J Hum Ecol, 60(2,3): 106-116 (2017)
DOI: 10.31901/24566608.2017/60.2-3.06

Statistical Assessment of Malaria Risk Factors Using Cox Proportional Hazard Approach

Ruffin Mutambayi*, James Ndege, Adeboye Azeez, Yong Song Qin and Akinwumi Odeyemi

Department of Statistics, University of Fort Hare, Private Bag X1314, Alice 5700, South Africa

KEYWORDS Cox Function. Kaplan-Meier. Log-Rank Test. Malaria. Modelling. Semi-Parametric. Survival Function

ABSTRACT The Cox proportional hazards (PH) model with time-dependent covariates has been widely used in medical and health related studies to examine the impacts of time-varying risk factors on survival. The study aimed to model the relapsing time using biographical, sanitation, environmental and preventive information as covariate risk factors. The Kaplan-Meier method, log-rank test and the Cox proportional hazards (PH) model to examining the covariates. The results indicates that the model $h(t) = h_0(t)\exp\left(1.91613X_{dump} - 0.49633X_{spr} + 0.81466X_{inft}\right)$ was found to fit the data better, as confirmed by the result of the global test that present reasonable and significant results (Likelihood Ratio: 18.2264, $p<0.0004$; Score: 17.6569, $p<0.0005$ and Wald: 19.3975, $p<0.0002$). In conclusion it was found that, 'dumping site' ($p<0.0106$; 95% C.I: 1.545-29.451), 'spray used' ($p<0.$ 0.0220; 95% C.I: 0.391- 0.915), and 'information related to source of malaria' ($p<0.$ 0.0012; 95% C.I: 1.380-3.725), have a significant effect on the relapsing time of patients under investigation.

INTRODUCTION

The typical viewpoint to model covariates impacts on survival and to manipulate censored data is the Cox proportional hazard model (Li and Owzar 2015). Mainly regression analysis is used for identifying the risk covariates; hence due to the presence of censoring in survival observations, standard regression models cannot be applied. Also simple logistic regression analysis has the restriction of strictly permitting an outlook of survival likelihood over the entire investigation period as a unique time interval and it presumes that each patient is at risk over the whole study period. But for longitudinal studies or other cases where patients have different and not stable time at risk, the concept is not valid. For this purpose, in survival analysis, Cox's regression model is widely applicable.

The unique attribute of Cox PH model is its power to approximate the association between the hazard function and covariates without formulating any assumption about the distribution related to hazard function of the baseline (Abdelaal and Zakria 2015). Hence sometimes the Cox model is introduced as a semi-parametric model.

As it requires few assumptions, the Cox PH model is the most extensively used method to analyze censored observations in clinical trials (Collet 2003). The model was suggested by Cox in 1972 to investigate survival observations with and without censoring for distinguishing dissimilarities of survivorship in treatment therapy and prediction components in clinical trials. In Cox model, the major assumption (known as PH assumption) is that the hazard ratios between two individuals that are non-time-related. However, Hosmer and Lemeshow (1990) observed that this assumption is not always appropriate as sometimes the assumptions maybe violated. Many at times, researchers do not test the assumptions before fitting the model.

Furthermore, the Cox model is considered to be robust to violation of model assumptions and the evaluation and description of approximated hazard ratios are always positive (Hosmer and Lemeshow 1990; Ponnuraja and Venkatesan 2010). The Cox proportional hazards associates predictors to the hazard function through regression model as indicated below:

$$h\left(t/x\right) = h_0(t)c\left(\beta' x\right) \quad (1)$$

*Address for correspondence:
Ruffin Mutambayi
Department of Statistics
University of Fort Hare, Private Bag X1314
Alice, Eastern Cape, South Africa, 5700
Telephone: +27 (0) 40 602 2460;
Cell: +27(0) 79 242 2277
E-mail: mmutambayi@ufh.ac.za,
you2me4you@gmail.com

where is known as baseline hazard function obtained when all the individual covariates incorporated in the model are fixed and equal to zero; $\beta' = (\beta_1, \beta_2 \ldots \beta_p)$ is a parameter vector of the fitted regression coefficients; $x = (x_1, x_2 \ldots x_p)'$ is the value of the vector of predictor variables for a specific individual, and $c(.)$ is a fixed, known scale functions (Bieszk-Stolorz 2015).

This is a semi-parametric model where the baseline hazard $h_0(t)$ is approximated non-parametrically, while the covariates impact is restricted by the parametric representation c(B/x). Where c(.), take an exponential form:

$$c(\beta' x) = e^{(\beta' x)} = e^{\left(\sum_{j=1}^{p} \beta_j x_{ji}\right)} \quad (2)$$

Which persuades that the hazard is non-negative and presumes that predictor impacts on the hazard are multiplicative. So

$$h(t/x) = h_0(t)c(\beta' x) = h_0(t)e^{(\beta' x)} = h_0(t)e^{\left(\sum_{j=1}^{p} \beta_j x_{ji}\right)} \quad (3)$$

Proportional Hazards

The Cox Proportional hazards is a technique for examining the effect of various covariates over the time for a clarified event to occur. In the case of a result such as death this is known as Cox regression for survival analysis. The Cox model is named 'proportional hazards model' since the ratio of hazard rates of two individuals with predictor values χ_1 and χ_2, at time t is:

$$\frac{h(t/x_1)}{h(t/x_2)} = \frac{h_0(t)e^{(\beta' x_1)}}{h_0(t)e^{(\beta' x_2)}} = \frac{e^{(\beta' x_1)}}{e^{(\beta' x_2)}} = e^{[\beta'(x_1 - x_2)]} \quad (4)$$

In Cox proportional model, the hazard ratio is time-independent as the ratio does not depend on the time t_4. Hence the hazard function at time given by: $h(t/\chi) = h_o(t)e(\beta'_x)$. Therefore, the hazard, cumulative hazard, survival, and the probability density function can be expressed as follows:

$$H(t/x) = \int_0^t h(u/x)\,du = \int_0^t h_0(u)e^{(\beta' x)}du = H_0(t)e^{(\beta' x)} \quad (5)$$

$$S(t/x) = e^{-[H(t/x)]} = e^{-\left[H_0(t)e^{(\beta' x)}\right]} \quad (6)$$

$$f(t/x) = h_0(t)e^{(\beta' x)}e^{-\left[H_0(t)e^{(\beta' x)}\right]} \quad (7)$$

The substantial benefit of Cox Proportional Hazards regression is that you can still fit censored regression models without regard to the distribution of the baseline hazard function.

METHODOLOGY

Parametric PH Models

The parametric proportional hazard model is a special case version of the Cox proportional hazards regression model were some special assumptions about $h_0(t)$, such as the exponential and Weibull distributions are presumed. But the advantage of Cox regression model is that such assumptions can be ignored. The main dissimilarity of the two types of regression models is that the $h_0(t)$ is presumed to have a specific distribution when a complete parametric proportional hazard regression model is fitted to the observations, while the Cox regression model has not such assumption. A number of different parametric PH regression models can be extracted by selecting dissimilar hazard functions.

Weibull PH Model

The Weibull model will be used if plotting the hazards function against time gives a hazard rates curve that is non-constant but monotonic that either increases or decreases exponentially. $h_0(t)$ For this model the baseline hazards function $h_0(t)$ will be $\lambda\gamma\tau^{(\gamma-1)}$, hence the PH failure rate of a particular subject with covariates $(x_1, x_2, \ldots x_p)$ when assumed a *Weibull PH model* is determined by:

$$h(t/x) = \lambda\gamma t^{(\gamma-1)}e^{(\beta_1 x_1 + \beta_2 x_2 \ldots + \beta_p x_p)} = \lambda\gamma t^{(\gamma-1)}e^{(\beta' x)} \quad (8)$$

Where the parameters λ and γ are defined as scale parameter and shape parameter, respectively.

Exponential PH Model

Under this model, the hazard function is presumed to be constant over time when plotted against time to event. Hence, the hazard function $h_0(t) = \lambda$ of a given subject is determined as:

$$h(t/x) = \lambda e^{(\beta_1 x_1 + \beta_2 x_2 \ldots + \beta_p x_p)} = \lambda e^{(\beta' x)} \quad (9)$$

Likelihood Estimation for the Cox PH Model

Derivation of an estimator of β cannot be based on an ordinary likelihood function since $h_0(t)$ is not specified parametrically in the Cox regression model. Instead, partial likelihood function has been proposed by Cox (Kalbfleisch and Prentice 2002). The Statistical Analysis of Fail-

ure Time for the estimation of regression parameters which is a function depending on β only.

Cox Partial Likelihood

Consider $t_1,t_2,...t_n$ as survival time related to n individual observations, and the ranked censored time of r individuals at risk at t_j is denoted by $R(t_j,)$ that is $R(t_j)$, that is $R(t_j)$ the number of individuals that are alive and uncensored at a time just prior to (t_j). The conditional probability that the individual experiments the event at (t_j) given that one individual from the risk set on $R(t_j)$ dies at (t_j) is defined as:

$$= \frac{h_i(t_i)}{\sum_{k\in R(t_j)} h_k(t_j)} = \frac{h_0(t_j)e^{(\beta' x_j)}}{\sum_{k\in R(t_j)} h_0(t_j)\,e^{(\beta' x_k)}} = \frac{e^{(\beta' x_j)}}{\sum_{k\in R(t_j)} e^{(\beta' x_k)}} \quad (10)$$

The product of the probabilities over r death times expressed as:

$$L(\beta) = \prod_{j=1}^{r} \frac{e^{(\beta' x_j)}}{\sum_{k\in R(t_j)} e^{(\beta' x_k)}} \quad (11)$$

The partial likelihood function for the model is:

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{e^{(\beta' x_i)}}{\sum_{k\in R(it)} e^{(\beta' x_k)}} \right]^{\delta_i} \quad (12)$$

where $R(t_j)$ can be determined as a risk set at time t_j and $(\delta_i=0)$ is the phenomenon index if (i^{th}) survival time is right censored and unity otherwise as defined by Cox. The estimates of β from Cox procedure using the partial likelihood are steady and coherent nonetheless of the shape of $h_o(t)$. The partial likelihood is well founded when there is absence of ties in the observations.

The Score Function and Information Matrix

The regression coefficient β are estimated with that maximize the partial likelihood. Presuming no ties, the log partial likelihood is:

$$l(\beta) = \log L(\beta) = \log \prod_{i=1}^{n} \left[\frac{e^{(\beta' x_i)}}{\sum_{k\in R(t_i)} e^{(\beta' x_k)}} \right]^{\delta_i} = \sum_{i=1}^{n} \delta_i \left[\beta' x_i - \log\left(\sum_{k\in R(t_i)} e^{(\beta' x_k)} \right) \right]$$

$$\text{for} \quad h = 1,2,...p. \quad (14)$$

The maximum partial likelihood estimate $\hat{\beta}$ can be obtained uniquely by solving the partial likelihood equation:

$$U_h(\beta)=0$$

Whereas the second derivative of the partial likelihood is given by:

$$\frac{\partial^2 l(\beta)}{\partial\beta_g\beta_h} = -\sum_{i=1}^{n} \left[\left(\frac{\sum_{k\in R(t_i)} x_{kh} x_{kg} e^{(\beta' x_k)}}{\sum_{k\in R(t_i)} e^{(\beta' x_k)}} \right) - \left(\frac{\sum_{k\in R(t_i)} x_{kh} e^{(\beta' x_k)}}{\sum_{k\in R(t_i)} e^{(\beta' x_k)}} \right) \left(\frac{\sum_{k\in R(t_i)} x_{kg} e^{(\beta' x_k)}}{\sum_{k\in R(t_i)} e^{(\beta' x_k)}} \right) \right] \quad (15)$$

$$\frac{\partial^2 l(\beta)}{\partial\beta_g\beta_h} = -\sum_{i=1}^{n} \left[\left(\frac{\sum_{k\in R(t_i)} x_{kh} x_{kg} e^{(\beta' x_k)}}{\sum_{k\in R(t_i)} e^{(\beta' x_k)}} \right) - \left(\frac{\left(\sum_{k\in R(t_i)} x_{kh} e^{(\beta' x_k)}\right)\left(\sum_{k\in R(t_i)} x_{kg} e^{(\beta' x_k)}\right)}{\left(\sum_{k\in R(t_i)} e^{(\beta' x_k)}\right)^2} \right) \right] \quad (16)$$

This matrix for g=1,2,... p and h=1,2,...p and, is a sum over $i=1,2...,n$ for loaded covariance matrices for the x vector within the populations at risk at the time t_j. The negative of the second partial derivatives provide the observed information matrix which estimates the covariance matrix of the estimated regression coefficients (Klein and Moeschberger 1997).

Consequently,

$$I(\beta) = -\frac{\partial^2 l(\beta)}{\partial\beta_g\partial\beta_h} \quad (17)$$

$I(\beta)$ is defined as the observed information matrix. The score vector $U(\beta_o)$assessed at the real estimate of β will be asymptotically spread as a multivariate normal with mean vector zero and covariance matrix which can be unbiased estimated by $I(\beta_o)$.

$$U(\beta_0) \square N\left(0, I(\beta_0)\right) \quad (18)$$

The estimate β will also be asymptotically normal

$$\hat{\beta} \square N\left(\beta_0, I^{-1}(\beta_0)\right) \quad (19)$$

Model Checking in Cox Regression Model

After fitting the regression model to the dataset, the adequacy of the fitted regression model needs to be evaluated which is usually performed using model residuals.

Cox-Snell Residuals

Proposed by Cox and Snell, the concept is used to examine the general fit of Cox regression model (Collet 2015).

The Cox-Snell residual for the i^{th} individual is described as:

$$r_{ci} = \exp\left(\hat{\beta}' x_i\right) \hat{H}_0(t_i) \quad (20)$$

where $\hat{H}_0(t)$ stands as estimate of the baseline cumulative hazard function at time t_j. In practice the Nelson-Aalen estimate is generally used. If the final PH regression model is perfect and the $\hat{\beta}$ are near to the correct values of the β, then r_{ci} should approximate a censored sample from a unit exponential distribution. Therefore, a plot of the Nelson-Aalen cumulative hazard estimate of residuals $\hat{H}(r_{ci})$ versus residuals r_{ci} should be a straight line through the origin with a slop of 1, if the fitted model is correct.

Proportional Hazard Assumption Checking

One of the basic assumptions of the Cox regression model is the proportionality of hazards

within covariates. Exceptionally, the model presumes that each predictor has a multiplicative impact in the hazards function that is steady over time. Several proposals for examining the PH assumption are accessible for quality cohort studies, including both statistical tests and graphical procedures. There are various technics for checking if a model fulfills the assumption of proportionality (Graphical Method, Scaled Schoenfeld Residuals, Adding Time Dependent Covariate) (Han et al. 2016).

Graphical Method

The survival function for Cox regression model is defined as:

$$S_i(t) = [S_0(t)]\exp\{\beta' x_i\}$$

where $x=(x_1,x_2,...,x_p)'$ is the values of the vector of covariates for a given individual. By using the logarithm twice in the equation related to survival function, the authors get:

$$\ln[-\ln S_i(t)] = \beta' x_i + \ln[-\ln S_0(t)] \quad (21)$$

The log-log curves corresponding to two distinct covariates $x_1=(x_{11},x_{12},...,x_{1p})$ and $x_2=(x_{21},x_{22},...,x_{1p})$ which does not depending on t is given as:

$$\ln[-\ln S_i(t,x_1)] - \ln[-\ln S_i(t,x_2)] = \sum_{i=1}^{n} \beta(x_{1i} - x_{2i}) \quad (22)$$

This provides the basis for assessing the validity of PH assumption. The estimated plot of log (-log*(survival)*) against survival time for two groups would show the curves to be parallel if the hazard functions are proportional (Xiaonan et al. 2013). The technique does not provide good result for categorical covariates with many levels because the graph becomes jumbled.

Scaled Schoenfeld Residuals

As one of the methods used to check the proportionality, Schoenfeld residuals are described as the product of the inverse of the approximated variance-covariance matrix of the Schoenfeld residual and the Schoenfeld residual (Kabtamu and Sharma 2014; Buchholz et al. 2014).

The scaled Schoenfeld residual can be used to assess time trends and lack of proportionality.

$$r_{pji}^{*} = \left(V^{-1}\right) r_{pji} \quad (23)$$

Where r_{pji} is the Scaled Schoenfeld residual and is r_{pji} the Schoenfeld residual. Under the null hypothesis, the authors expect to observe a steady function over time. For the proportional hazards assumption to hold, a straight line with zero gradient is foreseen.

The Stratified Cox Regression Model

This is an improved Cox regression model that accommodates a dominance by stratifying the covariates that violate the PH assumptions. The covariates that does not violate the assumptions are added to the model. However, the predictors being stratified out are not incorporated in the model (Gonzalez et al. 2013). In the stratified Cox regression model, models are defined based on their interaction and no-interaction. (Abdelaal and Zakria 2015; Jones et al. 2015).

No-interaction Model

In no-interaction model, the strata split the individuals into disjoint categories with baseline hazard $h_{ok}(t)$ and a collective estimate for the regression parameter coefficients $\beta_1, \beta_2,...,\beta_p$ are alike for each stratum. The hazard function for the failure time is expressed by:

$$h_k(t/x) = h_{0k}(t)\exp(\beta' x) \quad (24)$$

where k denotes the particular stratum $(k=1,...K)$, β is a vector of unknown regression parameters, are k unknown baseline hazard functions, k indicates that each stratum has its particular baseline hazard function and β are the same across strata.

The proportional hazards for two individuals in the same stratum expressed as:

$$\frac{h_k(t/x_1)}{h_k(t/x_2)} = \exp(x_1 - x_2)\beta' \quad (25)$$

On the other hand, individuals from different groups can have for non-proportional hazards functions may differ

$\frac{h_{0k}(t)}{h_{0k'}(t)}$, comparing strata k to k'.

Since these functions are unrestricted any relationship for this hazard ratio overtime is possible. The partial likelihood for the stratified Cox model is the product of partial likelihoods in each stratum:

$$L(\beta) = \prod_{k=1}^{K} L_k(\beta) \quad (26)$$

Interaction Model

The data set can be classified into k strata following that the variable does not meet the proportional hazards assumption; for instance, the interaction model is determined as:

$$h_k(t/x) = h_{0k}(t)\exp\left[\beta_{1k}x_1 + \beta_{2k}x_2 + ... + \beta_{pk}x_p\right] \quad (27)$$

In this model, the individual coefficient has the k subscript, which indicates the k^{th} stratum and characterizes that the regression coefficients are dissimilar for distinct strata. So if there is no interaction the stratified Cox regression model will include regression coefficients that were constant over the strata. In a case where interaction is allowed, distinct coefficients for each of the stratum will be derived. Likelihood ratio test statistics is exercised to inspect the no-interaction assumption.

$$LR = -2\log L_{(\text{no int.})} - \left[-2\log L_{(\text{no int.})}\right] \quad (28)$$

The likelihood ratio (LR) test collates log likelihood statistics for the interaction model and the no-interaction model.

Significance of the Research

This research will result in the following boosts to the society (health practitioners, educators, researchers, community)

- ✓ It will dispense a statistical approach skilful in dictating the risk determinants affecting the relapsing from malaria.
- ✓ It will dispense the risk determinants themselves.
- ✓ Stimulate legitimate organization and policy skilful on prevention, control and intervention strategies of malaria.
- ✓ Conclusions of this exploration will be utilized as a starting point for further researches on malaria regularity in communities with the similar attributes as the one include in this investigation.

Objectives of the Research

The following are the study objectives:

- ✓ To determine the risk determinants that influence the regularity of malaria.
- ✓ To establish the pre-imminent cox proportional hazards model for the observations.
- ✓ To evaluate the established model for adequacy utilizing existing tools.

Data Set

Run in Lubumbashi (Democratic Republic of Congo), the study consisted of 109 patients who qualified based on the entry conditions such as the minimum age, the type of malaria that is supposed to be uncomplicated falciparum of malaria, a pregnancy test for female patients, among others. From the 109 patients, 98 (89.9 %) relapsed from the uncomplicated falciparum malaria while the rest of 11 patients (10.1 %) did not experience the event till the end of the study that lasted 180 days from the day the patient is declared free from malaria till he comes back as a result of a relapse. The majority of patients were males (61.5 %). 81.7 percent of the patients were aged from 18 to 59 years, 51.4 percent were single, 41.3 percent did not complete secondary school and a majority of 67.9 percent of patients were living in formal brick as type of dwelling.

In the following analysis, the time to relapse in days was captured as dependent variable and the sanitation, environmental; prevention and biographical information of patients were also captured as covariates.

The results of Cox PH regression model is obtained by using (SAS 9.1.3) statistical packages.

Description of the Variables

In the analysis, the following were used as variables:

- ***Survival Time:*** time to relapse or censoring time, measured in days.
- ***Death Status:*** the event indicator, equal to 1 for those who relapsed during the period of the study and 0 for those who did not relapse or censored.
- ***Patient Age:*** a categorical age variable, [1: <18 years; 2: 18 years-35 years; 3: 36 years-59 years; 4: >= 60 years];
- ***Tdwel:*** type of dwelling: a categorical variable, [1: formal brick; 2: backyard shack];
- ***Source:*** source of drinking water: [1: unprotected spring; 2: capped spring; 3: unprotected dug well; 4: protected dug well; 5: tub well; 6: surface water; 7: cart with small tank; 8: public tap; 9: piped into yard; 10: piped into dwelling];
- ***Toilet:*** kind of toilet [1: no facility; 2: shared pit latrine; 3: own pit latrine; 4: own pit latrine with slab; 5: shared pit latrine with slab; 6: own pit latrine with cement slab and vent pipe; 7: shared pit latrine with cement slab and vent pipe];
- ***Maint:*** I have a well maintained toilet in my yard [1: yes; 2: no];
- ***Dump:*** There are dumping places in the area where I am living [1: yes; 2: no];

- ***Pit:*** is your pit toilet well covered [1: yes; 2: no];
- ***Water:*** Is your compound having water dummy [1: yes; 2: no];
- ***Cult:*** Are you cultivating in your compound [1: yes; 2: no];
- ***Spr:*** indicate the type of mosquito spray used [1: doom; 2: Baygon];
- ***Net5:*** type of net being used [1: untreated net; 2: locally treated net; 3: long-lasting insecticide treated net; 4: don't know];
- ***Net6:*** Has the net been treated with chemical to kill mosquitoes? [1: yes; 2: no; 3: don't know];
- ***Inf1:*** in your opinion, what is the main cause of malaria? [1: drinking dirty water; 2: eating immature sugarcane; 3: eating other dirty food; 4: mosquito bites; 5: getting soaked with rain; 6: cold of changing weather; 7: don't know].

RESULTS

According to the Univariate Cox PH analysis, that all the covariates are statistically significant and selected as significant risk factors related to relapsing after being treated for uncomplicated falciparum malaria. The result indicates that all the variables were found to be a significant risk factors of malaria relapse (Table 1).

A full multivariate Cox PH analysis was conducted (by using stepwise selection process) including risk factors that had a p-value less than or equal 0.99 in univariate Cox PH analysis. The result indicates that maint, pit, spr and infl were significant factors contributing to malaria relapse. Having a well maintained toilet in my yard is 16.7 percent reduced rate of malaria relapse, having pit toilet well covered is 14 percent increased the rate of malaria relapse, the type of mosquito spray used was having 38.9 percent reduced rate of malaria relapse and opinion on patients' main cause of malaria was 2 percent increased the rate of malaria relapse (Table 2). The prior and posterior steps were used to add and remove variables in the model. The variable selection in the model was set at a threshold of p-value = 0.99 (SLENTRY = 0.99) and variable removal threshold from the model was set at p-value = 0.995 (SLSTAY = 0.995). Thus, variables with $p < 0.99$ will be included in the model testing and to the variables that will be subsequently tested should be $p < 0.995$.

Furthermore, in Cox model optimization, the variables are removed from the predictive model if they have the highest p-value and over threshold until they show a significance in the prediction of the hazard rate. In Table 2, variable "net6" has highest p-value and was removed from the regression test. The variable "source" was removed next because it has highest p-value. This strategy was applied to every analysis and the results is displayed in Table 3.

Hence after the selection process, the variables "tdwel", "maint", "dump", "pit", "spr" and "infl" were kept in the modelling of the Cox proportional hazards as they are statistically significant.

The predictors' interaction with each other was used for prediction in the model and all possible combinations of interactions of all raw variables are included for proportional hazard regression analysis test for goodness of model fit. However, from the variable interaction, variable "main_infl" was removed based on highest p-value = 0.9515. Subsequently, interaction terms "pit_infl", "spr_infl" were removed from the

Table 1: Univariate analysis

Variable	*DF*	*Parameter estimate*	*Standard error*	*Chi-square*	*Pr>ChiSq*	*Hazard ratio*
ge	1	0.38279	0.20680	3.4262	0.0591	1.466
tdwel	1	-0.24122	0.12181	3.9217	0.0477	0.786
source	1	-0.43590	0.21541	4.0949	0.0430	0.647
toilet	1	0.66826	0.20534	10.591	0.0011	1.951
maint	1	0.45680	0.22680	4.0566	0.0440	1.579
dump	1	-0.47928	0.20578	5.4246	0.0199	0.619
pit	1	0.54369	0.22269	5.9607	0.0146	1.722
water	1	-0.55062	0.20633	7.1215	0.0076	0.577
cult	1	-0.47447	0.22486	4.4525	0.0349	0.622
spr	1	-0.49974	0.21480	5.4129	0.0200	0.607
net5	1	0.62993	0.20617	9.3357	0.0022	1.877
net6	1	0.26114	0.20519	1.6197	0.2031	1.298
infl	1	0.19605	0.06906	8.0602	0.0045	1.217

Table 2: Multivariate Cox PH regression analysis

Variable	*DF*	*Parameter estimate*	*Standard error*	*Chi-square*	*Pr>ChiSq*	*Hazard ratio*
ge	1	0.21818	0.23903	0.8331	0.3614	1.244
tdwel	1	-0.26596	0.14190	3.5131	0.0609	0.766
source	1	0.19347	0.73937	0.0685	0.7936	1.213
toilet	1	0.47602	0.39822	1.4289	0.2319	1.610
maint	1	-1.78825	0.90454	3.9084	0.0480*	0.167
dump	1	1.49187	1.06044	1.9792	0.1595	4.445
pit	1	2.70705	1.36582	3.9283	0.0458*	14.985
water	1	0.99367	1.46904	0.4575	0.4988	2.701
cult	1	0.92311	0.77378	1.4232	0.2329	2.517
spr	1	-0.49244	0.23536	4.3775	0.0364*	0.611
net5	1	0.09009	0.38700	0.0542	0.8159	1.094
net6	1	0.02072	0.25607	0.0065	0.9355	1.021
infl	1	0.89399	0.39811	5.0425	0.0247*	2.445

Table 3: The final Cox PH model

Variable	*DF*	*Parameter estimate*	*Standard error*	*Chi-square*	*Pr>ChiSq*	*Hazard ratio*
tdwel	1	-0.25284	0.12691	3.969	0.0463	0.777
maint	1	-1.97742	0.8212	5.7983	0.016	0.138
dump	1	1.93978	0.78225	6.1491	0.0131	6.957
pit	1	2.07766	0.81411	6.5131	0.0107	7.986
spr	1	-0.47452	0.22161	4.5847	0.0323	0.622
infl	1	0.78281	0.26109	8.9895	0.0027	2.188

regression analysis with *p*-values at 0.3910 and 0.8086 respectively. After optimization and routine check, there was no other interaction terms that was significant in the model. The optimization of the final led to the results display in the Table 4.

After going through all the selection process, it was found that none of the interaction terms was statistically significant in the model; furthermore some of the variables such as "pit", "maint" and "tdwel" that were significant on the univariate selection process were dropped out.

Hence the final proportional hazard model is: $h(H=h_o(t)\ _{exp}(1.91613\ x_{dump}-0.49633x_{spr}+\ 0.81466x_{infe})$

The proportionality model assumption was satisfied for the categorical variables like "dump", "spr" and "infl" that were found to be significant in the selection process and these three categorical variables were treated as fixed covariates, since they never change over time that is, the time-dependent variables were not detected in the final model. Therefore, the model is suitable to predict the time to event observations with fixed certain measurements or characteristics.

Model Checking

Validation of Assumption of Proportionality

The final model was built based on the proportionality assumption test for hazards between groups. The Kaplan-Meier Curve *(log(-log (survival)))* was used to establish the assumption of proportionality, and the curves between groups were parallel. However, the overall proportionality of the final regression model for the categorical variables is observed. The assumption of proportionality of the three categorical variables that are in the final model was confirmed by the results displayed in Table 5 where all the three p-values are less than the threshold of 0.05.

Table 4: Optimized Final Cox PH Model

Variable	*DF*	*Parameter estimate*	*Standard error*	*Chi-square*	*Pr> ChiSq*	*Hazard ratio*	*Hazard ratio 95% Confidence limits*
dump	1	1.90887	0.75198	6.4437	0.0111	6.745	1.545 ; 29.451
spr	1	-0.49633	0.21664	5.2487	0.0220	0.609	0.391 ; 0.9150
infl	1	0.81466	0.25233	10.4232	0.0012	2.258	1.380 ; 3.7250

Table 5: Test for proportional assumption

Variable	*DF*	*Parameter estimate*	*Standard error*	*Chi-square*	*Pr>ChiSq*	*Hazard ratio*
dump	1	1.91613	0.74945	6.5368	0.0106	6.795
spr	1	-0.49633	0.21664	5.2487	0.0220	0.609
infl	1	0.81466	0.25233	10.4232	0.0012	2.258

Linear Hypotheses Testing Results

Label	*Wald Chi-square*	*DF*	*Pr>ChiSq*
test_proportionality	18.7619	3	0.0003

From the result as presented in Table 5, the test labelled "test_proportionality" is significant (p-value = 0.0003), which means that all the covariates considered in the final model are significant, hence the proportionality of hazards is met.

Statistical Significance of Covariates

While examining statistical significance of covariates three tests are being used: partial likelihood ratio test, score test and Wald test. Null hypothesis is always the same and assumes that parameters for all covariates are equal to 0. The first test is founded on the partial likelihood function that is used in estimation process. The Proc Phreg SAS function enables to obtain results of below described tests of statistical significance of covariates (Table 6).

As it can be read from the first part of the output, all tests of joint statistical significance are coherent and lead to conclusion that all variables retained in the final model are statistically significant. Considering each of the three covariates separately, the conclusion is similar - each of the covariates is significant at the level 0.05.

DISCUSSION

Malaria is a foremost root of death and sadness in the Democratic Republic of Congo (Global Burden Disease Compare 2010). To comprehend the risk factors of malaria relapse, many efforts have been made to identify such factors. In studies conducted in Sub-Saharan Africa, many risk factors were identified to be associated with progression of the disease (Jamison et al. 2006; Prathiba et al. 2012), and the risk factors such as gender, type of dwelling, source of drinking water and so on (Prathiba et al. 2012).

This study is building a predictive model of malaria risk factors of relapsing. Survival analysis has been widely used to determine the risk factors in various fields, such as biomedicine, engineering, social science, and so on. In the model, the univariate analysis is conducted to test the association between variable and the time to event through the Kaplan-Meier plots. However, further analysis is tested by Cox proportional hazards regression, hence the model fit is expressed as:

$$h(H=h_o(t)\ _{exp}(1.91613\ x_{dump}-0.49633x_{spr}+\ 0.81466x_{infl})$$

From the Cox PH model, the different relative risk related to each risk factor in the model can

Table 6: Test of significance of Covariate and adequacy of the model

Test	*Chi-square*	*DF*	*Pr>ChiSq*
Likelihood Ratio	18.2264	3	0.0004
Score	17.6569	3	0.0005
Wald	19.3975	3	0.0002

Analysis of Maximum Likelihood Estimates

Variable	*DF*	*Parameter estimate*	*Standard error*	*Chi-square*	*Pr>ChiSq*	*Hazard ratio*
dump	1	1.90887	0.75198	6.4437	0.0111	6.745
spr	1	-0.51389	0.21711	5.6027	0.0179	0.598
infl	1	0.81871	0.25233	10.4467	0.0012	2.268

be determined as follows: Firstly the risk factor "dump" with a coefficient of 1.916 has been found significant in the final model (p<0.0106; 95% HR C.I: 1.545-29.451). Fixing other covariates from the model, patients who were exposed to dumping sites after being treated for malaria have short expected relapsing time than those who were not. The corresponding relative risk (hazard ratio) is:

$$\text{Relative Risk} = \frac{h_0(t)\exp(1.916 \times 1)}{h_0(t)\exp(1.916 \times 0)} = \exp(1.916) = 6.79$$

Hence, patients who were exposed to dumping sites or living in area with dumping sites are 6.79 times (an increase of 579 %) likely than those who were not exposed to relapse from malaria. The results are confirmed by a study done in Abuja, Nigeria by Nasir et al. (2015) where it was found that there was statistical relationship between malaria parasitemia among residents proximal to environmental waste dumpsites, more so with their age distribution ($P < 0.05$). The same conclusion was reached in a study done in Freetown, Sierra Leone, by Sankoh et al. (2013); the authors find that open dump sites do not facilitate environmental protection; they can constitute crucial public health problems and surrounding impacts in general and malaria in particular in urban areas. Moreover a study done in Accra, Ghana by Ramatta et al. (2014) where it was confirmed that about 83 percent of the respondents thought that inappropriate refuse management could lead to malaria infection. Last but not least a study done in Osun State, Southwestern Nigeria by Bamidele et al. (2012), revealed that 81.1 percent of malaria cases are related to hygiene and sanity which results once more confirm the authors' findings.

Secondly the covariate "spr" has been found also to be statistically significant in the final model (p<0. 0.0220; 95% C.I: 0.391-0.915). Considering other risk factors as constants, malaria patients who were using the first choice of spray recommended as one of preventive measures have a long expected relapsing time than those who were not using it. The corresponding relative risk is:

$$\text{Relative Risk} = \frac{h_0(t)\exp(-0.496 \times 1)}{h_0(t)\exp(-0.496 \times 0)} = \exp(-0.496) = 0.609$$

Hence a relapsing time decrease of 39.1 percent is expected for patients who were using the first choice of mosquito spray. Thus patients who were using the first choice of mosquito spray will have a longer expected relapsing time than those who did not. This finding confirms the result from the study done in East Shoa Zone, Ethiopia, by Hamusse et al. (2012) where it was found that after the focal spray, there was significant depletion in malaria prevalence in the villages and spraying was mingled with a 62 percent reduction in malaria prevalence. Furthermore in a research done by Fullman et al. (2013), it was found that the usage of mosquito spray and net have reduced significantly malaria infection in sub-Saharan Africa, which result confirms the findings based on the usage of spray as a measure of prevention of malaria.

Lastly the covariate "inf1" was also proved significant in the study (p<0. 0.0012; 95% C.I: 1.380-3.725). Fixing other risk factors in model, malaria patients who are not well informed about malaria have short expected relapsing time than those who are aware of the disease and well informed about it. The associated relative risk is:

$$\text{Relative Risk} = \frac{h_0(t)\exp(0.815 \times 1)}{h_0(t)\exp(0.815 \times 0)} = \exp(0.815) = 2.259$$

Thus for a lack of information about source of malaria, the risk of relapsing of a given patient will increase by 125.9 percent (2.259-1.00). Interestingly, knowing the main cause of malaria seems to be involved in impacting the time to relapsing from uncomplicated falciparum malaria. It makes sense somehow, since once aware of the cause of the disease one will be able to prevent it. This result is confirming the study done in Osun State, Southwestern Nigeria by Bamidele et al. (2012) where it was found that 65.5 percent of malaria cases were attributed mosquito bites, and the same result was confirmed by a study done by Singh et al. (2014) in Aliero, Northern Nigeria where it was discovered that awareness of the role of mosquitoes in malaria transmission and the causes of malaria can lower the risk of contracting the disease, and a comprehensive knowledge about malaria safeguards was high (90%), but not bounced back in their application (16%). Furthermore a study done by Katherine et al. (2013) done in northern Sri Lanka, showed that of the (99.0%) of investigated people who were well informed about the term 'malaria', 78.7 percent thought that malaria was a life-and-death illness and 71.3 percent of people were aware that malaria was transmitted via mosquito bites did not get the disease as they were aware of the disease and they took some preventive measures. Moreover, 80 per-

cent of surveyed people were aware that not all types of mosquitoes could pass on malaria; this was significantly related to level of education, with those having only primary education level having a mediocre malaria's knowledge.

Other factors, such as mosquito spray used as prevention method, dumping sites in the area where patients are living are well known disease-related risk factors as confirmed by Prathiba et al. (2012). It is not surprising to see their influence on the time to relapse from uncomplicated falciparum malaria in this study. In other terms, the higher values (categories) of the predictors are, the relapsing time will be shorter, while the hazard rate will be higher.

CONCLUSION

The objective of this research was to model the time to relapse from uncomplicated falciparum malaria, hence Cox PH approach was used to model the biographical, environmental, sanitation and preventive risk factors related to the relapsing time of uncomplicated falciparum malaria patients. Based on the results, the Cox PH model was found to fit the data better, as confirmed by the result of the global test that present reasonable and significant results (Likelihood Ratio: 18.2264 with $p<0.0004$; Score: 17.6569 with $p<0.0005$ and Wald: 19.3975 with $p<0.0002$).

The Cox PH model was significantly dependent to the relapsing time of uncomplicated falciparum malaria patients, hence the mathematical model below was found adequately fitting the data. It was found also in the study that none of the covariates was related to the time and all the possible interactions of the covariates retained in the model were not significant. In overall, the model fits the dataset well and the result has been confirmed by the goodness of fit for the linearity of the covariates.

RECOMMENDATIONS

In the light of the above, the researchers wishes to make some recommendations, which, if taken into consideration, might bring some positive changes to the current approach.

1. Peculiar edification of the community members, the planning of more public waste bins, and the removal of waste could assist prevents exposing the community to diseases such as malaria.
2. This study demonstrated that use of mosquito spray is effective in diminishing malaria incidence, relapsing from the disease but a larger scale study should evaluate the effectiveness of the type of spray used in the study in reducing malaria prevalence against its surrounding effect and auxiliary master plan for malaria intercepting.
3. Misunderstandings about malaria transferal and its sources still exist. Awareness about preventive and protective measures are not definitely convert into improvement in implementations. There is a necessity for educational and informative programs to expand the communities' efforts to broaden desirable attitude and practices regarding malaria and their involvement for malaria control.
4. Health edification on malaria predominance and control should be strengthened particularly through training and retraining of community health volunteers and to attain a long term control and management of malaria, importance should be laid on eradication of the breeding sites of mosquitoes in the community.

REFERENCES

Abdelaal MAA, Zakria SHEA 2015. Modeling survival data by using Cox Regression Model. *American Journal of Theoretical and Applied Statistics*, 4(6): 504-512. doi: 10.11648/j.ajtas. 20150 406.21.

Bamidele JO, Ntaji MI, Oladele EA, Bamimore OK 2012. Community participation in malaria control in Olorunda Local Government Area, Osun State, Southwestern Nigeria. *African Journal of Infectious Diseases*, 6(2): 24–28.

Bieszk-Stolorz B 2015. Geometric method of determining hazard for the continuous survival function. *Folia Oeconomica Stetinensia*, 15(1): 22-33. doi: 10.1515/foli-2015-0031

Buchholz A, Sauerbrei W, Royston P 2014. A measure for assessing functions of time-varying effects in survival analysis. *Open Journal of Statistics*, 4: 977-998. doi: 10.4236/ojs.2014.411092.

Collet D 2003. *Modeling Survival Data in Medical Research.* 2nd Edition. London: Chapman and Hall, CRC.

Collet D 2015. *Modelling Survival Data in Medical Research, Chapman & Hall/CRC Texts in Statistical Science.* 3rd Edition. Illustrated, Revised. Boca Raton London New York Washington, D.C.: A CRC Press Company.

Fullman N, Burstein R, Lim S, Medlin C, Gakidou E 2013. Nets, spray or both? The effectiveness of insecticide-treated nets and indoor residual spraying in reducing malaria morbidity and child mortality in sub-Saharan Africa. *Malaria Journal,* 12: 62. DOI: 10.1186/1475-2875-12-62.

Global Burden Disease Compare 2010. From <http://viz.healthmetricsandevaluation.org/gbd-compare> (Retrieved on 5 January 2016).

Gonzalez C, Dupuy J, López M, Luaces P, Rodríguez C et al. 2013. Stratified Cox Regression Analysis of Survival under CIMAvax®EGF Vaccine. *Journal of Cancer Therapy*, 4: 8-14. doi: 10.4236/jct.2013.48A002.

Hamusse SD, Balcha TT, Belachew T 2012. The impact of indoor residual spraying on malaria incidence in East Shoa Zone, Ethiopia. *Glob Health Action*, 5(2): 11619. doi: 10.3402/gha.v5i0.

Han K, Song K, Choi BW 2016. How to develop, validate, and compare clinical prediction models involving radiological parameters: Study design and statistical methods. *Korean J Radiol,* 17(3): 339-350.

Hosmer D, Lemeshow S 1990. *Applied Survival Data Analysis: Regression Modelling of Time to Event Data.* Wiley-Interscience Publication.

Jamison DT, Feachem RG, Makgoba MW, Eduard RB, Florence KB et al. 2006. *Disease and Mortality in Sub-Saharan Africa.*2nd Edition. Washington DC: The International Bank for Reconstruction and Development/The World Bank. ISBN-10: 0-8213-6397-2ISBN-13: 978-0-8213-6397-3. 2006.

Jones E, Sweeting MJ, Sharp SJ, Thompson SG, EPIC-InterAct Consortium 2015. A method making fewer assumptions gave the most reliable estimates of exposure–outcome associations in stratified case–cohort studies. *Journal of Clinical Epidemiology*, 68(12), 1397–1405. http://doi.org/10.1016/j.jclinepi. 2015.04.007

Kabtamu T, Sharma MK 2014. Application of Cox Proportional Hazards Model in Case of Tuberculosis Patients in Selected Addis Ababa Health Centres, Ethiopia. *Tuberculosis Research and Treatment,* 2014: Article ID 536976, 11 pages. doi:10.1155/2014/536976.

Kalbfleisch JD, Prentice RL 2002. *The Statistical Analysis of Failure Time Data,* 2nd Edition. New York: Wiley.

Katherine K, Gawrie NL, Galappaththy JJ, Kurinczuk SR, Deepika F 2013. Knowledge, attitudes and practices relevant to malaria elimination amongst resettled populations in a post-conflict district of northern Sri Lanka. *Transactions of The Royal Society of Tropical Medicine and Hygiene*, 107(2): 110-118. https://doi.org/10.1093/trstmh/trs015

Klein JP, Moeschberger ML 1997. *Survival Analysis Techniques for Censored and Truncated Data.* New York: Springer.

Li Z, Owzar K 2015. Fitting Cox Models with doubly censored data using Spline-Based Sieve Marginal Likelihood. *Scand J Statist*. Doi: 10.1111/sjos.12186. 2015.

Nasir IA, Muhammad MA, Emeribe AU, Babayo A, Shehu MS 2015. Prevalence of malaria parasitaemia among residents proximal to environmental waste dumpsites in Gwagwalada metropolis, Abuja, Nigeria. *J Med Trop*, 17: 91-96.

Ponnuraja C, Venkatesan P 2010. *Survival Models for Exploring Tuberculosis Clinical Trial Data-An Empirical Comparison.* Chennai: Tuberculosis Research Centre (ICMR).

Prathiba M, De Silva, John MM 2012. Factors contributing to urban malaria transmission in sub-Saharan Africa: A systematic review. *Journal of Tropical Medicine*, 2012: Article ID 819563, 10 pages. doi:10.1155/2012/819563

Ramatta MY, Dennis C, Philip BA 2014. Domestic waste disposal practice and perceptions of private sector waste management in urban Accra. *BMC Public Health*, 14(1): 697.

Rodrýguez G 2010. *Parametric Survival Models.* Spring 2001; Revised Spring 2005. Princeton, New Jersey 08544, USA: Princeton University Press.

Sankoh FP, Yan X, Tran Q 2013. Environmental and health impact of solid waste disposal in developing cities: A case study of Granville Brook Dumpsite, Freetown, Sierra Leone. *Journal of Environmental Protection*, 4(7): 665-670. doi: 10.4236/jep. 2013. 47076.

Singh R, Musa J, Singh S, Ebere UV 2014. Knowledge, attitude and practices on malaria among the rural communities in Aliero, Northern Nigeria. *Journal of Family Medicine and Primary Care*, 3(1): 39–44. http://doi.org/10.4103/2249-4863.130271

Xiaonan X, Xianhong X, Marc G, Thomas ER, Sylvia WS 2013. Testing the proportional hazards assumption in case-cohort analysis. *BMC Medical Research Methodology*, 13: 88. doi:10.1186/1471-2288-13-88.

Paper received for publication on September 2016
Paper accepted for publication on December 2016