Africa Journal of Physical Sciences Vol. 5, pp. 20-38, December 2020
http://journals.uonbi.ac.ke/index.php/ajps/index
ISSN 2313-3317

Understanding the Risks Factors of Under-Five Child Mortality in Kenya: Random Survival Forest and Accelerated Failure Time Shared Frailty Models

Khaoya Moses Mutaki[1,a*], Nelson Owuor Onyango[2,b], and Rachel Sarguta[3,c]

[1,2,3]School of Mathematics, University of Nairobi, P.O. Box 30197-00100, Nairobi, Kenya.
[a]mutaki1988@gmail.com*, [b]onyango@uonbi.ac.ke, [c]rsarguta@uonbi.ac.ke

ARTICLE INFO

Article History:
Submitted: 26 October 2018
Accepted: 30 November 2020
Available online: December 2020

Keywords:
Risks factors
Shared Frailty
Random Survival Forest
Accelerated Failure Time

ABSTRACT

Under-five mortality rates is one of the health indicators of great importance to any country. Kenya is among the countries in the Sub-Saharan Africa with high Under-Five Child Mortality (U5CM) rates. It is therefore important to apply best statistical approaches to establish which factors influence child mortality. This will go a long way to inform the optimal design of health intervention strategies within the country and globally. In this study, Random Survival Forest (RSF) and Accelerated Failure Time Shared Frailty Models have been used to analyze U5CM based on the Kenya Demographic Health Survey (KDHS, 2014) dataset. Akaike Information Criterion (AIC) statistics was used to select the model of best fit. Results obtained from fitting the AFT-shared frailty model, showed that there was presence of unobserved heterogeneity at community level. However, there was no evidence to conclude the existence of unobserved heterogeneity at the household level. Among the variants of the AFT Shared Frailty models analysed, the Log-logistic AFT- model showed that "the sons who have died," "daughters who have died," "duration of breastfeeding," and "months of breastfeeding" had significant influence on the U5CM (p <0.05). The Log-logistic AFT model with Gaussian frailty emerged to be the best model for the U5CM since it had the least Akaike Information Criterion (AIC) statistic. On the other hand, the results from Random Survival Forest, "sons who have died," "daughters who have died," "living children plus current pregnancy," "sex of child," "duration of breastfeeding," "number of living children," and "months of breastfeeding" were ranked as important factors that have influence on the under-five mortality. Furthermore, this study also found out that there was presence of unobserved heterogeneity at community level of clustering. At the household level however, there was no unobserved heterogeneity, hence there was no need for household frailty term.

1. Introduction

Based on World Health Organization (WHO) findings, the Under-Five Child Mortality (U5CM) rate has declined by a margin of 56% globally, from an estimated rate of 93 deaths per 1000 live births in 1990 to 41 deaths per 1000 live births in 2016. Approximately 20,000 fewer children died every day in 2016 than in 1990 [1]. In Africa back in 1970, the child mortality rate was 229 from every 1,000 live births. By 2010, this rate had reduced to 111 deaths per 1,000 live births [2]. The under-five child mortality has reduced by 39% in Sub-Saharan Africa between 1990 and 2011. Many studies in Sub-Saharan Africa show that the under-five population is growing rather fast [3]. According to the UNDP-Kenya reports despite significant declines in under-five mortality over the last 3 decades, Kenya did not achieve Millennium Development Goal 4 (MDG 4 on Reduced Child Mortality) by 2015 [32].

In the developing nations, the study of U5CM has always been a very important issue in health programs. A nation's level of wealth index growth and quality of life are imitated by its U5CM rates. Besides, to monitor and assess population and healthiness programs and guidelines, the U5CM rates are used. In Kenya, the infant mortality rate was reported to be 52 deaths per 1,000 live births, and U5CM rate is 74 deaths per 1,000 live births according to the KDHS 2014. All early childhood mortality rates declined between the 2003 and 2014, based on statistics from the KDHS surveys [4]. However, the MDG 2015 Target was 33 deaths per 1000 for the U5CM and 26 deaths per 1000 for the Infant Mortality Rates, which was not achieved by Kenya. Guided by the National Health Sector Strategic Plan II (NHSSPII) and the Vision 2030 Medium Term Plan, the Kenya government has embarked on efforts to fast track the attainment of the Sustainable Development Goals (SDG's) on Child survival and development.

The primary objective of this study is to identify the factors responsible for the Under-Five Child Mortality (U5CM), and to examine whether or not there is presence of unobserved heterogeneity on under-five mortality both at household and community levels. The statistical challenges include variable selection problem when determining the determinants of U5CM based on KDHS data with over 700 such possible covariates. Besides, this work attempts to account for possible clustering in data at household or community level, something that has been ignored in many similar studies.

The statistical problem involved in studies aimed at establishing the determinants of U5CM lies in variable selection. This is a major statistical problem when dealing with large datasets (especially in terms of establishing possible covariates that could be included in a prognostic model) KDHS 2014 for instance provides 1000 plus variables that may qualify as determinants of U5CM. One of the main approaches gaining prominence in statistics is the machine learning paradigm. This is because of the lack of over-reliance of such techniques on statistical distributions. Therefore, the use of machine learning technique (in this case Random Survival Forest) can assist in selection of these risk factors in a reliable manner [5]. For instance, Random Survival Forests [6] was used to determine U5CM in a study carried out in Uganda [5]. However, in this study, the variables included in the original model were selected based on previous literature. This is often the practice, although it is a very subjective way of deciding which covariates are to be included into the predictive model. Some studies have applied the Accelerated Failure Time (AFT) model assuming that effects of the covariates accelerate or decelerate the survival lifetime of the U5CM by some constant [8,9]. But these determinants do not at all times take into consideration the actual differences in the risk particularly in clustered survival data. So, inclusion of the unobserved random factor (frailty term) on the model improves correct measure of the determinants effect, thereby evading the problem of overestimation or underestimation of the model parameters. To cater for clustering at community or household levels, frailty type models have been used to model U5CM [7]. AFT-shared frailty model has received much attention recently. W. Pan [7] proposed the AFT frailty model by assuming the AFT gamma frailty model. P.K. Swain [8] used AFT – shared frailty models to study HIV/AIDS patients on Anti-Retroviral Therapy. P. Lambert [9] used parametric AFT with random effects on a kidney transplant survival data. Vaupel [12] introduced the term frailty in order to account for the unobserved heterogeneity, random effects and association in univariate (survival) models. In understanding the determinants of U5CM using shared frailty model [13], a study conducted in Uganda found out that there was existence of unobserved heterogeneity at household level but there was no enough evidence to conclude existence of unobserved heterogeneity at community level. Sex of the household head, sex of the child, and number of births in the past one year were found to be having significant influence on mortality [13]. A study on infant mortality in India it was found that there was presence of unobserved heterogeneity both at individual and community levels. In the same study, it was also established that child mortality was higher among women married before 18 years of age compared to those married after 17 years of age [14].

Various studies have been conducted to estimate the effect of prognostic factors on the U5CM in Kenya [21,22], but very few studies have considered use of machine learning algorithms and the effect of clustering This study attempts to investigate the factors influencing child survival by using Random Survival Forest for variable selection and AFT-shared frailty model to account for clustering. Gaussian frailty model with baseline distributions as exponential, Weibull, log-normal, and log-logistic have been used to estimate the effect of prognostic factors on the U5CM. The result of this model has been compared to the results of a model without frailty model. In order to compare the overall performance of these models, we have used Akaike Information Criterion (AIC) statistics.

The remaining part of this paper is organized as follows. In the Methodology section, the data, the theory behind Random Survival Forest, AFT and AFT-shared frailty models have been discussed. In the results section, the outcome obtained from RSF analysis on KDHS-2014 dataset as well as from AFT and AFT-shared frailty model have been shown, followed by a discussion section and a conclusion.

2. Materials and Methods

a. Data

The dataset used is the child records from Kenya Demographic Health Surveys (KDHS 2014). The KDHS 2014 dataset was drawn from a master sampling frame, Fifth National Sample Survey, and Evaluation Programme (NASSEPV). This is a kind of structure that Kenya National Bureau of Statistics (KNBS) currently uses to carry out household surveys in Kenya. Kenya as a country is divided into a total of forty-seven counties. In the process concerning this development of NASSEPV, each of these forty-seven counties was stratified in two categories; urban and rural strata, resulting into 92 strata. This sample had a total of 40,300 households from 1612 community clusters that were spread across the entire country, with 995 clusters from the rural zones and 617 from urban zones. The samples were selected independently in each sampling stratum, using a two-stage sample design. In the first stage, the 1612 EAs were chosen with equal likelihood from NASSEPV frame. The households from listing operations served as the sampling frame for the second stage of selection, from where a total of 25 households were selected from each cluster [4].

The KDHS-2014 Child record dataset includes women of ages between 15 to 49 years. This study includes only children of between 1-59 months old, accounting for a total observation of 20354.

The study variables:

The response variable;

Under-five mortality is described as the mortality from the age of 1month to the age of 59months. The outcome of interest in this study is "risk of death occurring in an age interval of 1-59month period." The dependent variable was therefore, survival time in months for children under five years of age.

The independent variables and variable selection;

The original data has 1099 variables excluding survival time and event variables. Out of these, 313 variables had 100% missing data and were deleted. Therefore, random forest for survival regression and classification was applied to the remaining 786 covariates to select those variables that had influence on under-five mortality, ranking the variables according to their importance. The split rule used was logrank and its variants, based on an overall proportional hazards assumption. The split rules could be varied in future work, to account for possible cases on non-proportional hazard, while applying a classification algorithm to a dataset with event time as data outcome.

b. Random Survival Forest:

Random forest for survival regression and classification is implemented in the R-package **randomForestSRC**. Random survival forest is a simple but robust approach that has been considered as an attractive alternative model choice for survival data. This approach is an extension of the random forest [6]. This method is fully non-parametric, has fewer assumptions and can deal with data of high dimension easily [5]. Random survival forest does not impose a restrictive structure on how the variables should be combined. If the relationship between the predictor variables and the response variable is complex with non-linear patterns and interactions then RSF is capable of incorporating these complexities robustly [15, 16].

Steps to develop algorithm for random survival forests;

- *B* bootstrap samples are drawn from the original data. Each bootstrap sample excludes 37% of the data on average, this excluded data is referred to as Out-Of Bag data (OOB data).
- Grow a survival tree for each bootstrap sample. At every node of the tree, *p* candidate variables are randomly selected. The node is split using the candidate variable that maximizes the survival difference between daughter nodes.
- Tree is grown to its full size under the constraint that a terminal node must have not less

than d_0>0 unique deaths.

- Compute a cumulative hazard function (CHF) for each tree. Average to obtain ensemble CHF.
- Use the OOB data to calculate prediction error for the ensemble CHF.

Log-rank split rule

Suppose a node *h* can be split into two daughter nodes *α* and *β*. The best split at a node *h*, on a covariate *x* at a split point *c∗* is the one that gives the largest log-rank statistic between the two daughter nodes. The log-rank statistic for a split on *x* at a given covariate value *c∗* is defined as;

$$i(x, c\,*) = \frac{\sum_{j=t_1}^{t_N} (d_{\alpha,j} - E(D_{\alpha,j}))}{\sqrt{\sum_{j=t_1}^{t_N} v\,ar\left(D_{\alpha,j}\right)}} \tag{1}$$

where $d_{\alpha,j}$ is the number of events in daughter node *α* at time point *j*. The expected number of events in daughter node *α*, $E(D_{\alpha,j})$ and its variance are given by;

$$E\left(D_{\alpha,j}\right) = d_j\left(\frac{R_{\alpha,j}}{R_j}\right) \tag{2}$$

$$var\left(D_{\alpha,j}\right) = \frac{R_{\alpha,j}}{R_j}\left(1 - \frac{R_{\alpha,j}}{R_j}\right)\left(\frac{R_j - d_j}{R_j - 1}\right)d_j \tag{3}$$

Where d_j is the total number of observed events at time point *j*. $R_{\alpha,j}$ is the number of individuals at risk in node *α* at time point *j* and R_j the combined number at risk in daughter nodes *α* and *β*.

c. Accelerated Failure Time (AFT) and AFT-shared frailty models:

AFT-model is a parametric approach that follows distributional assumptions. It assumes that the effects of the covariates are either to accelerate or decelerate the survival time by some constant.

Accelerated failure time (AFT) model

The probability density function is expressed as follows;

$$f(t) = (\sigma t)^{-1} f_o\left(\frac{logt - log\psi(X)}{\sigma}\right) \tag{4}$$

whereby σ is the scale parameter, and $\psi(X)$ is some function of covariates.

$$\psi(X) = exp(\mu + X'\beta) \tag{5}$$

Therefore, corresponding Accelerated Failure Time (AFT) model can be written in a regression form as;

$$logT = \mu + X'\beta + \sigma\epsilon \tag{6}$$

whereby μ is an intercept, ϵ is a random variable with a density function $f_o(\epsilon)$ and the corresponding baseline survival function $S_o(\epsilon)$. AFT models do allow a wide range of parametric forms for the density function. The survival function of the AFT models is expressed in the following form;

$$S(t) = S_o^*[(\frac{t}{\psi(X)})^{\frac{1}{\sigma}}] = S_o\left(\frac{logt - log\psi(X)}{\sigma}\right) \tag{7}$$

Where S_o^* is the baseline survival function. Since $\psi(X) = exp(\mu + X'\beta)$, the survival function can be rewritten as;

$$S(t) = S_o\left(\frac{logt-\mu-X'\beta}{\sigma}\right) \tag{8}$$

Inference for AFT models

For random lifetime $T_i$for the subjects $i = 1, \dots, n$, the likelihood function under model (8) is expressed as;

$$L(\beta,\sigma) = \prod_{i=1}^{n}(\frac{1}{\sigma}f_o(\frac{logt_i - \mu - X'\beta}{\sigma}))^{\delta_i}S_o(\frac{logt_i - \mu - X'\beta}{\sigma})^{1-\delta_i}. \tag{9}$$

Using $\epsilon_i = \frac{logt_i-\mu-X\prime\beta}{\sigma}$, the log-likelihood function takes the form;

$$l(\beta,\sigma) = -rlog\sigma + \sum_{i=1}^{n}[\delta_i logf_o(\epsilon_i) + (1-\delta_i)logS_o(\epsilon_i)] \tag{10}$$

Where $r = \sum\delta_i$ refers to the number of events. The first partial derivatives of $l(\beta,\sigma)$will give the maximum likelihood estimators $\hat{\beta}$and $\hat{\sigma}$.

The distributions mostly used in AFT models are exponential, Weibull, lognormal, and log-logistic distributions.

Exponential distribution

It has only one parameter, it's pdf is expressed by;

$$f(t) = \lambda e^{-\lambda t}, t > 0, \lambda > 0 \tag{11}$$

The survival function, $S(t)$ which refers to as the chances of an individual living up to or past time t can be obtained from;

$$S(t) = -\int_0^t \lambda\, e^{-\lambda u}du \tag{12}$$

$$= e^{-\lambda t}$$

Cumulative distribution function, $F(t)$ is expressed as;

$$F(t) = 1 - S(t) = 1 - e^{-\lambda t} \tag{13}$$

Hazard function,$h(t)$ is given by;

$$h(t) = \frac{f(t)}{S(t)} = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda \tag{14}$$

Cumulative hazard function, $H(t)$ is given by;

$$H(t) = \int_0^t \lambda \, du = \lambda t \tag{15}$$

Weibull distribution

This distribution has two parameters; its pdf can be given by;

$$f(t) = \lambda k t^{k-1} e^{-\lambda t^k}, \lambda > 0, k > 0 \tag{16}$$

Its survival function, $S(t)$ would be expressed b;

$$S(t) = e^{-\lambda t^k} \tag{17}$$

The cumulative distribution function, $F(t)$ can be gotten from;

$$F(t) = 1 - S(t) = 1 - e^{-\lambda t^k} \tag{18}$$

The hazard function,$h(t)$ would be given by;

$$h(t) = \frac{f(t)}{S(t)} = \frac{\lambda k t^{k-1} e^{-\lambda t^k}}{e^{-\lambda t^k}} = \lambda k t^{k-1} \tag{19}$$

Cumulative hazard function, $H(t)$ would be given by;

$$H(t) = \int_0^t \lambda \, k u^{k-1} du = \frac{\lambda k u^k}{k} |_0^t = \lambda t^k \tag{20}$$

Lognormal distribution

General formula for probability density function of the lognormal distribution is;

$$f(t) = \frac{e^{-((\ln((t-\theta)/m))^2/(2\sigma^2))}}{(t-\theta)\sigma\sqrt{2\pi}} \tag{21}$$

where $t > \theta; m, \sigma > 0$, σ refers to shape parameter, θ represents parameter for location and m refers to scale parameter (and is also median for the distribution). When $t = \theta$, $f(t) = 0$. In a scenario whereby $\theta = 0$, and $m = 1$ we will have a standard distribution. The probability density function (pdf) for the standard form is

$$f(t) = \frac{e^{-((lnt)^2/2\sigma^2)}}{t\sigma\sqrt{2\pi}} \tag{22}$$

where $t > 0$, $\sigma > 0$, since the general form of the pdf can be written in standard form, succeeding formulars have been written in standard form. Cumulative distribution function, $F(t)$ is expressed as

$$F(t) = \Phi\left(\frac{\ln(t)}{\sigma}\right) \tag{23}$$

where $t \geq 0$; $\sigma > 0$. Φ refers to the cumulative distribution function of normal distribution. Survival function expressed as

$$S(t) = 1 - F(t) = 1 - \Phi\left(\frac{\ln(t)}{\sigma}\right) \tag{24}$$

Hazard function is expressed as;

$$h(t) = \frac{\left(\frac{1}{t\sigma}\right)\phi\left(\frac{lnt}{\sigma}\right)}{\Phi\left(\frac{-lnt}{\sigma}\right)} \tag{25}$$

where $t > 0$, $\sigma > 0$. ϕ refers to pdf of normal distribution. The cumulative hazard function is expressed as;

$$H(t) = -ln\left(1 - \Phi\left(\frac{\ln(t)}{\sigma}\right)\right) \tag{26}$$

Log-logistic distribution

Log logistic distribution is a parametric model which can be applied in survival analysis for those events whose rates increase initially and then decrease over time. It is a probability distribution of a random variable whose logarithm has a logistic distribution. It's probability density function is expressed as;

$$f(t) = \frac{(\beta/\alpha)(t/\alpha)^{\beta-1}}{(1 + (t/\alpha)^{\beta})^2} \tag{27}$$

where $t > 0$, $\alpha > 0$, $\beta > 0$. α and β are scale and shape parameters respectively. The cumulative distribution function (CDF), $F(t)$, is given by;

$$F(t) = \frac{1}{1 + (t/\alpha)^{-\beta}} \tag{28}$$

It's survival function, $S(t)$ is;

$$S(t) = 1 - F(t) = [1 + (t/\alpha)^{\beta}]^{-1} \tag{29}$$

It's hazard function, $h(t)$ is expressed as;

$$h(t) = \frac{f(t)}{S(t)} = \frac{(\beta/\alpha)(t/\alpha)^{\beta-1}}{1 + (t/\alpha)^{\beta}} \tag{30}$$

Best fitting model selection

According to Akaike [26], the principle in statistical model building dictates that the increase in the number of parameters should be stopped the moment it has been discovered that more increase does not give significant improvement of the fit of the model to the data. He suggested the use of Akaike Information Criterion (AIC) which is expressed as;

$$AIC = -2(log - likelihood) + 2k \tag{31}$$

Where k refers to the number of parameters. AIC value will always increase if an unnecessary variable has been included in the model. Therefore, this means that the smaller the AIC value the better the model.

Accelerated failure time (AFT) shared frailty model

Frailty is an unobserved random effect that is shared by subjects within a group or it can be referred to as an unobserved random factor that modifies multiplicatively the hazard function of an individual or cluster of individuals. It is known that subjects in the same cluster are more alike than the subjects in different clusters since they share similar environment. In this study we have applied a shared frailty model to study the cluster variation effect on the child survival which cannot be accounted for by the covariates itself. Shared frailty model is a mixture model since the common risk in each cluster (the frailty) is assumed to be random [8]. This model assumes that all the event times in a cluster are independent given the frailty terms. This is to say, it is a conditional independence model where the frailty is common to all subjects in a cluster and so responsible for creating dependence between event times. This is the reason a shared frailty model can be expressed as a mixed (random effects) model in survival analysis with cluster variation (frailty) and individual variation described by the hazard function [10].

Thus, frailty models try to account for correlations within groups. It is worth noting that just applying Cox-proportional hazards model or accelerated failure time (AFT) model directly to a cluster data while ignoring the possible correlations in each cluster may lead to incorrect conclusions [11].

AFT-shared frailty model is suitable when subjects within a cluster share a common unobserved heterogeneity. It explicitly takes into account the possible correlation among failure times [8]. Frailty term gets into the AFT-model as random effects. AFT models with shared frailty is expressed in the following form

$$logT_{ij} = \mu + X'_{ij}\beta + w_i + \sigma\epsilon_{ij} \tag{32}$$

whereby μ is an intercept, β is a vector of regression coefficients, X_{ij} is the vector of fixed-effect covariate, σ is a scale parameter, $\epsilon_{ij}'s$ are independent and identically distributed random errors,T_{ij} is the event time for the j^{th} subject in the i^{th} cluster and $w_i's$ are the frailty terms which are assumed to be independent and identically distributed with density function $f(w_i)$.The survival function for an AFT-shared frailty model at time t is expressed as;

$$S(t) = S_o^*[(\frac{t}{\psi_{ij}})^{\frac{1}{\sigma}}] = S_o\left(\frac{logt - log\psi_{ij}}{\sigma}\right) \tag{33}$$

whereby σ refers to the scale parameter, S_o^* is a survival function defined on $(0, \infty)$, and S_o is the baseline survival function that satisfies the relationship $S_o^*(\omega) = S_o(log\omega)$, ψ_{ij} is some function of the covariates. $\psi_{ij} = exp(\mu + X'_{ij}\beta + w_i)$. Conditional survival function is given by;

$$S_{ij}(t|w_i) = S_o\left(\frac{logt-\mu-X\prime_{ij}\beta-w_i}{\sigma}\middle|w_i\right) \tag{34}$$

where $S_o(.)$ is the survival function of ϵ_{ij} and μ is an intercept, β is vector of regression coefficients, X_{ij} is a vector of fixed-effect covariate of the j^{th} subject in the i^{th} cluster. We assume that the frailty term, w_i follows Gaussian distribution with mean and variance of μ and θ respectively. With $\epsilon_{ij} = \frac{logT_{ij}-\mu-X'_{ij}\beta-w_i}{\sigma}$, the conditional survival and hazard functions are expressed as;

$$S_{ij}(t|w_i) = S_o(\epsilon_{ij}|w_i) \tag{35}$$

$$h_{ij}(t|w_i) = \frac{1}{\sigma t} h_o(\epsilon_{ij}|w_i) \tag{36}$$

respectively, whereby $h_o(.)$ is the hazard function of ϵ_{ij}.

Let G be the number of clusters, $i = 1, \dots, G$, and n_i be the number of subjects within the i^{th} cluster. The conditional likelihood for the observed data is;

$$L_c = \prod_{i=1}^{G}\prod_{j=1}^{n_i}[\frac{1}{\sigma t_{ij}} h_o(\epsilon_{ij}|w_i)]^{\delta_{ij}} S_o(\epsilon_{ij}|w_i) \tag{37}$$

Integrating out the unobserved frailties w_i, the marginal likelihood function for all the clusters is given by;

$$L_m = \prod_{i=1}^{G} \int \prod_{j=1}^{n_i}[\frac{1}{\sigma t_{ij}} h_o(\epsilon_{ij}|w_i)]^{\delta_{ij}} S_o(\epsilon_{ij}|w_i) f(w_i) dw_i \tag{38}$$

Estimates of the parameters (σ, β, θ) can be found by maximizing the likelihood function (38).

Gaussian frailty

The Gaussian frailty probability density function is given by

$$f(w) = \frac{1}{\sqrt{2\Pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(w-\mu)^2} \tag{39}$$

whereby μ is the mean of the distribution, σ is the standard deviation, $w\epsilon(-\infty, \infty)$ and σ^2 is the variance. Laplace transformation is given by

$$L(s) = e^{-s\mu + \frac{s^2\sigma^2}{2}} \tag{40}$$

Mean and variance can therefore be obtained from the first and second derivatives of the Laplace transformation;

$$L^1(s) = (-\mu + s\sigma^2) e^{-s\mu + \frac{s^2\sigma^2}{2}} \tag{41}$$

$$L^2(s) = (-\mu + s\sigma^2)(-\mu + s\sigma^2) e^{-s\mu + \frac{s^2\sigma^2}{2}} + \sigma^2 e^{-s\mu + \frac{s^2\sigma^2}{2}} \tag{42}$$

equating s to 0, therefore the mean and variance from laplace becomes;

$$E(W) = (-1)L^1(0) = \mu \tag{43}$$

$$Var(W) = L^2(0) - (-L^1(0))^2 = \mu^2 + \sigma^2 - \mu^2 = \sigma^2 \qquad (44)$$

Chi-square test

The Chi-square test is a non-parametric test that is used to determine whether there is an association between categorical variables. This test is applied in this context to establish whether or not there is an association within clusters. Presence of association within clusters implies existence of unobserved heterogeneity, while no association within clusters implies no existence of unobserved heterogeneity. The null and alternative hypothesis are stated as;

$$H_0 : No\ association\ within\ clusters$$
$$H_1 : There\ is\ association\ within\ clusters$$

The Chi-square statistic is given by;

$$\chi = \sum \frac{(O_i - E_i)^2}{E_i} \qquad (45)$$

Where O_i and E_i refers to the observed data, and expected values respectively. Degrees of freedom, which is $df = (r-1)(c-1)$, where r and c are the number of rows and columns respectively is used to read corresponding critical value at 5% level of significance on a Chi-square table. The critical and statistic values are compared, if the statistic value is greater than the critical value then it implies that the p-value is < 0.05 hence the H_0 is rejected and conclude that there is association within clusters, therefore, there is presence of the unobserved heterogeneity. If the statistic value is less than the critical value then it implies that the p-value is > 0.05 hence the H_0 is not rejected and conclude that there is no association within clusters, therefore, there is no presence of the unobserved heterogeneity.

d. Analysis approach

The analysis was done using STATA and R statistical software. STATA was used for data cleaning purposes, while the R packages survival and randomForestSRC were used for data analysis.

e. Ethics approval and consent to participate

We did not require any approval to conduct this study. A full ethics statement on data collection and handling of human subjects is given via the link:https://dhsprogram.com/What-We-Do/Protecting-the-Privacy-of-DHS-Survey-Respondents.cfm

3.Results and Interpretation

The results section is structured in two parts. The first part illustrates the variable selection exercise, using RSF methodology. Variables are selected based on the variable importance statistic. The second part analyses the determinants of U5CM and their effect using AFT models. The variables used in the AFT models are those derived from the variable selection exercise based on the RSF results.

a. Variable selection using Random Survival Forest

Table 1: Characteristics of the fitted Random Forest for Survival Regression and Classification

Sample size	20354
Number of deaths	428
Was data imputed	Yes
Number of trees	1000
Forest terminal node size	3
Average number of terminal nodes	290.723
Number of variables tried at each split	29
Total number of variables	786
Analysis	Random Survival Forest
Family	Survival
Splitting rule	Log-rank *random*
Number of random split points	10
Error rate	0.31%

From Table 1, the characteristics of the RSF model are displayed. The log-rank splitting rule was used in the classification. It is worth noting that this model is built using 786 covariates. To identify the most important covariates that determine U5CM in Kenya, permutation importance was applied to measure variable importance [15, 17, 18]. Results for the ranking of variables according to their level of influence on under-five mortality are summarized in table 2.

Table 2: Variable importance (VIMP) based on 1000 trees under log-rank splitting rule

Variable	Importance
Sons who have died	0.0167
Daughters who have died	0.0112
Living children + current pregnancy	0.0036
Sex of child	0.0034
Duration of breastfeeding	0.0028
Number of living children	0.0021
Months of breastfeeding	0.0020

Variable importance

Variable importance is used as a means for thresholding variables, since any variable with a VIMP less than 0.002 is likely to be noise [17]. Using this rule, the variables found to be having high influence on U5CM based on their level of importance are thus displayed in table 2 and these are the variables that we have chosen to use as our explanatory variables in the AFT models.

b. Summary of Socio-Demographic variables involved in the study

Table 3: Demographic and socioeconomic characteristics summarized by child survival (N= 20, 354 observations)

Variable	Mortality, n (percentage mortality)	Percentage (out of N = 20354)
Sons who have died		
None	18237(0.8)	89.6
1 son	1775(13.1)	8.7
2 sons	275(12.4)	1.4
3 sons	58(12.1)	0.3
4 sons	7(14.3)	0.0
5 sons	1(0)	0.0
6 sons	1(0)	0.0
Sex of child		
Male	10302(2.3)	50.6
Female	10052(1.9)	49.4
Births in last five years		
One	9531(1.2)	46.8
Two	8707(2.4)	42.8
Three	1999(4.6)	9.8
Four	107(7.5)	0.5
Five	10(10)	0.0
Contraceptives use		
Currently using	9760(1.6)	48.0
Used since last birth	6233(2.7)	30.6
Used before last birth	1091(2.8)	5.4
Never used	3270(2.1)	16.1
Currently breastfeeding		
No	14841(2.4)	72.9
Yes	5513(1.2)	27.1
Living children + current pregnancy		
None	26(100)	0.1
1 child	3051(2.6)	15.0
2 children	4618(2.2)	22.7
3 children	3876(1.8)	19.0
4 children	2983(1.8)	14.7
5 children	2075(2.2)	10.2
6children and above	3725(1.4)	18.3
Daughters who've died		
None	18503(1.1)	90.9
1 daughter	1580(11.6)	7.8
2 daughters	212(17)	1.0
3 daughters	44(20.5)	0.2
4 daughters	14(21.4)	0.1
5 daughters	1(0)	0.0

Variable	Mortality, n (percentage mortality)	Percentage (out of N = 20354)
Region		
Coast	2565(2.2)	12.6
North eastern	1556(1.7)	7.6
Eastern	2930(1.6)	14.4
Central	1371(1.7)	6.7
Rift valley	6651(1.5)	32.7
Western	1926(2.8)	9.5
Nyanza	2845(3.7)	14.0
Nairobi	510(3.1)	2.5

Table 3 displays the distribution of deaths of the under-five children at each factor level included in the analysis. It shows that among mothers who had male children, out of 10,302 children born, 2.3% died before their fifth birthday. Of the 10,052 female children born, 1.9% of them died before the 5th birthday. The table has also summarized the distribution of deaths and births of children for the rest of the covariates considered involved in the study.

c. Modeling for determinants using AFT model and its variants

Data analysis was done using four AFT-models; Exponential, Weibull, Lognormal, and Log-logistic distributions. The four models were checked for the parametric model assumptions (including assumptions such as linearity of covariates, normality of residuals, covariate independence among others), using residual plots. The predictive model assumptions were met fairly well [19]. The results from Log-logistic AFT model were reported since the model had the least Akaike Information Criterion (AIC) statistic.

Table 4: AIC values of the AFT and AFT-shared frailty models (Community and household)

Baseline distributions	No frailty	Gaussian Frailty	
		Community	***Household***
	AIC	AIC	AIC
Exponential	2630.088	2487.14	2471.722
Weibull	2566.057	2568.274	2696.672
Log-logistic	2462.352	2485.497	2461.255
Lognormal	2499.53	2508.123	2501.959

Table 4 shows results from the AIC values of the AFT and AFT-shared frailty models. We have assumed Exponential, Weibull, Log-logistic and lognormal distributions for baseline, while the frailty distribution is the Gaussian. The AIC values of the different AFT, and AFT models with Gaussian shared frailty model are displayed in table 4 for community and household clusters respectively. The AIC values of Log-logistic AFT, and Log-logistic AFT with Gaussian frailty model have been found to be minimum among all other considered models in all cases, indicating that it is the most efficient model.

The results of Log-logistic AFT and Gaussian shared frailty model with Log-logistic baseline distribution has been displayed in table 5. The estimated coefficients, p-values, parameter estimates of baseline distributions and frailty variance have also been displayed in table 5.

The Log-logistic AFT-model shows that i) the sons who have died, ii) daughters who have died, iii) duration of breast feeding and months of breastfeeding, were found to be having significant influence on the under-five mortality ($p<0.05$). Increase in the number of sons and daughters who have died in the households reduced the risk of death. Sex of child, number of living children, and living children plus current pregnancy were found to be non-significant factors for child mortality.

Based on a chi-square test on a null hypothesis that the variance of the community frailty term is zero ($\vartheta = 0$), the test statistics yielded a *p-value* of 0.028. At 5% level of significance, it means that there was enough evidence of existence of unobserved heterogeneity at community level. This implies that there are other factors affecting under-five child mortality at community level that are not explained by the observed covariates included in the model. The sources of the unobserved heterogeneity at the community level may be attributed to access to food and probably other factors that cannot be easily measured at community level. This is an area that needs further research in order to explain the reasons for unobserved heterogeneity at community level.

In the case of household frailty term, the test statistic returned a *p-value* of 0.28. At 5% level of significance, this implies that there was not enough evidence to show the existence of unobserved heterogeneity at household level. This statement means that the survival times of children under the age of five within the same household can be well explained by the observed covariates considered in the study without the inclusion of household frailty term. In this scenario one can apply Log-logistic AFT-model without frailty since the outcome suggests that there is no difference on the conclusions that would be drawn.

Table 5: AFT and AFT-shared frailty models for the under-five mortality

	Log-logistic (no frailty)		**Log-logistic(G) (community)**		**Log-logistic(G) (household)**	
Parameters	***Beta***	***P-value***	***Beta***	***P-value***	***Beta***	***P-value***
Intercept	4.6843	0.000	4.0527	0.000	4.3908	0.000
Sons who have died	-1.9436	0.000	-1.4136	0.000	-1.6941	0.000
Daughters who've died	-1.6536	0.000	-1.2341	0.000	-1.4572	0.000
Sex of child						
Male	Ref		Ref		Ref	
Female	0.3298	0.132	0.2739	0.067	0.3015	0.110
Number of living children	0.4986	0.125	0.3587	0.110	0.4347	0.120
Living children+	0.0516	0.873	0.0388	0.860	0.0439	0.870
Duration of breastfeeding	0.2275	0.000	0.1713	0.000	0.2011	0.000
Months of breastfeeding	-0.0402	0.002	-0.0317	0.001	-0.0364	0.001
Frailty				0.028		0.280
Variance			0.717		0.337	

4. Discussion

The study of U5CM is crucial especially in the Low and Middle-Income Countries including Kenya because of the relatively high rates. Kenya has however witnessed a significant decline in under-five child mortality recently, except that this rate is still higher than the globally targeted figures for U5CM of less than 33 deaths per 1000 births [4]. In this paper, an effort has been made to determine the possible determinants of the under-five child mortality in Kenya using Random Survival Forest and Accelerated Failure Time (AFT)-shared frailty models. The Random Survival Forests is a machine learning algorithm that was used to conduct variable selection. The Accelerated Failure Time models were fit using the selected variables and used to determine the effect of the covariates.

Random Survival Forests has increasingly become popular alternative way for analyzing time to event data [20]. This approach provides a classification algorithm that enables establishing which variables have an influence on the mortality outcome. The method is robust to deviations from statistical assumptions that bog parametric models. In this study, this approach was used to identify and select important covariates to U5CM based on variable importance. Other methods of variable selection exist in literature including methods for variable selection in Partial Least Squares Path Models [29], Bayesian variable selection methods [30], and many other variable selection approaches [31]. In this context for event time outcome data, covariates such as i) number of living children, ii) living children plus current pregnancy, and iii) sex of child emerged as important covariates in explaining the under-five child mortality in Kenya. However, these covariates did not appear to be significantly associated with the under-five child mortality rate in the AFT model. Random Survival Forest is fully non-parametric whereas AFT model is fully parametric.

Our study considered two levels of clustering, including community and household levels. Gaussian shared frailty assuming a Log-logistic baseline distribution was used to estimate effect of risk factors on child survival. In understanding the determinants of U5CM using shared frailty model [13], a study conducted in Uganda found out that there was existence of unobserved heterogeneity at household level but there was no enough evidence to conclude existence of unobserved heterogeneity at community level. A study on infant mortality in India it was found that there was presence of unobserved heterogeneity both at individual and community levels. In the same study, it was also established that child mortality was higher among women married before 18 years of age compared to those married after 17 years of age [14]. The output of the shared frailty model was compared to the output of the model without frailty, and the results were that there was no presence of the unobserved heterogeneity at household clusters. On the other hand, there was presence of the unobserved heterogeneity at the community level. Similar results to our study was also reported elsewhere [28], however [13, 27] which found out that there was presence of the unobserved heterogeneity at household level. Their findings could be attributable to the fact that country dynamics differ. These studies may have disagreed with ours as they were conducted in different countries and at different time settings.

In this study, variables relating to siblings and mother characteristics were key determinants of U5CM. i) Sons who have died, ii) daughters who have died were some of the factors associated with decreased risks of death in Kenya. In a similar study conducted using DHS data in Uganda, parent characteristics, sibling/mother characteristics were found to have an effect on U5CM. The study identified the following factors among others: i) sex of the household head, ii) sex of the child, and iii) number of births in the past one year were found to be having significant influence on U5CM [13].

The problem of high dimensional data especially associated with DHS datasets possess a big challenge to many statistical analysis approaches. One such problem is the issue of variable selection. It may not be possible to use over 700 variables in a predictive model and still make sense of the effect of important covariates. Random Survival Forest provided an effective approach to the problem of variable selection, through variable importance. With RSF, classification of risks factors for the under-five child mortality was accomplished through variable importance ranking.

5. Conclusions

Our study found out that breastfeeding and sibling/mother characteristics were key determinants of U5CM. Variables such as

i. sons who have died,
ii. daughters who have died,
iii. duration of breastfeeding, and
iv. months of breastfeeding

were strongly associated with U5CM from both Random Survival Forest model and the AFT model.

As far as clustering is concerned, it was also found out that there was presence of unobserved heterogeneity at community level, implying that there are other factors affecting U5CM at community level other than those explained by the observed covariates that were included in the model. On the other hand, there was no presence of the unobserved heterogeneity at household level, implying that the survival times of children under the age of five within the same household can be explained by the observed covariates that were considered in the study.

Due to the high dimensionality of the DHS datasets, variable selection problem is often faced while conducting an analysis based on such data. Random Survival Forest came in handy for classification of variables that were later on used in the AFT model for parameter estimation.

These findings do suggest that health care interventions could focus on communities and not necessarily at household level, in order to achieve outcomes on a wider scale. The heterogeneity across communities seems to play a dominant role in determining U5CM. Such disparities could be associated with socio-economic differences, including eating habits, religious inclinations, socio-political inclinations, education levels among other factors.

Acknowledgements

I am grateful to Demographic Health Surveys for giving me access to the KDHS-2014 dataset. I would also like to acknowledge DELTAS SSACAB Project and VLIR team project for support I have gained from them during my work on the manuscript, especially due to the short courses that were often conducted on the statistical methodology and research publication skills.

References

[1] WHO, Statistics on under-five child mortality, World Health Organization, Tech. Rep., 2017.

[2] Statista, Mortality rate of children under 5 in Africa, Statista, Tech. Rep.,2017.

[3] USAID, under-five mortality in sub-saharan Africa, United States Agency for International Development, Tech. Rep.,2017.

[4] KDHS, infant and under-five mortality rate, Demographic Health Surveys, Tech. Rep., 2014.

[5] Nasejje, J. B., & Mwambi, H. (2017). Application of random survival forests in understanding the determinants of under-five child mortality in Uganda in the presence of covariates that satisfy the proportional and non-proportional hazards assumption. *BMC research notes*, *10*(1), 459.

[6] L. Breiman, Random forests, *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[7] Pan, W. (2001). Using frailties in the accelerated failure time model. *Lifetime Data Analysis*, *7*(1), 55-64.

[8] Swain, P. K., & Grover, G. (2016). Accelerated failure time shared frailty models: application to HIV/AIDS patients on anti-retroviral therapy in Delhi, India. *Turkiye Klinikleri J Biostat*, *8*(1), 13-20.

[9] Lambert, P., Collett, D., Kimber, A., & Johnson, R. (2004). Parametric accelerated failure time models with random effects and an application to kidney transplant survival. *Statistics in medicine*, *23*(20), 3177-3192.

[10] Wienke, A. (2010). *Frailty models in survival analysis*. CRC press.

[11] Chen, P., Zhang, J., & Zhang, R. (2013). Estimation of the accelerated failure time frailty model under generalized gamma frailty. *Computational Statistics & Data Analysis*, *62*, 171-180.

[12] Vaupel, J. W., Manton, K. G., & Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, *16*(3), 439-454.

[13] Nasejje, J. B., Mwambi, H. G., & Achia, T. N. (2015). Understanding the determinants of under-five child mortality in Uganda including the estimation of unobserved household and community effects using both frequentist and Bayesian survival analysis approaches. *BMC public health*, *15*(1), 1003.

[14] Yadav, A. K., & Yadav, R. J. (2016). A Study on Childhood Mortality Using Shared Frailty Modeling Approach. *JApSc*, *16*(1), 11-17.

[15] T. J.M, Random survival forests, Journal of Thoracic Oncology, 6(12), Tech. Rep.

[16] J. Ehrlinger, gg random forests: Exploring random forest survival, *arXiv preprint arXiv:1612.08974*,2016.

[17] Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. The annals of applied statistics, 2(3), 841-860.

[18] Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. BMC bioinformatics, 9(1), 307.

[19] Mulera. B, Survival analysis of cardiovascular diseases with an extension to competing risks, Master's thesis, Minessota State University, Mankato, 2017.

[20] Segal, M. R., & Bloch, D. A. (1989). A comparison of estimated proportional hazards models and regression trees. Statistics in Medicine, 8(5), 539-550.

[21] Ettarh, R., & Kimani, J. (2012). Determinants of under-five mortality in rural and urban Kenya.

[22] Mutunga, C. J. (2011). Environmental determinants of child mortality in Kenya. In Health inequality and development (pp. 89-110). Palgrave Macmillan, London.

[23] Ayiko, R., Antai, D., & Kulane, A. (2009). Trends and determinants of under-five mortality in Uganda. East African journal of public health, 6(2), 136-140.

[24] Bailey, M. (1988). Factors affecting infant and child mortality in rural Sierra Leone. Journal of tropical

pediatrics, 34(4), 165-168.

[25] Gyimah, S. O., Ezeh, A., & Fotso, J. C. (2012). Frailty models with applications to the study of infant deaths on birth timing in Ghana and Kenya. Quality & quantity, 46(5), 1505-1521.

[26] Akaike, H. (1987). Factor analysis and AIC. In Selected papers of hirotugu akaike (pp. 371-386). Springer, New York, NY.

[27] Mani, K., Dwivedi, S. N., & Pandey, R. M. (2012). Determinants of under-five mortality in Rural Empowered Action Group States in India: An application of Cox frailty model. International Journal of MCH and AIDS, 1(1), 60.

[28] Khan, J. R., & Awan, N. (2017). A comprehensive analysis on child mortality and its determinants in Bangladesh using frailty models. Archives of Public Health, 75(1), 58.

[29] Mehmood, T., Liland, K. H., Snipen, L., & Sæbø, S. (2012). A review of variable selection methods in partial least squares regression. Chemometrics and Intelligent Laboratory Systems, 118, 62-69.

[30] O'Hara, R. B., & Sillanpää, M. J. (2009). A review of Bayesian variable selection methods: what, how and which. Bayesian analysis, 4(1), 85-117.

[31] Claeskens, G., Croux, C., & Van Kerckhoven, J. (2006). Variable selection for logistic regression using a prediction-focused information criterion. Biometrics, 62(4), 972-979.

[32] 1 https://www.ke.undp.org/content/kenya/en/home/post-2015/mdgoverview/overview/mdg4.html