MODELING THE RISK FACTORS OF MISCARRIAGE USING ADVANCED SURVIVAL ANALYSIS TECHNIQUES: CURE MODEL

ABSTRACT

Background: A miscarriage is an abrupt and distressing occurrence that might have adverse consequences for the individual experiencing it. However the risk factors that lead normal pregnancy to a miscarriage are fairly well established. That is, its prevalence and causes is still a subject for continuing investigation. Cox's model and the accelerated failure time model are commonly utilized statistical models to assess the factors associated with spontaneous abortion. These models assume that all patients under surveillance would inevitably experience the event of interest, assuming they are monitored for a sufficiently long period of time. However, it is imperative to recognize that certain subjects may not undergo a comprehensive manifestation of the phenomenon, irrespective of the length of their investigation.

Objectives: To estimate cure fraction of miscarriage and evaluate the effects of covariates on the cure rate, using cure models.

Method: This study used secondary data collected from 6077 pregnant women who were enrolled for antenatal care in Kakamega County General Teaching and Referral Hospital (KCGTRH) in Kakamega county, western Kenya. The study period was from 1 January, 2019 up to 31 October, 2020 and miscarriages were regarded as failures. Cox model and a proportional hazards mixture cure model (PHMC) were utilized to analyze the dataset.

Results: The cure model showed that place of residency (P = .003), ethnicity: kalenjin (P = .001), kikuyu (P = .014) and luo (P = .040), number of prior miscarriages (P = .000), number of previous stillbirths (P = .000), and number of ANC visits (P = .000) statistically affect cure fraction. However these factors did not affect survival time, apart from the number of ANC visits (P = .001). The number of previous miscarriages, stillbirths, ANC visits, site of residency, and ethnic groups (Kalenjin, Kikuyu, and Luo) had cure probabilities of 54.9%, 47.1%, 76.1%, 87.2%, 39.3%, 44.9%, and 58.3%, respectively. All analysis was carried out using R software. The level of significance was 5%.

Conclusion: The cure model in this study showed that these factors had effect on long-term survivors. On short-term trends there were little changes. Using the cure model to investigate miscarriage data provided more insights than Cox model analysis.

1. INTRODUCTION

A large number of pregnancies progress without major complications, however a significant percentage of women will still undergo adverse pregnancy outcomes. One of adverse pregnancy outcome is miscarriage. Miscarriage or spontaneous abortion is defined as involuntary pregnancy loss at a time before a fetus would be viable outside of mother's womb (Cai & Feng, 2005) or pregnancy terminated at less than 28 weeks of gestation due to non- human factors (Kline, 1984). Global prevalence is about 10% to 15% of all clinically recognized pregnancies (Simpson&Carson, 1993; Simpson& Mills, 1986; Zinaman et al., 1996). Empirical estimates of prevalence vary from as low as 2%-3% to as high as 30% (Cai&Feng, 2005). In Kenya prevalence is about 12.2% (Dellicour et al., 2007, 2013; Stanton, 2006).

The cause of miscarriage remains uncertain. Nevertheless, some studies have identified several factors that contribute to and elevate the likelihood of spontaneous abortion. Common risk factors cited are: woman's age, gravidity, length of pregnancy interval, pregnancy history(number of previous life births and fetal losses), chromosomal abnormalities and mutant genes, drug use, caffeine intake, smoking and drinking, obesity, place of residence and race. However the risk factors that lead normal pregnancy to a miscarriage are fairly well established. Moreover, determining the occurrence of miscarriage poses a significant challenge, particularly in low-income countries where there is less data is available on its prevalence(Dellicour et al 2016).

Cox's model and the accelerated failure time model are commonly utilized statistical models to assess the factors associated with spontaneous abortion. These models assume that all patients under study would experience the event of interest (miscarriage) provided they are followed a long period of time. In reality this is not the case, the event (miscarriage) may not occur totally in some subjects, no matter how long they are followed. A cure model is a class of models for censored survival data for subjects when some of them will not develop the event of interest however long they are followed (Peng and Taylor, 2014).

The introduction of cure models in the literature can be attributed to the influential works of Boag (1949) and Berkson and Gage (1952). In the domain of literature, it is

possible to discern two separate classifications of cure models. The initial introduction of the combination cure model can be attributed to Farewell (1977), with significant advancements made (1982). The second group comprises non-mixture cure models, as proposed by Yakovlev et al. (1996). For a thorough comprehension of several categories of these models, it is advisable to consult the publications of Maller and Zhou (2001) or Ibrahim, Chen, and Sinha (2001). A number of scholars have undertaken a comprehensive examination of survival analysis models, as demonstrated by the scholarly contributions of Peng and Taylor (2014) and Amico and Van Keilegon (2018). The models that examine cure rates in the field of economics are usually known as split population models, as explored in the seminal work of Schmidt and Whitte (1989). However, it is worth noting that reliability engineers frequently categorize these models as limited-failure population life models, as previously explored by Meeker (1997). Traditional survival models make the assumption that the probability of miscarriage decreases to zero as time progresses indefinitely. Consequently, these models are not directly applicable in situations where there exists a subset of pregnant women who will not experience a miscarriage. Cure rate models are an additional modeling tool that is gaining prominence in the field of survival analysis. In instances of miscarriages when certain pregnant individuals may not experience a miscarriage within the duration of the study and exhibit no discernible indications of miscarriage, employing cured models would be suitable for characterizing and examining the survivability of miscarriage in expectant mothers. Cure models have gained significant attention in the field of statistics and have been extensively discussed in academic literature. However, their use in certain domains of biostatistics has been relatively limited. Therefore, it is imperative to incorporate cure models alongside conventional survival analysis techniques when examining the survival durations of miscarriages. This study implemented cure models to develop prognostic risk factor model for miscarriages and use the model to assess key demographic, socioeconomic, medical and lifestyles factors that predispose pregnant women to miscarriage.

2. MATERIAL AND METHODS

2.1 Study Design

A retrospective cohort 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, hence secondary data from records of pregnant women was collected. 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.

2.2 Study Area

The research was carried out inside Kakamega County, western Kenya and limited to pregnant women enrolled for antenatal care services in Kakamega County General and Teaching hospital (KCGTRH). The county is located on the western part of Kenya. According to (Kenya Bureau of Statistics, 2019), the county under discussion ranks as the fourth most populated in Kenya, boasting a population above 1.8 million individuals. The KCGTRH has been chosen to host the study as it is a level 5 referral hospital receiving patients from the entire Kakamega county, with good laboratory and clinical infrastructure and therefore suitable for recruiting pregnant mothers all over the county. The study population included all pregnant women living in Kakamega county, western Kenya during the period from 1, January 2019 up to 31 October 2020, where cases of spontaneous abortion were identified.

2.3 Sampling Procedure

All pregnant women enrolled in antenatal care from January 1, 2019 up to 31 October 2020 in KCGTRH in Kakamega County, western Kenya and met the inclusion criteria were included in the sample. Decision was made to consider all of them as sample to get the accurate and precise findings which will be inferred to county population as the KCGTRH serve the whole county. All pregnancy records with the results of the pregnancy outcome within the stated period was selected and reviewed for this study.

Inclusion criteria

- Subjects under study must be a pregnant woman enrolled in antenatal care in KCGTRH

- Must have a confirmed pregnancy of a gestational age of 6 to 12 weeks or earlier.
- Subjects whose documentation of medical, gynaecological history and demographical data are available
- The result of the pregnancy outcome of the woman is documented.

Exclusion criteria

- Miscarriages from unrecognized pregnancy.
- Pregnant women not under antenatal care. This because of difficultyobtaining full information about them, such as their total number.
- Pregnancy of gestational age beyond 12 weeks
- Threatened miscarriage

2.4 Study Variables

The factors under consideration in this investigation were obtained from the existing standard medical records. The analysis utilized the covariates that were recorded during the initial antenatal care visit due to a scarcity of available secondary data.

The factors that are being measured or observed in a study are referred to as outcome variables. The event being examined was a spontaneous termination of pregnancy. The dependent variable will be operationalized as the time interval between the last menstrual period and either the incidence of a miscarriage or the termination of the study period, which aligns with the ending of the pregnancy. Participants who were unable to experience the event being studied before the end of the designated research period were categorized as censored, along with those who were no longer available for further observation. Miscarriage, commonly referred to as spontaneous abortion, denotes the involuntary cessation of a pregnancy before to attaining a gestational age of 28 weeks.

Independent variables included;

Socio-demographic characteristics: ethnicity, maternal age, marital status, , length of pregnancy interval, history of previous pregnancy (number of previous live births and foetal losses),parity, gravidity, induced abortion and use of certain contraceptives and number of antenatal care visits.

Socio-economic characteristics: Education level, employment status and place of residence (urban/rural).

Comment [DYA1]: Insert "in"

Comment [DYA2]: delete

Comment [DYA3]: delete

Clinical characteristics: signs and symptoms of pregnancy loss (blood discharge), chronic diseases (HIV, TB, diabetic, hypertension), infections (malaria, STDs, UTIs), haemoglobin status (HB) and folic intake.

Lifestyle characteristics: caffeine intake, exercises, stress, smoking, exposure to cigarette smoke and alcohol consumption, obesity and drug use. Pregnant mothers that were eliminated from the sample were those whose folders lacked essential information. The researcher utilized a checklist to gather data from these folders, focusing on various demographic and medical factors. These factors included ethnicity, gravidity, maternal age, parity, marital status, history of miscarriages and stillbirths, educational level, profession, HIV status, frequency of antenatal care visits, presence of malaria infection, urinary tract infection, sexually transmitted diseases, gestational age, place of residence, survival time, and survival status.

Miscarriage was the study endpoints with mothers follow up to go on as up to date of transferring out from health facility, death or until end of follow up period (end of pregnancy).

2.5. Mixture cure model

One can identify whether a data set has a fraction of cured, by looking at Kaplan Meier survival curve. If the survival curve has along and stable plateau at the end of the study, then the population has a proportion of cured subjects. A comprehensive evaluation of the existence of the cure models can be found in Maller and Zhou (1994). An additional avenue for expanding time-to-event models entails the integration of a cured percentage π, which denotes the probability of an individual achieving a state of cure. The survival function$(S(t))$, hazard function$(h(t))$, cumulative hazard function $(H(t))$ and probability density function $(f(t))$ within the cure model are deemed inappropriate in their inherent characteristics.

$$\lim_{t\to\infty} s(t) > 0 \qquad (1)$$

$$\lim_{t\to\infty} H(t) = \lim_{t\to\infty} \int_0^t h(t)\delta t < \infty \qquad (2)$$

And

$$\int_0^\infty f(t)\delta t < 1 \qquad (3)$$

Two classes of cure models have been proposed; mixture and non-mixture cure models. This study used mixture cure model. It models survival function as a mixture of two types of populations, those that are cured and those are not cured. The mixture cure model was first proposed by Farewell (Farewell, 1977 and 1982) and has the form:

Comment [DYA4]: You need to be specific. For instance, time from dash to dash

Comment [DYA5]: recast

Comment [DYA6]: a long

Comment [DYA7]: delete and put ", since"

Comment [DYA8]:

Comment [DYA9]: align your equation numbers

Comment [DYA10]: a not A

Comment [DYA11]:

Comment [DYA12]: contradiction with what you earlier mentioned

Probability (alive at time t) = probability (cured) + probability (not cured) × probability (alive at time t if not cured)

$$S(t) = \pi + (1 - \pi)S_n(t) \quad (4)$$

where $S_n(t)$ is the survivor function of the non-cured subjects, π the probability of being cured and $S(t)$ is the overall survival function consisting of cured and non-cured individuals. The probability of an individual achieving a cure is modeled using a logistic regression approach. The second aspect of the model relates to the analysis of patient survival in cases where a cure has not been attained. The utilization of Weibull and Cox models is a common practice for the modeling of this particular component.

If the proportional hazards model is assumed for the survival time of the non-cured individuals, the hazard at time t in such an individual is

$$h_{ni}(t) = h_{n0}(t)\exp(\boldsymbol{\beta}'\boldsymbol{x_i}) \quad (5)$$

Where $\boldsymbol{\beta}'\boldsymbol{x_i} = \beta_{1x_{1i}} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}$ is a linear combination of the values of p explanatory variables $X_1, X_2 \ldots X_p$ measured on this individual and $h_{n0}(t)$ is the baseline hazard function of the non-cured subjects. The survival function for the susceptible is then

$$S_{ni}(t) = [S_{n0}(t)]^{exp(\boldsymbol{\beta}'\boldsymbol{x_i})} \quad (6)$$

Where $S_{n0}(t) = exp\left\{-\int_0^t h_{n0}(u)\delta u\right\}$ is the baseline survivor function.

The mixture cure model's parameters can be estimated using EM algorithm to obtain maximum likelihood estimates of the parameters in the mixture cure models (Dempster et al., 1977). Let the data consist of n survival times $t_1, t_2, \ldots, t_n$ and δ_i is the event indicator for theithindividual such that $\delta_1 = 1$ if the ith individual dies and zero otherwise, then the likelihood function is

$$L(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \prod_{i=1}^{n}\{h_i(t_i)\}^{\delta_i} S_i(t_i) \quad (7)$$

Where $h_i(t_i)$ and $S_i(t_i)$ are replaced by $h_{ni}(t_i)$ and $S_{ni}(t_i)$ in the case of cure models.

The models with different explanatory variables in either cured or for non-cured individuals can be compared using the values of the statistic $-2logL(\hat{\beta}, \hat{\gamma})$, Akaike's information criteria (AIC) and the corrected Akaike's information criteria in normal way (AIC_c).The AIC_c is defined by $AICc = -2logL(\hat{\beta}, \hat{\gamma}) + 2q\frac{n}{n-q-1}$ where q is the number of parameters in the model and n is the sample size. The lower value of the selection criterion indicates the better fit.

Comment [DYA13]:
Comment [DYA14]: delete
Comment [DYA15]:
Comment [DYA16]: delete and write "$[S_{n0}(t)]^{exp(\boldsymbol{\beta}'\boldsymbol{x_i})}$"
Comment [DYA17]: reduce the space
Comment [DYA18]: delete and put ith
Comment [DYA19]: ith not ith
Comment [DYA20]: is:
Comment [DYA21]: where not Where
Comment [DYA22]:
Comment [DYA23]: delete
Comment [DYA24]: you need to also provide the formula for AIC

3. RESULTS AND DISCUSSION

The data used was a secondary data with a total of six thousand and seventy seven (6077) records of pregnant women who started antenatal care at Kakamega County Teaching and Referral Hospital (KCGTRH) in Kakamega County, were considered for the study. The data spanned a period ranging from 1st January 2019 to 31st October 2020. The response variable of interest for this study was the time until miscarriage in weeks. Covariates were ethnicity, gravidity, maternal age, parity, marital status, number of previous miscarriages, number of previous still birth, educational level, profession, HIV status, antenatal care visits(ANC), malaria infection, urinary tract infection(UTI), sexually transmitted diseases(STDs) and place of residence.

Comment [DYA25]: delete and put A

Comment [DYA26]: delete

Comment [DYA27]: delete and put "in thi"

Comment [DYA28]: the time of what? Specify the starting time please

Comment [DYA29]:

Comment [DYA30]: reduced the space

3.1. Descriptive results and analysis

Table 1 indicates number and percentage of those who experienced event (miscarriage) and who did not (censored). A total of 248 female participants, constituting 4.1% of the sample, reported experiencing miscarriages. Table 1 displays the descriptive statistics for the variables of interest.

Comment [DYA31]: those who give birth, are not censored and I guess they are part of those who did not experienced miscarriage

From Table 1, the distribution of the number of women who experienced miscarriages, varied across different levels of some categorized variables. This was indicated by the computed Pearson chi-square statistic with P-value below 5%.

Table 1 Descriptive Statistics

Variable	(N=248)	(N=5829)	Chi Square Test	
	Miscarriage n(%)	Not miscarriage n(%)	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)		

The Figure 1 shows the overall Kaplan-Meier survival plot. The graph demonstrates a plateau beyond the threshold of 0.95. It indicates that the length of the follow-up period was sufficient and there exists a noticeable cured fraction.

Comment [DYA32]:

Comment [DYA33]: please check the totals for each covariate because some are not correct, for instance, the totals for educational level and marital status are not correct

Comment [DYA34]: delete

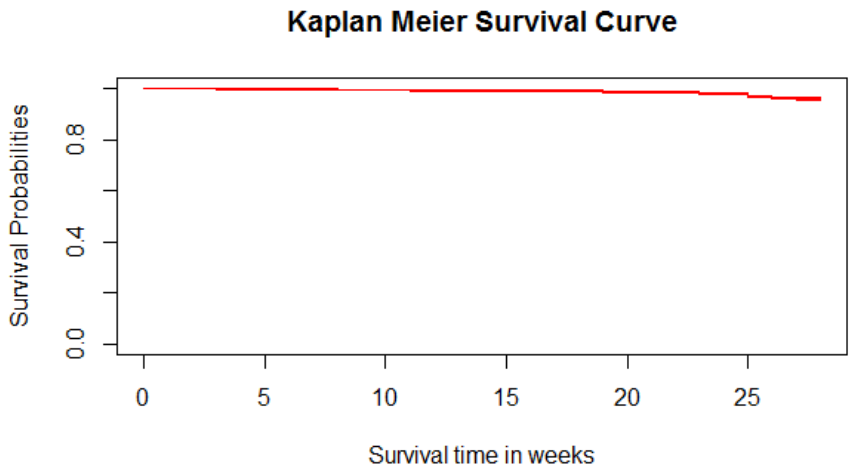

Figure 1 The overall Kaplan Meier survival plot of time to miscarriage in weeks

3.2. Model development

The relationship between survival time and covariates was modeled usingproportional hazards mixture cure model. Using the smcure R package a bivariate model, for single factor was estimated. The parameter estimates and standard errors obtained are displayed in Table 2. The bivariate investigation indicates that seven parameters displayed statistical significance at a significance level of 10 percent in connection to the cure fraction. The following variables, namely HIV status, gravidity, parity, age, marital status, employment status, and secondary level of education, were not statistical significance at a significance level of 10 percent in relation to the cure fraction. Seven covariates were statistical significance at a 10 percent level of significance and are candidates for the initial multivariable cure rate model utilizing the intentional selection method of covariates at a 10 percent significance level.

Comment [DYA35]: it should be 10% not 10 percent

Comment [DYA36]: you earlier mentioned that the analysis will be at 5% why now at 10%?

Table 2. Parameter estimates, standard errors, Wald statistic and associated P-values for the bivariate models.

Comment [DYA37]: p not P

	Cure Parameters				Survival Parameters			
Coefficient	Estimate	Standard Error	Z-value	P-value	Estimate	Standard Error	Z-value	P-value
Intercept	-3.10	0.16	-19.3	0.000				
Marital status	-0.06	0.19	-0.34	0.736	-0.014	0.156	-0.088	0.93
Intercept	-2.71	0.1	-27.7	0.000				
Residence	-0.61	0.14	-4.24	0.000	-0.105	0.128	-0.821	0.412
Intercept	-3.28	0.39	-84.57	0.000				
Employment	0.34	0.11	3.56	0.000	0.132	0.116	1.14	0.256
Intercept	-3.18	0.01	-265.0	0.000				
Kalenjin	1.12	0.36	3.15	0.002	-0.261	0.266	-0.980	0.327
Intercept	-3.18	0.01	-239.0	0.000				
Kikuyu	0.94	0.38	2.45	0.014	0.470	0.309	1.52	0.127
Intercept	-3.2	0.03	-122.7	0.000				
Luo	0.38	0.21	1.88	0.061	0.177	0.172	1.025	0.305
Intercept	-3.03	0.11	-28.4	0.000				
Secondary	-0.17	0.14	-1.18	0.236	-0.24	0.158	-1.52	0.128
Intercept	-3.20	0.02	-139.0	0.000				
Tertiary	0.49	0.20	2.39	0.017	0.188	0.188	1.0	0.318
Intercept	-3.21	0.02	-200.7	0.000				
No. of Previous miscarriages	0.6	0.13	4.61	0.000	0.550	0.103	5.36	0.000

Intercept	-3.22	0.02	-188.6	0.000				
No. of Previous stillbirths	0.97	0.2	4.88	0.000	0.557	0.115	4.85	0.000
Intercept	-3.23	0.06	-55.78	0.000				
Parity	0.06	0.04	1.33	0.183	0.021	0.038	0.557	0.578
Intercept	-3.63	0.31	-11.72	0.000				
Age	0.018	0.012	1.56	0.119	0.013	0.008	1.605	0.108
Intercept	-1.10	0.324	-3.40	0.000				
ANC Visits	-0.698	0.117	-5.96	0.000	-0.283	0.071	-3.997	0.000
Intercept	-3.29	0.110	-29.86	0.000				
Gravidity	0.059	0.048	1.24	0.217	0.021	0.035	0.596	0.551
Intercept	-3.17	0.010	-309.4	0.000				
HIV	0.303	0.472	0.644	0.52	-0.597	0.195	-3.07	0.002

These seven factors selected for inclusion in the initial multivariable model were : place of residence, employment status, ethnicity, level of education, count of prior miscarriages, count of prior stillbirths, and count of antenatal care visits. The multivariable model obtained through a planned selection of factors is presented in Table 3.

Table 31Parameter estimates, standard errors, Wald statistic and associated P-values for the initial multivariable model

	Cure Parameters				**Survival Parameters**			
Coefficient	Estimat	Standar	Z-	P-value	Estimat	Standar	Z-	P-

	e	d Error	valu e		e	d Error	valu e	value
Intercept	-1.250	0.321	- 3.90	0.00009 6				
Number of ANC visits	-0.628	0.11	- 5.73	0.00000 0	-0.24	0.087	- 2.76 3	0.005 7
Place of residence	-0.435	0.142	- 3.06	0.00225	-0.046	0.119	- 0.38 6	0.699 8
Kalenjin	1.20	0.348	3.46	0.00054	-0.135	0.296	- 0.45 8	0.647 0
Kikuyu	0.888	0.379	2.34	0.01913	0.542	0.472	1.14 9	0.250 4
Luo	0.380	0.190	1.99 8	0.04574	0.127	0.194	0.66 8	0.504 0
Number of previous miscarriage	0.479	0.135	3.54 8	0.00389	0.29	0.154	1.88 6	0.059 2
Number of previous stillbirths	0.788	0.1888	4.17 5	0.00003	0.316	0.180	1.75 0	0.080 0
College level	0.223	0.2518	0.88 5	0.376	0.082	0.235	0.34 9	0.726 9
Employment status	0.175	0.1480	1.18 3	0.2369	-0.046	0.133	- 0.34 4	0.731 0

The multivariable model obtained in Table 3 indicates that the covariates college level and job status do not exhibit statistical significance at a 5 percent level of significance with respect to the cure proportion. These covariates were omitted

from the final multivariable model. The final multivariable proportional hazards mixture cure model is presented in Table 4.

Thus the fitted proportional hazards mixture cure (PHMC) model is therefore

Incidence part;

$$\ln\left(\frac{\pi(z_i)}{1-\pi(z_i)}\right) = -1.141 - 0.634\text{ANC visits} - 0.467Residence + 1.22Kalenjin + 0.913Kikuyu + 0.397Luo + 0.503Previous\ miscarriage + 0.804Previous\ stillbirths$$

Latency part;

$$h(t, \boldsymbol{X}) = h_0(t)\exp(-0.244\text{ANC visits} - 0.037Residence - 0.145Kalenjin + 0.54Kikuyu + 0.109Luo + 0.28Previous\ miscarriage + 0.315Previous\ stillbirths) \quad (8)$$

The cure rate is determined using

$$cure\ rate = 1 - \pi(z_i) \quad (9)$$

Comment [DYA38]: please use the proper notation

Table 4Parameter estimates, standard errors, Wald statistic associated P-values for the final multivariable model

Comment [DYA39]: your tables shoud not run over two pages

	Cure Parameters				**Survival Parameters**			
Coefficient	Estimate, β_i	Standard Error	Z-value	P-value	Estimate, β_i	Standard Error	Z-value	P-value
Intercept	-1.141	0.309	-3.69	0.000				
Number of ANC visits	-0.634	0.114	-5.57	0.000	-0.244	0.074	-3.3	0.001
Place of residence	-0.467	0.16	-2.93	0.003	-0.037	0.116	-0.32	0.748
Kalenjin	1.22	0.363	3.37	0.001	-0.145	0.255	-0.57	0.57
Kikuyu	0.913	0.372	2.46	0.014	0.54	0.391	1.38	0.168
Luo	0.397	0.193	2.06	0.04	0.109	0.164	0.664	0.506

Number of previous miscarriage	0.503	0.112	4.47	0.000	0.28	0.154	1.81	0.07
Number of previous stillbirths	0.804	0.181	4.45	0.000	0.315	0.165	1.92	0.055

As seen in Table 4, two separate sets of estimates are obtained, each corresponding to a certain component of the model. Upon conducting an initial analysis of the cure component, it becomes evident that the p-values linked to all parameters are below the established threshold of 0.05. This indicates a statistically significant influence of these variables on the proportion of individuals who achieve a cure. To interpret these effects, we employ the same methodology as that employed for the classical logistic regression model.

The covariate of ethnicity demonstrates a statistically significant influence on the odds of being uncured, suggesting that the likelihood of being uncured varies among women of different ethnic backgrounds. The covariate of ethnicity was classified as Luhya in the reference group. The probability of encountering a miscarriage for a Kalenjin woman, relative to a Luhya woman, is 3.3963. In a comparative analysis, it is seen that the likelihood of remaining uncured for a Kikuyu woman, in relation to a Luhya woman, is calculated to be 2.4933. Similarly, the probability of an uncured condition for a Luo woman, compared to a Luhya woman, is determined to be 1.4875. The results of this study suggest that women who are not affiliated with the Luhya ethnic group have the highest probability of sustaining a miscarriage at any given moment. The cure model allows for the computation of the cure rate for each category of ethnicity by utilizing the parameter estimates. The cure rate is determined using equation (9).

The cure rate for a woman whose ethnicity is Kalenjin adjusting for the other covariates is 0.3933 suggesting 39.33 percent of women whose ethnicity is Kalenjin will never experience miscarriage not matter how long they are followed in the cohort. For the case of kikuyu and luo their cure fractions are 0.4478 and 0.5825

respectively, suggesting that 44.78 percent and 58.25 percent of women from this ethnicity are cured from miscarriage. Hence, it can be deduced with certainty that there exists a substantial disparity in cure rates between the Luhya ethnic group and other ethnicities. The likelihood of achieving a cure for the confounders, specifically the number of previous miscarriages, number of previous stillbirths, number of antenatal care (ANC) visits, and place of residency, are 54.88%, 47.09%, 76.09%, and 87.16% correspondingly.

The statistical analysis reveals that there is a significant association between the number of antenatal care (ANC) visits and the survival time of women who have not received a cure. This is supported by a p-value of 0.00097, which is below the predetermined level of significance of 5%. In this section, we make the assumption that a Cox proportional hazards (PH) model is utilized. The hazard ratio corresponding to a single-unit increment in the number of antenatal care (ANC) visits is determined to be exp(-0.2443), yielding a value of 0.7833. Hence, a rise of one unit in the quantity of antenatal care (ANC) visits is linked to a reduction of 21.67 percent in the probability of miscarriage among expectant mothers.

In the context of uncured patients, the hazard function associated with the covariate of ethnicity is unaffected by the particular ethnic background of a woman. The covariate of ethnicity was classified as Luhya in the reference group. The hazard ratio of female individuals from the Kalenjin ethnic group, in comparison to those from the Luhya ethnic group, is 0.8651. In a similar vein, the hazard ratio for a female individual affiliated with the Kikuyu ethnic group, in comparison to an individual affiliated with the Luhya ethnic group, is calculated to be 1.7153. Finally, the hazard ratio for a female individual of the Luo ethnic group, in comparison to an individual of the Luhya ethnic group, is 1.1151. Nevertheless, the p-values corresponding to these groupings fail to achieve statistical significance at the 5% threshold. Therefore, it is not feasible to conclusively determine the existence of a significant discrepancy in the hazards encountered by the Luhya ethnic group in relation to other ethnicities with regards to the duration of survival among uncured individuals. The results suggest that there is no statistically significant change in the likelihood of experiencing a miscarriage among different ethnic groups.

Furthermore, there is an observed increase of 32.3% in the probability of a woman encountering a miscarriage prior to the 28th week for each additional previous miscarriage, and a 37.1% rise for each new past stillbirth. Nevertheless, it is imperative to acknowledge that a conclusive determination regarding the significance of these aspects cannot be made. However, the p-values that approach significance suggest a potential positive effect on the incidence of miscarriage in the uncured population.

With regards to the covariate pertaining to place of residence, it is apparent that the hazard ratio for the urban category is 0.9634 at a significance level of 5 percent. The hazard ratio suggests that there is a 3.7% decrease in the probability of experiencing a miscarriage among mothers living in urban areas in comparison to those dwelling in rural areas. Mothers who live in urban settings demonstrate a higher probability of surviving miscarriages and experience extended periods of survival in comparison to their counterparts residing in rural locations. The mixture cure model demonstrates that odds ratios (ORs) beyond one signify a rise in the proportion of cured individuals, while hazard ratios (HRs) below one indicates an enhancement in the likelihood of survival for individuals who remain uncured.

The estimated survival probability for categorized covariates can also be obtained by using the mean values of the continuous variables. The survival curves obtained from the predictions are illustrated in Figures 2, 3, 4 and 5, respectively.

From the predicted survival curves in the Figure 2, those pregnant women who reside in the urban area have a better predicted survival probability, meaning they are less likely to experience miscarriages than those who do reside in rural areas. More so, the fitted survival probability curves show that Kalenjin, Kikuyu and Luo women tend to have shorter predicted survival of miscarriage than Luhya women. This indicates that the survival of the miscarriages by the luhya tribe is better than that of Kalenjins', kikuyus' and luos', particularly from week ten of the gestational period. Before that week, their predicted survival experience seems to be the same. The cure model indicates that gains in survival of miscarriage were mainly as a result of more women being long term survivors of miscarriage as short term survival trends have not shown any drastic improvements.

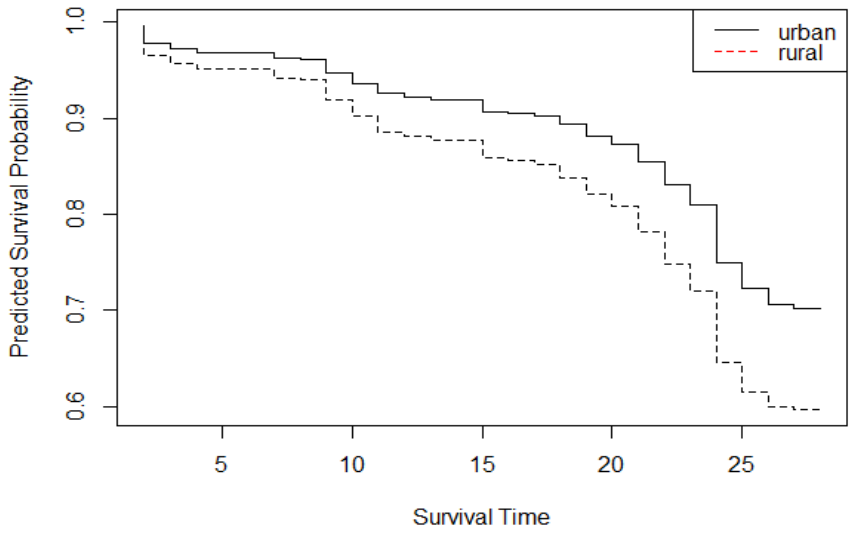

Figure 2. Fitted survival curves for the place of residence.

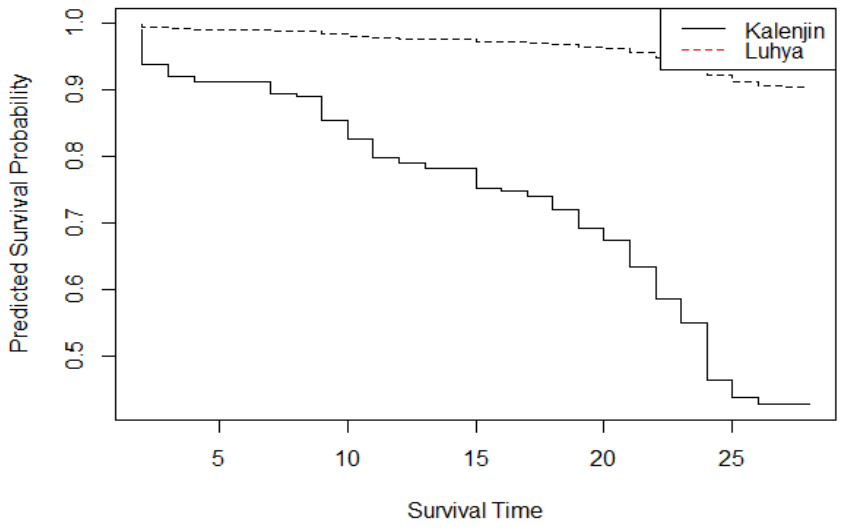

Figure 3. Fitted survival curves for the Kalenjin's

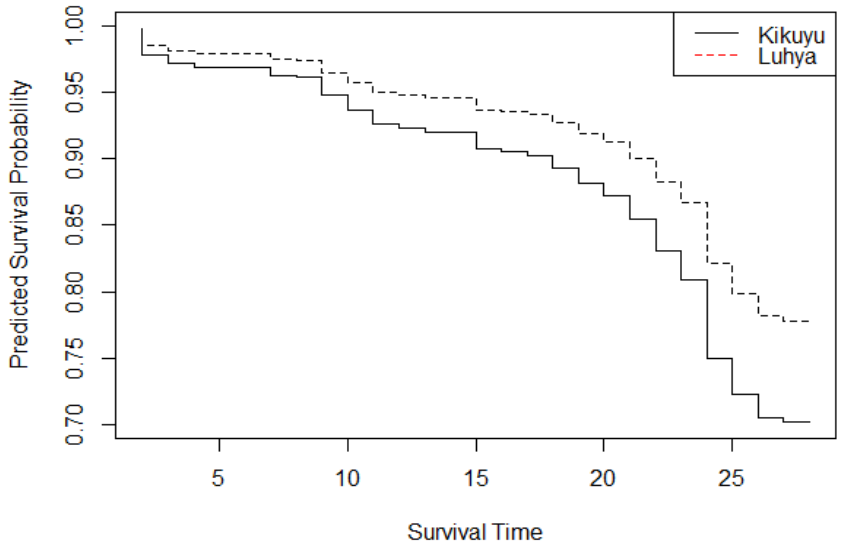

Figure 4. Fitted survival curves for the Kikuyu's

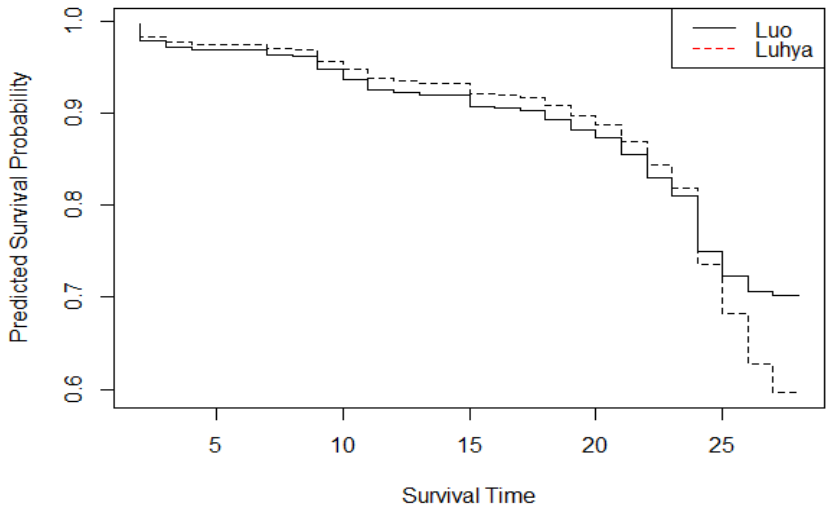

Figure 5. Fitted survival curves for the Luo's.

4.CONCLUSION

This paper estimated cure fraction of miscarriage and the effects of covariates on the cure rate, using cure models. The cure model showed that place of residency ($P = .003$), ethnicity: kalenjin ($P = .001$), kikuyu ($P = .014$) and luo ($P = .04$), number of prior miscarriages ($P = .000$), number of previous stillbirths ($P = .000$), and number of ANC visits ($P = .000$) statistically affect cure fraction. However these factors did not affect survival time, apart from the number of ANC visits ($P = .001$). The number of previous miscarriages, stillbirths, ANC visits, site of residency, and ethnic groups (Kalenjin, Kikuyu, and Luo) had cure probabilities of 54.88%, 47.09%, 76.09%, 87.16%, 39.33%, 44.78%, and 58.25%, respectively. Classical methodologies lack the ability to differentiate between curative treatments and treatments that only extend the lifespan (Bonadonna and Valagussa, 1981; Gamel et al., 1993; Gamel and Vogel, 1997). In addition, their statistical power decreases as the length of the follow-up period rises, as noted by Frankel and Longmate (2002) and Maetani et al. (2013). The cure model in this study showed that these factors had effect on long-term survivors. On short-term trends there were little changes. Using

the cure model to investigate miscarriage data provided more insights than classical models analysis

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.

Bibliography

Amico, M & Van keilegom, I. (2018). Cure models in survival analysis. *Ann., 5*, 311-342.

Berkson, J. & Gage, R. (1952). Survival curve for cancer patients following treatment'. *Journal of the American Statistical Association, 47*, 501-515.

Boag, J. (1949). Maximum Likelihood Estimates of the Proportion of Patients Cured by Cancer Therapy. *Journal of the Royal Statistical Society, 11*(1), 15-53.

Bonadonna,G and Valagussa, P. (1981). Dose-Response Effect of Adjuvant Chemotherapy in Breast Cancer. *The New England Journal of Medicine, 304*, 10-15.

Cai, C., Zou, Y., Peng, P. & J., Z. (2012). *smcure*, r package version 2.0. Retrieved from URL: https://CRAN.R-project.org/package=smcure

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

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*.

Farewell, V. (1977). A model for binary variable with time-censored observations. *Biometrika, 64*, 43-46.

Farewell, V. (1982). The use of a mixture model for the analysis of survival data with long-term survivors. *Biometrics, 38*, 1041-1046.

Frankel, P. and Longmate, J. (2002). Parametric Models for Accelerated and Long-Term Survival: A Comment on Proportional Hazards. *Statistics in Medicine, 21*, 3279-3289.

Gamel, J.W., and R. L. Vogel, R.L. (1997). Comparison of Parametric and Non-Parametric Survival Methods Using Simulated Clinical Data, , Vol. 16, No. 14. *Statistics in Medicine, 16*(14), 1629-1643.

Gamel, J.W., Vogel, R.L. and McLean, I.W. (1993). Assessing the Impact of Adjuvant Therapy on Cure Rate for Stage 2 Breast Carcinoma. *British Journal of Cancer,, 68*, 115-118.

Ibrahim, J. G., Chen, M.-H. and Sinha, D. (2001). 'Bayesian semiparametric models for survival data with a cure fraction. *Biometrics, 57*, 383-388.

Kenya Bureau of Statistics. (2019). *"2019 Kenya populayion and Housing census volume IV: Distribution of popualtion by Socio-economics charateristics"*.

Retrieved may 2, 2021, from https://www//housingfinance africa.org/appuploads/VOLUME-IV-KPHC-2019.PDF

Kline and Stein. (1984). Spontaneous abortion. *Perinatal Epidemiology. M.B. Bracken,* 23-51..

Maetani, S. and Gamel, J. (2013). Parametric Cure model versus Proportional Hazards Model in Survival Anlysis of Breast Cancer and Other Malignancies. *Advances in Breast Cancer Research, 2*, 119-125.

Maller, R. A., and Zhou, X. (2001). *Survival Analysis with Long-term Survivors.* New York: Wiley.

Meeker, W. (1987). Limited failure population life test: Application to intergrated circuit reliability. *Technometrics, 29*, 51-65.

Peng, Y. & Taylor, J.M.G. (2014). Cure models. In J. v. Klein (Ed.), *Handbook of Survival Analysis, Hand-books of Modern Statistical Methods series* (pp. 113-134). Chapman &Hall, Boca R.

Schmidt, P. &. (1989). Predicting criminal recidism using split populatio survival time models. *J. Econometric, 40*, 141-159.

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

World Health Organization. (1979). *Manual of the International Statistical Classification of Diseases Injures and Causes of Death.* Geneva, Switzerland: World Health Organization.

Yakovlev, A.Y., Tsodikov, A.D. (1996). *Stochastic models of tumor lattercy and their biostatistical application.* Singapore.: World Scientific.

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