Scientific Research Publishing

Open Journal of Statistics, 2021, 11, 931-962
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

The Cox Proportional Hazard Regression Model Vis-à-Vis ITN-Factor Impact on Mortality Due to Malaria

Anthony Joe Turkson*, John Awuah Addor, Francis Ayiah-Mensah

Department of Mathematics, Statistics and Actuarial Science, Takoradi Technical University, Takoradi, Ghana
Email: *anthony.turkson@ttu.edu.gh

How to cite this paper: Turkson, A.J., Addor, J.A. and Ayiah-Mensah, F. (2021) The Cox Proportional Hazard Regression Model Vis-à-Vis ITN-Factor Impact on Mortality Due to Malaria. *Open Journal of Statistics*, **11**, 931-962.
https://doi.org/10.4236/ojs.2021.116055

Received: September 28, 2021
Accepted: December 3, 2021
Published: December 6, 2021

Abstract

This study has provided a starting point for defining and working with Cox models in respect of multivariate modeling. In medical researches, there may be situations, where several risk factors potentially affect patient prognosis, howbeit, only one or two might predict patient's predicament. In seeking to find out which of the risk factors contribute the most to the survival times of patients, there was the need for researchers to adjust the covariates to realize their impact on survival times of patients. Aside the multivariate nature of the covariates, some covariates might be categorical while others might be quantitative. Again, there might be cases where researchers need a model that has the capability of extending survival analysis methods to assessing simultaneously the effect of several risk factors on survival times. This study unveiled the Cox model as a robust technique which could accomplish the aforementioned cases. An investigation meant to evaluate the ITN-factor vis-à-vis its contribution towards death due to Malaria was exemplified with the Cox model. Data were taken from hospitals in Ghana. In doing so, we assessed hospital in-patients who reported cases of malaria (origin state) to time until death or censoring (destination stage) as a result of predictive factors (exposure to the malaria parasites) and some socioeconomic variables. We purposefully used Cox models to quantify the effect of the ITN-factor in the presence of other risk factors to obtain some measures of effect that could describe the relationship between the exposure variable and time until death adjusting for other variables. PH assumption holds for all three covariates. Sex of patient was insignificant to deaths due to malaria. Age of patient and user status were both significant. The magnitude of the coefficient (0.384) of ITN user status depicts its high contribution to the variation in the dependent variable.

Keywords

Baseline Hazard, Cox Model, Hazard Function, Hazard Ratio, Survival Function

1. Introduction

Event history analysis is an omnibus term for the collection of statistical methods that focuses on the timing and occurrence of events. Reference [1] posits that survival analysis techniques model the probability of a change in a dependent variable Y_t from an origin state j to a destination state k, as a result of predictive factors, and that the duration of time between states is referred to as event time. Survival analysis models are used to examine the survival and hazard rates for some events of interest which are probabilistic in nature. One of the goals of survival analysis is to obtain some measures of effect that can describe the relationship between a predictor variable of interest and time to failure, after adjusting for the other variables, we have identified in the study and included in the model, this measure of effect is the hazards ratio [2]. Reference [3] has argued that survival analysis examines the effect of changes in the covariates on the duration of time preceding the event as well as the probability that the event will occur.

Standard procedures for survival and event history analysis involve modelling time to death or failure, often as a function of covariates, using either parametric or semiparametric approaches. Various parametric families of models are used in the analysis of lifetime data, including the exponential and the Weibull, with the latter being popular due to its flexibility. The Cox regression model is a cornerstone of modern survival analysis and is widely used in many other fields as well. The model is used to investigate the impact of various explanatory or predictor variables on the outcome or response variable with the view of identifying, salient and crucial variables which have telling effect on the study [4]. In mathematical terms, we can equally say that the Cox proportional hazards model is used to model survival data as a function of covariates. The purpose of the model is to evaluate simultaneously the effect of several factors on survival. In other words, it allows us to examine how specified factors influence the rate of a particular event happening at a particular point in time. This rate is commonly referred as the hazard rate [5]. Predictor variables (or factors) are usually termed covariates in the survival-analysis literature. The Cox model is expressed by the hazard function denoted by $h(t)$. Briefly, the hazard function can be interpreted as the risk of dying at time t. The response variable is the hazard function $h(t)$, which assesses the probability that the event of interest (in this case, death) occurred before time t. The equation models this hazard as an exponential function (exp) of an arbitrary baseline hazard h_0 when all covariates are null, and β is the regression coefficient of the covariate, x.

Though ordinary regression analysis (ORA) could be used to achieve the same purpose, statistician flown on its use due to the problem of incomplete data associated with most prospective studies. ORA cannot take into consideration partial or incomplete information from the entire study group. The Cox's proportional hazards regression model has the ability of taking into consideration partial information from censored data as well as full information from uncensored data and is therefore more appropriate in such situations. The Cox proportional hazard model is a statistical method that finds the cumulative probability of an event, it also accounts for impact of covariates on that probability. The model works for both quantitative predictor variables and for categorical variables. Furthermore, the Cox regression model extends survival analysis methods to assess simultaneously the effect of several risk factors on survival time [6].

In the Cox model, we take cognizance of two types of covariates; those that depend on time, and those which are time-independent. In this paper, we limited our discussion on time-independent covariates. We also looked at some aspects of the Cox proportional hazards regression model. Special emphasis was placed on the following areas: how to develop the model; popularity of the model; hypothesis testing for proportional hazards model; the stratified Cox model; meaning of the proportional hazard (PH) model; failure of the PH model; testing of the proportional hazards model; alternative method for assessing the PH model; hazard ratio; the likelihood ratio test; and graphical approach to the lol-log plot. In order for researchers to apply the model to life and properties, it is deemed expedient to subject covariates to empirical survival analytic studies. Applying the theories behind Cox models is particularly useful in examining treatment comparisons based on the time to some events while adjusting for the effect of concomitant variables. It is useful for predictions to maintain optimal maintenance policies in engineering, medical and biomedical studies. The Cox model has the advantage of preserving the variable in its original quantitative form, and of using a maximum of information [7] [8]. The paper is organized as follows; it provides:

1) Theoretical framework which underpins the study.

2) Theories by which Cox model could be laid out.

3) Simulation study on the Cox model.

4) Real case empirical studies with apt interpretation of the outcome.

This study has the propensity of supporting enquirers in understanding and interpreting the hazard ratio as a measure of effect that describes the relationship between the predictor variables and time to failure (time to obtaining the event of interest).

1.1. Theoretical Framework Underpinning the Study

Reference [9] employed Cox proportional hazard regression as a less parametric alternative to generalized linear model (GLM) and ordinary least squares model (OLS) even when there was no need to correct for censoring. They examined how well the alternative estimators behaved econometrically in terms of bias when the data were skewed to the right. Specifically, they provided evidence on the performance of the Cox model under a variety of data generating mechanisms and compared it to the estimators studied recently in [10]. They noted that the gamma regression model with a log link seemed to be more robust to alternative data generating mechanisms than either OLS on ln(y) or Cox proportional hazards regression. In conclusion they found out that the proportional hazard assumption was an essential requirement to obtaining consistent estimate of the $E(y|x)$ using the Cox model.

Reference [11] proposed the use of the Cox proportional hazard model (CPHM) for the analysis of early-failure data associated with power cables. They alluded to the fact that the CPHM analyses simultaneously a set of covariates and identifies those which have significant effects on the cable failures. In order to demonstrate the appropriateness of the model, they obtained relevant historical failure data related to medium voltage (MV, rated at 10 kV), distribution cables and high voltage (HV, 110 kV and 220 kV). The transmission cables' data were collected from a regional electricity company in China. It was revealed in the study that the model was more robust than the Weibull distribution, again, it was demonstrated that the method provided could single out a case of poor manufacturing quality with a particular cable joint by using a statistical hypothesis test. They proposed an approach which could potentially help resolve any legal dispute that may arise between a manufacturer and a network operator. Reference [12] underscored the fact that the Cox proportional hazard model was one of the most common methods used in time to event data analysis. He disclosed that the model was based on several restrictive assumptions one of which concerned tied events, he presented and compared five ways for handling tied events. On the basis of the calculations performed, it could be stated that exact expression and the discrete model gave the best results in terms of fit statistics; even though they were the most time-consuming. Efron and Breslow approximations were much faster but had the worst model fit. They performed a simple method which was based on subtracting a tiny random value from each tied survival time. It was revealed that, if estimation precision was not as important as estimation time, then Breslow or—more preferably—Efron approximations might be used. But if time was of the essence, then one should consider choosing an exact method or discrete model that can provide better fit statistics and more efficient parameter estimates.

Reference [13] posited that the survival of patients after surgery depended much on identifying risk factors through the use of appropriate tools and dealing with the risk factors. They obtained data from 330 gastric cancer patients diagnosed at the Iran cancer institute during the 1995-99 period, the patients were followed up to the end of 2011. The survival status of these patients in 2011 was determined by reopening the files as well as phone calls and the effect of various factors such as demographic, clinical, treatment, and post-surgical on patients' survival were studied. Based on Cox-Snell Residuals and Akaike Information Criterion (AIC), it was revealed that the exponential (AIC = 969.14) and Gompertz (AIC = 970.70) models were more efficient than other accelerated failure-time models. The results of the Cox proportional hazard model as well as the analysis of accelerated failure-time models showed that variables such as age (at diagnosis), marital status, relapse, number of supplementary treatments, disease stage, and type of surgery were among factors affecting survival ($p < 0.05$).

Reference [14] examined the interactive effects of Cyclooxygenase-2 (COX-2) expression, Federation of Gynecology and Obstetrics (FIGO) stage, Vascular Endothelial Growth Factor (VEGF) expression and histological grade prognostic factors using the multivariate Cox proportional hazards model. The result revealed that the risk of death for the patients with COX-2 positive expression was 2.8 times than that with COX-2 negative expression. For FIGO stage, VEGF expression and histological grade, their risk of death were 2.2, 2.1, and 1.84 times respectively. It was concluded that COX-2 expression, FIGO stage, VEGF expression and histological grade were the most important prognostic factors for Emergency Operations Center (EOC) after curative resection. Reference [15] utilized data from the Bangladesh demographic and health survey (BDHS), 2014 to identify the determinants of neonatal, infant and under-five mortality in Bangladesh. They performed Log-rank test for bivariate analysis and applied Cox proportional hazard model. The results revealed that maternal education, region, exposure to NGO activities were significant determinants of under-five and infant mortality, whereas region of descent, gender of child, child's size at birth played significant role in reducing neonatal mortality. It was concluded that policymakers should give priority to maternal education, delve into regional issues that affect neonatal mortality, and consider issues about child's size at birth as well as engage non-governmental organizations (NGO) to assist in reducing neonatal, infant and under-five mortality in Bangladesh.

Reference [16] noted that the Cox proportional-hazards regression model had achieved widespread use in the analysis of time-to-event data with censoring and covariates. They noted that the covariates may change their values over time and therefore discussed the use of such time-dependent covariates. They further noted that the interrelationships between the outcome and variables over time could lead to bias unless the relationships were well understood. They indicated that the form of a time-dependent covariate was much more complex than in Cox models with fixed (non-time-dependent) covariates and that constructing it involves a function of time. In the study [16], child mortality was considered as the dependent variable. Child Mortality was deemed to measure the probability of dying between the age of one and four years (expressed per 1000 live births). They also considered several important socioeconomic and demographic predictors which included the following: Age of women (15 - 19, 20 - 24 and 25 - 49 years); education of women (illiterate, literate but below primary, primary but below middle, middle but below high school and high school and above); place of residence (rural and urban); child's gender (Female and Male); mass media exposure (no exposure and any exposure); wealth quintile (poorest, poorer, middle, richer and richest); religion (Hindu, Muslim and others); caste (Scheduled Caste (SC), Scheduled Tribe (ST), Other Backward Class (OBC) and others); birth order (1, 2 - 3 and 4 or more); birth Interval (less than 2 years and greater than 2 years); parity (1 - 2, 3 - 4 and ≥5); working status of women (Not working, working at home and working away from home); women empowerment (not empowered, partially empowered and Fully empowered); and region.

Reference [17] conducted a study on factors influencing women's waiting time to first birth in Bangladesh, they applied the Cox proportional hazard model. In their study, the event of interest was waiting time to first birth after marriage. The variable could not be obtained directly and therefore they used the difference between the age of the women at first birth and age at first marriage as the waiting time to first birth. Women who were still waiting for their first birth after termination of study were considered to be censored. The event of interest variable was measured in months. The censoring indicator was equal to 1 if the observation was found to have had their first child and 0 if they did not have any child. Some demographic and socio-economic variables were selected as explanatory variables—few of these were: Current working status; age of woman, region of descent, type of residence; religious affiliation; educational level; household head; media influence; ideal number of children; wealth quintile; partners level of education; and occupation of partner.

1.2. Developing the Cox Proportional Hazards Regression

The difficulties one encounters with parametric models can be resolved with the proportional hazard's models. For two individuals who differ only in the relevant membership (e.g., treatment verses control) their predicted log-hazard will differ additively by the relevant parameter estimate, which is to say, their predicted hazard rate will differ by e^{β}, *i.e.*, multiplicatively by the anti-log of the estimate. Thus, the estimate can be considered a hazard ratio, that is, the ratio between the predicted hazard for a member of one group and that for a member of the other group, holding everything else constant. For a continuous explanatory variable, the same interpretation applies to a unit difference. Other hazard rate models have different formulations and the interpretation of the parameter estimates differs accordingly.

Assuming that the value of the covariate *x*, is fixed and does not change over time, the regression model will be

$$y = \beta_0 + \beta_1 x + \sigma\varepsilon^* . \tag{1}$$

where $y = \ln(t)$ and ε^* is $\ln\varepsilon$.

Expressed on a time scale, the model becomes

$$t = \varepsilon^{\sigma}\mathrm{e}^{\beta_0+\beta_1 x} . \tag{2}$$

In survival analysis, the survival time is determined by a systematic component $\beta_0 + \beta_1 x$ and the error component ε. Choosing a parametric distribution for the error component ($\sigma = 1$). The hazard function will also assume a new parametric structure. The hazard function for a subject with covariate equal to x will be

$$h = (t, x, \beta) = \mathrm{e}^{-(\beta_0 + \beta_1 x)}. \tag{3}$$

Two points worth knowing here are that

1) The hazard function does not depend on time; its value is determined by the covariate x and the unknown parameters β_0 and β_1

2) The hazard function and systematic component in the regression model are inversely related.

The fact that the hazard does not depend on time means that the risk of failure is the same no matter how long the subject is followed. Models that are used to describe survival times in a comparative sense are often called semi-parametric regression models. Typically, when we want to compare the survival experience of sub-groups, we need to specify the hazard function as a function of time and covariates.

$$h(t, x, \beta) = h_0(t) \times r(x, \beta) \tag{4}$$

The hazard function in Equation (4) is the product of two functions. The function $h_0(t)$ shows how the hazard function changes with survival time. The other function $r(x, \beta)$ shows how the hazard function changes as a function of the subject covariates.

The functions must be selected such that $h(t, x, \beta) > 0$

When the $r(x, \beta) = 1$, $h_0(t)$ is referred to as the baseline hazard function.

Under the model in Equation (4), the ratio of the hazard functions of two subjects with covariate values denoted x_1 and x_0 is given as

$$HR(t, x_1, x_0) = \frac{h(t, x_1, \beta)}{h(t, x_0, \beta)}$$

$$HR(t, x_1, x_0) = \frac{h_0(t) r(x_1, \beta)}{h_0(t) r(x_0, \beta)} = \frac{r(x_1, \beta)}{r(x_0, \beta)} \tag{5}$$

The hazard ratio HR depends only on the function $r(x, \beta)$. If the ratio in Equation (5) is easily interpreted then the baseline function which is a function of time is of little importance.

Reference [18] was the first to propose that in the model of Equation (5)

$$r(x, \beta) = \exp(x\beta)$$

With this parameterization the hazard function is now denoted by

$$h(t, X, \beta) = h_0(t) \exp\left(\sum_{i=1}^{n} \beta_i x_i\right) \tag{6}$$

and the Hazard Ratio

$$HR(t, x_1, x_0) = \mathrm{e}^{\beta(x_1 - x_0)} \tag{7}$$

More generally, we can write Equation (6) as

$$h(t,X,\beta)=h_0(t)\exp(\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n)$$

where $h(t,X)$ is the hazard at time *t* for a subject with a set of predictors $X=x_1,x_2,\cdots,x_n$, $h_0(t)$ is the baseline hazard function, and $\beta=\beta_1,\beta_2,\cdots,\beta_n$ are the model parameters describing the effect of the predictors on the overall hazard. The interpretation of the Cox model is done using hazard ratios (*HR*), defined as the ratio of the predicted hazard function under two different values of a predictor variable. This model is referred to in the literature by a variety of terms such as the Cox model, the Cox proportional hazards model or simply the proportional hazards model. They are also called semi-parametric functions because the baseline hazard function $h_0(t)$ is not explicitly defined. Characteristics of the Cox model are outlined below:

- it is a product of a function in *t* and a function in *X*;
- *X* is time independent;
- the baseline hazard is an unspecified function, making it a semi-parametric model

Equation (7) can be interpreted as "relative risk".

The coefficients $\beta_1,\beta_2,\cdots,\beta_k$ are estimated by Cox regression, and can be interpreted in a similar manner to that of multiple logistic regressions.

Suppose the covariate (risk factor) is dichotomous and is coded 1 if present and 0 if absent. Then the quantity $\exp(\beta_1)$ can be interpreted as the instantaneous relative risk of an event at any time, for an individual with the risk factor present compared with an individual with the risk factor absent, given that both individuals are the same on all other covariates. Suppose the covariate is continuous, then the quantity $\exp(\beta_1)$ is the instantaneous relative risk of an event, at any time, for an individual with an increase of 1 in the value of the covariate compared with another individual, given both individuals are the same on all other covariate. For example when a covariate is dichotomous such as gender with a value of $x_1=1$ for males and $x_0=0$ for females, the hazard ratio in Equation (7) becomes $HR(t,x_1,x_0)=\mathrm{e}^{\beta}$, if the value of the coefficient is $\beta=\ln(1.5)$ then $HR(t,x_1,x_0)=\mathrm{e}^{\ln(1.5)}=1.5$ which means that males are failing one and a half times that of females. A hazard ratio of one (1) means that there is no effect. One (1) is the null value for the exposure-outcome relationship. The term proportional hazards refer to the fact that the hazard functions are multiplicatively related, that is to say, their ratios are constant over survival time. In assessing the validity of the model, this assumption is important. One way to specify the distribution of survival time is through the hazard function. If we use the relationship between the survival function and the hazard function $S(t)=\mathrm{e}^{-H(t)}$ where $H(t)=-\ln(S(t))$, then

$$S(t,x,\beta)=\mathrm{e}^{-H(t,x,\beta)} \tag{8}$$

where $H(t,x,\beta)$ is the cumulative hazard function at time *t* for a subject with covariate *x*.

One important decision in survival analysis is how to properly model the conditional hazard rate of failure given certain predictor variables (covariates); this is due to the fact that statisticians are interested in finding out whether the predictor variables are correlated or uncorrelated with the survival or failure times. The model provides a technique for exploring the association of predictor variables with failure times and survival distributions; it is also used for studying the effect of a primary covariate or a predictor of interest while adjusting for other variables. This model assumes that given an m-dimensional vector of covariates Z, the conditional hazard rate given by,

$$h(t \mid z) = \lim_{\Delta t \to 0+} \frac{1}{\Delta t} P\{t \le T < t + \Delta t \mid T \ge t, Z = z\} \tag{9}$$

is a function of the independent predictor variables,

$$h(t \mid z) = h_0(t)\Re(z). \tag{10}$$

The function $\Re(z) = \exp(\psi(z))$ is parameterized as follows $\psi(z) = \beta^{\mathrm{T}} z$ with $\beta = (\beta_1, \beta_2, \cdots, \beta_m)^{\mathrm{T}}$ being a vector of unknown parameters, $z = (z_1, z_2, \cdots, z_m)^{\mathrm{T}} = (\varphi_1(x), \varphi_2(x), \cdots, \varphi_m(x))^{\mathrm{T}}$, is a vector of specified functions φ_i. Also, $h_0(t)$ is an unknown baseline hazard function. Once the conditional hazard rate is given, the conditional survival function $S(t \mid z)$ and conditional density function $f(t \mid z)$ can also be determined. The relationship between the hazard rate, survival function and density function are given below

$$S(t \mid z) = \exp(-H(t \mid z)), \quad f(t \mid z) = h(t \mid z) \cdot S(t \mid z). \tag{11}$$

where $H(t \mid z) = \int_0^t h(t \mid z) \mathrm{d}t$ is the cumulative hazard function.

Not all the survival times $T_1, T_2, \cdots, T_N$ were fully observed, instead one observes for the ith subject an event time $X_i = \min(T_i, C_i)$, where T_i and C_i are respectively the failure and censoring times of the ith subject. The censoring indicator $\delta_i = I(T_i \le C_i)$, as well as an associated vector of covariates Z_i can be denoted as follows:

$$\{(Z_i, X_i, \delta_i) : i = 1, 2, \cdots, N\}.$$

Which is an Independent Identical Distribution sample from the population (Z, X, δ) with $X = \min(T, C)$ and $\delta = I(T \le C)$. If the random variable T and C are positive and continuous then

$$\Re(x) = \frac{E(\delta \mid Z = z)}{E\{H_0(X) \mid Z = z\}}. \tag{12}$$

where $H_0(X) = \int h_0(u) \mathrm{d}u$, is the cumulative baseline hazard's function. This function allows one to estimate the function $\Re$ using regression techniques if $h_0(X)$ is known. The likelihood function can also be derived.

When $\delta = 0$, all we know is the survival time $T_i \ge C_i$ and the probability of getting this is

$$P(T_i \ge C_i \mid Z_i) = P(T_i \ge X_i \mid Z_i) = S(X_i \mid Z_i). \tag{13}$$

When $\delta = 1$, the likelihood of getting T_i is $f(T_i \mid Z_i) = f(X_i \mid Z_i)$.

Therefore, the conditional likelihood for getting the data is

$$L=\prod_{\delta_i=1} f\left(X_i \mid Z_i\right) \prod_{\delta_i=0} S\left(X_i \mid Z_i\right)=\prod_{\delta_i=1} h\left(X_i \mid Z_i\right) \prod_{i} S\left(X_i \mid Z_i\right) \tag{14}$$

$$\begin{aligned} L &= \sum_{\delta_i=1} \log\left\{h\left(X_i \mid Z_i\right)\right\}-\sum_{i} H\left(X_i \mid Z_i\right) \\ &= \sum_{i} \delta_i \log\left\{h\left(X_i \mid Z_i\right)\right\}-\sum_{i} H\left(X_i \mid Z_i\right) \end{aligned} \tag{15}$$

From the proportional hazards model

$$L=\sum_{i} \delta_i \log\left\{h_0\left(X_i\right)\Re\left(Z_i\right)\right\}-\sum_{i} h_0\left(X_i\right)\Re\left(Z_i\right) \tag{16}$$

If both the functions $\psi(\cdot)$ and $h_0(\cdot)$ are parameterized, the parameters can be estimated by maximizing the likelihood in Equation (16)

1.3. Popularity of the Cox Proportional Hazard Model

There are several reasons why the Cox model is very popular, six of them are listed here:

- The Cox model is robust. It usually fits the data well no matter which parametric model is used;
- We can get the estimate of the effect without knowing $h_0(t)$;
- The estimated hazards are always non-negative;
- The β, s can be estimated and the hazard ratio calculated;
- The hazard function $h(t,X)$ and the survival function $S(t,X)$ can be estimated; and
- The Cox model is preferred over the logistic model which ignores survival time and censoring information.

1.4. Hypothesis Testing for Proportional Hazard Models

One way of finding out if the predictor variables really contribute to the risk or hazard function (after fitting the Cox model) is to conduct a test of hypothesis. There are two tests that will be very useful in testing this hypothesis. They are the Wald and the likelihood ratio tests: For models with multiple parameters, it is convenient to use the Wald test for one parameter at a time. When fitting different nested models, the likelihood ratio test is most convenient.

For a test of a single parameter being equal to 0, the Wald test statistic is:

$Z^2=\dfrac{\beta^2}{V^2(\hat{\beta})}$, If H_o is true, $Z^2 \approx \chi^2$ (or, equivalently, $z \approx N(0,1)$ Large values of Z^2 support the alternative hypothesis. For multivariate models, a version of the Wald test exists, which comes from a χ^2 distribution with more degrees of freedom, but we will rarely need this. The likelihood ratio test statistic for the hypothesis that a single parameter is equal to zero is

$G=-2\left[L_p(\text{reduced})-L_p(\text{full})\right]$ If H_o is true $G \approx \chi^2$. For tests of multiple parameters being equal to zero, the degrees of freedom increase as explained below

Let us consider the hypothesis:

$$H_k : \beta_{k1} = \cdots = \beta_{kl} = 0, \quad \left(1 \le k_1 \le k_2 \le \cdots \le k_l\right). \tag{17}$$

The hypothesis in Equation (17) is rejected at a significance level of α if

$$LR_k > \chi^2_{1-\alpha}\left(l\right), \text{where } \; k = k_1, k_2, \cdots, k_l$$

If H_k is satisfied, the variables $x_{k1}, x_{k2}, \cdots, x_{kl}$ will be excluded from the model, therefore the hypothesis becomes

$$H_{1,2,\cdots,m} : \beta_1 = \cdots = \beta_m = 0 \tag{18}$$

Equation (18) means that none of the predictor variables identified contributed to the hazard or risk of death. The hypothesis in (17) is rejected if;

$$LR_{1,2,\cdots,m} > \chi^2_{1-\varepsilon}\left(m\right)$$

$$\text{The hypothesis: } \; H_j : \beta_j = 0, \quad \left(j = 1, 2, \cdots, m\right). \tag{19}$$

This implies that the model with or without the predictor variables gives the same results.

The hypothesis in Equation (19) is rejected if; $LR_j > \chi^2_{1-\varepsilon}\left(1\right)$

1.5. Stratified Cox (SC) Procedure

The stratified Cox proportional hazard model allows the underlying hazard function to vary across the strata variables. The procedure demands that the Cox proportional hazards (PH) model is modified to make provision for control by stratifying a variable that fails to satisfy the PH assumption. Variables that satisfy the PH assumption are included in the model, whereas the variables that fail to satisfy the PH assumption are stratified by their non-inclusion in the model.

Let's assume that $Z_1, Z_2, \cdots, Z_K$ do not satisfy PH, and $X_1, X_2, \cdots, X_P$ do satisfy the PH assumption. We will define a new variable $\overset{*}{Z}$ from the *Z*'s which will be used for the stratification. If race and sex do not satisfy the PH assumption then we can form combinations from the categories as follows

From Table 1, we note therefore that $\overset{*}{Z}$ has $k = 6$ categories or strata.

The hazard function for the stratified Cox model is given below

$$h_g\left(t, X\right) = h_{0g}\left(t\right)\exp\left[\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_P\right], \quad g = 1, 2, \cdots, k \left(k = 6\right) \tag{20}$$

$\overset{*}{Z}$ is not included in the model but $X_1, X_2, \cdots, X_P$ are included in the model In this model there will be different baseline hazards functions

$h_{0g}\left(t\right), g = 1, 2, \cdots, k$ but same coefficients $\beta_1, \beta_2, \cdots, \beta_p$. The fitted SC model will yield different estimated survival curves for each stratum because the baseline hazards are different for each stratum. However, since the coefficients of the *X*'s

Table 1. Table exemplifying how to stratify by the variable race.

	White 1	Black 2	Asian 3
Male 1	1.1	1.2	1.3
Female 2	2.1	2.2	2.3

are the same for each stratum, estimates of the hazard's ratio will be the same. This feature of the SC model is referred to as the no-interaction assumption. The no interaction assumption implies that the hazard ratios are the same for each stratum.

If the only predictor that failed to satisfy the PH assumption is sex and the covariates are Age(X_1) and user status (X_2) then the SC model becomes

$$h_g(t,X)=h_{0g}(t)\exp\left[\beta_1 X_1+\beta_2 X_2\right], \text{ with } g=1, 2 \text{ [male = 1, female = 2]}$$

or

$$h_1(t,X)=h_{01}(t)\exp\left[\beta_1\text{Age}+\beta_2\text{User Status}\right] \quad \text{for males} \tag{21}$$

$$h_2(t,X)=h_{02}(t)\exp\left[\beta_1\text{Age}+\beta_2\text{User Status}\right], \text{ for females} \tag{22}$$

In the models above age and user status are in the model whereas sex is not in the model Sex is therefore controlled by stratification. Since the age and user status variables are included in the model, we can estimate the effect of each variable adjusted for the effect of the other variable and sex. The estimated hazard ratio for the effect of age adjusting for user status and sex is given by e^{β_1}, and that for 'user status' adjusting for age and sex is given by e^{β_2}

We use the stratified Cox model to control for the sex variable which does not satisfy the PH assumption. The implication here is that the sex variable is being adjusted for stratification, we have also included the age and preventive measure variable (which do satisfy the PH assumption) into the model. In other words, the age and preventive measure variable have been adjusted by their inclusion into the model. In the model we can infer that the hazard ratio for the effect of the preventive measure variable adjusted for age and sex is given by the value 1.452, this value can be interpreted to mean that the exposed group (that is the group that do not use the insecticide treated net) has 1.5 times the hazard of death as compared to the less exposed group (group that use ITN) reference [19].

1.6. The Meaning of the PH Assumption

The PH assumption requires the hazard ratio (*HR*), defined as the ratio of the predicted hazard function under two different values of a predictor variable to be constant over time. In other words, the hazard function for one individual should be proportional to the hazard function for any other individual. Moreover, the proportionality constant should be independent of time, that is to say, at any time *t*, $\frac{h_i(t)}{h_j(t)}=C$. where C is a constant. C may depend on the explanatory variables but not on time**.** Graphically, the hazards for different individuals on the same graph should not cross paths. The rule is that if the hazards cross paths, then the PH assumption is violated, resulting in the inappropriateness of the use of the Cox PH model. It should be noted that a bit of crossing at early time points may be a product of noise in the survival estimates and may not constitute a violation of the proportional hazard's assumption.

There are a variety of techniques, both graphical and test-based, for assessing the validity of the proportional hazard's assumption. One technique is to simply plot Kaplan-Meier survival curves to compare two groups with no covariates. If the curves cross each other, the proportional hazards assumption is violated. If on the other hand the curves do not cross the path of each other, then the PH assumption is satisfied. An important caveat to this approach must be kept in mind for small studies. There may be a large amount of error associated with the estimation of survival curves for studies with a small sample size; therefore, the curves may cross even when the proportional hazards assumption is met. The complementary log-log plot is a more robust test that plots the logarithm of the negative logarithm of the estimated survivor function against the logarithm of survival time. If the hazards are proportional across groups, this plot will yield parallel curves. Another common method for testing the proportional hazards assumption is to include a time interaction term to determine if the *HR* changes over time, since time is often the culprit for non-proportionality of the hazards. If the group time interaction term is not zero, it is evidence against proportional hazards.

1.7. Failure of the Proportional Hazards Assumption

If the PH assumption does not hold, there are options for improving the non-proportionality in the model. Other new covariates can be included in the model, again, non-linear terms for existing covariates, or interactions among covariates can be incorporated. Alternatively, the model could be stratified in the analysis on one or more variables. This approach will lead to estimates of a model in which the baseline hazard is allowed to be different within each stratum, but the covariates effects are equal across strata. Other options include dividing time into categories and using indicator variables to allow hazard ratios to vary across time, and changing the analysis time variable (e.g., from elapsed time to age or *vice versa*).

1.8. Alternative Method for Assessing the PH Assumption

The goodness of fit approach is appealing because it provides a test statistics and p-value for assessing the PH assumption for a given predictor of interest. This approach was originally proposed by Schoenfeld but has been modified in [20] and is based on the residuals defined by Schoenfeld now known as the Schoenfeld residuals. For each predictor in the model Schoenfeld residuals are defined for every subject who has an event [19]. The steps for running the test are based on the null hypothesis that; 'The correlation between the Schoenfeld residuals and the ranked failure time is zero, that is $H_0 = \rho = 0$' The outline follows below:

- Run a Cox PH model and obtain Schoenfeld residual for each predictor;
- Create a variable that ranks the order of failure. The subject who had the first event gets a value of 1; the next gets a value of 2 and so on;

- Save the Schoenfeld residuals of the model and the scaled Schoenfeld residuals;
- For persons censored, the value of the residual is set to missing; and
- Test the correlation between the variables created in the first and second steps.

If the null hypothesis is rejected, we will conclude that the PH assumption is violated, otherwise, it is not violated.

1.9. Hazard Ratio

A hazard ratio is a measure of how often an event of interest happens in one group compared with another group. Hazard ratio (*HR*) is a measure of an effect of an intervention on an outcome of interest over time. The outcome could be an adverse/negative outcome or a favorable/positive outcome. Hazard Ratio (*i.e.*, the ratio of hazards) = Hazard in the treatment group divided by Hazard in the control group. It is also termed as the instantaneous relative risk. If the predictor variable is continuous, then the quantity $\exp(\beta_1)$ is the instantaneous relative risk of an event, at any time. For an example, if a predictor variable is dichotomous such as presence/absence of virus with a value of $x_1 = 1$ for presence and $x_0 = 0$ for absence, the hazard ratio becomes $HR(t, x_1, x_0) = \exp(\beta)$. If the value of the coefficient is $\beta = \log(2.5)$ then $HR(t, x_1, x_0) = \exp\{\log(2.5)\} = 2.5$ which is interpreted to mean that with the presence of the virus patients are failing two and a half times that of absence of a virus.

A hazard ratio greater than one (1) indicates that the covariate is positively associated with the probability of the event and negatively associated with the length of survival time.

In summary,

- *HR* = 1.0, implies equal risk rates. (No effect, differences are likely due to chance);
- *HR* > 1.0, implies increased risk rate in control group (increase in Hazard); and
- *HR* < 1.0, implies decreased risk rate in control group (reduction in the hazard).

In studies on cancer patients:

- A covariate with hazard ratio > 1 (*i.e.*: $\beta > 0$) is called bad prognostic factor; and
- A covariate with hazard ratio < 1 (*i.e.*: $\beta < 0$) is called good prognostic factor.

The computation of the hazard ratio assumes that the ratio is consistent over time; therefore, if the survival curves cross, the hazard ratio statistic should be ignored. The term proportional hazards refer to the fact that the hazard functions are multiplicatively related, that is to say their ratios are constant over survival time [19]. In assessing the validity of the model, this assumption is important.

While a hazard ratio (*HR*) and relative risk (*RR*) are similar in some aspects, there is a slight difference between the two. For instance, in a clinical trial, a researcher might investigate the Hazard rates and Relative risk for two types of drug users: user X and user Y. Assuming that both the hazard ratios (*HR*) and relative risk (*RR*) were 3.0, then the interpretation of the results is as follows:

- The relative risk (*RR*) tells us that the risk of death is three times higher with user X than with user Y over the entire period of the study. *RR* does not care about the timing of the event.
- The hazard ratio (*HR*) tells us that the risk of death is three times higher with user X than with user Y at any particular point in time. *HR* cares about the total number of events and also the timing of the events.

The distinguishing feature is the timing or time period under consideration. In evaluating Hazard ratios, it is imperative that we support our results with other measures like the median survival time, overall survival, or time to progression.

1.10. The Likelihood Ratio Test

To help choose between two alternatives; random verses systematic variation based on the observed difference between two log-likelihood values generated from two statistical models, the application of a theorem from theoretical statistics has been proposed [21]. The theorem states that the difference between two log-likelihood values multiplied by −2 has an approximate chi-square distribution when three conditions hold. The first condition is that the two models generating the log-likelihood values must be calculated from exactly the same data. The second is that the compared models must be nested (That is, one model is a special case of the other). The third condition is that the two log-likelihoods must differ only because of random variation. When the first two conditions apply, a test statistic with a chi-square distribution produces an assessment of the plausibility of the third condition.

The likelihood ratio test can be used to perform several tests. For instance, it could be used to test the significance of an interaction term in a model and the significance of a covariate in a model after adjusting for the other covariates. To test the significance of a covariate like usage of ITN, we need to compute the difference between the log likelihood statistics of the reduced model which does not contain the covariate and the likelihood statistics of the full model containing the covariate. The formula is given below;

$$LR = -2\log L_R - \left(-2\log L_F\right). \tag{23}$$

where, R denotes the reduced model and F the full model.

It has been indicated in [19] that the LR statistics is a chi-square statistic χ^2 with p degrees of freedom where p is the number of covariates or predictors being assessed (in this example $p = 1$) under the null hypothesis that the covariate is not significant.

1.11. Graphical Approach to Log-Log Plot

This plot is simply a transformation of an estimated survival curve that results from taking the natural logarithm of an estimated survival probability twice, that is $-\log\left(-\log\hat{S}\right)$.

Under the Cox model the survivorship function is

$$S(t,x,\beta)=\left[S_0(t)\right]^{\exp(x\beta)}. \tag{24}$$

Alternatively, we can write it as $S(t,X)=\left[S_0(t)\right]^{\exp\sum_{j=1}^{p}\beta_j x_j}$, where $S_0(t)$ is the baseline survival function.

Taking the log of the expression twice we shall obtain

$$\log S(t,X)=\exp\left(\sum_{i=1}^{p}\beta_i x_i\right)\times\log S_0(t)$$

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

$$-\log\left[-\log S(t,X)\right]=-\left[\sum_{i=1}^{p}\beta_i x_i\right]-\log\left[-\log S_0(t)\right]$$

Note: $\log S(t,X)$ and $\log S_0(t)$ are negative but $-\log S(t,X)$ is positive. For two subjects

$$X_1=\left(X_{11},X_{12},\cdots,X_{1P}\right),\quad X_2=\left(X_{21},X_{22},\cdots,X_{2P}\right)$$

$$-\log\left[-\log S(t,X_1)\right]=-\left[\sum_{i=1}^{p}\beta_i x_{1i}\right]-\log\left[-\log S_0(t)\right]$$

$$-\log\left[-\log S(t,X_2)\right]=-\left[\sum_{i=1}^{p}\beta_i x_{2i}\right]-\log\left[-\log S_0(t)\right]$$

$$\begin{aligned}&-\log\left[-\log S(t,X_1)\right]-\left(-\log\left[-\log S(t,X_2)\right]\right)\\&=-\left[\sum_{i=1}^{p}\beta_i x_{1i}\right]-\log[-\log S_0(t)-\left\{-\left[\sum_{i=1}^{p}\beta_i x_{2i}\right]-\log\left[-\log S_0(t)\right]\right\}\end{aligned}$$

$$\begin{aligned}&\therefore\log\left[-\log S(t,X_1)\right]-\left(-\log\left[-\log S(t,X_2)\right]\right)\\&=\left[-\sum_{i=1}^{p}\beta_i x_{1i}\right]+\left[\sum_{i=1}^{p}\beta_i x_{2i}\right]=\sum_{i=1}^{p}\beta_i\left(x_{2i}-x_{1i}\right)\end{aligned}\tag{25}$$

Alternatively, if the predictor variables are time independent, then the PH model is given by

$$h(t\mid z)=h_0(t)\Re(z). \tag{26}$$

where $\Re(z)=\exp(\psi(z))$ and $\psi(z)=\beta^{\mathrm{T}}z$, with $\beta=\left(\beta_1,\beta_2,\cdots,\beta_m\right)^{\mathrm{T}}$.

Then we can render Equation (26) in terms of the exponential function we obtain

$$h(t\mid z)=h_0(t)\exp\beta^{\mathrm{T}}z \tag{27}$$

And taking the natural logarithm at both sides of Equation (27) gives

$$\log h(t \mid z) = \log h_0(t) + \beta^{\mathrm{T}} z$$

For different units of the vector z say x_1 *and* x_2, with corresponding coefficients β_1 and β_2, we obtain

$$\log h(t \mid x_1) = \log h_0(t) + \beta_1 x_1 \tag{28}$$

$$\log h(t \mid x_2) = \log h_0(t) + \beta_2 x_2 \tag{29}$$

We see from Equations (28) and (29) that the baseline hazards are constant in both cases, they do not contribute to the predictions. If the PH assumption is satisfied for this data, then the graphs of the time functions Equations (28) and (29) will be approximately parallel. The graph of the differences between Equations (28) and (29) does not involve *t*. The formula says that if we use the Cox PH model and plot the estimated log-log survival curves for two subjects on the same graph, the two graphs would be approximately parallel and the distance between the two curves is the linear expression involving the differences in predictor values which does not involve *t*. This parallelism of the log-log survival plots for the Cox PH model provides us with a graphical approach for assessing the PH assumption [22].

2. Materials and Method

2.1. Target Population

The target population was all the residents of Sekondi-Takoradi district in Ghana. Data on malaria cases was obtained from three hospitals in the district, using observational studies, interviews and records from the records department. {Malaria accounts for about 1 million deaths in Africa annually and has slowed economic growth in African countries by up to 1.3% per year. Insecticide-treated nets (ITNs) undergo a series of tests to obtain listing by World Health Organization (WHO) prequalification. These tests characterize the bio-efficacy, physical and chemical properties of the ITN. ITN procurers assume that product specifications relate to product performance [23] [24]. The observational studies were carried out on patients who had been diagnosed of severe malaria and were on admission at the hospital. The study spanned over a 4-month period beginning from 1 September 2009 to 31 December 2009. The patients were enrolled into the study at different times as and when they were diagnosed and admitted. Within the study period patients who were discharged were treated as censored, those who died from a different ailment besides malaria were treated as censored, those who died from malaria were treated as patients who obtained the event of interest. At the end of the study period, all patients who were still on admission were considered as censored. In all a total of 1793 patients were enrolled into the study. The patients were made up of males, females, young and old, rich and poor, those from the country side and those from the cities. For each patient, data on the following were obtained, age on admission, gender, level of exposure indicative by type of mosquito net used at home, date of admission, date of discharge/death, cause of death and censoring status. The assumptions made about the patients were that all of them received the same treatment once they were on admission; a further assumption was that those on admission were considered as first-time in-patients who had had no previous admission records.

2.2. Data Analysis

The extracted data for each person was coded as follows: Gender: Male = 1; Female = 2.

Type of preventive measure used: Insecticide treated net = 0; mosquito nets plus others = 1.

Censoring status: if patient died it is coded as = 1; If patient was discharged, died from a different sickness or was alive at the end of study, the code was = 0

Difference between date of discharge/death and date of admission = survival time in days. Age was seen as a continuous quantitative variable. The coded variables were keyed into SPSS version 20 and analyzed using survival analysis models. The survival experiences of the two exposure groups were compared and contrasted using the Kaplan-Meier survival curves. The Cox proportional hazards model was used firstly to assess the risk of malaria-death for the two exposure groups; secondly, it was used to explore the relation between the baseline risk factor (malaria-death) and the predictor variable of interest (level of exposure), after adjusting for possible interaction effect of sex. The log rank test was used to test whether Kaplan-Meier curves for the two exposure groups in the entire population were statistically equivalent. The likelihood ratio test was used to ascertain the significance of the preventive measure variable (ITNs) that was used to lessen the exposure level to the mosquito parasites. The log-minus-log method was used to fit the biomedical data to assess whether the exposure data satisfies the proportional hazards assumption (Figure 1).

Table 2 gives a pictorial view of the survival analysis scheme from the origin state, that is, arrival on admission to the destination point, that is, end of study period.

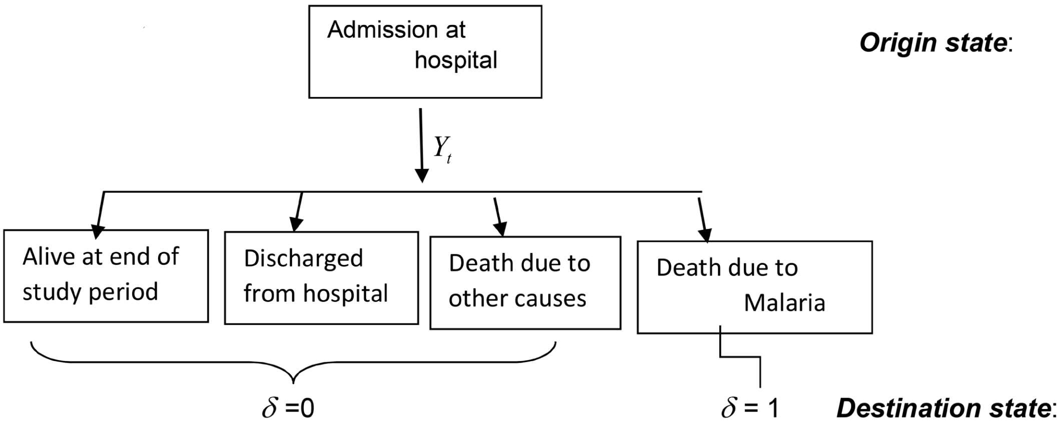


Figure 1. Survival model for biomedical data.

Predictor variables: age, sex, preventive measures (level of exposure)

Y_t = dependent variable; Y_t (age, sex, level of exposure)

Event of interest = death due to malaria

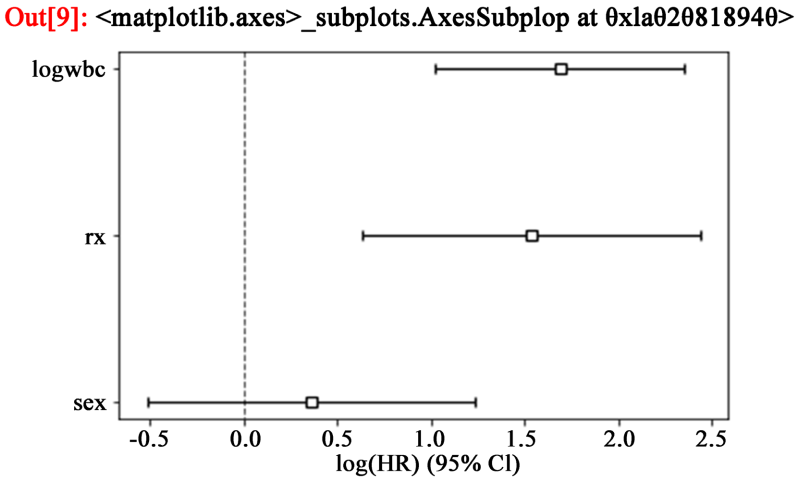
$$\delta = \begin{cases} 0, & \text{Censored}\left(\text{did not obtain the event ofinterest}\right) \\ 1, & \text{Failure}\left(\text{obtained the event ofinterest i.e death due to malaria}\right) \end{cases}$$

2.3. Simulation

Simulation studies present an important statistical tool to investigate the performance, properties and adequacy of statistical models in pre-specified situations. One of the most important statistical models in medical research is the Cox proportional hazards model. In this paper, techniques to generate survival times for simulation studies regarding Cox proportional hazards models are presented. We simulated the data set called "anderson.dat", which consisted of survival times on 42 leukemia patients [5]. Table 2 represents a truncation of "anderson.dat". the simulation studies were performed using the first four subjects.

Figure 2 gives the probability value (p-value) for the three covariates. The p-value for $\log wbc$ (0.00) is less than 0.005 and the hazard ratio (*HR*) is 5.40 indicating a strong relationship between $\log wbc$ value and increase risk of relapse. Holding the other covariates constant, a higher value of $\log wbc$ is associated with poor survival. Here, a person with higher $\log wbc$ has a higher risk of death.

The p-value for treatment status (*Rx*) = 0.002) which is less than 0.005 and *HR* is 4.64 indicating a strong relationship between *Rx* value and increase risk of relapse. Holding other covariates constant, a higher value of *Rx* is associated with poor survival. A person with higher *Rx* value has a higher risk of death. The p-value for sex (0.42) is greater than 0.005 and the *HR* for sex is 1.43. Figure 3 displays the curves of the survival probability for the first 4 persons in our dataset (Table 2). We notice that the first and second persons (person-0 and person-1) both have a high survival chance with their curve lying above the other carves (two curves walking the same path). The third person (Person-2) has the lowest



```
Out[9]: <matplotlib.axes>_subplots.AxesSubplop at 0xla020818940>
```

Figure 2. Plot to identify which of the three covariates (factors) affected the subjects the most.

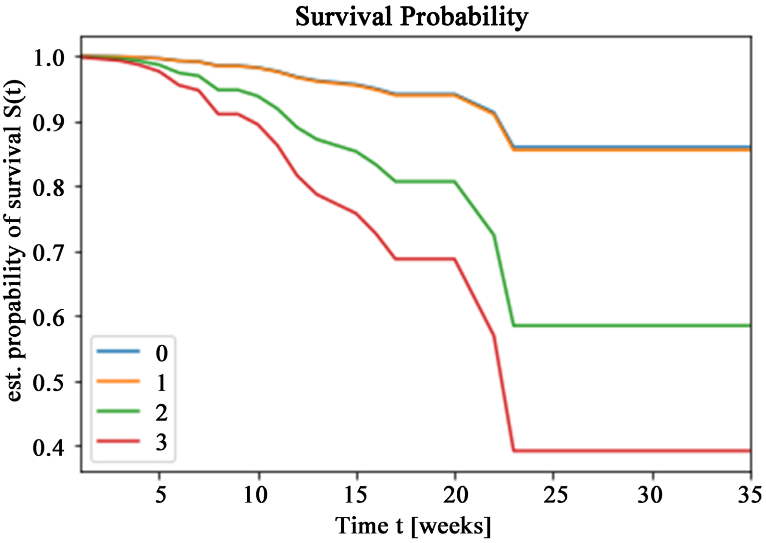


Figure 3. Plots of the survival probability for the first 4 persons in our dataset (**Table 2**). We notice that person-0 and person-1 both have a high survival chance (two curves walking the same path). Person-2 has the lowest survival chances. Person 3 has a high $\log wbc$ value (2.53).

Table 2. Remission survival times on 42 leukemia patients.

Subj	Surv	Relapse	Sex	logwbc	Rx
1	35	0	1	1.45	0
2	34	0	1	1.47	0
3	32	0	1	2.2	0
4	32	0	1	2.53	0
5	25	0	1	1.78	0
6	23	1	1	2.57	0
7	22	1	1	2.32	0
8	20	0	1	2.01	0
9	29	0	0	2.05	0

Table 3. Output after Fitting the Cox Regression Model (CoxPHFitter) by considering all parameters "Surv", "Relapse", "Sex", "Logwbc", "Treatment Status (Rx)".

Out put	Further Details
model	lifelines.CoxPHFitter
duration col	"Surv"
event col	"relapse"
baseline estimation	breslow
number of observations	42
number of events observed	30
partial log-likelihood	−69.81
time fit was run	2021-09-23 04:31:45 UTC

Table 4. Results of the Cox simulated values based on the output variables and the data set.

	β	exp (β)	s.e (β)	Coef. lower 95%	Coef. upper 95%	Exp (coef.) lower 95%	Exp (coef) upper 95%	z	p	log2p
Sex	0.36	1.43	0.45	−0.52	1.23	0.60	3.43	0.80	0.42	1.25
Logwbc	1.69	5.40	0.34	1.03	2.35	2.79	10.48	4.99	<0.005	20.69
Rx	1.54	4.64	0.46	0.63	2.44	1.88	11.45	3.34	<0.005	10.20

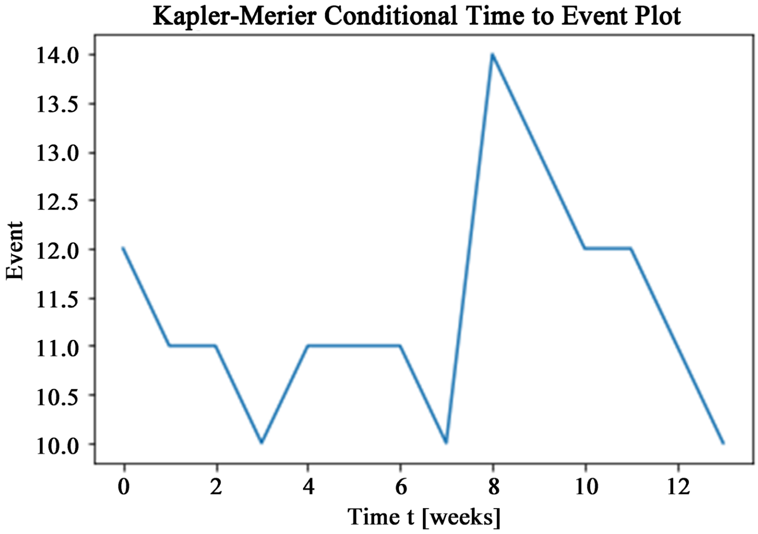


Figure 4. Plotting of Median Conditional Time to Event-using Kaplan Meier. As time passed, the median survival time decreases, increased again, remained stable for a while and decreased sharply, increased again and finally decreased.

survival chances. The fourth person (person 3) has a high $\log wbc$ value (2.53) (Table 3). Figure 4 resents the graph of the median conditional time to event-using Kaplan Meier. As time passed, the median survival time fluctuates (decreases, increases, remained stable for a while, decreases sharply, increases again and finally decreases). Table 4 shows the results of the Cox model simulated values based on the output variables (Table 3) and the data set (Table 2), we note that the $\log wbc$ and Rx variables were significant but that of age was not significant (Tables 5-8).

3. Results

3.1. Hypothesis Testing for Various Conjectures

3.1.1. Hypothesis One

H_o: The survival curves of the exposed and the less exposed groups are equivalent.

Test statistic: Log-Rank test (Mantel-Cox)

Decision criteria: Reject the H_o if the p-value is less than $\alpha = 0.05$

Table 5. Shows the output after fitting data using Kaplan-Meier-Fitter with duration variable, or time “Surv” and event observed as “relapse”. Fitted with 42 total observations with12 right-censored observations.

Event_at	Removed	Observed	Censored	Entrance	At_risk
0.0	0	0	0	42	42
1.0	2	2	0	0	42
2.0	2	2	0	0	40
3.0	1	1	0	0	38
4.0	2	2	0	0	37
5.0	2	2	0	0	35
6.0	4	3	1	0	33
7.0	1	1	0	0	29
8.0	4	4	0	0	28
9.0	1	0	1	0	24
10.0	2	1	1	0	23
11.0	3	2	1	0	21
12.0	2	2	0	0	18
13.0	1	1	0	0	16
15.0	1	1	0	0	15
16.0	1	1	0	0	14
17.0	2	1	1	0	13
20.0	1	0	1	0	11
22.0	2	2	0	0	10
23.0	2	2	0	0	8
25.0	1	0	1	0	6
29.0	1	0	1	0	5
32.0	2	0	2	0	4
34.0	1	0	1	0	2
35.0	1	0	1	0	1

Footnotes: #at_risk—it stores the number of current patients; at-risk = current patient at risk + entrance removed; #event_at—It stores the value of the timeline for the dataset (*i.e.*, time the patient was observed in the experiment or time the experiment was conducted); # Removed = observed + censored; # Censored = Persons that did not relapse; and #Observed = Persons that relapsed (died).

Table 6. Summary of data on malaria cases for users and non-users of ITN.

Event	ITN Users	Other Nets	Total
Died	66	325	391
Not Die	339	1063	1402
At Risk	405	1388	1793

Table 7. Variables in the Cox proportional hazards model.

Predictor variables	β	SE	Wald	df	Sig.	Exp (β)
Sex	0.091	0.102	0.791	1	0.374	0.913
Preventive Measure	0.384	0.135	8.041	1	0.005	1.468
Age	0.013	0.002	48.229	1	0.000	1.013

Table 8. Variables in the stratified Cox model (Stratified by sex).

Predictor variables	β	SE	Wald	df	Sig.	Exp (β)
Preventive Measure	0.373	0.135	07.607	1	0.006	1.452
Age	0.013	0.002	49.112	1	0.000	1.014

Table 9. Summary of Test results for testing the equality of Survival curves for exposed and less exposed groups.

Test	Chi-Square	dof	Significant
Log Rank (Mantel-Cox)	9.323	1	0.002
Breslow (Generalized Wilcoxin)	10.310	1	0.001
Tarone-Ware	9.720	1	0.002

From the test results presented as Table 9, the p-value of the two exposure groups was less than 0.05, that is (p-value = 0.002 < 0.05). We therefore reject H_o (Hypothesis 3.1.1) and conclude that the survival curves of the exposed group and the less exposed group were significantly different.

3.1.2. Hypothesis Two

H_O: The method of protection adopted to prevent exposure to parasite was not significant.

Test statistic: Likelihood ratio test $LR = -2\log L_R - \left(-2\log L_F\right)$.

Decision criteria; Reject the H_O if the p-value is less than $\alpha = 0.05$.

The −2 Log Likelihood $LR = -2\log L_R - \left(-2\log L_F\right)$ value for the full model (that is, the one containing both age and preventive measure variables [Table 10 and Table 11]) was 5217.642, while the reduced model (that is, the one containing only the age variable) was 5226.168. $LR = 5226.168 - 5217.64 = 8.526$. The LR statistics is a chi-square statistic χ^2 with one degree of freedom (because we are assessing only one predictor variable-usage of ITN) under the null hypothesis that the predictor variable was not significant. From a web based statistical calculator, the chi-square value of 8.526 translates to a p-value of 0.0035 < 0.05. Since the p-value was less than 0.05 we reject the H_O, (Hypothesis 3.1.2) and conclude that the method of prevention was significant.

3.1.3. Hypothesis Three

H_O: The survival experiences of the exposed and the less exposed groups is not significant after stratifying by sex.

Table 10. Variables in the full and reduced Cox models.

Predictor variables	β	SE	Wald	df	Sig.	Exp(β)
Full Cox model						
Preventive Measure	0.379	0.135	7.860	1	0.005	1.461
Age	0.013	0.002	47.424	1	0.000	1.013
Reduced Cox Model						
Age	0.013	0.002	48.635	1	0.000	1.013

Table 11. Log likelihood statistic for full and reduced model.

	−2 Log likelihood	Chi-square	df	Sig
Full model	5217.642	58.262	2	0.000
Reduced model	5226.168	50.304	1	0.000

Test statistic: Wald's Statistics.

Decision criteria; Reject the H_O, if the p-value is less than $\alpha = 0.05$.

We note from Table 5 that the p-value of the preventive measure variable (p-value = 0.006 < 0.05) was significant. We therefore reject the null hypothesis (Hypothesis 3.1.3) and conclude that the preventive measure variable is significant after stratifying by sex.

3.1.4. Hypothesis Four

H_o: The correlation between the ranked failure time and the Schoenfeld residuals is zero (not significant).

Test statistic: Schoenfeld residual test (Mantel-Cox).

Decision criteria: Reject the H_o, if the p-value is less than $\alpha = 0.05$ or 0.01.

From Table 9 (the results provided by the computer) the correlation between ranked failure time and Schoenfeld residual was significant at both the 0.05 and 0.01 levels of significance, thus, we have every evidence to reject hypothesis 3.1.4. If we consider the last row (rank of duration significant two-tailed), we note that the null hypothesis was rejected for the sex variable (p = 0.04) but not rejected for preventive measure (p = 0.21) and age variable (p = 0.85).

4. Discussions

We see from (Table 6) that out of the 1793 patients sampled 405 representing 22.6% were using insecticide treated nets while the majority (77.4%) was using other types of nets like window netting and ordinary treated nets. It was also established that of those who were using ITN 16% died within the four months study period while 84% survived, again out of the non-users of ITN 23% died while 77% survived within the same study period. The ratio of death of male to female was 1.2:1. This ratio indicates that death due to malaria for the period of observation was not gender related.

The plot of (Figure 5) gives a graphical picture of the survival curves of the two groups of users of ITN. We notice from the graph that the survival experiences at the first few days of the study appeared to be the same but thereafter the differences showed up clearly, we see also that the curve for the users of ITN consistently lies above that of the non-users, this characteristic shows that users

Table 12. Computer results of correlation between residual covariates and ranked survival time.

		PR for Sex	PR for Prev.meas	PR for Age	Rank of duration (days)
PR for Sex	P. correlation Sig. (2-tailed)	1.000	0.108* 0.033	0.144** 0.004	0.101* 0.044
PR for Prev.meas	P. correlation Sig. (2-tailed)	0.108* 0.033	1.000	0.141** 0.005	−0.062 0.219
PR for Age	P. correlation Sig. (2-tailed)	0.144** 0.004	0.141** 0.005	1.000	−0.010 0.847
Rank of duration (days)	P. correlation Sig. (2-tailed)	0.101* 0.044	−0.062 0.219	−0.010 0.847	1.000

Note: *PR* = Partial Residual, N = 391; P correlation = Pearson correlation. Prev.meas means preventive measure; Sig. means significant value; *Correlation is significant at the 0.05 level (2-tailed). **Correlation is significant at the 0.01 level (2-tailed).

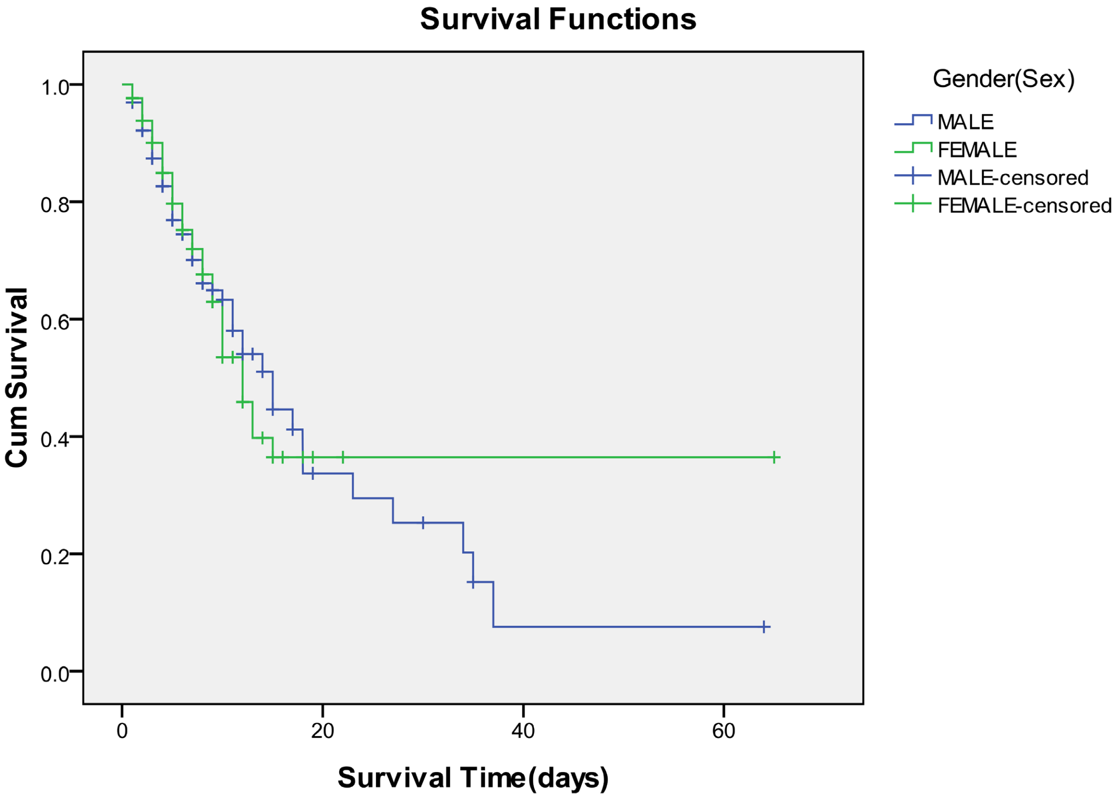


Figure 5. Survival functions of users of ITN and non-users. The curve for users of ITN lies above that of the non-users. We see also that for those who did not use ITN there were many steps within the curve with each step representing death.

of ITN have a better survival prognosis than non-users. The difference further means that ITN was effective at all points during the observational period. From the graph we could also estimate the median survival times for the two classes of users. This is done by locating 0.5 on the y-axis and proceeding horizontally till it meets the curves, once the horizontal line meets the curve, we draw a vertical line from the point of intersection of the curve and the horizontal line to meet the x-axis. From the graph the median survival time of the non-users of ITN was approximately 10 days while that of the users of ITN was close to forty (40) days. The median value further confirms our claim that users of ITN have better survival prognosis than non-users.

The failure potential of the users of ITN and non-users is presented as Figure 6. It is worth mentioning that while the survival function gives the probability of surviving, the hazard function or rate gives the risk of failing. The higher the hazard rates the worst the impact on survival. The curves in the figure depicts that non-users of ITN are at a higher risk of malaria deaths than users

From Figure 7, we could infer that the survival experiences of males and females were approximately the same, this implies that sex do not contribute significantly to death due to malaria. The difference between two plots is given by $\left[\sum_{i=1}^{p}\beta_i\left(x_{2j}-x_{1j}\right)\right]$, and what the expression is saying is that, if we use a Cox PH model and plot the estimated log-log survival curves for two groups on the same

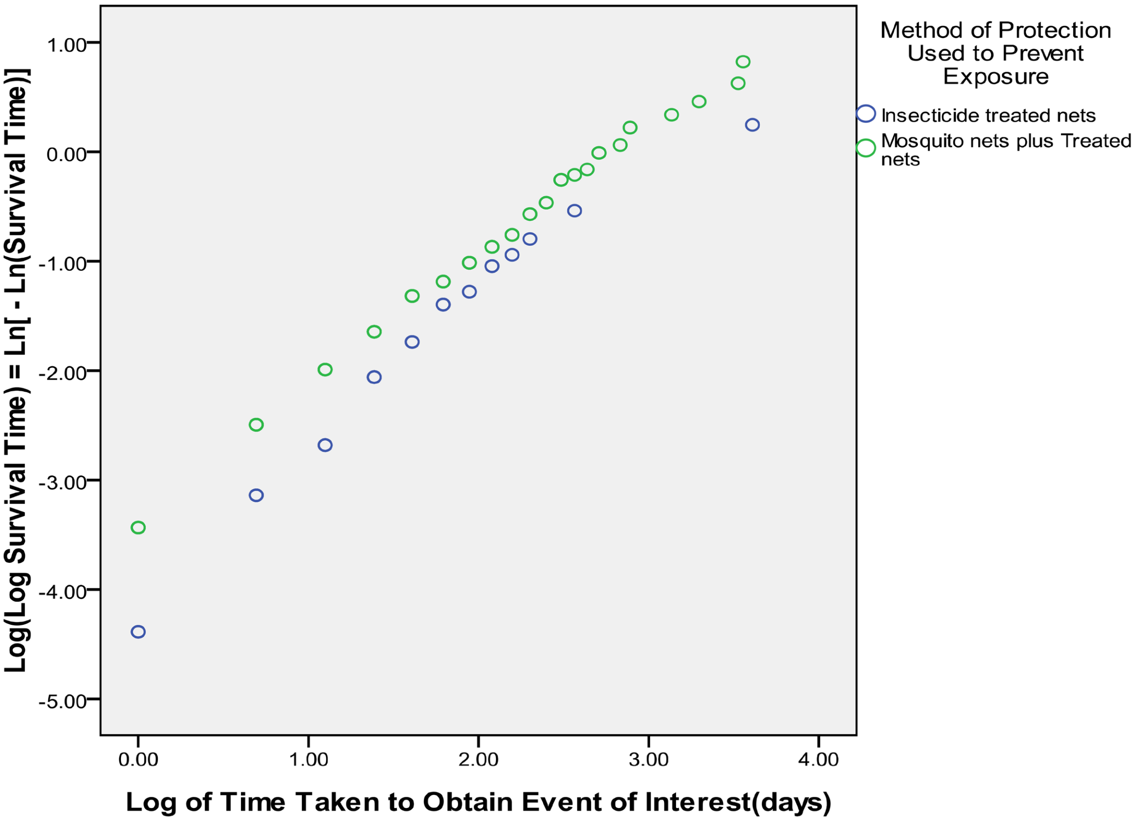


Figure 6. Hazard curves used for comparing the hazards for the users and Non-users of Insecticide treated nets.

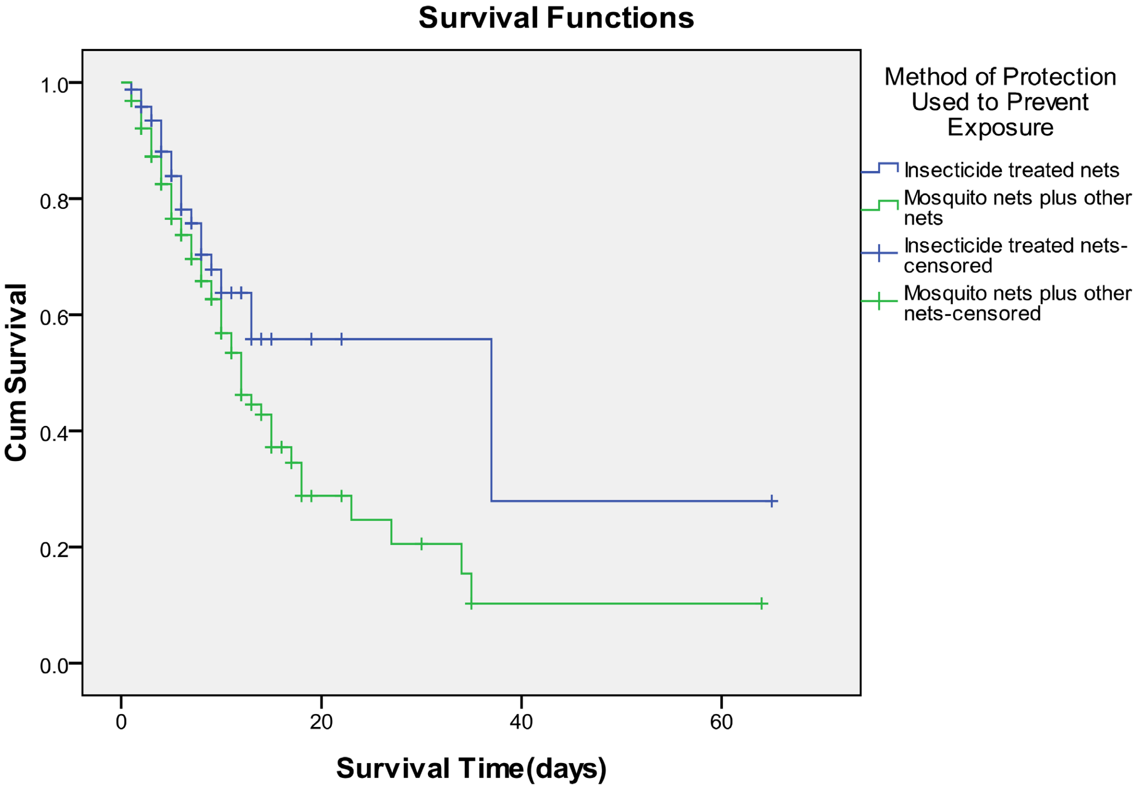


Figure 7. Survival curves used for comparing the survival experiences of males and females and their risk of malaria death. The curves for the male and the female cross each other at various points.

graph the curves will be approximately parallel, and the distance between them is the linear expression involving the difference in the predictor values (method of protection used for preventing mosquito bites) which does not involve *t*. The summary of the four possible results from the examination of the log negative log Kaplan-Meier survival estimates plotted against the log of time as shown in **Figure 8** for the two levels of protection against exposure to the mosquito parasite are given below.

- Parallel and straight lines imply that the Weibul model, (WM), Accelerated Failure time (AFT) and the Proportional hazard (PH) assumption hold.
- Parallel but not straight lines imply that the PH assumption holds but neither the WM nor AFT model holds.
- Non-parallel and non-straight lines suggest that PH, AFT and WM do not hold
- Non-parallel but straight lines imply that the WM holds but neither the PH nor the AFT hold.

Examining **Figure 8** critically, we notice that the plots for both the less exposed and the exposed as indicated by the use of ITN or otherwise are reasonably straight suggesting that the Weibul assumption reasonably holds. We notice again that the two curves are approximately parallel (their gradients ρ are approximately the same) implying that the PH and the AFT assumptions hold. This parallelism of the log-log Kaplan Meier survival curves for the Cox PH

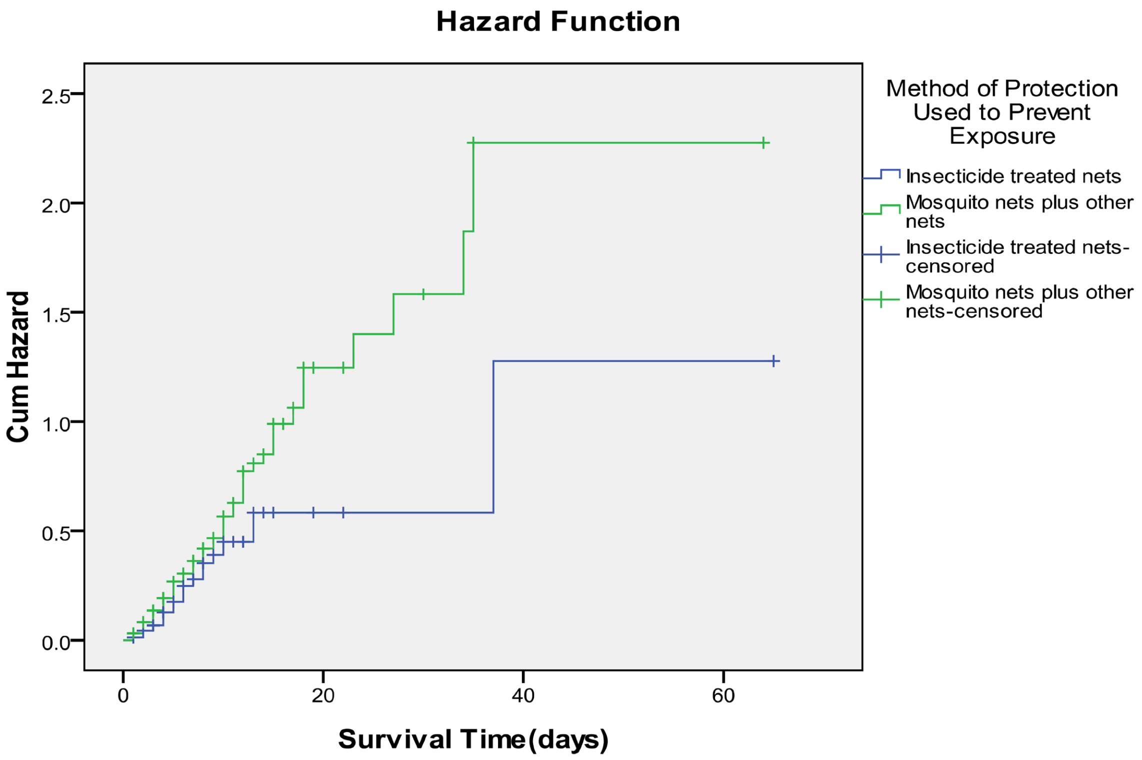


Figure 8. Log-Log Survival curves for assessing Weibull, Accelerated Failure Time (AFT) and Proportional Hazards (PH) assumptions.

provides us with a graphical approach for assessing the PH assumption. We could infer from the parallelism of the plots that once the plots are parallel, under no circumstance will the survival experiences of the two groups of users be the same.

We used the stratified Cox model (Table 8) to control for the sex variable which does not satisfy the PH assumption. The implication here is that the sex variable is being adjusted for stratification, we have also included the age and preventive measure variable (which do satisfy the PH assumption) into the model, in other words the age and preventive measure variables have been adjusted by their inclusion into the model. In the model we can infer that the hazard ratio for the effect of the preventive measure variable adjusted for age and sex is given by the value 1.452., this value can be interpreted to mean that the exposed group (that is the group that do not use the insecticide treated net as a means of preventing exposure to the mosquito parasite) has 1.5 times the hazard of death through malaria as the less exposed group (group that use ITN as a means of preventing exposure to the malaria parasite)

The variables in the stratified Cox model (Table 8) provides us with useful information to test whether there is any difference in the population survival curves for the two classes of users of ITN, after adjusting for sex (since sex did not contribute significantly to the risk of malaria death). The null hypothesis for this test was that there was no difference in the survival curves of the users of ITN and non-users. The p-value of the log-rank test ($0.002 < 0.05$) was highly significant, implying that there was a statistically significant difference between the population survival curves after adjusting for sex. This result states inter alia that if the whole population elements were included in this study the survival experiences of the users of ITN and non-users would have been different, this further means that the predictor variable under consideration does contribute significantly to the death due to malaria.

The Cox proportional hazards (PH) model is presented in Table 7, in this model the PH assumption was assumed to hold for all three covariates. The model used all 1793 patients observed in the study. The output variable was time in days until a patient die. The method of estimation used to obtain the coefficients was the maximum likelihood estimation (MLE). A p-value of 0.374 (from column 6) was obtained for the sex variable. This value indicates that the sex variable was not significant, that is to say, the sex of the patient plays no significant role in deaths due to malaria, however the p-values (0.000) of age of the patient and the p-value (0.005) of user status of ITN (Type of preventive measure used) were both highly significant telling us that the risk of malaria death was dependent on one's age and the method of prevention adopted (in this case users and non -users of ITN). From column 2, the magnitude of the coefficient (0.384) of ITN user status depicts that user status contributes largely to the variation in the dependent variable (that is death due to malaria), while the contributions from the remaining covariates age and sex were insignificant. The hazard ratio denoted by $\exp(\beta)$ in the Table 4 indicates that the ratio of the users of ITN and non-users was 1.468, which translates into saying that the non-users if ITN were 1.5 times at risk of malaria death than users of ITN. For age and sex variables the hazard ratios do not give any useful information. It should be recalled that a hazard ratio of one means that there was no effect.

The Cox adjusted log-log plots (Figure 7) were fitted using the mean values of age and sex and were used to evaluate the PH assumption for the preventive measure. From this figure, we noticed that the two graphs were approximately parallel which translates into saying that the survival experiences of the users of ITN and non-users can in no way be the same. Table 12 gives us the results for the Schoenfeld statistical test, in this test, the null hypothesis H_O is that the PH assumption was not violated. The p-values for testing whether the correlation was zero between the ranked survival time and the covariates (Schoenfeld residuals) are the p-values for the statistical test. From the computer output (Table 12), the following results were obtained:

Case A: p-value for sex-Schoenfeld residual and the ranked survival time = 0.044 < 0.05.

Case B: p-value for preventive measure-Schoenfeld residual and the ranked survival time = 0.219 > 0.05.

Case C: p-value for age-Schoenfeld residual and the ranked survival time = 0.847 > 0.05.

In case A, the null hypothesis was rejected, thus we conclude that for the sex variable the PH assumption was violated, which also means that in determining death due to malaria the sex variable was not a risk factor. In the cases B and C, we do not have enough evidence to reject the null hypothesis, so we conclude that the PH assumption was not violated for the age and preventive measure variables, by implication we can assert that in determining the risk of malaria deaths these variables (age and preventive measure) might play significant roles.

From the computer outputs labeled as Table 5 and Table 6 we could assess the significance of the preventive measure variable using the likelihood ratio test which was given as $LR = -2\log L_R - \left(-2\log L_F\right)$

The −2 Log Likelihood value for the full model (one containing both age and preventive measure variable) was 5217.642, while the value for the reduced model (one containing only the age variable) was 5226.168

$$LR = 5226.168 - 5217.642 = 8.526$$

The LR statistics is a chi square statistic χ^2 with one degree of freedom (because we are assessing only one predictor-usage of ITN) under the null hypothesis that the predictor is not significant.

From a web based statistical calculator, the chi square value of 8.526 translates to a p-value of 0.0035 < 0.05, thus we have enough evidence to reject the null hypothesis, and conclude that the predictor under investigation (type of preventive measure used by patients) was significant and therefore contributes significantly to the risk of malaria death.

5. Conclusions

At the onset, we sought to provide theoretical framework underpinning the Cox proportional hazards model; outline theories on which the Cox model could be laid out; do some simulation study on the Cox model and provide a real case empirical studies with apt interpretation of the outcome. In clinical investigations and medical researches, there may be many situations, where several known quantities potentially affect patient prognosis. One or two of these competing risk factors might predict one's predicament more than others, in seeking to find out which of the risk factors contribute or have the highest impact on the survival time of a patient, there is the need for researchers to adjust the covariates to realize the impact of each of them on the survival times of the patients. Aside the multivariate nature of the covariates, some covariates might be categorical while others might be quantitative. Again, there might be cases where we need a model that has the capability of extending survival analysis methods to assessing simultaneously the effect of several risk factors on survival time. A method of analysis that can accommodate all the enumerated situations is none other than the Cox proportional Hazards model. The discovery of a diagnostic key assessment indicator to diseases such as malaria has been on the ascendancy. Most of these methods focus on classification problems, that is, adopting a model that discriminates patients into distinct clinical groups. Few papers have been published on approaches that predict a patient's event risk or hazard of death due to a predictive risk factor. This study has effectively integrated data into multivariable Cox proportional hazard models for risk prediction in malaria. Subsequently, it is insightful, besides the main objective of the study to say that:

- All the three models, Weibull, accelerated failure time, and the Proportional hazards assumptions were satisfied;
- The method of protection adopted and age satisfied the proportional hazards assumption but sex did not;
- The hazard ratios of the exposed group were 1.5 times the hazards of the less exposed group; and
- Sex of residents did not contribute to the risk of malaria death, but the method of protection and age contributed towards the risk of malaria death.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Blossfeld, H.P. and Rohwer, G. (1995) Techniques of Event History Modeling. Lawrence Erlbaum Associates, Mahwah, NJ, 667-694.

[2] Tableman, M. and Kim, J.S. (2004) Survival Analysis Using S. Chapman and Hall, New York. https://doi.org/10.1201/b16988

[3] Lee, E.T. (1980) Statistical Methods for Survival Data. Wadsworth, Belmont, CA.

[4] Turkson, A.J. (2010) Understanding the Basic Concepts of Survival Analysis. *Journal of Biostatistics*, **4**, 63-80.

[5] Klein J.P. and Moeschberger, M.L. (1997) Survival Analysis Techniques for Censored and Truncated Data. Springer-Verlag, New York.

[6] Allison, P.D. (1995) Survival Analysis Using the SAS System: A Practical Guide. SAS Institute, Inc., Cary, NC.

[7] Diggle, P., Heagerty, P., Liang, K.Y. and Zeger, S. (2002) Analysis of Longitudinal Data. 2nd Edition, Oxford University Press, New York.

[8] Bugnard, F., Ducrot, C. and Calavas, D. (1994) Advantages and Inconveniences of the Cox Model Compared with The Logistic Model: Application to a Study of Risk Factors of Nursing Cow Infertility. *Veterinary Research*, **25**, 134-139.

[9] Basu, A., Manning, W.G. and Mullahy, J. (2003) Comparing Alternative Models: Log vs Cox Proportional Hazard? *Health Economics*, **13**, 749-765. https://doi.org/10.1002/hec.852

[10] Manning, W.G. and Mullahy, J. (2001) Estimating Log Models: To Transform or Not to Transform? *Journal of Health Economics*, **20**, 461-494. https://doi.org/10.1016/S0167-6296(01)00086-8

[11] Tang, Z., Zhou, C., Jiang, W., Jing, X., Yu, J., Alkali, B. and Sheng, B. (2014) Analysis of Significant Factors on Cable Failure Using the Cox Proportional Hazard Model. *IEEE Transactions on Power Delivery*, **29**, 951-957. https://doi.org/10.1109/TPWRD.2013.2287025

[12] Borucka, J. (2014) Methods of Handling Tied Events in the Cox Proportional Hazard Model. *Studia Oeconomica Posnaniensia*, **2**, 91-106.

[13] Zare, A., Hosseini, M., Mahmoodi, M., Mohammad, K., Zeraati, H. and Naieni, K.H. (2015) A Comparison between Accelerated Failure-Time and Cox Proportional Hazard Models in Analyzing the Survival of Gastric Cancer Patients. *Iranian Journal of Public Health*, **44**, 1095-1102.

[14] Wang, M., He, Y., Shi, L. and Shi, C. (2011) Multivariate Analysis by Cox Proportional Hazard Model on Prognosis of Patient with Epithelial Ovarian Cancer. *European Journal of Gynaecological Oncology*, **32**, 171-177.

[15] Nilima, S., Sultana, R., Ireen. S. (2018) Neonatal, Infant and Under-Five Mortality: An Application of Cox Proportional Hazard Model to Bdhs Data. *Journal of the Asiatic Society of Bangladesh*, *Science*, **44**, 7-14.

[16] Fisher, L.D. and Lin, D.Y. (1999) Time Dependent Covariates in the Cox Proportional Hazards Regression Model. *Annual Review of Public Health*, **20**, 145-157. https://doi.org/10.1146/annurev.publhealth.20.1.145

[17] Chowdhury, M.H., Hassan, M.Z. and Sabiha, M. (2018) Factors Influencing Women's Waiting Time to First Birth in Bangladesh: An Application of Cox Proportional Hazard Model. *Journal of Science and Technology*, **16**, 100-116.

[18] Cox, D.R. (1972) Regression Models and Life-Tables. *Journal of the Royal Statistical Society*: *Series B*, **34**, 187-202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

[19] Kleinbaum, D.G. and Klein, M. (2005) Survival Analysis: A Self-Learning Test. 2nd Edition, Springer, New York. https://doi.org/10.1007/0-387-29150-4

[20] Harrel, F.E., Lee, K.L. and Mark, D.B. (1996) Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumption and Adequacy and Measuring and Reducing Errors. *Statistics in Medicine*, **15**, 361-387. https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4%3C361::AID-SIM168%3E3.0.CO;2-4

[21] Selvin, S. (2008) Survival Analysis for Epidemiology and Medical Research: A Practical Guide. Cambridge University Press, New York.

[22] Hosmer, D.W. and Lemeshow, S. (1999) Applied Survival Analysis, Regression Modeling of Time to Event Data. John Wiley and Sons, New York.

[23] Sharp, B., Van Wyk, P., Sikasote, J.B., Banda, P. and Kleinschmidt, I. (2002) Malaria Control by Residual Insecticide Spraying in Chingola and Chililabombwe, Copperbelt Province, Zambia. *Tropical Medicine & International Health*, **7**, 732-736. https://doi.org/10.1046/j.1365-3156.2002.00928.x

[24] Skovmand, O., Dang, D.M., Tran, T.Q., Bossellman, R. and Moore, S.J. (2021) From the Factory to the Field: Considerations of Product Characteristics for Insecticide-Treated Net (ITN) Bio-Efficacy Testing. *Malaria Journal*, **20**, Article No. 363. https://doi.org/10.1186/s12936-021-03897-7