MODELING THE RISK FACTORS OF MISCARRIAGE USING SURVIVAL ANALYSIS TECHNIQUES

ABSTRACT

Background: Miscarriage, also known as spontaneous abortion, is a significant adverse outcome of pregnancy. The risk factors associated with the transition from a normal pregnancy to a complete miscarriage before 28 weeks of gestational age have not been exhaustively established. Use of logistic regression to assess the factors associated with spontaneous abortion excludes the longitudinal and incompleteness aspects of miscarriage data. However, miscarriage is dynamical process where time until it occurs may be of interest.

Objectives: This paper modeled the risk factors associated with miscarriage using survival analysis, estimated and compared survivorship of levels of categorized variables, fitted proportional hazards and accelerated failure time models and compared their inferential capacity and their efficacy in identifying risk factors.

Methods: A retrospective study was conducted for pregnant women who were enrolled for antenatal care in Kakamega County General Teaching and Referral Hospital (KCGTRH) in Kakamega county, western Kenya. Pregnant women with recognized pregnancy and enrolled in the pre-natal care during the period from 1 January, 2019 up to 31 October, 2020 was recruited into the study. For descriptive analysis and estimation of survival functions, the study used Kaplan-Meier (K-M) and chi-squared test for independence. Comparison of survivorships in categorized variables was done using Kaplan-Meier curves and log-rank test. The Cox proportional hazards (PH) model and parametric models were used to analyze miscarriage data. All analysis were carried out using R software and SPSSv20. The level of significance was 5%.

Results: Of the total sample 248 mothers (4.1%) miscarried, while 5729 (95.9%) were censored. The significant factors identified by log rank test were ethnicity ($P = .000$), levels of education ($P = .048$), place of residence ($P = .000$), employment status ($P = .004$), malaria status ($P = .000$) and UTI status ($P = .000$). The covariates in categorized form found significant by log rank were number of previous stillbirths ($P = .000$) and number of

ANC visits ($P = .000$). The factors ethnicity, place of residence, malaria status, number of previous miscarriages, number of previous stillbirths and number of ANC visits were identified as the risk factors associated with miscarriages using cox model, parametric proportional hazards model and accelerated failure time models. The study found equivalent hazard ratios for among Cox model, parametric proportional hazards (PH) models andaccelerated failure time (AFT) models.

Conclusion: From the findings of this studyit can be concluded thatthe Gompertz proportional hazards (PH) regression model demonstrates a more favorable level of conformity to the data in comparison to the Cox model and accelerated failure time (AFT) models and there is association between certain explanatory variables and the time to miscarriages. In this paper survival analysis was used to analyze miscarriage data.

1. INTRODUCTION

Miscarriage, referred to as spontaneous abortion in medical terminology, is characterized as the inadvertent termination of pregnancy before the point at which the fetus would be capable of sustaining life autonomously beyond the confines of the maternal uterus (Cai, Y. and Feng, WQ, 2005).An alternative definition is the cessation of a non-viable pregnancy weighing below 500g prior to attaining 22 weeks of gestation (WHO, 1977). However, the definition of miscarriage is a multifaceted issue that is shaped by varying legal systems, religious viewpoints, and cultural norms surrounding pregnancy termination, both on a global scale and within specific nations(Ganatra et al., 2015).

The exploration of the occurrence and causes of miscarriage or spontaneous abortion in various populations remains an ongoing area of research. This is due to the elevated incidence of pregnancy loss during the period when pregnancies can only be identified through clinical means. The incidence of miscarriages has demonstrated notable fluctuations among many groups and temporal epochs. The current prevalence of miscarriage is commonly claimed to be around 10% to 15% of all pregnancies that have been clinically identified(Simpson and Carson, 1993; Simpson, J.L.and Mills, J.L., 1986; Zinaman et al., 1996; Garcia-Enguidanos et al., 2002). The empirical estimation of the prevalence of the phenomenon being studied has yielded a range of

reported percentages, with the lower end estimates ranging from 2% to 3%, and the top end estimates reaching as high as 30% (Cai, Y. and Feng, WQ, 2005).Previous investigations undertaken by Dellicour et al. (Dellicour S, Desai M, Mason L, et al, 2007; Dellicour S, Desai M, Mason L, et al., 2013) and Stanton (Stanton et al., 2006) have found a frequency of 12.2% in Kenya.

The cause of miscarriage remains uncertain. Nevertheless, some studies have identified several factors that contribute to and elevate the likelihood of spontaneous abortion. Various risk factors are involved in reproductive health, including demographic parameters such as a woman's years of age, gravidity, duration of gestation gap, and pregnancy history (including the number of prior live births and fetal losses)(Andersen et al., 2002; Regan et al., 1989)(Regan et al., 1989; Coste et al., 1991; Reagan, 1991; Parazzin et al., 1997; Ogasawara et al., 2000).Genetic factors encompass chromosomal abnormalities and mutations in genes(Garcia-Enguidanos et al., 2002; Simpson and Carson, 1993). Social and environmental factors encompass various determinants that can influence individuals' health outcomes. These factors include drug usage, caffeine intake, smoking and drinking habits, obesity, place of residence, and race(Chatenoud, L., Parrazin, F., Di cintio, E., Zanconato, G., Benzi, G., Bortolus, R. and La Vecchia, C., 1998; Cnattingius et al., 2000; Lindbohm et al., 2002; Rasch, 2003; Peck et al., 2010). Additional factors that can contribute to adverse pregnancy outcomes encompass uterine anomalies, fibroids, cervical insufficiency, post-operative alterations, chronic diseases, starvation, trauma, infectious diseases, social upheaval, and induced abortion(Carlson, E. and Mourgova, M., 2003; Ellett, K., Buxton, E.J., and Luesley, D.M., 1992; Shapiro, S. and Bross, D, 1980) as cited in (Wood, 1994).

Logistic regression has been extensively utilized by a substantial number of researchers to assess the factors associated with spontaneous abortion, specifically focusing on the viability of pregnancy at the end of the first trimester. Logistic regression is a statistically appropriate regression analysis method that is commonly employed to investigate binary outcomes, such as mortality. While this methodology offers odds ratios for potentially influential factors related to the risk of miscarriages, it does not take into account the longitudinal dimension. However, miscarriage is a dynamic process in which the duration till its occurrence may be of importance. In certain cases, there may be uncertainty regarding the occurrence and timing of a miscarriage event for some women. In other words, individuals who did not encounter

a miscarriage at the conclusion of the study time, had their participation in the study discontinued, or became disconnected from the researchers. The study may introduce bias if it excludes certain participant categories, as these categories may exhibit unique characteristics that could provide useful insights in the research. Hence, the utilization of risk/odds ratios or logistic regression methodologies may be considered unsuitable in a subsequent inquiry, particularly when the duration of follow-up plays a vital role in elucidating the incidence of the event under consideration, as these techniques overlook the temporal aspect.

Longitudinal studies involve the systematic observation of outcomes over an extended period of time. During this time, each individual in the study group is closely observed until the specific event of interest takes place. The statistical methodology deemed appropriate for this objective is often known as time-to-event analysis or survival analysis. Survival time models is one of the main research techniques employed in many fields including but not limited to medicine, biology , epidemiology, demography and engineering. For deep comprehensive and treatment of the subject see texts by (Kalbfleisch and Prentice, 2002; Collet, 2003; Therneau and Grambsch, 2000; Cox and Oakes, 1984; Lee and Wang, 2003). Analysis of duration time data involves estimating the basic functions related to the data set such as survival functions and hazard rate functions. This is estimation is done through nonparametric, semi-parametric and parametric methods. For non-parametric life-table method fuller details of this application can be found in the books by (Armitage, Berry and Matthews, 2002;Pollard, Yusuf and Pollard, 1990; Woodward, 2014). A text by (Collet, 2003; Lee and Wang, 2013; Kaplan and Meier, 1958; Breslow and Growley, 1974; Meier, 1975) provides detailed information on the derivations of the Product limit or Kaplan and Meier non parametric estimator. Another important non parametric estimator is Nelson-Aalen for its derivation can be found in the texts by (Altshuler, 1970; Nelson, 1972; Aalen, 1978b). Formulation of the hypothesis testing

procedures can be found in books by(Altman 1991; Armitage, Bery and Matthews 2002). Log –rank test procedures can be found in work of (Mantel and Haenszel, 1959; Mantel, 1966; Peto and Peto, 1972). Semi-parametric method (Cox proportional hazards model) and parametric method (Acceleration failure time model) their construction estimation and implementation full details can be found in texts by (Cox and Oakes 1984; Cox 1972;Therneau and Grambsch2000; Kalbfleisch and Prentice 2002). It is imperative to recognize that studies investigating miscarriage and involving the inclusion of pregnant women may encounter censorship and truncation, as there is an indeterminate fraction of the study population who had experienced prior pregnancy losses prior to their enrollment. The utilization of unconditional logistic regression, a commonly employed statistical method in such investigations, does not incorporate the consideration of censoring and truncation. On the other hand, survival analysis possesses the ability to handle censored and truncated data, is therefore considered to be more appropriate for this objective.

2.MATERIAL AND METHODS

2.1. Subjects and Variables under study

The study utilized a secondary dataset of 6,077 records of pregnant women who initiated prenatal treatment at Kakamega County Teaching and Referral Hospital (KCGTRH) in Kakamega County. The dataset encompassed a time frame that extended from January 1, 2019, to December 31, 2020.

The primary variable under investigation in this study was the duration in weeks until the occurrence of miscarriage. The independent variables, or covariates, included factors such as ethnicity, gravidity (number of pregnancies), maternal age, parity (number of previous live births), marital status, history of previous miscarriages, history of previous stillbirths, educational level, profession, HIV status, frequency of

antenatal care visits, presence of malaria infection, presence of urinary tract infection, presence of sexually transmitted diseases, and place of residence. For easier analysis of the data, several continuous covariates, including gravidity, mother's age, number of previous miscarriages, number of previous stillbirths, ANC visits, and parity, were grouped into categories. However, despite this categorization, these variables were still utilized in the model fitting process. The research acquired ethical clearance from the Institutional Ethics and Research Committee (MUCHS-MTRH IERC) of Moi University College of Health Sciences and Moi Teaching and Referral Hospital. The office of the chief executive for health County hospitals granted permission or approval to utilize medical data of pregnant women from specific facilities.The data analysis was carried out using R software and SPSS. The level of significance was 5%.

2.2 Terminology and Notation

The duration between the commencement of participation in a research project and the occurrence of a specific event of interest is commonly known as time-to-event or survival time. The variable "the time of an event", denoted as T, is a continuous random variable with a probability density function (pdf) denoted as $f(t)$ and a cumulative density function (cdf) denoted as $F(t)$. The $F(t) = P(T \leq t)$, represents the likelihood that the event has occurred by time t. The designated value of interest for the stochastic variable T is represented by the symbol t.

The sign d shall be utilized to represent the state of failure or censorship. In this context, the symbol $d$represents a binary variable that takes the value of 1 when an event of interest (such as failure) happens during the study period, and takes the value of 0 when the survival time is censored by the end of the study period. The probability of the occurrence of event time T at a specific time t is formally defined as

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t)}{\Delta t} \quad (2.1)$$

Survival function $S(t)$ is the probability of an individual surviving at least until time t. It gives the probability that the random variable T exceeds the specified time t.

$$S(t) = P(T > t) = 1 - F(t) = \int_t^{\infty} f(u)\delta(u) \quad (2.2)$$

$$f(t) = -\frac{\delta S(t)}{\delta t} \quad (2.3)$$

Hazard function, is the rate of event occurrence per unit time. It is the probability that if a subject survives to time t, he will succumb to the event in the next instant. It is denoted by $h(t)$ and is given by the formula:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t / T \ge t)}{\Delta t} \quad (2.4)$$

If one knows the form of $S(t)$, one can obtain the corresponding $h(t)$ and vice versa. This relationship can be expressed as:

$$h(t) = \frac{f(t)}{S(t)} = -\frac{\delta S(t)}{S(t)\delta t} = -\frac{\delta}{\delta t} lnS(t) \quad (2.5)$$

2.3. The Likelihood Functions for Censored Data.

Suppose that we have n units with life times $t_1, t_2, \ldots, t_n$ with a survivor function $S(t)$ with corresponding density $f(t)$ and hazard function $h(t)$. Let unit i be observed for time t_i. Then if the unit died at t_i, its contribution to the likelihood function is the probability density at that duration;

$$L_i = f(t_i) = h(t_i)S(t_i) \quad (2.6)$$

If the unit is still alive at time t_i, the probability of this event is

$$L_i = S(t_i) \quad (2.7)$$

This is its contribution to the likelihood function. Hence the likelihood function for the two contributions together is

$$L = \prod_{i=1}^{n} L_i = \prod_{i=1}^{n} h(t_i)^{d_i} S(t_i) \quad (2.8)$$

where $d_i = \begin{cases} 1 & i^{th}\ unit\ is\ uncensored \\ 0 & otherwise \end{cases}$ (2.9)

The log-likelihood function then becomes

$$logL = \sum_{i=1}^{n} [d_i\, logh(t_i) + logS(t_i)] \quad (2.10)$$

2.4. Survival Models and Estimation

2.4.1 Non-parametric Survival Estimation (estimating survival curves)

Assume a scenario in which there exists a population of n individuals, each having a recorded survival time $t_1, t_2, \ldots, t_n$. Further suppose that there are r death times amongst the individuals, where $r \leq n$. The r sorted death times are given as $t_{(1)} < t_{(2)} < t_{(3)} \ldots < t_{(r)}$. Thus the jth sorted death time is denoted by $t_{(j)}$ for $j = 1, 2, \ldots . r$. Let the number of subjects who are still surviving just before time t_j be denoted by n_j and d_j to represent those experienced event at this time. For a very small time interval h that includes one event, the Probability of individual gets an event in the interval $(t_{j-h}, t_j) = \frac{d_j}{n_j}$ then

$$S(t) = 1 - \frac{d_j}{n_j} = \frac{n_j - d_j}{n_j} \quad (2.11)$$

The survivor function in the Kaplan-Meier estimation is computed by multiplying a sequence of calculated probability. That is

$$\hat{S}(t) = \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j}\right) \text{ for } t_k \leq t < t_{k+1}\ , k = 1, 2, \ldots, r \quad (2.12)$$

Where k denote the number of observations in which failure events are observed to occur, with k being less than or equal to n.

Standard error of the estimated survival function that is, Kaplan-Meier estimator is given as;

$$VarS(t) \cong (S(t))^2 \sum_{j=1}^{k} \frac{d_j}{n_j(n_j-d_j)} \quad (2.13)$$

and standard error,se

$$\big(S(t)\big) \cong \sqrt{(S(t))^2 \sum_{j=1}^{k} \frac{d_j}{n_j(n_j-d_j)}} \quad (2.14)$$

A $(1-\alpha)100\%$ confidence interval for $S(t)$ for a given value of t is

$$\hat{s}(t) \pm Z\alpha_{/2} se(S(t))$$

2.4.2 Semi-Parametric Model (Cox Proportional hazards model, PHM)

The model provides a mathematical representation of the risk or probability of an event occurring at a specified moment, known as the hazard, for a person with a predetermined set of explanatory variables$X_1, X_2 \ldots X_p$. Let $h_0(t)$ determine the hazard function for an individual when the values of all the explanatory variables are zero. The function $h_0(t)$ is commonly referred to as the baseline hazard function.

The general proportional hazards model is

$$h_i(t) = exp\big(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}\big) h_0(t)$$

$$= h_0(t) e^{\sum_{j=1}^{p} \beta_j X_{ij}} \quad (2.15)$$

The linear model of logarithm of the hazard ratio is

$$\log\left(\frac{h_i(t)}{h_0(t)}\right) = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \beta_1 x_{ip} \quad (2.16)$$

2.4.3 Fitting the Cox Regression Model

The construction of the maximum likelihood estimator for the Cox proportional hazards model is as follows: Consider a dataset, consisting of n individuals, in which there are $r$number of unique death times and $n-r$the number of censored survival

times. The r ordered death times will be denoted by $t_{(1)} < t_{(2)} < \cdots < t_{(r)}$ so that $t_{(j)}$ is the jth ordered death time and $R(t_{(j)})$ be the risk set. Cox (1972) gave likelihood function for the model as

$$L(\beta) = \prod_{j=1}^{r} \frac{exp(\boldsymbol{\beta' x_{(j)}})}{\sum_{l \in R(t_j)} exp(\boldsymbol{\beta' x_l})} \quad (2.17)$$

In which $x_{(j)}$ is the vector of covariates or the set of explanatory variables associated with the individual who experiences mortality at the jth ordered death time$t_{(j)}$. The denominator represents the aggregate of the values of $exp(\beta' x)$ (hazards) in consideration of the current circumstances, there is a population of persons who are still susceptible to potential risks $t_{(j)}$ the numerator represents the hazard experienced by the subject who experienced the event at a certain time $t_{(j)}$. The denominator for the term corresponding to time $t_{(j)}$ is the sum of the hazards for those subjects still at risk at time $t_{(j)}$, and numerator is the hazard for the subject who got the event at time $t_{(j)}$. The likelihood can be calculated as the multiplication of individual likelihoods associated with each of the failure periods in a specific order.

The provided expression is the partial log-likelihood function associated with the given context as

$$Log\big(L(\beta)\big) = \sum_{j=1}^{r} \{\beta' x_{(j)} - Log \sum_{l \in R(t_j)} \exp(\beta' x_{(l)})\} \quad (2.18)$$

The maximum likelihood estimates of the β-parameters can be found by maximizing this log-likelihood function using numerical methods (Newton-Raphson procedure).

2.4.4 Accelerated Failure Time Model

In the context of the Accelerated Failure Time (AFT) model, the hazard function of the ith individual at a given time is often described as $t, h_i(t)$ is

$$h_i(t) = e^{-\alpha' x_i} h_0(t/e^{\alpha' x_i}) \quad (2.19)$$

And

$$S_i(t) = S_0\{t/e^{\alpha' x_i}\} \quad (2.20)$$

Where $\boldsymbol{\alpha}' = (\alpha_1, \alpha_2, \dots, \alpha_p)$ is the vector of unknown coefficients of the values of p explanatory variables $X_1, X_2, \dots, X_p$ with values $x_{1i}, x_{2i}, \dots x_{pi}$ for the ith individual.

The log linear form of the AFT model for the random variable T_i the lifetime of the ith subject is

$$logT_i = \mu + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi} + \sigma\varepsilon_i \quad (2.21)$$

The parameters μ and σ are commonly referred to as intercept and scale parameters respectively. The random variable ε_i is utilized to represent the departure of the values of $logT_i$ from the linear component of the model.

The likelihood function can be derived in a straightforward manner, from which the estimates can be obtained by the application of iterative methods. The probability associated with the n recorded survival periods $t_1, t_2, \dots t_n$ is given as

$$L = \prod_{i=1}^{n}\{f_i(t_i)\}^{\delta_i}\{S_i(t_i)\}^{1-\delta_i} \quad (2.22)$$

Where $f_i(t_i)$ and $S_i(t_i)$ are the density and survival functions of the i[th] individual at t_i respectively and δ_i is the event indicator for the i[th] observation. Expressing the likelihood function in terms of the survivor and density function of ε_i yields

$$L(\alpha, \mu, \sigma) = \prod_{i=1}^{n}\{\sigma(t_i)\}^{-\delta_i}\left\{f_{\varepsilon_i}(Z_i)\right\}^{\delta_i}\left\{S_{\varepsilon_i}(Z_i)\right\}^{1-\delta_i} \quad (2.23)$$

Taking the log of the likelihood function, we have

$$\ln L(\alpha, \mu, \sigma) = \sum_{i=1}^{n}\{-\delta_i \ln \sigma(t_i) + \delta_i \ln f_{\varepsilon_i}(z_i) + (1 - \delta_{)i} \ln S_{\varepsilon_i}(z_i)\}$$

(2.24)

Where

$$Z_i = \frac{ln\, t - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \cdots - \alpha_p x_{pi}}{\sigma} \quad (2.25)$$

The estimates of the unknown parameters, $\mu, \sigma, \alpha_1, \alpha_2, \ldots \alpha_p$ can be obtained by maximizing the log-likelihood function using the Newton-Raphson procedure.

3. Results

3.1 Distribution of study subjects and description of variables

The study utilized a secondary dataset of 6,077 records of pregnant women who initiated prenatal treatment at Kakamega County Teaching and Referral Hospital (KCGTRH) in Kakamega County. The dataset encompassed a time frame that extended from January 1, 2019, to December 31, 2020.

The primary variable under investigation in this study was the duration in weeks until the occurrence of miscarriage. The independent variables, or covariates, included factors such as ethnicity, gravidity (number of pregnancies), maternal age, parity (number of previous live births), marital status, history of previous miscarriages, history of previous stillbirths, educational level, profession, HIV status, frequency of antenatal care visits, presence of malaria infection, presence of urinary tract infection, presence of sexually transmitted diseases, and place of residence. For easier analysis of the data, several continuous covariates, including gravidity, mother's age, number of previous miscarriages, number of previous stillbirths, ANC visits, and parity, were grouped into categories. However, despite this categorization, these variables were still utilized in the model fitting process. Table 1 displays the covariates together with their respective descriptions, categorizations, and distributions of the study individuals.

From Table1 the ethnicity distribution is as follows; eighty nine (1.5%) were Kalenjin's, one hundred and four (1.7%) were Kikuyu's, five thousand two hundred and thirty one (86.1%) were Luhya's, five hundred and sixty five (9.3%) Luo's and others tribes were eighty eight (1.4%). Distribution of study subjects in other variables is as shown in the Table 1.

Table1 Description and Categorization of Variables

Overall (N=6077)		
Variable	**No. of mothers**	**Percentage (%)**
Ethnicity		
Kalenjin	89	1.5
Kikuyu	104	1.7
Luhya	5231	86.1
Luo	565	9.3
Others	88	1.4
Marital status		
Single	815	13.4
Married	5262	86.6
Educational level		
Primary	878	14.4
Secondary	4715	77.6
College	484	8.0
HIV Status		
Positive	111	1.8
Negative	5966	98.2
Malaria infection		
No	6073	99.9
Yes	4	0.1
STDs status		
No	6073	99.9
Yes	4	0.1
UTI status		
No	6073	99.9
Yes	4	0.1
Profession		
Unemployed	4407	72.5
Employed	1670	27.5
Place of residence		
Rural	1296	21.3
Urban	4781	78.7
Age of mother		
<20	863	14.2
21-25	2055	33.8

26-30	1556	25.6
31-35	1016	16.7
>35	587	9.7
Number of previous miscarriage		
≤2	6058	99.7
3-4	16	0.3
5-6	3	0.0
Number of previous stillbirths		
≤2	6069	99.9
3-4	7	0.1
>41	1	0.0
Number of ANC visits		
≤2	1009	16.6
3-4	4794	78.9
5-6	246	4.0
>6	28	0.5
Gravidity		
≤2	4003	65.9
3-4	1651	27.2
5-6	373	6.1
>6	50	0.8
Parity		
≤2	5053	83.1
3-4	858	14.1
5-6	147	2.4
>6	19	0.3

Time of follow up was calculated by assessing the time of conception or last menstrual period and date of miscarriage or censoring. The survival status of miscarriage was coded as

$$Status(d) = \begin{cases} 1 & miscarried \\ 0 & not\ miscarried \end{cases}$$

3.2 Preliminary Results and Analysis

Descriptive statistics were employed to outline the overall characteristics of the cohort, encompassing measurements such as the median, mean, and proportions. The female participants in the cohort being studied had an average age of approximately 26 years, with a standard deviation of 5.967. The age range of the female participants

varied from 13 to 47 years old. A total of 248 female participants, constituting 4.1% of the sample, reported experiencing miscarriages. Table 2 displays the descriptive statistics for the variables of interest.

According to the findings presented in Table 2, the distribution of events, specifically the number of women who experienced miscarriages, varied across different ethnic groups. Among the ethnic groups examined, the Kalenjin's had ten cases (11.2%), the Kikuyu's accounted for ten cases (9.6%), the Luhya's had one hundred and ninety-two cases (3.7%), the Luo's accounted for thirty-two cases (5.7%), and women from other tribes accounted for only four cases (4.5%). It has been observed that the Luhya ethnic group, despite constituting the bulk of the county's population, had the lowest incidence of miscarriages. The computed Pearson chi-square statistic for the ethnicity of women suggests a significant connection between survival status and ethnicity ($X^2 = 25.694, p - value < 0.001$). Distribution of events in other levels of categorized variables was as indicated in the table 2. Significance level was 5%.

Table2 Descriptive Statistics

Variable	(N=248) Miscarriage n(%)	(N=5829) Not miscarriage n(%)	Chi Square Test	
			X^2	p-value
Ethnicity				
Kalenjin	10(11.2)	79(88.8)	25.7	0.0000
Kikuyu	10(9.6)	94(90.4)		
Luhya	192(3.7	5039(96.3)		
Luo	32(5.7)	533(94.3)		
Others	4(4.5)	84(95.5)		
Marital status				
Single	35(4.3)	780(95.7)	0.110	0.741
Married	213(4.0)	504(96.0)		
Educational level				
Primary	33(3.8)	845(96.2)	6.075	0.048
Secondary	185(3.9)	4530(96.1)		
College	30(6.2)	452(93.8)		
HIV Status				
Positive	6(5.4)	105(94.6)	0.507	0.477
Negative	242(4.1)	5724(95.9)		
Malaria infection				
No	244(4.0)	5829(96.0)	94.078	0.0000
Yes	4(100.0)	0(0.0)		
STDs status				
No	248(4.1)	5825(95.9)	0.172	0.680
Yes	0(0.0)	4(100.0)		
UTI status				
No	246(4.1)	5827(95.9)	21.561	0.000
Yes	2(50.0)	2(50.0)		
Profession				
Unemployed	160(3.6)	4247(96.4)	8.310	0.004
Employed	88(5.3)	1582(94.7)		
Place of residence				
Rural	81(6.2)	1215(93.8)	19.799	0.000
Urban	167(3.5)	4614(96.5)		
Age of mother				
<20	26(3.0)	837(97.0)	3.506	0.469
21-25	83(4.0)	1972(96.0)		
26-30	71(4.6)	1485(95.4)		
31-35	43(4.2)	973(95.8)		
>35	25(4.3)	562(95.7)		
Number of previous miscarriage				
≤2	246(4.1)	5812(95.9)	3.031	0.220
3-4	2(12.5)	14(87.5)		
5-6	0(0.0)	3(100.0)		
Number of previous stillbirths				
≤2	246(4.1)	5823(95.9)	10.78	0.005

3-4	2(28.6)	5(71.4)		
>4	0(0.0)	1(100.0)		
Number of ANC visits				
≤2	85(8.4)	924(91.6)	59.466	0.000
3-4	157(3.3)	4637(96.7)		
5-6	6(2.4)	240(97.6)		
>6	0(0.0)	28(100.0)		
Gravidity				
≤2	148(3.7)	3855(96.3)	7.483	0.058
3-4	81(4.9)	1570(95.1)		
5-6	19(5.1)	354(94.9)		
>6	0(0.0)	50(100.0)		
Parity				
≤2	199(3.9)	4854(96.1)	2.970	0.396
3-4	43(5.0)	815(95.0)		
5-6	6(4.1)	141(95.9)		
>6	0(0.0)	19(100.0)		

3.3 Estimation and Comparison of Survivorship

The Figure 1 shows the overall Kaplan-Meier survival plot.The K-M plot revealed that about one percent of the pregnant women experienced the event within week twelve. It was also observed that more of the women were censored throughout the time period of study.

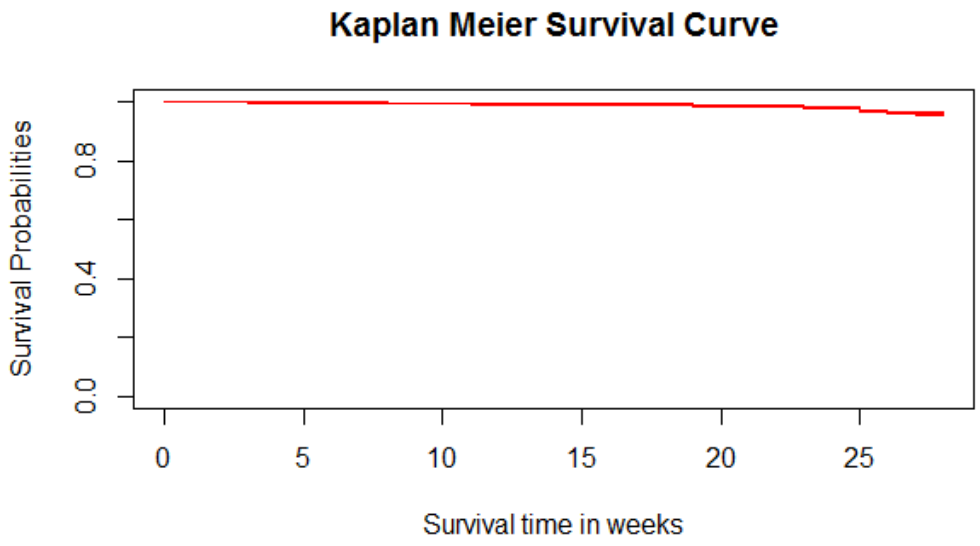

Fig .1

Figures 2 to 4 present the Kaplan Meier curves for some of the covariates, encompassing ethnicity, place of residence, maternal age and antenatal care visits (ANC).

A survival function characterized by a steep decline and an extended tail towards the right indicates a higher likelihood of experiencing early miscarriages. If the survivorship function does not converge to zero, it indicates that the highest recorded survival time in the research is a censored observation. In addition, a level whose curve is above curves of the other levels in the category indicates that its estimated survivor function is always greater than those of the other levels. This means that at any time t the estimated probability of survival past t is greater for this level.

The observation can be made from Figure 2 that the Kaplan Meier curves representing the survival of Kalenjin and Kikuyu women exhibit a tendency towards shorter

survival compared to the other groups. The curve representing the performance of the Luhya ethnic group continuously remains higher than the curves representing other categories, with the exception of the curve representing other tribes. This suggests that the Luhya tribe exhibits higher probability of survival for miscarriages compared to the Kalenjins', Kikuyus', and Luos', particularly starting from the tenth week of the gestational period. Prior to that particular week, it appears that their experiences in terms of survival were similar. Upon closer examination of the graph, it becomes apparent that the disparities among the curves are minimal.

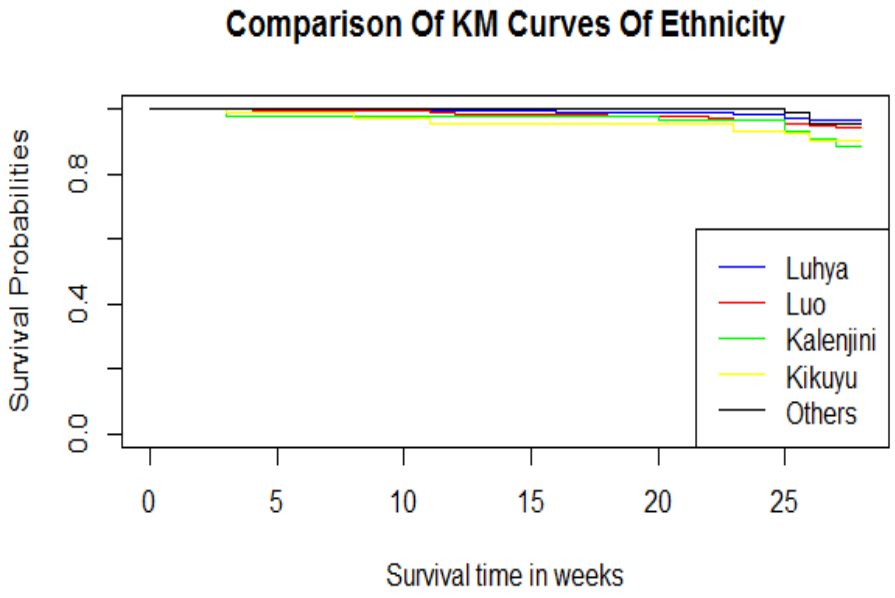

Figure 1 Kaplan Meier survival estimate curve by ethnicity

Figure 3 shows that women from the rural areas have lower survival probability estimates as compared to women from the urban areas for miscarriage, this because the curve for rural area lies below that of urban area

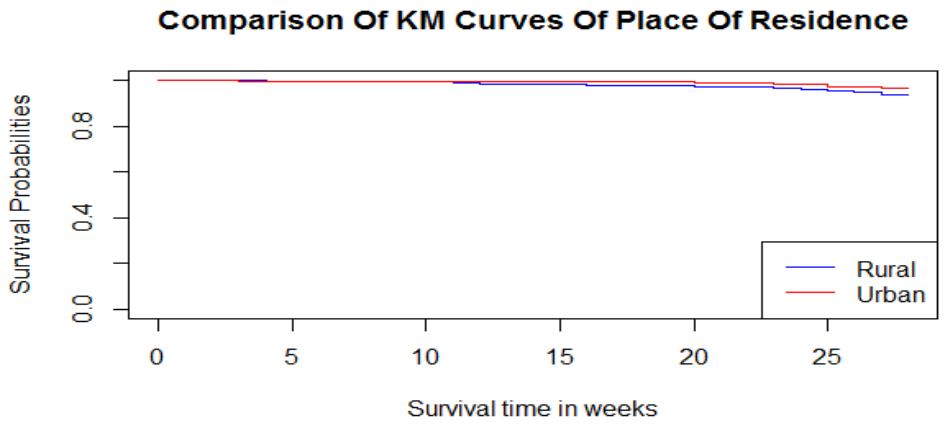

Figure 3 Kaplan Meier survival estimate curve by place of residence

Figure 4 shows the curves which are overlapping and thus indicating that there is no significant difference in the prospects of survival for the age of women categories.

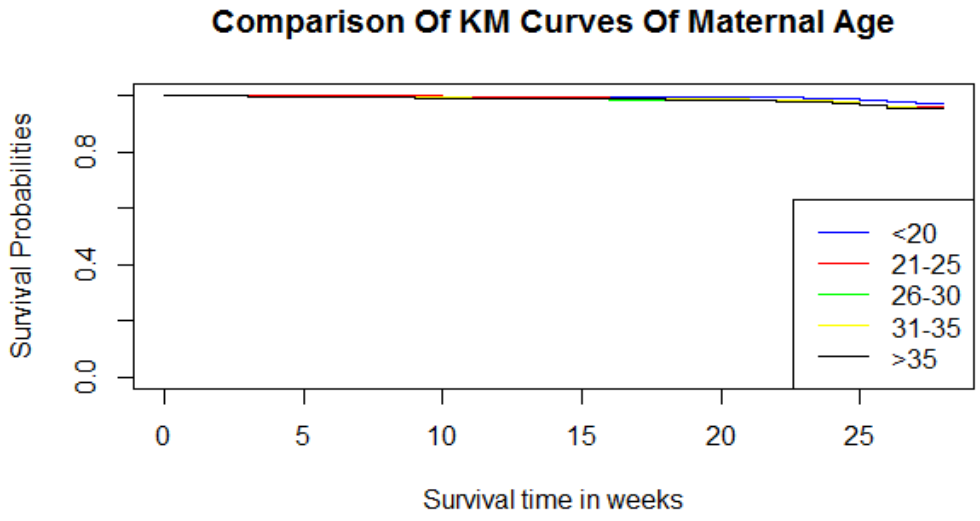

Figure 4 Kaplan Meier survival estimate curve by age of woman

Furthermore from Figure 5 women with number of ANC visits of above six have longer survival compared to those with ANC visits below six. Women with ANC visits of two and below have the lowest Survival curve over the whole period under study

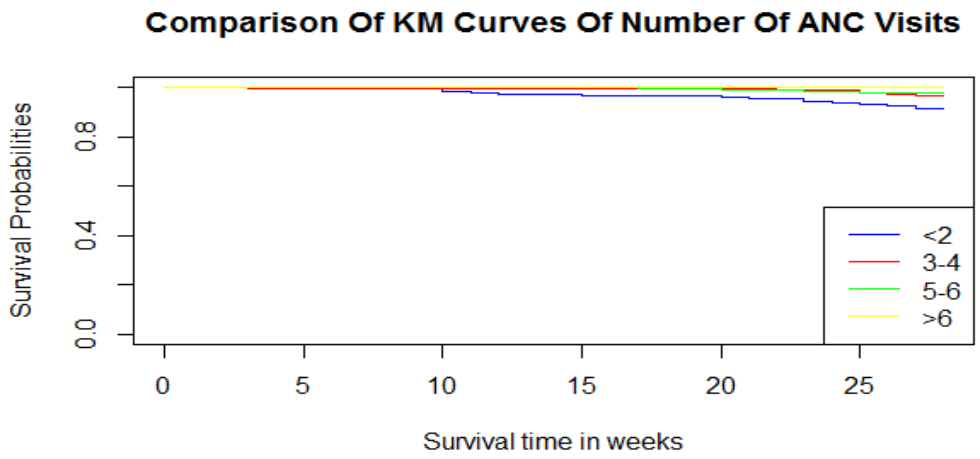

Figure 5 Kaplan Meier survival estimate curve by number of ANC visits

3.4 Comparison of survival functions using formal tests

In general, it can be noticed that the majority of the KM curves exhibit graphical distinctions among the covariate categories, with the exception of marital status, HIV status, and age of the woman. However, the visualization of estimated Kaplan-Meier survival plots does not clearly indicate any major differences among the estimated survival curves. As a result, a comprehensive analysis was performed using the Log rank (Mantel-Cox), Breslow (Wilcoxon), and Tarone-Ware tests to evaluate differences among the different categorical variables. The general hypothesis suggests that there is a lack of disparity rather than the presence of distinctions among the different groupings. Hence, the objective is to perform a test to substantiate the hypothesis that:

The null hypothesis (H_0) posits that there is no statistically significant variation in the survival times of miscarriage across the various categories whereas alternative hypothesis (H_1): A notable disparity exists in the survival durations of miscarriage across the various populations.

The statistical analysis conducted on ethnicity, as displayed in Table 3, reveals a statistically significant difference ($P = .000$) in survival experiences of miscarriage among the different ethnic groups. This conclusion is derived from the rejection of the null hypothesis. In the same way the, Log rank, Wilcoxon, and Tarone-Ware tests were employed to evaluate the influence of educational attainment, residential location, and occupational or employment status on survival of miscarriage. The tests revealed a statistically significant disparity in survival of miscarriage among the various categories.

In contrast, the Log rank, Wilcoxon, and Tarone-Ware tests when were employed to examine the influence of marital status, HIV status, age of a woman, and STDs on the duration of survival, the results revealed that there were no statistically significant disparities seen in survival outcomes across the diverse groups. As a result, the null hypothesis was not rejected, that there was no significant difference in survival times among the groups of women.

In addition, a series of statistical tests, namely the log-rank test, Wilcoxon test, and Tarone-Ware test, were conducted in order to examine the influence of urinary tract infections (UTIs), number of past stillbirths, and number of antenatal care (ANC) visits on survival outcomes. The findings of the study demonstrated a statistically

significant disparity in survival outcomes among various categories, hence resulting in the rejection of the null hypothesis. The log rank, Wilcoxon, and Tarone-Ware tests were employed to evaluate the association between the number of previous miscarriages, gravidity, and parity, and the diverse survival outcomes.

The findings of the conducted tests revealed that there were no statistically significant disparities identified across the various categories. Hence, the null hypothesis cannot be rejected, indicating that there is no substantial variation in the failure times of women among the various categories.

Table3: **Formal Test of Survival Functions**

Variable	**Log Rank**			**Wilcoxon**			**Tarone-Ware**			**Decision**
	X^2	**d.f**	**p-value**	X^2	**d.f**	**p-value**	X^2	**d.f**	**p-value**	
Ethnicity	26.444	4	0.000	26.541	4	0.000	26.492	4	0.000	Reject H_o
Marital status	0.113	1	0.735	0.117	1	0.732	0.115	1	0.735	Fail to reject H_o
Education level	6.243	2	0.044	6.271	2	0.043	6.257	2	0.044	Reject H_o
Place of residence	20.225	1	0.000	20.298	1	0.000	20.262	1	0.000	Reject H_o
Employment status	8.466	1	0.004	8.534	1	0.003	8.500	1	0.004	Reject H_o
HIV status	0.452	1	0.501	0.417	1	0.519	0.434	1	0.510	Fail to reject H_o
Malaria infection	152.070	1	0.000	149.404	1	0.000	150.735	1	0.000	Reject H_o
STDs infection	0.167	1	0.683	0.167	1	0.683	0.167	1	0.683	Fail to reject H_o
UTIs	32.397	1	0.000	32.577	1	0.000	32.489	1	0.000	Reject H_o
Age of a woman	6.943	4	0.139	67.043	4	0.137	6.993	4	0.136	Fail to reject H_o
Number of previous miscarriage	3.318	2	0.190	3.365	2	0.186	3.341	2	0.188	Fail to reject H_o

s										
Number of previous stillbirths	13.339	2	0.001	13.467	2	0.001	13.405	2	0.001	Reject H_o
Number of ANC visits	62.112	2	0.000	62.848	2	0.000	62.481	2	0.000	Reject H_o
Gravidity	7.483	3	0.058	7.495	3	0.058	7.491	3	0.058	Fail to reject H_o
Parity	3.024	3	0.338	3.066	3	0.382	3.045	3	0.385	Fail to reject H_o

3.5 Fitting Cox proportional model

The analysis commenced by including all covariates in the model (referred to as the complete model). Subsequently, a backwards stepwise regression procedure was employed, with the Akaike Information Criterion (AIC) serving as the metric for assessing the model's fit. Through this process, a more parsimonious model was identified based on the AIC criterion. The analysis of the full model yields the results reported in Table5 for the case where continuous variables are categorized and in Table6 for the case where they are not categorized.

The regression coefficients (Betas), estimated hazard ratios (exponentiated coefficients), standard errors, Wald statistics, matching p-values, and 95% confidence ranges for the hazard ratios are presented in Tables 5 and 6. The following covariates were selected using backwards stepwise AIC criterion when continuous variables are in categorized scheme: ethnicity (Kalenjin, kikuyu, luo), age (<=20), number of ANC visits (1 and 2-3), number of previous miscarriages (1-2), previous stillbirths (1-2 and >2), place of residence and malaria status.

In the case of continuous covariates when they are in the continuous form seven variables in full model were selected by this approach to be included in initial multivariable model. These seven variables were: ethnicity, number of ANC visits, number of previous miscarriages, previous stillbirths, resident place, employment status and malaria status. Finally, the covariates that were selected for inclusion in the initial multivariable model through method of backward stepwise using AIC criterion were: ethnicity, number of ANC visits, number of previous miscarriages, previous stillbirths, place of residence, and malaria status. The explanatory variables that were chosen have also been determined to be statistically significant at a significance level of 5-10 percent. As a result, they are potential candidates for inclusion in the initial multivariable model.

The multivariable model used in this investigation was determined by a backward stepwise selection process, which included continuous variables in their original continuous form. The resulting model, as presented in Table 7, was chosen as the model of interest for this study. Using the backward stepwise Akaike Information Criterion (AIC) criterion, it was seen that the covariates that were not initially considered for selection were likewise determined to lack statistical significance at a 5 percent level.

Table4 The Cox PH Full Model with Continuous variables in Categorized form

Covariate	Coef.	Exp(coef.)	s.e(coef.)	Z-statistic	P-value	95% Conf. interval	
Marital status(single)	0.109	1.115	0.2059	0.527	0.598	0.745	1.6688
Residence(urban)	-0.491	0.612	0.1433	-3.423	0.000	0.462	0.810
Employment(yes)	0.166	1.180	0.145	1.113	0.266	0.882	1.58
HIV(+)	0.251	1.286	0.423	0.594	0.552	0.561	2.945
Malaria(+)	3.173	23.88	0.758	4.185	0.000	5.402	105.533
STDS(+)	-14.40	0.000	4225	-0.003	0.997	0.000	infinity
UTI(+)	-0.927	0.396	1.085	-0.853	0.394	0.047	3.324
Ethnicity							
Luhya(Ref.)							
Kalenjin	1.136	3.115	0.33	3.449	0.001	1.633	5.942
Kikuyu	0.641	1.899	0.338	1.895	0.058	0.978	3.686
Luo	0.404	1.498	0.194	2.082	0.037	1.024	2.193
0thers	0.31	1.363	0.508	0.610	0.542	0.504	3.689
Education level							
Primary(Ref.)							
Secondary	0.020	1.02	0.198	0.101	0.92	0.692	1.505
Tertiary	0.037	1.038	2.847	0.131	0.895	0.594	1.814
Age (21-25)							
<=20	-0.401	0.67	0.216	-1.854	0.064	0.438	1.023
26-30	0.078	1.081	0.176	0.444	0.657	0.766	1.526
31-35	-0.023	0.978	0.22	-0.103	0.918	0.636	1.504
>35	-0.009	0.991	0.278	-0.033	0.974	0.574	1.709
Parity(1-2)							
0	0.044	1.045	0.165	0.268	0.788	0.7562	1.445
3-4	0.192	1.212	0.194	0.992	0.321	0.8291	1.771
5-6	-0.086	0.918	0.435	-0.197	0.844	0.3913	2.153
>6	-14.39	0.000	1633	-0.009	0.993	0.0000	Infinity
No.ofANC visits(4-5)							
1	2.929	1.87	0.332	8.821	0.000	9.7569	35.851
2-3	0.556	1.744	0.1944	2.86	0.004	1.1912	2.552
6-7	-0.13	0.878	0.7327	-0.178	0.859	0.2088	3.691
>7	-14.11	0.000	1826	-0.008	0.994	0.0000	Infinity
No. of Previous miscarriages(0)							
1-2	0.983	2.671	0.2075	4.735	0.000	1.7787	4.013
>2	0.986	2.681	0.7349	1.342	0.18	0.6349	11.32
No. of Previous stillbirths(0)							
1-2	1.397	4.044	0.228	6.129	0.000	2.5868	6.322
>2	1.588	4.894	0.7669	2.071	0.038	1.0886	22.001
Global test							
Wald test				293.3	0.000		
Likelihood ratio test				192.1	0.000		

Score (logrank)	502.5	0.000	

Table5The Cox PH Full Model with Continuous variables in Continuous form

Covariate	Coef.	Exp(coef.)	s.e(coef.)	Z-statistic	P-value	95% Conf.interval	
Marital status(single)	0.047	1.048	0.1975	0.236	0.81356	0.7114	1.5429
Residence(urban)	-0.446	0.64	0.141	-3.162	0.002	0.485	0.844
Employment(yes)	0.192	1.212	0.149	1.295	0.195	0.906	1.622
HIV(+)	0.120	1.128	0.424	0.284	0.776	0.491	2.589
Malaria(+)	3.501	33.15	0.711	4.924	0.000	8.226	133.6
STDS(+)	-12.48	0.000	1427	-0.009	0.993	0.000	infinity
UTI(+)	0.346	1.413	1.02	0.339	0.735	0.191	10.428
Ethnicity Luhya(Ref.)							
Kalenjin	1.158	3.182	0.326	3.553	0.000	1.681	6.026
Kikuyu	0.888	2.431	0.326	2.727	0.001	1.284	4.603
Luo	0.379	1.46	0.194	1.954	0.051	0.999	2.135
0thers	0.257	1.294	0.507	0.508	0.612	0.479	3.495
Education level Primary(Ref.)							
Secondary	0.099	1.103	0.198	0.497	0.619	0.749	1.626
Tertiary	0.282	1.325	0.278	1.011	0.312	0.768	2.286
No. of Previous miscarriages	0.407	1.502	0.102	4.009	0.000	1.231	1.833
No. of Previous stillbirths	0.623	1.864	0.114	5.433	0.000	1.489	2.333
Parity	0.040	1.041	0.058	0.701	0.484	0.93	1.166
Age	0.003	1.003	0.014	0.225	0.822	0.976	1.031
ANC Visits	-0.663	0.515	0.096	-6.94	0.000	0.427	0.621
Global test							
Wald test				222.8	0.000		
Likelihood ratio test				159.4	0.000		
Score (logrank) test				362.4	0.000		

Table6Cox Proportional Hazard model with Included Covariates in the Best Selection of Covariates

Covariate	Coef.	Exp(coef.)	s.e(coef.)	Z-statistic	P-value	95% conf. interval	
						Lower	Upper
Ethnicity Luhya(Ref.)							
Kalenjin	1.168	3.216	0.324	3.598	0.000	1.702	6.077
Kikuyu	0.905	2.472	0.325	2.788	0.005	1.308	4.672
luo	0.404	1.499	0.192	2.105	0.035	1.028	2.184
Residence (urban)	-0.482	0.617	0.139	-3.468	0.001	0.470	0.811
No. of Previous miscarriages	0.443	1.557	0.099	4.474	0.000	1.283	1.891
No. of Previous stillbirths	0.662	1.939	0.106	6.243	0.000	1.575	2.387
Malaria(+)	3.456	31.682	0.517	6.687	0.000	11.506	87.235
No.of ANC visits	-0.664	0.515	0.096	-6.936	0.000	0.427	0.621

3.6 Comparison of Proportional hazards models to parametric models

This section presents a comparative analysis of parametric proportional hazards (PH) models and the semi-parametric Cox regression model. Furthermore, the research also provided a comparative analysis of accelerated failure time models, with a specific focus on the AFT model having survival time distributions, exponential, Weibull, log-logistic, and log-normal distributions. The study also encompassed a comparative analysis of the best model obtained from the proportional hazards (PH) models and the accelerated failure time (AFT) models.

Table 8 presents the coefficient estimates for the Cox regression model and the parametric models of proportional hazards, namely the Exponential, Weibull, and Gompertz models. The evaluation focused on the consistency of the coefficient estimates for the proportional hazards (PH) models. Overall, there exists a discernible resemblance in the values of the coefficient estimates when comparing the Cox model and parametric proportional hazards (PH) models.

Table8 Comparison of Hazard Ratios and coefficient estimates for the PH models

Covariate	**Cox**		**Exponential**		**Weibull**		**Gompertz**	
	Coeff. (Beta)	HR	Coeff.	HR	Coeff.	HR	Coeff.	HR
Kalenjin	1.168	3.22	1.164	3.2	1.17	3.21	1.17	3.22
Kikuyu	0.905	2.47	0.896	2.45	0.906	2.47	0.906	2.48
luo	0.404	1.5	0.405	1.5	0.409	1.51	0.408	1.5
Residence (urban)	-0.482	0.62	-0.465	0.63	-0.478	0.62	-0.481	0.62
No. of Previous miscarriages	0.443	1.56	0.394	1.48	0.423	1.53	0.433	1.54
No. of Previous stillbirths	0.662	1.94	0.598	1.82	0.631	1.88	0.646	1.91
Malaria(+)	3.456	31.68	3.01	20.3	3.24	25.63	3.38	29.22
No.of ANC visits	-0.664	0.51	-0.632	0.53	-0.656	0.52	-0.661	0.52

Based on the data provided in Table 9, it is apparent that the Gompertz model demonstrates the lowest Akaike Information Criterion (AIC) value of 3411 when compared to the alternative proportional hazards (PH) models. The log-likelihood value generated from the analysis also serves as evidence supporting the superiority of the Gompertz model, as its log-likelihood value is comparatively smaller than that of the other proportional hazards (PH) models.

Table97 Comparison of the PH models using Log-likelihood and AIC

Model	**Number of parameters**	**Log-likelihood**	**AIC**
Cox	0	-2078.959	4173
Exponential	1	-1794.4	3606
Weibull	2	-1718.2	3456
Gompertz	2	-1695.602	3411

Table 10 presents a comprehensive comparison of coefficient estimates and standard errors pertaining to the accelerated failure time (AFT) models.

However, it is apparent that the standard errors for all the models demonstrate a similar degree of uniformity.

Table10Comparison of Standard Errors for AFT models

Covariate	**Exponential**		**Weibull**		**Lognormal**		**Loglogistic**	
	Coef.	Std. error	Coef.	Std. error	Coef.	Std. error	Coef.	Std. error
Intercept	4.513	0.274	3.782	0.117	4.344	0.159	3.753	0.12
Kalenjin	-1.164	0.325	-0.478	0.136	-0.682	0.197	-0.491	0.143
Kikuyu	-0.897	0.325	-0.371	0.136	-0.567	0.189	-0.388	0.141
Luo	-0.405	0.192	-0.168	0.079	-0.206	0.104	-0.173	0.081

Residence (urban)	0.465	0.14	0.196	0.058	0.290	0.048	0.201	0.06
No. of Previous miscarriages	-0.394	0.105	-0.173	0.043	-0.235	0.081	-0.180	0.053
No. of Previous stillbirths	-0.598	0.112	-0.259	0.047	-0.44	0.097	-0.298	0.063
Malaria(+)	-3.010	0.517	-1.329	0.224	-1.849	0.572	-1.276	0.3
No.of ANC visits	0.632	0.094	0.269	0.042	0.290	0.048	0.268	0.042

The AIC values were employed to evaluate the adequacy of the model fit, with lower values suggesting a higher level of fit. Based on the findings shown in Table11, it can be observed that the Weibull accelerated failure time (AFT) model demonstrates the highest suitability for data fitting, as evidenced by its possession of the lowest Akaike Information Criterion (AIC) value in comparison to the other AFT models

Table11Comparison of the AFT models using Log-likelihood and AIC

Model	**Number of parameters**	**Log-likelihood**	**AIC**
Exponential	1	-1794.4	3606.815
Weibull	2	-1718.2	3456.448
Log-logistic	2	-1720.1	3460.212
Lognormal	2	-1740.9	3501.851

The log-likelihood ratios and Akaike Information Criterion (AIC) were utilized to ascertain the most suitable model for the provided dataset. Table 12 displays the log-likelihood ratios and Akaike's Information Criterion (AIC) values pertaining to the most favorable Proportional Hazards (PH) and Accelerated Failure Time (AFT)

models.After conducting a comparison of the AIC values, it becomes apparent that the Gompertz model exhibits the most superior level of fit to the data in its entirety.

Table12Comparison of the Weibull AFT and Gompertz PH models

Model	Number of parameters	Log-likelihood	AIC
Gompertz PH	2	-1695.2	3411
Weibull	2	-1718.6	3456

Therefore, the best fitting model to this data is the Gompertz PH model whose estimates are presented in Table 13.

Table13The Full Model of the Gompertz PH)

Covariate	Coef.	Exp(coef.)	Stand. Error	p-value (LRT)	95% interval	Conf.
Kalenjin	1.17	3.221	0.325	0.0022	1.7044	6.0851
Kikuyu	0.906	2.475	0.325	0.0141	1.3098	4.6773
Luo	0.408	1.504	0.192	0.0428	1.0324	2.1915
Residence(urban)	-0.481	0.618	0.139	0.0008	0.4704	0.8117
No. of Previous miscarriages	0.433	1.543	0.1	0.0005	1.2683	1.8764
No. of Previous stillbirths	0.646	1.908	0.107	0.0000	1.546	2.355
Malaria(+)	3.38	29.37	0.517	0.0000	10.669	80.8537
No.of ANC visits	-0.661	0.516	0.096	0.0000	0.428	0.6225

3.8 Discussion of the Analysis Results Above

The current study utilized the Cox proportional hazard model to assess the factors that impact the duration of survival in cases of miscarriage among pregnant individuals. The present study has successfully identified a number of noteworthy factors that can serve as predictors of miscarriage. These factors encompass ethnicity, prior history of miscarriages, prior history of stillbirths, malaria status, area of residence, frequency of antenatal care visits. Previous scholarly investigations in the field of literature have examined the influence of race or ethnicity as a potential confounding variable in the assessment of miscarriage risk (Wen et al., 2001; Guendelman et al., 1990), as well as its examination as a major determinant (Mukherjee et al., 2013). This and other studies agree in acknowledging ethnicity as a separate predictor when assessing the likelihood of miscarriage. The collected data are consistent with other research findings that suggest maternal ethnicity plays a substantial role in determining the likelihood of miscarriage.

The prognostic value of the age of the pregnant lady in relation to the duration of survival in cases of miscarriage was shown to be insignificant. In contrast to the findings of the present study, prior research undertaken by Andersen et al. (2002) and Regan et al. (1989) has suggested that age does indeed exert a significant influence. Nevertheless, while examining survivorship functions across various age groups, it was observed that pregnant women within the age bracket of 21 to 25 years had a greater probability of surviving miscarriage compared to the remaining cohorts. The survival function for females under the age of 20 is shown to have the lowest value. This finding aligns with a multitude of other researches that have identified age as a

key predictor of miscarriage, explaining a considerable fraction of the observed occurrences. An illustrative investigation conducted by Andersen et al. (2000) employed registry linkage to examine the likelihood of miscarriage in pregnancies that were formally acknowledged, with a particular focus on age-related factors. The research revealed varying rates of miscarriage among different age groups of women. Specifically, the likelihood of experiencing a miscarriage was found to be 13% for women aged 12-19 years, 11% for those aged 20-24 years, 12% for those aged 25-29 years, 15% for those aged 30-34 years, 25% for those aged 35-39 years, 51% for those aged 40-44 years, and 93% for those aged 44 years and above. Based on a study conducted by de la Rochebrochard et al. (2002), there is a notable increase in the probability of encountering a miscarriage among couples where the female partner is aged 35 years or older, and the male partner is aged 40 years or older. The study to examine the risk factors linked to first trimester miscarriage among Asian women using a prospective methodology, conducted by Leong et al. (2013) revealed a significant association between maternal age and the incidence of miscarriage. The research findings suggested that there is an increased probability of miscarriage (hazard ratio HR=1.95) among women aged 34 years and above, as compared to women aged between 20 and 30 years.

The present investigation revealed that women with a prior history of pregnancy loss exhibited a heightened susceptibility to miscarriage in comparison to those without such a history. This suggests that the presence of previous stillbirths and miscarriages exerts an adverse influence on the subsequent fertility outcomes of pregnancies. This discovery aligns with prior research undertaken by Ogasawara et al. (2009), Regan et

al. (1989), Nilsson et al. (2014), and Dellicour et al. (2016). The conclusions of the investigation done by Kagereki (2019) were consistent with those previously reported.

Moreover, this study has provided evidence that the number of antenatal care visits plays a crucial role in determining the occurrence of miscarriage. This finding is consistent with the studies conducted by Ortiz et al. (2021).

The findings derived from the Cox regression model also demonstrated a statistically significant correlation between place of residency and the occurrence of spontaneous abortion. The results of the study indicate that women living in rural areas experienced a shorter estimated duration until pregnancy termination in comparison to their counterparts live in urban settings. These results are consistent with the conclusions drawn from earlier investigations undertaken by Carlson and Mourgova (2003), Ellett Buxton and Luesley (1992), and Shapiro and Bross (1980 as quoted in Wood, 1994).

The findings of this study have showed that the Cox proportional hazards (PH) model exhibited a higher information criterion in comparison to the parametric survival models. This observation indicates that the parametric models being examined had superior goodness of fit in comparison to the Cox proportional hazards model. Additionally, the research findings indicated that parametric proportional hazards (PH) models displayed a higher level of competitiveness in comparison to parametric accelerated failure time (AFT) models.

4. CONCLUSION

In this study, different survival analysis and modelling techniques; product limit estimator, log-rank tests, the proportional hazards models and acceleration failure time models were presented.The objective of this study was to analyze the time patterns of survival of miscarriage and ascertain risk factors that have influence on these patterns within the context of a healthcare facility in Kenya.The study documented a decreased overall prevalence of miscarriage in contrast to other research conducted in Kenya, which revealed a proportion of 12.2% (Dellicour, Desai Mason, et al., 2013; Dellicour, Desai, Mason, et al, 2007; Stanton et al., 2006).From the results, some factors were identified by formal tests and survival plots that response variable changed with the changes in the levels of these variables. The studies of (Carlson and Mourgora, 2003; Ellett, Buxton, and Luesley, 1992; Shapiro and Bros, 1980) had similar results.Thus the study found some evidence for association between some explanatory variables and the survival time of miscarriages.The factors ethnicity, place of residence, malaria status, number of previous miscarriages, number of previous stillbirths and number of ANC visits were identified as the risk factors associated with miscarriages when cox model, parametric proportional hazards model and accelerated failure time models were implemented. From the findings of this study it can be concluded that the Gompertz proportional hazards (PH) regression model best fits the dataset.

ETHICAL APPROVAL

The research acquired ethical clearance from the Institutional Ethics and Research Committee (MUCHS-MTRH IERC) of Moi University College of Health Sciences and Moi Teaching and Referral Hospital. The office of the chief executive for health County hospitals granted permission or approval to utilize medical data of pregnant

women from specific facilities. Prior to conducting the review, authorization or permission was acquired from the appropriate authorities at the hospitals and Antenatal clinic. The utilization of encrypted hard disks was employed for the purpose of gathering datasets from various health facilities. The names of mothers were obtained from databases consisting of computer-based or paper-based records. These names were de-identified and no data was shared, ensuring the confidentiality and privacy of the information were upheld.

REFERENCES

Aalen, O. (1978b). Nonparametric inference for a family of counting processes. *Annals of Statistics, 6*, 701–726.

Altman, D. (1991). *Practical Statistics for Medical Research.* London: Chapman & Hall/CRC.

Altshuler, B. (1970). Theory for the measurement of competing risks in animal experiments. *Mathematical Biosciences, 6*, 1-11.

Andersen et al. (2002). Fever in pregnancy and risk of fetal death: a cohort study. *Lancet, 360*(9345), 1552-6.

Armitage, P., Berry, G. and Matthews, J.N.S. (2002). *Statistical Methods in Medical Research* (4th ed. ed.). Oxford: Blackwells Science Ltd.

Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics*, 89-99.

Breslow, N.E. and Crowley, J. (1974). A large Sample Study of the Life Life Table and Product Limit Estimates under random Censorship. *Annals os Statistics, 2*, 437-453.

Cai, Y. and Feng, WQ. (2005). Famine, social disruption, and involuntary fetal loss: Evidence from Chinese survey data. *Demography, 42*, 301-322.

Carlson, E. and Mourgova, M. (2003). Demographic Consequences of Social Inequality in Pregnancy Outcomes.". *Genus LIX, 2*, 11-28.

Chatenoud, L., Parrazin, F., Di cintio, E., Zanconato, G., Benzi, G., Bortolus, R. and La Vecchia, C. (1998). Paternal and maternal smoking habits before conception and during the first trimester: relation to spontaneous abortion. *Am Epidemiol, 8*(8), 520-6.

Cnattingius et al. (2000). Caffeine intake and the risk of first trimester spontaneous abortion. *N Engl J Med, 343*(25), 1839-45.

Collet, D. (2003). *Modelling Binary data* (2nd ed. ed.). Boca Raton, Florida: Chapman & Hall/CRC.

Coste et al. (1991). Risk factors for spontaneous abortion: a case control study in France. *Hum Reprod, 6*(9), 1332-7.

Cox, D. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, 34*, 187-220.

Cox, D.R. and Oakes, D. (1984). *Analysis of Survival Data.* London: Chapman & Hall/CRC.

De la Rochebrochard, E. and Thonneau, P. (2002). Paternal age and maternal age are risk factors for miscarriage; results of multicenter European study. *Hum Reprod , 17*(6), 1649-56.

Dellicour S, Desai M, Mason L, et al. (2007). Epidemiology and burden o fmalaria in pregnancy. Lancet Infect Dis 2007;7:93–104. *Lancet Infect Dis 2007;7:93–104, 7*, 93-104.

Dellicour S, Desai M, Mason L, et al. (2013). Exploring risk perception and attitudes to miscarriage and congenital anomaly in rural Western Kenya. *PLoS ONE, 8*.

Dellicour, S. A. (2016). Weekly miscarriage rates a community-based prospective cohort study in rural western Kenya.. . *BMJ 6:e011088*.

Ellett, K., Buxton, E.J., and Luesley, D.M. (1992). The Effect of Ethnic Origin on the Rate of Spontaneous Late Mid-Trimester Abortion." 2(1). *Ethnicity and Disease, 2*(1), 84-86.

Ganatra et al. (2015). Expanding access to medical abortion challenges and opportunities. 22(44 Suppl 1):. *Reprod Health Matters, 22*(44 Suppl1), 1-3.

Garcia-Enguidanos et al. (2002). Risk Factors in Miscarriage: a Review.102(2):111-119. *European Journal of Obstetrics, Gynecology and Reproductive Biology, 102*(2), 111-119.

Grambsch, P.M. and Therneau, T.M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika, 81*, 515-526.

Guendelmn et al. (1990). Generational differences in perinatal health among the Mexican population: fingings from HHANES 1982-84. *AM J Public Health, 80*(suppl), 61-65.

Kagereki, E.K., Wanjoya, A. and Mageto, T. (2019). Modelling Cases of Spontaneous Abortion Using Logistic Regression. *International Journal of Data Science and Analysis, 5*(6), 143-147.

Kalbfleisch, J.D. and Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data* (2nd ed. ed.). New York: Wiley.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association, 53*, 457-481.

Lee, E.T. and Wang, J.W. (2013). *Statistical Methods for Survival Data Analysis* (4th ed. ed.). New York: Wiley.

Leong, J.K., Ghim, H., Rahul, M., John, C.A., Suan, T.B., Thiam Chye, T. and Truls Østbye. (2013). A prospective study of risk factors for first trimester miscarriage in Asian women with threatened miscarriage. *Singapore Med J, 54*(8), 425-431.

Lindbohm et al. (2002). effects ofeExposure to environmental tobacco smoke on reproductive health. *Scand J Work Eviron Health, 28*(2), 84-96.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports, 50*, 163-170.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719–748.

Meier, P. (1975). *Estimation of a distribution function from incomplete observations. In: Perspectives in Probability and Statistics.* (J. Gani, Ed.) london: Academic Press.

Mukherjee et al. (2013). Risk of miscarriage among Black and White women in a US: Prospective cohort study. *AM J Epidemiol, 177*(11), 1271-1278.

Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics,, 14*, 945–965.

Nilsson et al. (2014). Risk factors of miscarriage from prevention perspective: a nation wide follow up study. *B JOG, 121*(11), 1375-84.

Ogasawara et al. (2000). Embryonic karyotype of abortuses in relation to the number of previous miscarriages. *Fertil Steril, 73*(2), 300-4.

Ortiz et al. (2021). Metabolomics in endometriosis: challenges and perspective for future studies. *Reproduction Fertil, 2*(2), R35-R50.

Parazzin et al. (1997). Determinants of risk factors of spontaneous abortions in the first trimester of pregnancy. *Epidemiology, 8*(6), 681-3.

Peck et al. (2010). A review of the epidemiologi evidence concerning the reproductive health effects of caffeine consumption: a 200-2009 update. *Food Chem Toxicol, 48*, 2449-2576.

Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant procedures. *Journal of the Royal Statistical Society,A, 135*, 185–207.

Pollard, A.H., Yusuf, F. and Pollard, G.N. (1990). *Demographic Techniques* (3rd ed. ed.). Sydney: Pergamon Press.

Rasch, V. (2003). Cigarette, alcohol and caffeine consumption: risk factors for spontaneous abortion. *Acta Obstet Gynecol Scand, 82*(2), 182-8.

Reagan, L. (1991). Recurrent Miscarriage. *Br Med J, 302*, 543-544.

Regan et al. (1989). Influence of past reproductive performance on risk of spontaneous abortion. *Bmj, 299*(6698), 541-5.

Shapiro, S. and Bross, D. (1980). Risk Factors for Fetal Death in Studies of Vital Statistics Data: Inference and Limitations. (I. P. Hook, Ed.) *Human Embryonic and Fetal Death*, 89-106.

Simpson and Carson. (1993). Biological Causes of Fetal Loss. (R. Gray, Ed.) *Biomedical and Demographic Determinants of Reproduction,*, 287-315.

Simpson, J.L.and Mills, J.L. (1986). Methodologic Problems in Determining Fetal Loss Rates. In G. S. B. Brambati (Ed.), *Chorionic Villus Sampling: Fetal Diagnosis of Genetic Diseases in the First Trimester* (p. 227). New York: M. Dekker.

Stanton et al. (2006). Still Birth rates: delivering estimates in 190 countries. *Lancet, 367*, 1487-1494.

The R Foundation for Statistical computing. (2011). R version 2.13.1. (The R Foundation for Statistical computing) Retrieved from The R Foundation for Statistical Computing website: http://www.r-pro-ject.org/foundation

Therneau, T.M. and Grambsch, P.M. (2000). *Modelling Survival Data: Extending the Cox Model.* New York: Springer.

WHO. (1977). Recommended Definitions, Terminology and Format for Statistical Tables Related to the Perinatal Period and Use of a New Certificate for Cause of Perinatal Deaths. Modifications Recommended by FIGO as Amended October 14, 1976. *Acta Obstet Gynecol Scand, 56*(3), 247-253.

Wood, J. (1994). *Dynamics of human Reproduction: Biology, Biometry, Demography.* New York: Aldine de Gruyter.

Woodward, M. (2014). *Epidemiology: Study Design and Data Analysis* (3rd ed. ed.). Boca Raton, Florida: Chapman & Hall/CRC.

Zinaman et al. (1996). "Estimates of Human Fertility and Pregnancy Loss. *Fertility and Sterility, 65*(3), 503-509.