JIMMA UNIVERSITY
Community Oriented Education
ጅማ ዩኒቨርሲቲ

MODELLING TIME-TO-MALARIA: A COMPARISON OF COX PROPORTIONAL HAZARDS AND SHARED GAMMA FRAILTY MODELS

A case study of children under the age of 10 years residing near Gilgel Gibe hydroelectric power dam, Jimma Ethiopia.

By:

Belay Birlie Yimer

A Thesis Submitted to the Department of Statistics, School of Graduate Studies, College of Natural Science, Jimma University
In Partial Fulfilment for the Requirements of Master of Science (MSc) Degree in Biostatistics

October, 2011

Jimma, Ethiopia

MODELLING TIME-TO-MALARIA: A COMPARISON OF COX PROPORTIONAL HAZARDS AND SHARED GAMMA FRAILTY MODELS

A case study of children under the age of 10 years residing near Gilgel Gibe hydroelectric power dam, Jimma Ethiopia.

M.Sc. Thesis

Belaya Birlie

October 2011

Jimma University

DEPARTMENT OF STATISTICS, SCHOOL OF GRADUATE STUDIES

JIMMA UNIVERSITY

As thesis research advisors, we herby certify that we have read and evaluated the thesis prepared by *Belaya Birlie* under our guidance, which is entitled "MODELLING TIME-TO-MALARIA: A COMPARISON OF COX PROPORTIONAL HAZARDS AND SHARED GAMMA FRAILTY MODELS". We recommend that the thesis be submitted as it fulfils the requirements for the degree of Master of Science (MSc) in Biostatistics.

Yehenew Getachew (Asst Prof, PhD scholar) ---------------------- ---------------

Major advisor Signature date

Wondwosen Kassahun (Asst Prof, PhD scholar) -------------------- --------------

Co-advisor signature date

As the members of the board of examiners of MSc thesis open defence examination of **Belay Birlie Yimer**, we certify that we have read and evaluated the thesis and examined the candidate. We recommend that the thesis be accepted as it fulfils the requirements for the degree of Master of Science in Biostatistics.

Name of chairman Signature Date

Name of first Advisor Signature Date

Name of second Advisor Signature Date

Name of internal Examiner Signature Date

Name of External Examiner Signature Date

Department head Signature Date

Statement of the Author

First, I declare that this thesis is a result of my genuine work and all sources of materials used for writing it have been duly acknowledged. I have submitted this thesis to Jimma University in partial fulfilment for the Degree of Master of Science. The thesis can be deposited in the library of the university to be made available to borrowers for reference. I solemnly declare that I have not so far submitted this thesis to any other institution anywhere for that award of any academic degree, diploma or certificate.

Brief quotations from this thesis are allowed without requiring special permission provided that an accurate acknowledgement of the source is made. Requisites for extended quotations for the reproduction of the thesis in whole or in part may be granted by the head of the department of statistics when in his or her judgment the proposed use of the material is for a scholarly interest. In all other instances, however, permission must be obtained from the author.

Name**: Belaya Birlie** Signature ------------------------

Place: Jimma University, Jimma

Date of submission: ---------------------

Abstracts

In many epidemiological studies time to event data are grouped into strata (or clusters), such as, geographic region, districts, villages and so on. Consequently, cluster specific effects on survival times may cause an extra variation. Under such circumstance, it is substantive importance to draw inference on the nature and magnitude of these effects albeit the primary focus being on survival times. In model based analysis the aforesaid effect (called frailty) are usually accommodated by the use of frailty survival models.

The objective of this thesis is to model the time to first malaria infection due to p. falcuiprum in children living near to the Gilgel Gibe dam using Cox proportional hazards and shared gamma frailty models with an attempt to compare these two modelling approaches.

We apply the two modelling approaches to the analysis of malaria dataset. The dataset comprise time to first malaria infection of 2040 under 10 children observed during the period from July 2008 to June 2010.

This study revealed that, Cox PH model estimates the risk of malaria infection for children residing in proximity to the dam is significantly lower than children's living in distant from the dam. However, when we take the clustering of children within locality into account (using frailty model) there was no statistical significant difference in hazard of contracting malaria between the two groups, namely at risk and control. The likelihood ratio test of the heterogeneity parameter (theta) in all the fitted frailty models, however, showed that theta is significantly different from zero (P<0.000), indicating that there is a clear clustering of study subjects (children) with in their localities.

In the future, it is better to see also the result by including a frailty term at least in a pair-wise manner and also spatial distance of households, in the modelling of time-to-malaria.

Acknowledgement

This thesis grew out of a research project provided by my advisor Mr. Yehenew Getachew. I am deeply indebted to Mr. Yehenew, who opened my eyes for survival analysis and guided me through out this thesis. This thesis could not have been written without his constant help and support.

I am sincerely grateful to my co-advisor, Mr. Wondosen Kasshun, for his continues advice and guidance.

My further appreciation goes to Mr. Delenasaw Yewhalaw, for giving me the malaria data set and also advices whenever necessary, inline with the epidemiology of malaria..

Last but not least, I want to thank my family and friends for their unreserved support and encouragement.

Table of Contents

List of tables

List of Figures

1. Introduction

1.1. Background of the study

Globally, more than two billion people live in areas where they are at risk of contracting malaria (Breman, 2001). Annually about 300 to 500 million new cases of malaria, primarily due to *P.falcipurm*, are observed in the world with 90% in sub-Saharan Africa and these accounts for an estimated one million children's deaths (snow et al, 1999). In 2001, the disease accounts for an estimated loss of 44.7 million disability adjusted life years (DALYs) with a DALY loss of > 87% occurring in the sub-Saharan Africa (WHO, 2003); in 2002 the estimated malaria burden has increased to 46.5 million DALYs (WHO, 2004). An estimated 90% of this burden is related to environmental factors (WHO, 1977). Reliable analysis of these environmental risks to health is therefore fundamental for the prevention and control of the disease, for evidence-based guidance for health policy and planning. However, analysis of how changes in the environmental risk factors and the incidence and prevalence of malaria are related is sparse.

The distribution of malaria is governed by a large number of factors relating to the parasite, the vector and the host. Many of these factors affect the interactions between parasite, vector and host in some way. Within this context, the development of water projects and their operation has a history of facilitating increased transmission of vector borne diseases (service, 1991). This risk factor comprises a number of different components that are related to the transmission dynamics of malaria, which collectively influence morbidity and mortality and hence the malaria burden. Various studies have been done to investigate malaria incidence and prevalence in dam sites compared to a distant site, however consistent result is not obtained yet.

This study is intended to model the time-to-first malaria infection of children's living around Gilgel Gibe Hydroelectric Power Dam using survival modelling framework. Despite, survival models have a long history in the biostatistical and medical literature (Cox and Oakes 1984), there are very few literatures regarding the use of survival analysis in modelling the spread and incidence of malaria as compared to other statistical models, such as, General Liner Model (GLM) and Generalized Liner Mixed Effect Model (GLMM).

As opposed to modelling disease incidence and mortality, survival modelling provides a slightly different perspective with regard to the nature of the disease. It focus upon how many are expected to survive after a certain period of time, how fast is the rate of failure, and what derives shortened or prolonged survival; all of these may be influenced by several factors such as prevalence of the disease in the population, physiological characteristics of a person, environmental, socio-economic, access to health care facility etc.

Survival analysis is a statistical method for data analysis where the outcome variable of interest is the time to the occurrence of an event (Klembaum, D. G., 1996). Hence, survival analysis is also referred to as "time-to-event analysis", which is applied in a number of applied fields, such as medicine, public health, social science, and engineering. In medical science, time to event can be time until recurrence in a cancer study, time to death, or time until infection. In the social sciences, interest can lie in analyzing time to events such as job changes, marriage, birth of children and so forth. The engineering sciences have also contributed to the development of survival analysis which is called failure time analysis since the main focus is in modelling the lifetimes of machines or electronic components (Lawless, J. F., 1982). The developments from these diverse fields have for the most part been consolidated into the field of survival analysis. Because these methods have been adapted by researchers in different fields, they also have several different names: event history analysis (sociology), failure time analysis (engineering), duration analysis or transition analysis (economics). These different names do not imply any real difference in techniques, although different disciplines may emphasize slightly different approaches. Survival analysis is the name that is most widely used and recognized (Lee, E. T., and Wang, J. W., 2003).

The complexities provided by the presence of censored observations led to the development of a new field of statistical methodology. The methodological developments in survival analysis were largely achieved in the latter half of the 20th century. Although

Bayesian methods in survival analysis (IBRAHIM, J. G. et al., 2001) are well developed and are becoming quite common for survival data, our application will focus on frequentist methods. There have been several textbooks written that address survival analysis from a frequentist perspective. These include Lawless, J. F. (1982), Cox and Oakes (1984), Fleming and Harrington (1991), and Klein and Moeschberger (1997).

One of the oldest and most straightforward non-parametric methods for analyzing survival data is to compute the life table, which was proposed by Berkson and Gage (1950) for studying cancer survival. One important development in non-parametric analysis methods was obtained by Kaplan and Meier (1958). While non-parametric methods work well for homogeneous samples, they do not determine whether or not certain variables are related to the survival times. This need leads to the application of regression methods for analyzing survival data. The standard multiple linear regression models are not well suited to survival data for several reasons. Firstly, survival times are rarely normally distributed. Secondly, censored data result in missing values for the dependent variable (survival time) (Klembaum, D. G. 1996).

The Cox proportional hazards (PH) model is now the most widely used for the analysis of survival data in the presence of covariates or prognostic factors. This is the most popular model for survival analysis because of its simplicity, and not being based on any assumptions about the survival distribution. The model assumes that the underlying hazard rate is a function of the independent covariates, but no assumptions are made about the nature or shape of the hazard function. In the last several years, the theoretical basis for the model has been solidified by connecting it to the study of counting processes and martingale theory, which was discussed in the books of Fleming and Harrington (1991) and of Andersen et al (1993). These developments have led to the introduction of several new extensions to the original model. However the Cox PH model may not be appropriate in many situations and other modifications such as stratified Cox model (Klembaum, D. G. 1996) or Cox model with time-dependent variables (Collett, D., 2003) can be used for the analysis of survival data.

Implicitly, most of the statistical models and methods for time-to-event data (and here especially the Cox PH model) were developed under the assumption that the observations from subjects are statistically independent of each other. While this is sensible in many applications, it has become obvious that this assumption does not hold in other situations which are not common as originally thought. In many epidemiological studies time to event data are clustered and event times among members of the same cluster may not be independent. In this case, conventional survival analysis may yield consistent estimates of the marginal hazard if the marginal hazard is incorrectly modelled (Huster et al., 1989). However, variance estimates overestimate the true variance when the independent variables vary within a nit, and underestimate when the independent variables are constant within a unit, leading to incorrect inferences.

A commonly used and very general approach to the problem of modelling multivariate data is to specify independence among observed data items conditional on a set of unobserved or latent variables. A multivariate model for the observed data is then induced by averaging over assumed distribution for the latent variables. The dependence structure in the multivariate model arises when dependent latent variables enter into the conditional models for multiple observed data items, and the dependence parameters often may be interpreted as variance components. Frailty models for multivariate survival data are derived under a conditional independence assumption by specifying latent variables which act multiplicatively on the base line hazard (Wienke 2003).

Extensive research has been devoted to the frailty issue in survival analysis and generalized linear model (GLIM). Recently, investigators have recognized that ignoring individual heterogeneity may lead to inaccurate conclusions. Models for heterogeneity have been proposed by Vaupel et al. (1979), who introduced frailty as an unobserved quantity in population mortality. Oakes (1989) proposed frailty models for bivariate survival times and introduced several possible frailty models. Flinn and Heckman (1982) also introduced heterogeneity into their model for analyzing individual event histories. They believed that improper modelling of heterogeneity will result in biased estimates since the covariates in the model fail to explain the true effect of the covariates on a response variable. Keyfitz and Littman (1979) showed that ignoring heterogeneity will lead to an incorrect calculation of the life expectancy from known death rates. A similar conclusion was reached by Vaupel et al. (1979) using a continuous mixture model in which an unobserved non-negative random frailty represents all individual differences in endowment for longevity.

1.2 Statement of the problem

In the literature the use survival analysis in modelling malaria disease dynamics is not that much as compared to the general liner modelling framework, yet survival approaches has the capacity to incorporate many of the features of this approaches. Particularly, the traditional Cox PH model has the potential to deal with aspects such as non-normality, censoring as well as to investigate the effect of explanatory variables directly on survival time.

The standard situation of the application of survival methods in most clinical researches projects assume a homogeneous population being studied. That is, all individuals sampled in to the study are subjects in principle under the same risk factor. The appropriate survival (and here especially the Cox PH) then assumes that the survival data of different individuals are independent of each other. While this is sensible in many applications, it has become obvious that this assumption does not hold in other situations which are not common as originally thought. In many epidemiological studies time-to-event data are clustered and event times among members of the same cluster may not be independent.

Frailty modelling approach accounts for this problem by specifying independence among observed data items conditional on a set of unobserved or latent variables. Whereas, the Cox proportional hazards model has no such term and dependence of the event times is not accounted for. This lack of accountability can lead to biased estimates of both the regression coefficients and hazard rates and the magnitude of bias that will be committed depend on various factors.

In this study, we used and compare for their efficiency, the Cox PH model (without taking in to account the clustering in the data) and its extension shared gamma frailty model to investigate the pattern of malaria incidence among children's living around Gilgel Gibe area using important covariates.

1.3 Objectives of the study

1.3.1 General objective

- The main objective of this study is to model time-to-first malaria infection of children living around Gilgel Gibe hydroelectric power dam using Cox PH and gamma shared frailty modelling approaches and then after to compare their performance.

1.3.2 Specific objective

- To investigate the pattern of malaria incidence among children's living around Gilgel gibe area while taking into and (not into) account the clustering of study subjects with in villages.

- To identify important risk factors or covariates that are significantly associated with time to first malaria infection, and

- To compare the two commonly used modelling approaches in survival analysis; namely Cox PH and gamma shared frailty models using malaria data set.

2. Literature review

2.1 Global burden of malaria

Globally, more than two billion people live in areas where they are at risk of contracting malaria (Breman, 2001). Annually about 300 to 500 million new cases of malaria, primarily due to *P. falcipurm*, are observed in the world with 90% in sub-Saharan Africa and these accounts for an estimated one million children's deaths (snow et al, 1999). In 2001, the disease accounts for an estimated loss of 44.7 million disability adjusted life years (DALYs) with a DALY loss of > 87% occurring in the sub-Saharan Africa (WHO, 2003); in 2002 the estimated malaria burden has increased to 46.5 million DALYs (WHO, 2004). An estimated 90% of this burden is related to environmental factors (WHO, 1977). Reliable analysis of these environmental risks to health is therefore fundamental for the prevention and control of the disease, for evidence-based guidance for health policy and planning. However, analysis of how changes in the environmental risk factors and the incidence and prevalence of malaria are related is sparse.

The discovery of an interactive effect between HIV infection and malaria morbidity (Whitworth *et al.* 2000; Chandramohan and Greenwood 1998; Verhoef et al. 1999) exacerbates the potential for devastating health consequences in populations with large numbers of individuals who are co-infected. In resource-poor countries in Africa, malaria prevention and treatment consume large proportions of health budgets, and since it poses a threat to indigenous populations as well as visitors, it acts as a deterrent to tourism and foreign investment in these countries. Malaria therefore not only affects the health status of Africa's population, but also has far-reaching economic consequences inhibiting economic development (Wernsdorfer and Wernsdorfer 1988). The impact of malaria on the region has been recognized by the convening of the first African summit of heads of state on malaria in Abuja, Nigeria in April 2000. A report to the summit meeting calls, amongst other things, for more research on trends in incidence and prevalence, epidemic outbreaks and clinical epidemiology (Sachs 2000). A better understanding of the distribution of malaria has been identified as an important tool in its control (Snow *et al*. 1996).

2.2 Biology of malaria

2.2.1 Transmission of malaria

Malaria is caused by the parasite of genus *Plasmodium*. The four species of *Plasmodium* are *P. falciparum*, *P.malariae*, *P.ovale* and *P.vivax*. In Africa the predominant species of the disease causing-parasite is *P. falciparum*. Infection of the human host occurs when a person is bitten by a female *Anopheles* mosquito which has previously become infected. The parasite, called sporozoite at this stage of its cycle, enters the human body via the saliva of the mosquito which is injected into the blood. The parasites multiply in the liver, and re-invade the blood via red blood cells as merozoites. These develop into a stage known as the trophozoite, which is the one visible in blood films, and subsequently divide by the process of schizogony to produce further merozoites, which invade non-infected blood-cells. Some of the merozoites develop into new trophozoites whilst others develop into male micro- or female macrogametocytes. Uninfected *Anopheles* mosquitoes become infected if they feed on a person with mature gametocytes in their peripheral blood. In the mosquitos the microgametozytes exflagellate into gametes before fertilising the macrogametocytes, thereby forming zygotes. The zygote changes into an ookinete and then into an oocyst, which is found in the mid-gut wall of the mosquito. Large numbers of sporozoites are formed within the oocyst. The sporozoites leave the oocyst to invade the mosquito.s salivary glands, from where they can infect another human host when the mosquito takes a blood meal. The incubation period of the parasite in the vector takes 13 days to complete at 24û C. for *P.falciparum*. The vector will only become infective if it survives this *sporogonic* cycle (Gilles and Warrell 1993, chapter 2).

Malaria as a disease is therefore closely bound to conditions which favor the survival of the anopheles mosquito and the life cycle of the parasite. These conditions are predominantly determined by climatic factors, by vegetation coverage and by the vector’s access to water surfaces for breeding requirements (Molineaux, 1988; Gillies and De Meillon, 1968; Ghebreyesus, 1999). In the absence of any human intervention these conditions are predominantly determined by climatic and environmental factors.

The most important vectors of malaria in Africa are members of the *An. Gambiae* complex and *An. funestus*. Identification of the distribution of particular species is important since malaria vector control measures may have to take account of behavioural differences between species to be effective (Coetzee *et al*. 2000; Gillies and De Meillon 1968). For example, indoor biting and indoor resting habits (endophagy and endophily respectively), make mosquitoes more susceptible to control by residual insecticide on interior walls of houses, and to other insecticide treated materials such as bednets.

Five species of the *An. gambiae* complex are vectors of malaria. The two species which are the most efficient vectors of malaria parasites, *An. gambiae sensu strict* and *An. arabiensis,* are also the most widely distributed throughout most of sub- Saharan Africa. They often occur together, but *An. arabiensis* predominates in drier areas, whilst *An gambiae* predominates in more humid areas. *An gambiae* generally has a higher vectorial capacity than any of the other species, in part due to it being highly anthropophilic. It is also mainly endophagic and endophilic, making it amenable to control by indoor house-spraying of residual insecticide, at least in areas of moderate transmission intensity. *An. arabiensis,* on the other hand, is partly zoophagic and mainly exophagic and exophilic. It is generally considered a less efficient vector of malaria than *An gambiae*, but it is nevertheless the principal malaria vector in many areas (White 1974). *A. bwambae* is found only in the Semliki forest area in Uganda. It is partially endophagic and partially endophilic. The two saltwater species of the *An. gambiae* complex are *An. melas* and *An. merus* which are found in West Africa and in East Africa respectively. *An. merus* is exophilic and mainly zoophagic, whereas *An. melas* displays a more mixed resting and biting behaviour. *An. funestus* of the *An funestus* group, the other major vector of malaria in many parts of tropical and sub-tropical Africa (Armah *et al.* 1997; Gillies and De Meillon, 1968) bites humans; it is exophagic and endophilic. Since it breeds mainly in permanent water bodies, it is associated with all-year as opposed to seasonal malaria transmission (Sharp *et al*. 2000).

2.2.2 Clinical manifestations

Clinical malaria manifests itself in its mild form as a febrile illness associated with other non-specific symptoms (Bruce-Chwatt 1980, ch.3). The first clinical signs will only appear after the incubation period, which varies between nine and fourteen days for *falciparum* malaria. Clinical diagnosis is usually confirmed by a blood test, involving microscopic evidence of parasites in the blood, or by rapid diagnostic kit (Craig and Sharp 1997). However, in endemic countries infected individuals are often asymptomatic, so that parasitological evidence does not necessarily prove that the symptoms are due to malaria in a particular patient (Bruce-Chwatt 1980, pp. 35-51; Snow *et al*. 1997).

Severe life threatening malaria is usually due to *P.falciparum* malaria. In non-endemic areas cerebral malaria is the sequel that often sets in after the initial general symptoms. In such areas death due to malaria in both children and adults is usually due to cerebral malaria. In highly endemic areas severe malaria affects mainly young children and women during pregnancy. In such areas infants may enjoy a period of inherited immunity of up to 6 months. As this declines, clinical attacks become more severe, and often take the form of severe anaemia which is responsible for most deaths due to malaria in these areas. Depending on the intensity of exposure to the parasite, these children develop relative tolerance to malaria infection in their first few years of life. As a result of this older children and adults usually exhibit mild, non life-threatening clinical symptoms, if any.

2.3 Factors that affect malaria transmission

The distribution of malaria is governed by a large number of factors relating to the parasite, the vector and the host. Many of these factors affect the interactions between parasite, vector and host in some way. Within this context, the development of water projects and their operation has a history of facilitating increased transmission of vector borne diseases (service, 1991). This risk factor comprises a number of different components that are related to the transmission dynamics of malaria, which collectively influence morbidity and mortality and hence the malaria burden. The underlying reason is that, through the generation of new water bodies, new mosquito larval and adult habitats are created. The hydrological system and probably to a lesser degree, the atmospheric system might also be altered. Consequently this will have an effect on the development of malaria vector species and plasmodia, their survival rates and longevity, and most likely will result in increased mosquito densities. Without accompanying vector control strategies this is likely to result in a higher risk of disease transmission. Factors such as economic benefit from the water resource development project, personal protective measures, health seeking behavior and acquired immunity must also be taken in to account, as this factors might counterbalance negative impacts (Jennifer K. et al, 2005).

Various studies have been done to investigate malaria incidence and prevalence in dam sites compared a distant site, however consistent result is not obtained yet. For example, in India, an over four-fold increase in annual parasite incidence among children were recorded in villages closer to the Bargi dam (head end) compared to more distant villages (tail end) (Singh et al., 1999; Singh and Mishra, 2000). Similarly, in Tigray in northern Ethiopia, numerous small dams and irrigation systems were put in place at altitude above 1800 m with the broad aim of reducing dependence on rain fed agriculture, improving overall food production. Comparative appraisal of a series of cross-sectional malaria surveys among children carried out in close proximity to these newly constructed small dams and in villages farther away, revealed a seven-fold increase in malaria risk for those residing near dams (Ghebreyesus et al., 1999).

However, it was also found that dam areas displayed a lower malaria transmission compared with distant setting when integrated vector management or other control interventions have been applied. For example, in Uttaranchal, India, a study, which compared the parasitological indices in dam area to forest or plain areas, recorded a prevalence and annual parasite incidence of zero in dam area. Better economic status, insecticide spraying and more awareness towards health maintenance were described to be the main factors accounting for the lack of malaria transmission at the dam area (Shukla et al., 2001). In addition, in Thailand no increase of malaria incidence was observed near the Nong Wai dam and Ubol Ratana dam (Bunnag et al., 1979; Harinasuta et al., 1970).

Increasing rainfall and vegetation density generally have a favourable impact on malaria transmission through the provision of breeding sites and habitat for the vector. However, the differing breeding habits of different species of Anopheles complicate the relationship between rainfall and malaria transmission. Flooding, for example, may flush out larvae pools and lead to a temporary reduction in vector populations. Forest vegetation may inhibit *An. gambiae* because of the lack of sunlight. Nevertheless, insufficient annual rainfall, or seasonal rainfall, constitutes a distinct limitation to malaria transmission in areas where temperature is not a limiting factor. Rainfall of about 80mm per month for at least five months of the year has been identified as a minimum requirement for stable transmission to occur (Craig *et al*. 1999).

The relationship between the pattern of age-specific malaria morbidity and malaria transmission intensity has been well documented (Molineaux, 1988; Snow *et al.* 1997; Snow and Marsh, 1998b). In areas of high transmission intensity this generally shows that the incidence of clinical attacks peaks in early childhood and then declines rapidly with increasing age due to the acquisition of clinical immunity in such populations. In areas of moderate transmission intensity the age of peak transmission occurs at a later age, whereas in populations exposed to very low levels of transmission or to epidemic malaria, the risk of infection remains constant across all ages. This has been shown to be the case for both mild as well as severe clinical malaria (Snow and Marsh, 1998b).

3. Malaria data set and its study setup

The data set used for this study was generated by one of the VLIR IUC-JU project entitled "Malaria incidence and transmission among childern near Gilgel Gibe Hydroelectric Power Dam" with the principal investagtor (Mr. Delenasaw Yewhalaw). The study was undertaken with the over all aim of determining malaria incidence and patterns of its transmission among childerns living close to the newely constructed Gilgel Gibe hydroeletric dam.

3.1. Study area and study population

The study area is found in Jimma Zone,Southwestern part of Ethiopia, which is located 55 Km away from Jimma town, with an approximate latitude of 7^0 48' to 7^0 50' North and longitude 37^0 17' to 37^0 20' East. The area extends over 1,607 Km^2, with an altitude of about 1760 m.a.s.l. At an altitude of more than 1,600 m, malaria epidemics are frequent, and clinical immunity to malaria is low in the population. 26The study area is administratively structured into 4 districts (locally known as *woreda*) and 17 villages (locally known as *kebele*). All children who are less than 10 years of age and lived for at least 6-months in study area constitute the study population.

3.2. Study subjects

Inclusion criteria of children in the study were as follows: All children who are less than 10 years of age and continuous residence for at least 6-months in study area household since July 8, 2008, and intention to remain in the study area for the duration of study follow-up.

At the very early stage of the study, each household in the study area was visited and numbered and a baseline survey was conducted with the aim of collecting relevant information on individual and household characteristics. After having all the necessary information from the baseline survey, all villages surrounding the dam within a 10km radius were classified into two groups mainly based on maximum flaying ablity mosquto.

- at risk vllages:- villages within 3 km distance from the dam
- control villages:- villages which are 5 to 10 km away from the dam

From these two groups of villages, 8-pairs of villages were selected and paired based on various comparability factors, including similarity on ,eco-topography/altitude, population size, socio-economic activity, cropping area, health facility, with out major impounded water around them, etc. With such selection procedures, a total sample size of 2080 (130 under ten children/village * 8 villages * 2) study subjects were identified and included.

3.3. Data collection

3.3.1. Follow-up

After recruitinged a total of 2,080 under 10 children for the study, The *Parasitological study* was carried out longitudinally by trained health workers and the incidence of new malaria cases was continuously recorded through passive surveillance and an active case detection system based on house-to-house visits at fortnightly interval in the already identified at risk and control villages during July, 2008 to june,2010 along with demographic variables. Due to many factors, in the course of 2-years, about 40 children were lost to follow-up and the final data set used for analysis consists of 2040 children of which 548 became infected.

3.3.2.Household survey

At the very early stage of the study, a baseline survey was conducted with the aim of collecting relevant information on individual and household characteristics. During the baseline survey, each household in the study area was visited and numbered; Information on demographic, environmental characteristics of the households, age, sex, number of children (under 10) in each household, use of bed nets, whether the house is human dwelling, mixed dwelling or there is cattle shelter nearby were collected; knowledge, attitude and practice related to malaria were collected from parents or guardians of children; Each house was also assessed and recorded in terms of its structure/design, accessibility and proximity to health facilities. Using a hand-held global postioning system (GPS), the geographic coordnates of all households were also measured and mapped (see fig 3.1). Continious varible ware created to define the distance of each household to the dam.

3.4. Outcome

In this study, the primary outcome is the time from the start of the study (from July 8, 2008) to the time of first malaria infection due to *P. falciparum* in children or to the end of study (to June 4, 2010). An episode of *P. falciparum* malaria was defined as temperature greater than 37.50^{0}C, with confirmed P.falciparum asexual stages by microscopy. The individual survival times for first *p. falcuiparum* malaria infection outcome are therefore the actual time in days from the start of the study to the date of his first *p. falcuiparum* malaria infection. Censoring was caused by death, dropout or end of the study.

4. Basic topics in survival analysis

In general, survival techniques can be applied to a wider range of different situations, subject to the three necessary requirements as stated by Cox and oaks (1984); firstly a well defined time origin must be determined, then a scale for measuring the progress of time must be defined, and finally the exact definition of failure must be clear.

4.1 Survival functions

For most statistical application it is usual to describe models for probability distribution in terms of either the probability density function f(x) or the distribution function F(x). For survival analysis it is usually more appropriate to work with other functions which characterize the probability distribution. Let T be a positive random variable from a homogeneous population, representing the time until the relevant event occurs. In order to characterize the distribution of T one of the most often used functions is survivor function.

The survivor function, S (t), is defined for both discrete and continuous distribution as the probability that an individual survivors beyond time t i.e.

$$S(t) = p(T \geq t) \qquad 0 < t < \infty \qquad (4.1)$$

Here 0<S(t)<1 since s(0) =1 and

For continuous random variable T, the density function, this is unconditional probability of the events occurring at time t, f(t), is given by

Where the cumulative distribution function $F(t) = 1 - S(t)$; So that $S(t) \quad {}_{t} \; f(u)du$. Note that *f(t)dt* may thought of as the “approximate ” probability that the event will occur at time t and that f(t) is a non negative function with the area under f(t) being equal to one.

Many types of survival curves can be shown but the important point to note is that they all have the same basic properties. They are monotone, non increasing function equal to one at zero and zero as the time approaches infinity, their rate of decline, of course, varies according the risk of experiencing the event at time t but it is difficult to determine the essence of a failure pattern by simply looking at survival curve. A basic quantity fundamental in survival

Under the null hypothesis, the probability of experiencing an event at $t_{(j)}$ does not depend on the group, i.e., the probability of experiencing an event at $t_{(j)}$ is $\frac{d_j}{r_j}$. So that the expected number of deaths in group one is $E(d_{1j}) \quad e_{1j} \quad \frac{r_{1j}d_j}{r_j}$

The test statistic is given by the difference between the total observed and expected number of deaths in group one

(4.18)

Since d_{1j} has the hypergeometric distribution, the variance of d_{1j} is given by

(4.19)

So that the variance of U_L is

Under the null hypothesis, statistic (4.18) has an approximate normal distribution with zero mean and variance V_L. This then follows $\frac{U^2{}_L}{V_L} \sim x^2{}_1$

There are several alternatives to the log-rank test to test the equality of survival curves, for example, the Wilcoxon test (Gehan, E. A., 1965). These tests may be defined in general as follow

[redacted]

$$Q \quad \frac{\overset{m}{}\; w_j(d_{1j} \quad \hat{e}_{1j})}{\overset{m}{}\; w^2{}_j \hat{v}_{1j}}_{\;j\;0}$$ (4.20)

Where w_j are weights whose values depend on the specific test

The Wilcoxon test uses weights equal to risk size at $t_{(j),}$ $w_j = r_j$ .This gives less weight to longest survival times Early failures receive more weight than later failures. The Wilcoxon test places more emphasis on the information at the beginning of the survival curve where the number at risk is large. This type of weighting may be used to assess whether the effect of

possible to estimate the effect parameter(s) without any consideration of the hazard function. The proportional hazards assumption refers to the fact that the hazard functions are multiplicatively related. That is, their ratio is assumed constant over survival time. In other words, the Cox proportional hazards model assumes that changes in the hazard of any subject over time will always be proportional to changes in the hazard of any other subject and to changes in the underlying hazard over time.

The beauty of the Cox approach is that this vagueness creates no problems for estimation. Even though the baseline hazard is not specified, we can still get a good estimate for regression coefficients, β, hazard ratio, and adjusted hazard curves.

From the representation in equation (4.21) one can notice a couple of features. First, if $Z_j : : 0$ then the hazard function for the jth individual is the baseline hazard function. It's the hazard function in the absence of covariates or when all of the coefficients of the covariates are assumed to be zero. Second, if we divide both sides by $h_0(t)$, we get equation (4.22) which shows where the term proportional comes from. Since for each individual, is constant across time, equation (4.22) shows that at every value of t, the j^{th} individual's log hazard function is constant proportion of the baseline hazard. Very loosely speaking, this implies that each individual's hazard function is “parallel” to the $h_0(t)$.

$$\frac{h(t, \text{[redacted]}}{h_0(t)} \qquad \frac{}{h_0(t)} \qquad \exp(\boldsymbol{\beta}\ \mathbf{Z}_j) \qquad (4.22)$$

This is called a semi parametric model because a parametric form is assumed only for the covariate effect and the baseline hazard rate is treated non-parametrically.

If we look at two individuals with covariate values Z and , the ratio of their hazard rates is

$$= \frac{h(t/\mathbf{Z})}{h(t/\mathbf{Z}^*)} \quad \frac{h_0(t)\exp \overset{p}{} (\boldsymbol{\beta}_k\ \mathbf{Z}_k) \ \lfloor_k \text{[redacted]} \ \rfloor}{h_0(t)\exp \overset{p}{} (\boldsymbol{\beta}_k\ \mathbf{Z}_k^*) \ \text{[redacted]}} \quad \exp\left[\overset{p}{\underset{k\ \text{[redacted]}}{}} {}_k(\mathbf{Z_k} - \mathbf{Z_k}^*) \right] \qquad (4.23)$$

which is a constant with respect to time. So, the hazard rates are proportional. The quantity given in equation (4.23) is called the relative risk (hazard ratio) of an individual with risk factor Z having the event as compared to an individual with risk factor Z^*

The Cox proportional hazards model can equally be regarded as linear model, as a linear combination of the covariates for the logarithm transformation of the hazard ratio given by:

$\log$ [redacted]

Where $= (z_1, z_2, ..., z_p)$ is the values of the vector of explanatory variables for a particular individual and $\boldsymbol{\beta}= (\beta_1, \beta_2, ..., \beta_p)^t$ is a vector of coefficients.

Again the cumulative hazard function is given by: ; The corresponding survival functions are related as .

4.5.1 Fitting Cox PH model

Fitting the Cox proportional hazards model, we wish to estimate $h_0(t)$ and β. One approach is to attempt to maximize the likelihood function for the observed data simultaneously with respect to h0 (t) and β. A more popular approach is proposed by Cox (1972) in which a partial likelihood function that does not depend on h_0 (t) is obtained for β. Partial likelihood is a technique developed to make inference about the regression parameters in the presence of nuisance parameters (h_0 (t) in the Cox PH model

Let t_1; t_2, . . ., t_n be the observed survival time for n individuals. Let the ordered event experiencing time of r individuals be $t_{(1)} < t_{(2)} <$,. . ., $< t_{(r)}$ and let $R(t_{(j)})$ be the risk set just before $t_{(j)}$ and rj for its size. So that $R(t_{(j)})$ is the group of individuals who are alive and uncensored at a time just prior to $t_{(j)}$. The conditional probability that the i^{th} individual experiences the event at $t_{(j)}$ given that one individual from the risk set on $R(t_{(j)})$ dies at $t_{(j)}$ is

P(individual i experiences the event at $t_{(j\,)}$ one event from the risk set $R(t(_{j}))$ at $t_{(j)}$)

4.5.3 Cox proportional hazards model diagnostics

After a model has been fitted, the adequacy of the fitted model needs to be assessed. The model checking procedures below are based on residuals. In linear regression methods, residuals are defined as the difference between the observed and predicted values of the dependent variable. However, when censored observations are present and partial likelihood function is used in the Cox PH model, the usual concept of residual is not applicable. A number of residuals have been proposed for use in connection with the Cox PH model. We will describe three major residuals in the Cox model: the Cox-Snell residual, the deviance residual, and the Schoenfeld residual.

4.5.3.1 Cox-Snell residuals

The Cox-Snell residual is given by Cox and Snell (Cox, D. R., and Snell, E. J.,1968). The Cox-Snell residual for the i^{th} individual with observed survival time t_i is defined as

$$r_{ci} = \hat{H}_o(t_i) \exp(\overset{p}{\underset{i\,\text{[redacted]}}{}} (Z_{ik}\hat{b}_k)) \quad \hat{H}_i(t_i) \quad -\log(\hat{S}(t_i)) \tag{4.29}$$

Where $\hat{H}_0(t_i)$ is an estimate of the is Breslow's baseline cumulative hazard function at time t_i; which is given by

(4.30)

Let Y = H(T) be the transformation of T based on the cumulative hazard function. Then the survival function for Y is:

$$\begin{aligned} S_Y(y) \quad P(Y \quad y) \quad P(H(t) \quad y) \\ P(T \quad H_T^{-1}(y)) \quad S_T(H_T^{-1}(y)) \\ = \exp(-H_T(H_T\,\text{[redacted]} = \exp(-y) \end{aligned} \tag{4.31}$$

was derived by Kalbfleisch and Prentice (1973). This residual is motivated by the following result: Let T have continuous survival distribution S(t) with the cumulative hazard H(t) = -log(S(t)). Thus, $S_T(t)$ = exp(-H(t)).

Thus, regardless of the distribution of T, the new variable Y = H(T) has an exponential distribution with unit mean. If the model was well fitted, the value $\hat{S}_i(t_i)$ would have similar

4.5.4.1 Stratified Cox model

One method that we can use is the stratified Cox model, which strati.es on the predictors not satisfying the PH assumption. The data are stratified into subgroups and the model is applied for each stratum. The model is given by

$h_{ig}(t) = h_{og}(t)\ \exp($ [redacted] (4.34)

where g represents the stratum.

Note that the hazards are non-proportional because the baseline hazards may be different between strata. The coefficients β are assumed to be the same for each stratum g. The partial likelihood function is simply the product of the partial likelihoods in each stratum. A drawback of this approach is that we cannot identify the effect of this stratified predictor. This technique is most useful when the covariate with non-proportionality is categorical and not of direct interest.

4.5.4.2 Cox regression modelwith time -dependent variables

Until now we have assumed that the values of all covariates did not change over the period of observation. However, the values of covariates may change over time t. Such a covariate is called a time-dependent covariate. The second method to consider is to model non proportionality by time-dependent covariates. The violation of PH assumptions is equivalent to interactions between covariates and time. That is, the PH model assumes that the effect of each covariate is the same at all points in time. If the effect of a variable varies with time, the PH assumption is violated for that variable. To model a time-dependent effect, one can create a time-dependent covariate Z(t), ; where g(t) is a function of t such as t; log t or Heaviside functions, etc. The choice of time-dependent covariates may be based on theoretical considerations and strong clinical evidence.

The Cox regression with both time independent predictors Z_i and time-dependent covariates $Z_j(t)$ can be written

(4.35)

The hazard ratio at time t for the two individuals with different covariates Z and Z* is given by

$$\exp \quad {}^{p1}_{i\ 1} \quad \hat{}_{i} \ z_i^{*} \quad z_i \qquad {}^{p2}_{j\ 1} \quad \hat{}_{j}(z_j^{*}(t) \quad z_j(t)) \qquad (4.36)$$

Note that, in this hazard ratio formula, the coefficient $\hat{\alpha}_j$ is not time-dependent. , represents overall effect of $Z_j(t)$ considering all times at which this variable has been measured in this study. But the hazard ratio depends on time t. This means that the hazards of event at time t is no longer proportional, and the model is no longer a PH model.

In addition to considering time-dependent variable for analyzing a time-independent variable not satisfying the PH assumption, there are variables that are inherently defined as time-dependent variables. One of the earliest applications of the use of time-dependent covariates is in the report by Crowley and Hu (1977) on the Stanford Heart Transplant study.

4.6 Shared Frailty Models

The notation of frailty provides a covenant way to introduce random effect, association and unobserved heterogeneity into models for survival data. In its simplest form, a frailty is an unobserved random proportionality factor that modifies the hazard function of an individual, or of related individuals (Wienke, 2003). In essence, the frailty concept goes back to work of Greenwood and Yule (1920) on “accident proneness”. The term frailty was introduced by Vaupel et al (1979) in univariate survival models and the model was substantially promoted by its application to multivariate survival data in seminal paper by Clayton (1978) (without using the notation “frailty”) on chronic disease incidence in families.

Frailty models are extensions of the proportional hazards model which is best known as the Cox model (Cox, 1972), the most popular model in survival analysis. Normally, in most clinical application, survival analysis implicitly assumes a homogeneous population of individuals to be studied. This means that all individuals sampled into that study are subject in principle under the same risk (e.g., risk of death, risk of disease recurrence). In many applications, the study population cannot be assumed to be homogeneous but must be considered as a heterogeneous sample, i.e, a mixture of individuals with different hazards. For example, in many cases it is impossible to measure all relevant covariates related to the

A commonly used and very general approach to the problem of modelling multivariate data is to specify independence among observed data items conditional on a set of unobserved or latent variables. A multivariate model for the observed data is then induced by averaging over assumed distribution for the latent variables. The dependence structure in the multivariate model arises when dependent latent variables enter into the conditional models for multiple observed data items, and the dependence parameters often may be interpreted as variance components. Frailty models for multivariate survival data are derived under a conditional independence assumption by specifying latent variables which act multiplicatively on the base line hazard (Wienke 2003).

Extensive research has been devoted to the frailty issue in survival analysis and generalized linear model (GLIM). Recently, investigators have recognized that ignoring individual heterogeneity may lead to inaccurate conclusions. Models for heterogeneity have been proposed by Vaupel et al. (1979), who introduced frailty as an unobserved quantity in population mortality. Oakes (1989) proposed frailty models for bivariate survival times and introduced several possible frailty models. Flinn and Heckman (1982) also introduced heterogeneity into their model for analyzing individual event histories. They believed that improper modelling of heterogeneity will result in biased estimates since the covariates in the model fail to explain the true effect of the covariates on a response variable. Keyfitz and Littman (1979) showed that ignoring heterogeneity will lead to an incorrect calculation of the life expectancy from known death rates. A similar conclusion was reached by Vaupel et al. (1979) using a continuous mixture model in which an unobserved non-negative random frailty represents all individual differences in endowment for longevity.

From the above it is clear that the joint survivor function for one group is the Laplace transform of the frailty density function with parameter . In principle, any distribution on the positive numbers can be applied as a frailty distribution. In this thesis we will consider only the gamma distribution. For other distribution see Hougaard (2000), and ohman and Eberly (2001).

Gamma distributions have been used for many years to generate mixtures in exponential and Poisson models. From a computational point of view, gamma models fit very well into survival models, because it is easy to derive the formulas for any number of events. This is due to simplicity of the derivatives of the Laplace transform. This is also the reason why this distribution has been applied in most of the applications published until now.

The probability density function (pdf) of gamma distribution as

(4.41)

With [redacted]the shape parameter and the scale parameter. Furthermore we have

And

$$\mathrm{var}(U) \quad / \quad ^{2} \qquad (4.42)$$

In frailty modeling the typical choice of the parameters of the gamma distribution is . Using θ as notation for the variance of U, we have . This distribution with parameter is called one parameter gamma distribution with variance parameter θ. The density and Laplace transform of gamma distribution are respectively as follows.

(4.43)

From equation 3.39 it is easily seen that the marginal hazard is

(4.44)

4.6.2 Penalized partial likelihood for shared gamma frailty models

The addition of frailties in the Cox model leads to unobserved entities in the model which also prevail in the partial likelihood. It is however assumed that these frailties come from a gamma density with mean equal to 1 and unknown heterogeneity parameter θ. Therefore, a penalty is added to the partial likelihood that decreases with the distance of the frailty from one, the mean of the frailty density.

The penalty term on the log scale in the case of the gamma density is given by

(3.45)

The penalized partial likelihood for the frailty model is then given (McGilchrist, 1993) by

(4.46)

For fixed values of the heterogeneity parameter θ, maximization of the penalized partial likelihood criterion leads to the same parameter estimates for the fixed effects [redacted]and the frailties z_i as the EM-algorithm (Therneau et al., 2003). For a particular value of θ, estimates for the fixed effects, frailties and baseline hazards can thus be obtained by maximizing the panelized partial likelihood.

To make clear that we keep θ fixed in , we write $\overset{E}{A}_{ppl}$([redacted]; we further use to denote the values of β_0 *and* U that maximize, for the given value of θ, . We now consider the profile partial likelihood as a function of θ. However, the estimate of θ obtained from the EM-algorithm cannot be obtained by maximizing the profile penalized partial likelihood.

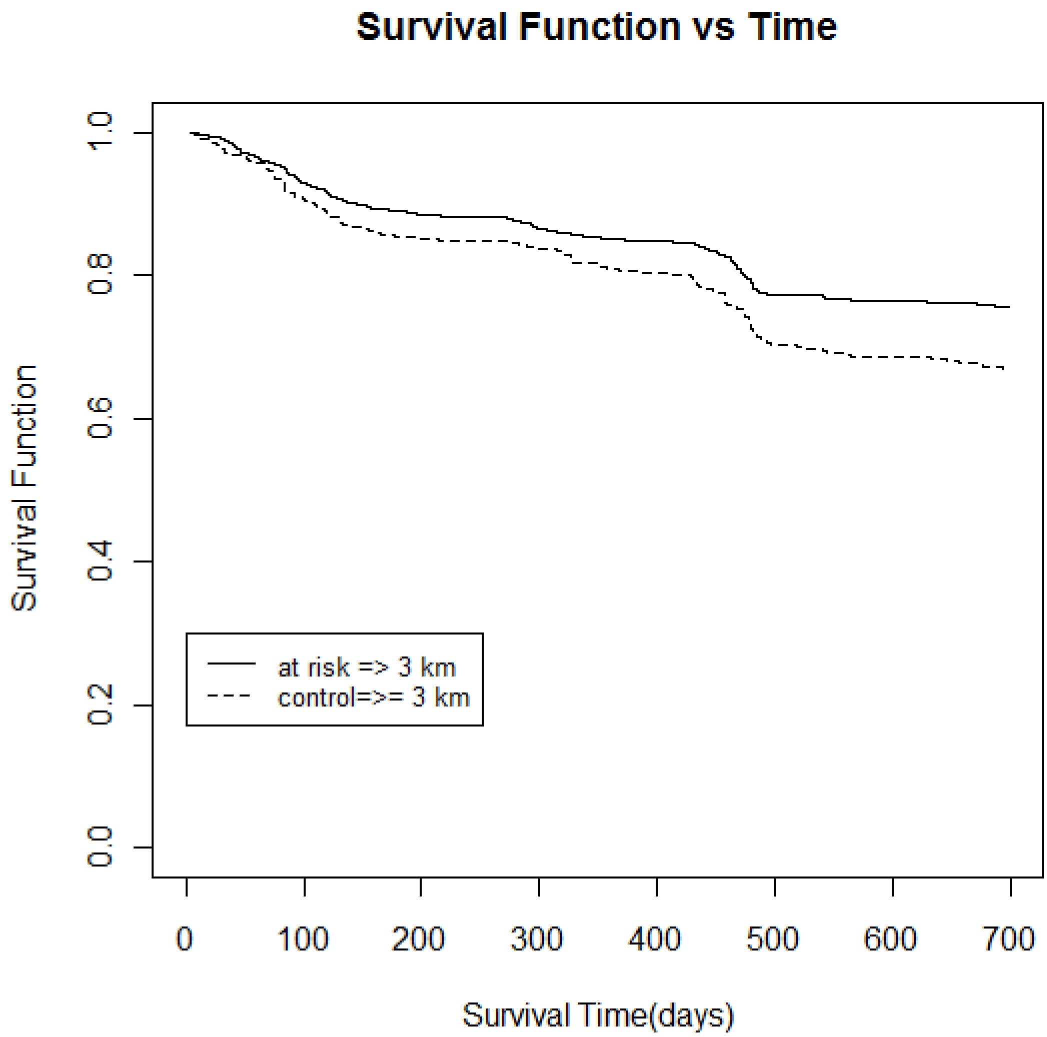
Table 5.1: Baseline characteristics in 2040 children

Variables	Mean (SD); count (%)
Distance	2.53(2.03)
age_start	4.95(2.05)
Distance group	
Less than 3 km	1,497(73.38)
Greater than or equal to 3 km	543(26.62)
Age_group	
<3 years	533(26.13)
4-7 years	1,272(62.35)
>=8 years	235(11.52)
Sex	
Female	981(48.09)
Male	1,059(51.91)
House structure	
Not corrugated	1,616(79.22)
Corrugated	424(20.78)

Table 5.2 log-rank test

Distance group	Events Observed	Events Expected	Total
< 3 km	368	407.16	1497
>= 3 km	180	140.84	543
Total	548	548	2040
	chi2(1)	14.68	
	Pr>chi2	0.0001	

Fig 5.1 Graph of the survival function of two groups for malaria dataset



The observed number of infection is 365 out of 1497 and 180 out of 543 in the at risk groups and control groups respectively. While the expected number of infection under the null hypothesis (the survival pattern of the two groups are the same) are 407.16 and 140.84 in the at risk groups and control groups respectively. The log rank test rejects the null hypothesis of equality of survival function with p-value<0.000 (chi-square statistic with 1 df 14.16). The results are consistent with what we saw from the graphical analysis.

5.2. Cox proportional model

The non-parametric methods that we used previously do not control for covariates and it requires categorical predictors. In order to determine demographic, climatic and environmental covariates which are associated with the observed time to first *p.falcuiprum* malaria infection we first use the Cox regression model.

We use univariate analysis to check all the risk factors before proceeding to more complicated models. We use a univariate Cox proportional hazards regression for every potential risk factor. The likelihood ratio test is considered in each univariate Cox PH model. Variables are identified as significant using a 0.1 significance level in the univariate model. We then fit the full multivariate Cox PH model including all the potential risk factors. In univariate (Table 5.3) and the full multivariate proportional hazards models (Table 5.4), distance group show a statistically significant association with time to first malaria infection. But other characteristics such as age at the start of follow up, age group, sex, house structure are not statistically significant, suggesting that these variables are not associated with the time to malaria infection due to *P.falcuiprum*. The uncategorized distance from the dam was significant in the univarite analysis but it become insignificant in the multivariable model.

Table 5.3 univariate Cox proportional hazards model

Covariates	**HR**	**Coef( )**	**Std. Err.**	**P-value**	**95% CI of**
Age group					
4-7 years	1.170	0.157	0.102	0.125	(-0.044,0.358)
>=8 years	1.033	0.033	0.156	0.831	(-0.273,0.339)
age_start	1.021	0.021	0.021	0.312	(-0.020,0.061)
Sex (male)	0.951	0.951	0.081	0.555	(0.804,1.124)
House structure					
(corrugated)	0.967	-0.032	0.106	0.764	(-0.240,0.177)
Distance	1.059	0.057	0.019	0.004	(0.018,0.096)
Distance group (>=3km)	1.414	0.347	0.090	0.000	(0.168,0.525)

Table 5.4. Multivariable Cox proportional hazards model

Covariates	**HR**	**Coef.**	**Std. Err.**	**P-value**	**95% CI**
Distance group(>=3km)	1.62	0.4849807	0.1815605	0.008	(0.129,0.840)
Sex(male)	0.95	-0.0560615	0.0855021	0.512	(-0.224,0.112)
age_group					
4-7 years	1.05	0.0476934	0.1261727	0.705	(-0.200,0.295)
>=8 years	0.82	-0.1929674	0.2202469	0.381	(-0.625,0.239)
House structure (Corrugated)	0.95	-0.048984	0.1064283	0.645	(-0.258,0.159)
Distance	0.96	-0.0339305	0.0402978	0.400	(-0.113,0.045)
Tvc					
Age at start	1.00	0.0001428	0.0000893	0.110	(-0.0000,0.0003)

Note: variables in tvc equation interacted with _t

As expected from non-parametric test, distance group is statistically significant also in a multivariate Cox PH model. The hazard ratio for time to first malaria infection is 1.62 in control group compared with at risk group.

After a Cox PH model is fitted, the adequacy of this model, including the PH assumption and the goodness of fit, needs to be assessed. We used $-\log(-\log(survival))$ plot to check the PH assumption for all the categorical variables.

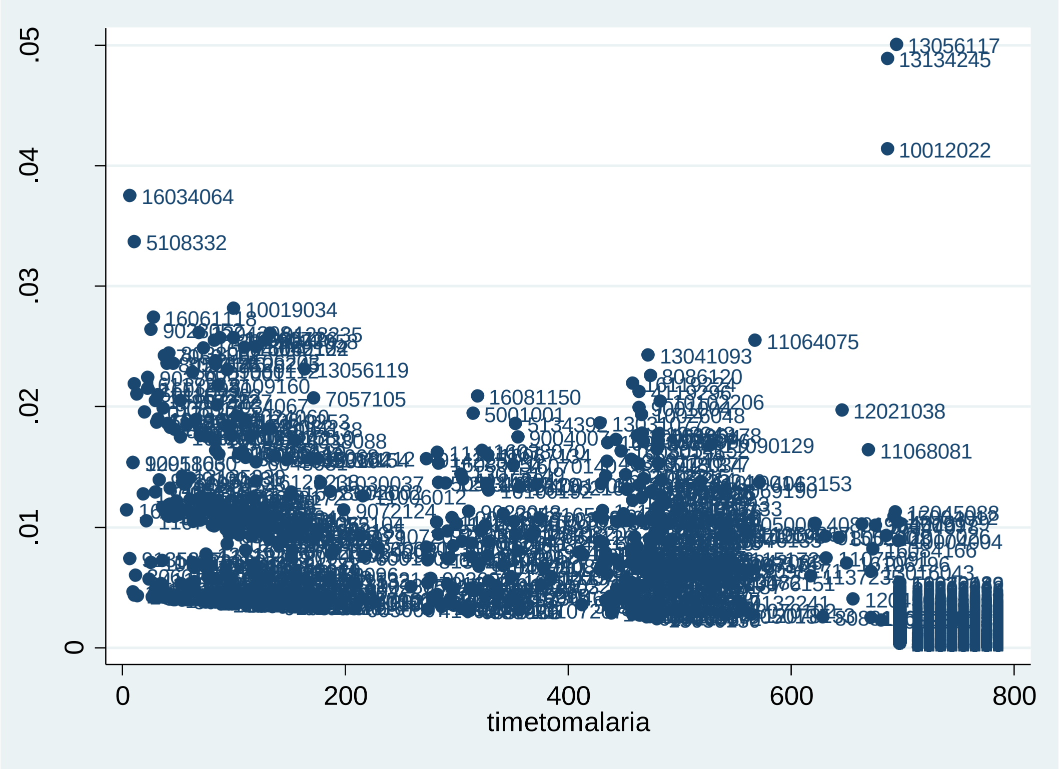
We assess goodness of fit using a plot of the Cox-Snell residuals against the cumulative hazard of Cox-Snell residuals (Figure 5.2). If the Cox regression model fits the data, these residuals should have a standard censored exponential distribution with hazard ratio 1. We can verify the model's fit by calculating—based, for example, on the Kaplan–Meier estimated survivor function or the Nelson–Aalen estimator—an empirical estimate of the cumulative hazard function, using the Cox–Snell residuals as the time variable and the data's original censoring variable. If the model fits the data, the plot of the cumulative hazard versus cs should approximate a straight line with slope 1. Comparing the jagged line with the reference 45^{o} line, we observe that the Cox model does not fit these data too badly. Because we use estimates , deviations from the 45^{0} line in the in the right-hand tail of the distribution of the above plots could be due in part to uncertainty about these estimates, since in this area the baseline hazard is more variable because of the reduced effective sample caused by prior failures and censoring.

The plot of deviance residual against the linear predictor shows that the deviance residuals seem not to be symmetrically distributed about zero. There are very high or very low deviance residuals which a signal for the presence of outliers (Figure 5.3). we used Likelihood displacement values to measure each subject's influence on the coefficient vector as a whole (fig 5.4). Likelihood displacement values measure influence by approximating what happens to the model log likelihood (more precisely, twice the log likelihood) when you omit subject i. The figure shows some subjects are influential.

Fig 5.2 Cox-Snell residuals plot Cox PH model

Fig 5.3 deviance residual plot Cox PH model

Fig 5.4 log-likelihood displacement plot for Cox PH mode



5.3. Shared gamma frailty modelling

In our data set children are clustered within a locality. It is recognized that individuals in the same community are more similar than the individuals in different communities because they shared similar (possibly unmeasured) environmental exposures. In Cox proportional hazard rate analysis, it is assumed that any differences in failure rates among individuals are picked up by the covariate structure which typically is assumed to act multiplicatively on a baseline hazard

Thus, a further extension the Cox model should be considered by taking in to account the hierarchical (clustered) structure of the data, ie., the nesting of children with in localities. This clustering can be taken in to account by adding a random effect as extra term. The locality is taken to be a random effect rather than a fixed effect because the individual locality is not of interest by itself; interest is rather in the heterogeneity between localities. Furthermore, introducing many fixed effects in a model might lead to convergence problems, especially if there is little variation in the covariates between localities (McGilchrist and Aisbett, 1991).

The relevant information for a child j (j= 1, ..., n$_i$) from locality i (i= 1, ..., s) is contained in a vector

With t_{ij} the time to first *p.falciprum* malaria infection or censoring, $_{ij}$ the censoring indicator is the vector of fixed effect covariates and $X_{ij}(t)$ is vector of time varying covariates.

We fit a shared gamma frailty model for this dataset. For child j, j= 1, 2, . . ., n$_i$, from cluster (locality) i, i= 1, . . ., s, defined as

$$h_{ij}(t) \quad h_0(t)u_i \exp(\beta^{'} X_{ij})$$

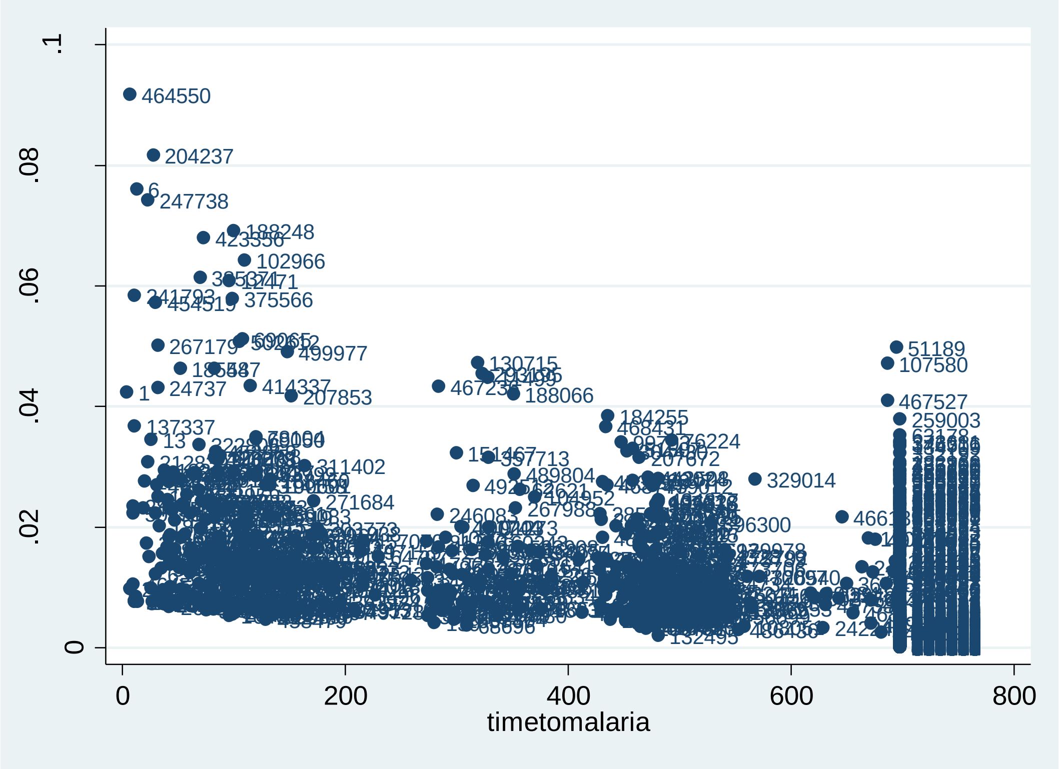
Where is the conditional hazards function for the jth child from the ith locality (conditional on u$_i$). the u$_i$'s are the actual values of a sample from a one parameter gamma distribution with mean equal to one and variance equal to . The parameter θ provides information on the variability (heterogeneity) in the population of clusters (locality).

The result of the Univariate and multivariable shared gamma frailty model and is given in table 5.5 and 5.6 respectively. From the out we can see that all tests of the likelihood-ratio test of, ,for all Univariate models is significant (p-value < 0.000), meaning that the correlation within localities cannot be ignored.

Unlike the Cox PH model in the Univariate shared gamma frailty model all covariates become statistically insignificant. The Cox PH model estimates the effect of distance and distance group on the log hazard of infection to be 0.057 (se: 0.019 and p-value: 0.004) and 0.347 (se: 0.090 and p-value: <0.000) respectively. whereas the shared gamma frailty model estimates the effect of distance and distance group on the log hazard of infection conditional on theta to be -0.009 (se: 0.048 and p-value: 0.848) and 0.204(se: 0.181and p-value:0.258) respectively. For both covariates the shared gamma frailty model estimates larger standard error for the coefficients as compared to the Cox PH model. Also, in the shared gamma frailty model the sign of the regression coefficient of distance is negative while it was positive in the Cox PH model.

Similarly, in the multivariate shared gamma frailty model the effect of all covariates on the log hazard became in significant except age at the start is significant at 10% significance level but not at 5%. The heterogeneity parameter θ is estimated to be 0.2818(se: 0.112) and the

Fig5.7 log-likelihood displacement plot for shared gamma frailty model



5.4 Comparison of Cox PH versus shared gamma frailty model

Table 5.7 gives the log-likelihood, AIC (akaki information criteria) and BIC (Bayesian information criteria) values of the two models. From the table we can see that the shared gamma frailty model has both a minimum AIC and BIC value, indicating that this model fit the data better than the Cox PH model which did not take in to account the clustering.

Table 5.7 comparison of Cox ph and frailty model

Model	Log-likelihood (null)	Log-likelihood (model)	df	AIC	BIC
Cox PH	-4095.401	-4085.069	7	8184.138	8223.483
Frailty	.	-4036.831	7	8087.662	8127.007

In both of the two models , other characteristics such as age at the start of follow up, age group, sex, house structure are not statistically significant, suggesting that these variables are not associated with the time to malaria infection due to *P.falcuiprum*.

Comparison of the two modelling approaches using different statistical techniques suggests that the shared gamma frailty model is better than the Cox PH model. This sis based on the significant result of the clustering parameter (theta).

6.2 Future lines of work

In this study we only considered modelling of time to first malaria using shared gamma frailty term at village level. That is, all study subjects (children) living in the same village shares the same frailty term. However, we feel that better result could have been obtained, if we include a frailty term at least in a pair wise fashion. i.e. In line with the epidemiology of malaria, spatial distance of households will have much more impact in the transmission of this disease. In other words, very close households will have higher chance of getting malaria (off course if there is a diseased person in one of the household) than those oriented far part. Hence, if in the future, if one includes this spatial distance in the modelling of time-to-malaria, a better result could be obtained using frailty modelling approaches.

11. Collett, D. Modelling Survival Data in Medical Research. Chapman and Hall, London, 2003.

12. Cox, D. R. Regression models and life-tables. Journal of the Royal Statistical Society Series B 34 (1972), 187.220.

13. Cox, D. R., and Oakes, D. Analysis of Survival Data. Chapman and Hall, London, 1984.

14. Cox, D. R., and Snell, E. J. A general de.nition of residuals with discussion. Journal of the Royal Statistical Society. Series B 30 (1968), 248.275.

15. Craig MH, Sharp BL (1997) Comparative evaluation of four techniques for the diagnosis of Plasmodium falciparam infections. Trans R Soc Trop Med Hyg 91(3): 279-82.

16. Crowder, J., and Hu, M. Covariance analysis of heart transplant survival data. Journal of American Statistical Association 78 (1977), 27.36.

17. Duchateau, L. and Janssen, P. (2008). The Frailty Model. Springer: New York.

18. Duchateau, L., Janssen, P., Lindsey, P., Legrand, C., Nguti, R. and Sylvester, R. (2002). The shared frailty model and the power for heterogeneity tests in multicenter trials. Computational Statistics and Data Analysis, 40, 603-20.

19. Fleming, T. R., and Harrington, D. P. Counting Processes and Survival Analysis. Wiley, New York, 1991.

20. Flinn, C.J. and Heckman, J.J. (1982). New methods for analyzing individual event histories. Socialogical Methodology, 99-140.

21. Gehan, E. A. A generalized Wilcoxon test for comparing arbitrarily singly censored samples. Biometrika 52 (1965), 203.223.

22. Ghebreyesus, T. A., Haile, M., Witten, K. H., Getachew, A., Yohannes, A. M., Yohannes, M., Teklehaimanot, H. D., Lindsay, S. W. & Byass, P. (1999) Incidence of malaria among children living near dams in northern Ethiopia: community based incidence survey. *BMJ*, 319, 663-666.

23. Gilles HM, Warrell DA (1993). Bruce-Chwatt.s Essential Malariology. Edward Arnold. London.

24. Gillies MT and De Meillon B (1968) The Anophelinae of Africa South of the Sahara. The South African Institute for Medical Research. Johannesburg, 212-219.

25. Greenwood, M. The natural duration of cancer. Reports on Public Health and Medical Subjects 33 (1926), 1.26.

26. Harinasuta, C., Jetanasen, S., Impand, P. & Maegraith, B. G. (1970) Health problems and socio-economic development in Thailand. *Southeast Asian J. Trop. Med. Public Health*, 1, 530-552.

27. Hougaard, P. (1986a). A class of multivariate failure time distributions. Biometrika, 73, 671-78.

28. Hougaard, P. (2000). Analysis of Multivariate Survival Data. Springer: New York.

29. Huster, W.J., Brookmeyer, R. and Self, S.G. (1989). Modelling paired survival data with covariates. Biometrics, 45, 145-56.

30. IBRAHIM, J. G., CHEN, M.-H., and SINHA, D. Bayesian Survival Analysis. Springer-Verlag, New York, 2001.

31. Kalbfleisch, J. D., and Prentice, R. L. Marginal likelihoods based on Cox's regression and life model. Biometrika 60 (1973), 267.278.

32. Kaplan, E., and Meier, P. Nonparametric estimation from incomplete observations. Journal of American Statistical Association 53 (1958), 457.481.

33. Keyfitz, N. and Littman, G. (1979). Mortality in a heterogeneous population. Population Studies, 33, 333-42.

34. Klein, J. P., and Moeschberger, M. L. Survival Analysis: Techniques for Censored and Truncated Data. Springer, New York, 1997.

35. Klein, J.P. (1992). Semi parametric estimation of random effects using the Cox model based on EM algorithm. Biometrics, 48, 795-806.

36. Klembaum, D. G. Survival Analysis: A Self learning text. Springer, New York, 1996.