number of Fisher Scoring iterations**: 6				

Table 3.10: Poisson regression model summary for SMA event counts. All the significant p-values are in bold.

We examined a range of variables, including demographic, clinical, genetic, and observational factors, to quantify their impact on SMA rates. The results revealed that certain variables, such as HIV status, cohort, sickle cell, and genetic variations in interferon gamma, had associations with the rate of SMA events. Furthermore, the inclusion of an offset variable, representing the logarithm of age in months for each individual, accounted for variations as they grew older and enhanced the accuracy of the analysis.

The Poisson rate regression model shown in 3.11 was used to investigate the count of anemia cases, with various independent variable variables, indicating Sex, cohort, sickle cell trait, alpha-thalassemia genetic variant, HIV status, and genetic variations in interferon-gamma. We continue to use the same offset, which is the logarithm of age in months, as in the previous model.

The Poisson regression model for anemia events offers valuable insights into the factors influencing the occurrence of anemia among the study population. Notably, the variable “Sex” demonstrates significance, with females exhibiting a lower risk of anemia compared to males. This highlights gender-based differences in susceptibility to anemia, potentially influenced by hormonal and genetic factors.

Variable	Estimate	Std. Error	z value	p-value
(Intercept)	-1.33643	0.02590	-51.607	$\mathbf{< 2 \times 10^{-16}}$
Sex-F	-0.11513	0.02330	-4.942	$\mathbf{7.72 \times 10^{-7}}$
cohort-2	-0.19992	0.02605	-7.674	$\mathbf{1.66 \times 10^{-14}}$
HbAS-AS	-0.10736	0.03344	-3.211	$\mathbf{0.00132}$
HbAS-SS	0.20156	0.10303	1.956	0.05042
ifnga-AG	0.06271	0.02970	2.111	$\mathbf{0.03474}$
ifnga-GG	0.10139	0.03508	2.890	$\mathbf{0.00385}$
Null deviance: 2273.9 on 836 degrees of freedom **Residual deviance**: 2169.0 on 830 degrees of freedom **AIC**: 5286.2 **Number of Fisher Scoring iterations**: 5				

Table 3.11: Poisson regression model summary for anemia event counts. All the significant p-values are in bold.

Cohort-related differences are observed, with cohort 2 showing a decrease in the risk of anemia compared to cohort 1. This result prompts further investigation into cohort-specific factors that may contribute to variations in anemia prevalence.

Examining hemoglobin genotypes, individuals with “AS” exhibit a reduced risk of anemia, emphasizing the protective effect of the sickle cell trait. Conversely, those with “SS” show an increased risk marginally significant, highlighting the diverse impact of different hemoglobin genotypes on anemia susceptibility. Furthermore, the genetic factors of interferon-alpha “GG” and “AG” demonstrate a significant positive association with anemia events, suggesting a potential role of this genetic variant in influencing anemia susceptibility.

The count of malaria cases was also analyzed with the Poisson rate regression and shown in Table 3.12. The offset variable considered in this model also represents the logarithm of age in months for each individual.

Variable	Estimate	Std. Error	z value	p-value
(Intercept)	-0.56789	0.02041	-27.827	$\mathbf{< 2 \times 10^{-16}}$
Sex-F	-0.06942	0.01670	-4.156	$\mathbf{3.24 \times 10^{-5}}$
cohort-2	-0.37307	0.01911	-19.517	$\mathbf{< 2 \times 10^{-16}}$
HbAS-AS	-0.15619	0.02445	-6.388	$\mathbf{1.68 \times 10^{-10}}$
HbAS-SS	-0.34944	0.09334	-3.744	$\mathbf{0.000181}$
alphathal-a/aa	0.01097	0.01922	0.571	0.568191
alphathal-a/a	-0.04286	0.02182	-1.965	$\mathbf{0.049450}$
HIV	-0.21072	0.05853	-3.600	$\mathbf{0.000318}$
ifnga-AG	0.02223	0.02074	1.072	0.283800
ifnga-GG	0.05215	0.02487	2.097	$\mathbf{0.035973}$
ifngg-GT	-0.02432	0.03466	-0.702	0.482928
ifngg-TT	0.40384	0.12179	3.316	$\mathbf{0.000914}$
Null deviance: 3941.7 on 836 degrees of freedom **Residual deviance**: 3417.0 on 825 degrees of freedom **AIC**: 7119.5 **Number of Fisher Scoring iterations**: 4				

Table 3.12: Poisson regression model summary for malaria count. All the significant p-values are in bold.

Individuals in cohort 2 demonstrated a significantly reduced risk of malaria compared to cohort 1. This observation could be indicative of improved preventative measures or changes in environmental conditions over time. The protective effect associated with being female suggests potential gender-specific differences in exposure or susceptibility to malaria, warranting further investigation.

The influence of hemoglobin genotypes was noteworthy. Both “AS” sickle cell trait and “SS” sickle cell disease were associated with a lower risk of malaria, aligning with existing literature on the protective role of hemoglobin variants against malaria infection. This reinforces the importance of genetic factors in modulating susceptibility to malaria within the studied population.

Surprisingly, individuals living with HIV exhibited a decreased risk of malaria. While this finding might seem counterintuitive, it could be linked to immune system interactions or variations in healthcare-seeking behavior among HIV-positive individuals. Understanding the mechanisms behind this association could contribute to better-targeted interventions for malaria prevention in the context of HIV.

Genetic variations in interferon-gamma and interferon-alpha genes also played a role. Individuals with the “GG” genotype had a higher risk of malaria, emphasizing the potential impact of host immune response pathways. Meanwhile, those with the “TT” genotype faced a significantly increased risk, highlighting the complex interplay between genetic factors and susceptibility to malaria.

3.3.2 Cox PH models

Survival analysis is an essential component of medical research and epidemiology, providing insights into the temporal dimension of health events. It delves beyond static cross-sectional data, offering valuable information on the manifestation of specific health conditions like anemia and severe malarial anemia (SMA) over time. When time is critical and events are recurrent, the Cox proportional hazards model emerges as a powerful statistical tool.

The Cox model allows us to examine the factors contributing to anemia and SMA, going beyond binary outcomes. We can explore how age impacts the onset of these conditions and the development of risk factors. The model’s ability to account for censoring in the data ensures a more accurate estimation of survival probabilities, especially when dealing with longitudinal studies. Time-varying covariates also provide a unique perspective, allowing us to assess the dynamic nature of risk factors in the context of anemia and SMA.

The Cox model offers us a comprehensive framework for understanding the progression of anemia in a population. It allows us to identify the critical factors influencing these health events and evaluate how they change over time, affecting the likelihood of experiencing anemia. Even though we have quite a few variables in the model, only a few of them seem to be significant statistically. cohort 2 seems to have a higher risk of attaining anemia. The variable “first.visit” marks the age of the individual at the first visit. The negative coefficient suggests individuals with higher age had a lower hazard for anemia. Males have a higher hazard of anemia, suggested by the positive coefficient. Subjects with malaria have an association with anemia, which can be seen here in the model. Malaria

Variable	Coefficient	Exp(Coeff)	SE(Coeff)	Robust SE	p-value
cohort-2	0.119239	1.126639	0.029387	0.038261	**0.001830**
first.visit	−0.036524	0.964135	0.002106	0.003201	$<$ **2** $\times$ **10**$^{-16}$
sex-M	0.101502	1.106832	0.023594	0.032131	**0.001583**
HIV	0.117435	1.124609	0.076308	0.095387	0.218270
HbAS-AS	−0.060591	0.941208	0.033501	0.042081	0.149909
HbAS-SS	0.323815	1.382392	0.104032	0.213507	0.129355
Malaria_all	−0.217252	0.804727	0.024145	0.031887	**9.54** $\times$ **10**$^{-12}$
IFNG_A_1616G-AG	0.062329	1.064313	0.029891	0.042167	0.139370
IFNG_A_1616G-GG	0.058301	1.060034	0.035457	0.048908	0.233236
num_anemia_evnt	0.018473	1.018645	0.003400	0.005088	**0.000283**
Concordance= 0.628 (se = 0.005) **Likelihood ratio test**= 936.4 on 10 df, p=$< 2 \times 10^{-16}$ **Wald test**= 632 on 10 df, p=$< 2 \times 10^{-16}$ **Score (log-rank) test**: 894.1 on 10 df, p=$< 2 \times 10^{-16}$ **Robust**:244.3, p=$< 2 \times 10^{-16}$					

Table 3.13: Cox proportional-hazards model summary for anemia. All the significant p-values are in bold.

usually tends to evolve into anemia because of a lack of healthy red blood cells. The number of events is statistically significant. In the context of the model, it signifies that subjects with more incidents of anemia were at higher risk of attaining anemia again in the future.

A similar Cox model is used to investigate SMA 3.14. A reminder that in our context, SMA is defined as Hgb less than 5. The condition is an extreme case of malarial anemia. We will use the age of the subjects as the time to event and model the multiple events. The following Table shows us the results of our Cox model for SMA.

Variable	Coefficient	Exp(Coeff)	SE(Coeff)	Robust SE	p-value
cohort-2	−0.967506	0.380030	0.196477	0.241430	**6.14e** $-$ **05**
first.visit	−0.032429	0.968091	0.009406	0.011701	**0.005582**
sexM	−0.184526	0.831499	0.117937	0.125815	0.142474
HIV	0.860578	2.364526	0.248387	0.260065	**0.000936**
HbAS-AS	−0.475905	0.621323	0.224375	0.257427	0.064501
HbAS-SS	0.890488	2.436318	0.364501	0.607996	0.143022
Malaria	0.915807	2.498790	0.122025	0.182467	**5.19e** $-$ **07**
num_SMA_evnt	0.533959	1.705671	0.042413	0.058455	$<$ **2e** $-$ **16**
Concordance: 0.814 (se = 0.012) **Likelihood ratio test**: 315.2 on 8 df, p < 2×10^{-16} **Wald test**: 210.6 on 8 df, p < 2×10^{-16} **Score (log-rank) test**: 624.5 on 8 df, p < 2×10^{-16} **Robust**: 127.1, p < 2×10^{-16}					

Table 3.14: Cox proportional-hazards model summary for SMA. All the significant p-values are in bold.

The cohort variable shows a statistically significant impact on the risk of experiencing SMA events. Cohort has a negative coefficient for cohort 2 signifying that hazard for cohort 2 experiencing SMA events is less compared to those in cohort 1. This could be because the individuals in cohort 2 tend to have higher age, which could give some protection from SMA.

A negative coefficient for the first visit suggests that individuals who are older during their first visit have a lower risk of having SMA compared to younger individuals. HIV has a positive coefficient, suggesting that individuals with HIV have a higher risk of getting anemia. If a person with SMA is also living with HIV, their overall health may be more fragile, and managing both conditions can be challenging due to the impact of HIV on the immune system.

Malaria-positive individuals have a higher hazard of having SMA. Naturally, malaria brings down the immune and pathological system of the body, lowering the Hgb and making a severe anemia risk higher. The number of SMA events is seen to have a positive coefficient, suggesting previous cases of SMA can increase the risk of SMA in the future.

Applying Cox proportional hazards models to explore severe malaria (SMA) and anemia events has yielded valuable insights into associated risk factors within our study population. For SMA, significant determinants include cohort membership, HIV status, and the number of previous SMA events. Cohort 2 individuals exhibit a reduced risk, while those with HIV face an increased risk. A history of multiple SMA events significantly heightens recurrence risk, emphasizing the need for its consideration in risk assessment. Regarding anemia, our findings highlight the roles of cohort 2, age at the first visit, and malaria status in influencing risk. Cohort 2 individuals display a lower risk of anemia, suggesting potential protective factors. Younger age at the first visit correlates with a decreased risk, emphasizing the importance of early healthcare access. Individuals with malaria face an elevated risk of anemia, emphasizing the intricate interplay between these health conditions.

3.3.3 Frailty models

Frailty models acknowledge that not all relevant risk factors may be measured or observable. These unmeasured variables can introduce substantial variations in the incidence of anemia and SMA across different study populations. By incorporating frailty, we can better capture this unobservable variation and make the analyses more robust.

We will run a frailty model for recurring events of anemia and SMA with our data. The Table 3.15 shows the result of the frailty model for anemia.

This variable represents the cohort membership. In this context, cohort 2 exhibited a statistically significant higher risk of anemia events compared to cohort 1. This suggests that individuals in cohort 2 may be exposed to unique risk factors or environmental conditions contributing to a greater likelihood of developing anemia.

As first. visit marks the age at the first visit and the negative coefficient for this variable indicates a higher age at the first visit, which was significantly associated with a reduced risk of anemia events. A positive coefficient for males suggests males have a higher risk of anemia. The presence of “SS” traits showed a statistically significant positive effect on the risk of anemia events. Sickle cell anemia is a genetic disorder characterized by abnormal hemoglobin production, leading to the formation of sickle-shaped red blood cells, which can reduce the oxygen-carrying capacity of the blood. This may result in a higher risk of anemia.

Variable	Coefficient	se(coef)	p-value
cohort-2	0.112817	0.045282	$\mathbf{1.3e-02}$
first.visit	-0.047702	0.002464	$\mathbf{1.7e-83}$
sex-M	0.102213	0.036427	$\mathbf{5.0e-03}$
HIV	0.188190	0.117188	$1.1e-01$
HbAS-AS	-0.056489	0.051191	$2.7e-01$
HbAS-SS	0.400632	0.172011	$\mathbf{2.0e-02}$
ALPHATHAL-a/aa	-0.007765	0.042378	$8.5e-01$
ALPHATHAL-a/a	0.006878	0.046693	$8.8e-01$
Malaria	-0.125739	0.026343	$\mathbf{1.8e-06}$
IFNG_A_1616G-AG	0.053549	0.046626	$2.5e-01$
IFNG_A_1616G-GG	0.040517	0.055136	$4.6e-01$
IFNG_G_183T-GT	-0.007729	0.076232	$9.2e-01$
IFNG_G_183T-TT	0.131687	0.329549	$6.9e-01$
frailty(StudyNo)			$\mathbf{1.6e-34}$
Variance of random effect = 0.1316258, l-likelihood = -46492.5			
Concordance: 0.668 (se = 0.005)			
Likelihood ratio test: 2097 on 413 df, p <2×10^{-16}			

Table 3.15: Frailty model summary for anemia. All the significant p-values are in bold.

The presence of malaria was strongly associated with a higher risk of anemia. Malaria is a known cause of hemolytic anemia, where red blood cells are destroyed by the malaria parasite. Effective malaria control measures, including vector control and antimalarial drugs, are vital in regions where both anemia and malaria are prevalent.

The statistical significance of frailty, represented by "StudyNo", as a random effect is noteworthy. Frailty accounts for unobserved differences among groups, indicating that variances in anemia risk are linked to differences in groups or our case individuals having multiple events. This suggests that there is heterogeneity in anemia risk among different individuals.

Other variables seen in the result are not statistically significant, but they contribute to a better model. They may or may not be relevant to our model, but we cannot say anything for sure. Further investigation is needed to look into these variables thoroughly.

Variable	Coefficient	se(coef)	p-value
cohort-2	-1.149311	0.192013	$\mathbf{2.16 \times 10^{-9}}$
first.visit	-0.046020	0.009509	$\mathbf{1.30 \times 10^{-6}}$
HIV	0.704409	0.245175	$\mathbf{0.00406}$
HbAS-AS	-0.674488	0.222045	$\mathbf{0.00238}$
HbAS-SS	0.800560	0.364068	$\mathbf{0.02788}$
Malaria	0.309696	0.128094	$\mathbf{0.01562}$
Concordance = 0.729 (se = 0.013)			
Likelihood ratio test: 153.1 on 6 df, p < 2×10^{-16}			
Wald test: 114.6 on 6 df, p < 2×10^{-16}			
Score (log-rank) test: 133.1 on 6 df, p < 2×10^{-16}			

Table 3.16: Frailty model summary for SMA. All the significant p-values are in bold.

In this frailty model exploring the risk of SMA, several key factors emerged as significant contributors. The second cohort exhibited a reduced risk of SMA events, indicating potential advancements or protective measures introduced during that timeframe. Age at the first visit played a pivotal role, with a higher age at the initial healthcare encounter associated with a decreased risk of SMA events. Notably, individuals with HIV demonstrated an increased risk, highlighting the heightened vulnerability of this population to complications such as SMA. Genetic factors further influenced the risk, with the sickle cell "AS" trait associated with a lower risk, while the sickle cell trait "SS" indicated an elevated risk of SMA. Malaria infection emerged as a significant risk factor, reaffirming the well-established connection between malaria and SMA. Despite the absence of a significant frailty term, the findings contribute to a nuanced understanding of the multifaceted determinants of SMA events.

3.4 Result summary

Several immune response-related variables exhibit significant associations with the likelihood of anemia, shown in Table 3.5. Lower levels of Interferon-Gamma (IFNg), Monokine Induced by Interferon-Gamma (MIG), and Monocyte Chemo attractant Protein-1 (MCP1) are linked to a reduced likelihood of anemia, potentially indicating variations in immune function, immune cell activation, and inflammatory responses that may influence anemia risk. On the other hand, higher levels of Interferon-Induced Protein 10 (IP10) are associated with an increased likelihood of anemia, suggesting heightened immune activation and inflammation in individuals with anemia.

Based on the Table 3.6, individuals with SMA exhibit a distinct immune response that differs in several ways from that observed in individuals with anemia. Moreover, the blood-related measurements of SMA patients also seem to display a more pronounced reaction compared to those of other individuals.

Table 3.7 reveals insights into the body's response to malaria. Lower red blood cell counts indicate increased susceptibility, and elevated IFNg levels suggest an intensified immune response. A decrease in IL4 signals a shift towards a more effective Th1 immune response. Reduced IL6 levels indicate inflammatory response modulation, while higher IL10 levels suggest a regulatory mechanism. Changes in other cytokines and chemokines illustrate a complex immune response, optimizing the host's defense against malaria and emphasizing the dynamic nature of the body's response.

The conducted linear regression analysis in Table 3.8 also revealed significant insights into the intricate interplay of clinical, hematological, and immunological factors in the context of malaria-induced anemia and immune response. These findings emphasize the multifaceted nature of malarial anemia, where variables such as red blood cell count, hematocrit, and mean corpuscular hemoglobin concentration (MCHC) play pivotal roles in reflecting the severity of anemia. Furthermore, specific immune factors, such as IFNa (Interferon Alpha), MIP1B (Macrophage Inflammatory Protein 1 Beta), and Eotaxin, exhibit significant associations with the immune response, highlighting their involvement in the complex host-pathogen interaction. While we observed a significant correlation between these factors and IFNg production, further investigation is warranted to decipher the precise mechanisms underlying this relationship.

The Poisson regression analysis delved into the nuanced rates of anemia 3.10, SMA events 3.11, and malaria events 3.12, revealing key contributors like participant cohorts, sex, sickle cell genotype variations, HIV status, and specific immune response-related gene variants. This comprehensive examination underscores the intricate interplay of genetic, environmental, and immune factors, collectively shaping the prevalence of anemia, SMA, and malaria events within the studied population.

The application of Cox proportional hazards models in analyzing anemia 3.13 and SMA 3.14 events over time. The models reveal significant risk factors for both conditions. In the anemia model, cohort 2, age at first visit, male gender, malaria, and the number of previous anemia events are significant predictors. For SMA, cohort membership, age at first visit, HIV status, malaria, and the number of previous SMA events are influential factors. Cohort 2 individuals have a lower risk of both conditions, while HIV-positive individuals face an increased risk of SMA. Recurrence of previous events significantly heightens the risk for both anemia and SMA. The findings underscore the complex interplay between various factors in influencing the risk of these health conditions.

The frailty model in 3.15 for anemia identifies significant risk factors, including cohort membership, age at the first visit, gender, sickle cell traits, and malaria infection. Cohort 2 exhibits a higher risk of anemia, and various genetic and environmental factors contribute to the risk. The frailty term indicates unobserved differences among groups.

For SMA in 3.16, the frailty model identifies key factors such as cohort, age at first visit, HIV status, sickle cell traits, and malaria infection. Cohort 2 shows a reduced risk of SMA, while HIV-positive individuals and those with the sickle cell trait “SS” have an increased risk. The findings underscore the multifaceted determinants of SMA events.

4 Conclusion

In our pursuit of understanding the complexities surrounding malaria-induced anemia and severe malarial anemia (SMA), our study has provided a nuanced examination of various factors contributing to these health challenges. By employing a comprehensive dataset covering both static and dynamic aspects, we've uncovered intricate details across clinical, hematological, and immunological domains.

Our cross-sectional analysis unveiled a symphony of immunological interactions during malaria infection. Notable correlations between immune markers such as IP10, MIG, and IFNg indicated the orchestration of T-cell responses. Conversely, negative associations of Eotaxin and RANTES with IFNg hinted at potential regulatory roles, possibly favoring alternative immune responses. The positive link between MCP1 and IFNg shed light on monocyte recruitment in the immune response against malaria. Anemia, a common complication, showed a positive association with IFNg production, suggesting an adaptive immune response to counteract both anemia and the infection.

Our longitudinal analysis delved into cohort-specific variations, unveiling patterns in demographics, clinical profiles, and genetic traits. Cohort 2 stood out with a higher median age and prevalent genetic variations. Poisson regression models, a statistical tool employed in our analysis, highlighted the impact of interferon-gamma gene variants on SMA rates. Specifically, we observed that individuals with certain genetic variations were more prone to developing severe malarial anemia. This finding adds a genetic dimension to our understanding, emphasizing the need for personalized approaches in malaria management.

Survival analysis, using Cox proportional hazards models, introduced a temporal perspective. Cohort dynamics significantly influenced SMA risk, with cohort 2 displaying a lower risk. Age at the first visit emerged as a crucial factor, with older individuals showing reduced risks of anemia and SMA. Individuals with HIV were more vulnerable to severe malarial anemia, emphasizing the need for targeted interventions in this population.

Integrating frailty models acknowledged unobservable variations contributing to anemia risk heterogeneity. The frailty term captured nuances often overlooked by conventional models, emphasizing the importance of unobserved factors. This nuanced approach revealed diverse pathways influencing anemia risk.

In conclusion, our study offers a comprehensive view of factors influencing malaria-induced anemia and SMA. From immunological intricacies to temporal dynamics and cohort-specific influences, each layer of exploration contributes valuable insights. The incorporation of Poisson regression models, elucidating the genetic underpinnings of severe malarial anemia, represents a significant stride in unraveling the complexities of host responses to malaria-induced complications. The path forward involves deeper molecular investigations and targeted studies to refine our understanding and inform tailored interventions.

Appendix A: R code for analysis

All the code given here is after the original data has been processed and cleaned properly so to maintain the quality of the data.

A.1 Crossectional Analysis

The following packages were used while working in R.

```
library(readr)
library(dplyr)
library(readxl)
library(psych)
library(stats)
library(effsize)
library(corrplot)
library(MASS)
library(gmodels)
library(epitools)
library(expss)
library(labelled)
library(descr)
library(data.table)
library(tidyverse)
library(gtsummary)
```

3.1 and 3.2 respectively can be obtained using the crossectional data as the dataset and using the following code.

```
crosstab(data_og$RPI_anaemic,data_og$‘Malaria All‘,
missing.include = FALSE,prop.r = T, prop.c = T,row.labels = F,
cell.layout = T, format = "SPSS",prop.t = T, prop.chisq = F ,
chisq = T,plot= FALSE, dnn = c("Anemia cases", "Malaria
cases"))
```

```
crosstab(data_og$RPI_SMA,data_og$‘Malaria All‘,
missing.include = FALSE, prop.r = T, prop.c = T, row.labels =
F , cell.layout = T,format = "SPSS", prop.t = T, prop.chisq =
F , chisq = T, plot= F, dnn = c("SMA cases", "Malaria cases"))
```

To procure the results of Table 3.3 the given code must be used.

```
stats_total<-data_og %>% select(c(RPI, ReticPer, RBCx1012L,
IFNg, IFNa, HgbgdL,Hct)) %>%
tbl_summary(statistic = list(all_continuous() ~ "{median}
({sd}) [{mean}] {IQR}"), missing ="no")

stats_group<-data_og %>% select(c(RPI,ReticPer,RBCx1012L,
IFNg, IFNa, HgbgdL, Hct, `Malaria All`)) %>%
tbl_summary(by=`Malaria All`, statistic =
list(all_continuous() ~ "{median} ({sd}) [{mean}]
{IQR}"),missing ="no")%>%add_p()

tbl_merge(tbls = list(stats_total,stats_group),tab_spanner =
c("Whole Sample", "Grouped Sample"))
```

The logistic regression results are displayed in Tables 3.5, 3.6, and 3.7. All tables are produced by similar code with varying input variables. Changing the response variable yields different results.

```
stepAIC(glm(RPI_anaemic ~ ., data = na.omit(data_immune),
family = binomial), direction='both',trace = 20)
```

Linear regression analysis shown in the Table 3.8 is produced by the code as follows.

```
stepAIC(lm(IFNg ~ ., data = na.omit(data_m)),
direction='both',trace = 20)
```

A.2 Longitudinal Analysis

The packages used for this part of the analysis in R are as follows.

```
library(readr)
library(dplyr)
library(readxl)
library(psych)
library(stats)
library(effsize)
library(corrplot)
library(MASS)
library(tidyverse)
library(survival)
library(coxme)
library(writexl)
library(report)
```

```
library(sjPlot)
library(ggplot2)
library(survminer)
library(gtsummary)
```

We have created a new data set for Poisson rate regression by selecting the count data of the event from the original dataset. We used a similar code to produce all three Poisson rate regression models but with different response variables. The offset used was Age in months and was consistent across all models. The results have been tabulated in Tables 3.10, 3.11, and 3.12 for SMA, Anemia, and Malaria counts, respectively.

```
stepAIC(glm(sma_counts ~ Sex + Cohort + HbAS + alphathal + hiv
+ ifnga + ifngg + offset(log(Age_mos)), data = rpi_counts,
family = poisson), direction='both',trace = 20)
```

For the Cox model, we have created a particular dataset for each model explaining a particular response variable. Table 3.13 shows us the results of the code given below. The code also includes the part used for creating the dataset. Replacing the response variable with the SMA event indicator in the following code will also give us the results of Table 3.14.

```
data_m$last.visit < unlist(lapply(split(data_m$Age_mos,data_m$StudyNo),
function(x) rep(x[length(x)],length(x))))

data_m$first.visit <-unlist(lapply(split(data_m$Age_mos,data_m$StudyNo),
function(x) rep(x[1],length(x))))

data_m11 <- data_m[data_m$Age_mos==data_m$last.visit |
data_m$RPI_anaemic=="1",]

data_m11$prev.event <- unlist ( lapply(split(data_m11$Age_mos,
data_m11$StudyNo),function(x)c(0,x)[1:length(x)]))

data_m11<- data_m11 %>% group_by(StudyNo) %>% mutate(num_anemia_evnt =
cumsum(RPI_anaemic)) %>%ungroup()

data_m11 <- data_m11 %>% mutate(num_anemia_evnt = case_when(num_anemia_evnt
== 0 ~ 0, num_anemia_evnt > 0 ~ num_anemia_evnt - 1))

cx_m11<- stepAIC (coxph(Surv (data_m11$prev.event, data_m11$Age_mos,
as.numeric(data_m11$RPI_anaemic), type = "counting") ~Cohort+ first.visit
+ sex+ HIV1 + HbAS+ALPHATHAL + Malaria_all + IFNG_A_1616G + IFNG_G_183T+
num_anemia_evnt, data = data_m11,id = cluster(StudyNo) ),direction =
'both', trace = 20 )
```

The frailty model shown in Tables 3.15 and 3.16 are produced by the same code but with different response variables. The data used in both is the longitudinal data after cleaning and processing the data.

```
stepAIC(coxph(Surv(data_m$prev.event, data_m$Age_mos, data_m$RPI_anaemic,
type = "counting")~Cohort+ first.visit+sex+ HIV1+HbAS+ALPHATHAL +
Malaria_all + IFNG_A_1616G + IFNG_G_183T +frailty(StudyNo), data = data_m
),direction = 'both', trace = 20 )
```

Bibliography

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

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

[3] Gabriela Loredana Popa, Mircea Ioan Popa, et al. Recent advances in understanding the inflammatory response in malaria: a review of the dual role of cytokines. *Journal of immunology research*, 2021, 2021.

[4] Douglas J Perkins, Tom Were, Gregory C Davenport, Prakasha Kempaiah, James B Hittner, and John Michael Ong'echa. Severe malarial anemia: innate immunity and pathogenesis. *International journal of biological sciences*, 7(9):1427, 2011.

[5] Jake Turner, Meghana Parsi, and Madhu Badireddy. Anemia. In *StatPearls [Internet]*. StatPearls Publishing, 2022.

[6] Greanious Alfred Mavondo and Mayibongwe Louis Mzingwane. Severe malarial anemia (sma) pathophysiology and the use of phytotherapeutics as treatment options. *Curr Top Anemia*, pages 189–214, 2017.

[7] Berenger Kaboré, Annelies Post, Mike LT Berendsen, Salou Diallo, Palpouguini Lompo, Karim Derra, Eli Rouamba, Jan Jacobs, Halidou Tinto, Quirijn de Mast, et al. Red blood cell homeostasis in children and adults with and without asymptomatic malaria infection in burkina faso. *Plos one*, 15(11):e0242507, 2020.

[8] Roger S Riley, Jonathan M Ben-Ezra, Rajat Goel, and Ann Tidwell. Reticulocytes and reticulocyte enumeration. *Journal of Clinical Laboratory Analysis*, 15(5):267, 2001.

[9] Hyeoun-Ae Park. An introduction to logistic regression: from basic concepts to interpretation with particular attention to nursing domain. *Journal of Korean Academy of Nursing*, 43(2):154–164, 2013.

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

[11] Melinda Mills. Introducing survival and event history analysis. *Introducing Survival and Event History Analysis*, pages 1–300, 2010.

[12] David G Kleinbaum and Mitchel Klein. *Survival analysis a self-learning text*. Springer, 1996.

[13] Jill C Stoltzfus. Logistic regression: a brief primer. *Academic emergency medicine*, 18(10):1099–1104, 2011.

[14] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[15] Julian J Faraway. *Linear models with R*. Chapman and Hall/CRC, 2004.

[16] B Winter and PC Bürkner. Poisson regression for linguists: A tutorial introduction to modelling count data with brms. language and linguistics compass, 15 (11), article e12439, 2021.

[17] Louise Kuhn, Leslie L Davidson, and Maureen S Durkin. Use of poisson regression and time series analysis for detecting changes over time in rates of child injury following a prevention program. *American journal of epidemiology*, 140(10):943–955, 1994.

[18] D Wayne Osgood. Poisson-based regression analysis of aggregate crime rates. *Journal of quantitative criminology*, 16:21–43, 2000.

[19] 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.

[20] 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.

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

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

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

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

[25] Usha S Govindarajulu, Haiqun Lin, Kathryn L Lunetta, and RB D'Agostino Sr. Frailty models: applications to biomedical and genetic studies. *Statistics in medicine*, 30(22):2754–2764, 2011.

[26] D Anderson and K Burnham. Model selection and multi-model inference. *Second. NY: Springer-Verlag*, 63(2020):10, 2004.

[27] Kenneth P Burnham, David R Anderson, and Kathryn P Huyvaert. Aic model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. *Behavioral ecology and sociobiology*, 65:23–35, 2011.

[28] Eric-Jan Wagenmakers and Simon Farrell. Aic model selection using akaike weights. *Psychonomic bulletin & review*, 11:192–196, 2004.

[29] 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.

Eidesstattliche Erklärung

Hiermit versichere ich – Arjit Basu – an Eides statt, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe.

Sämtliche Stellen der Arbeit, die im Wortlaut oder dem Sinn nach Publikationen oder Vorträgen anderer Autoren entnommen sind, habe ich als solche kenntlich gemacht.

Diese Arbeit wurde in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegt oder anderweitig veröffentlicht.

Mittweida, 30. November 2023	
Ort, Datum	Arjit Basu, M.Sc.