Health Services and Outcomes Research Methodology (2019) 19:145–174
https://doi.org/10.1007/s10742-019-00199-z

Modeling determinants of time-to-circumcision of girls: a comparison of various parametric shared frailty models

Daniel Biftu Bekalo[1]

Received: 23 January 2019 / Revised: 15 May 2019 / Accepted: 30 May 2019 / Published online: 7 June 2019

Abstract
Female genital mutilation (FGM), also known as female genital cutting or female circumcision, is one of the deeply rooted traditional practices, in which the external female genital organ is either partially or totally removed for non-medical reasons. In Ethiopia, FGM is widespread across the majority of regions and ethnic groups, having the highest national prevalence that leads them to various complications such as immediate urinary and genital tract infection, pain and hemorrhage, complications in childbirth and social, psychological and sexual complications. This study aimed to model and investigate the potential risk factors of time-to-circumcision of girls in Ethiopia using parametric shared frailty models where regional states of the girls were used as a clustering effect in the models. The data source for the analysis was the 2016 EDHS data collected from January 18, 2016 up to June 27, 2016 from which the survival information of 2930 girls on age at circumcision obtained. The gamma and inverse Gaussian shared frailty distributions with Exponential, Weibull and log-logistic baseline models was employed to analyze risk factors associated with age at circumcision using socio-economic and demographic factors. All the fitted models were compared by using AIC and BIC values from simulation study and actual dataset. The result revealed that about 22.4% of girls were circumcised and 77.6% were not circumcised. The median age at circumcision was 3 years. Based on AIC and BIC values from simulation experiment and graphical evidences, log-logistic model with inverse Gaussian shared frailty distribution preferred when compared with other models for age at circumcision dataset. The clustering effect was significant for modeling the determinants of time-to-circumcision of girls dataset. Based on the result of log-logistic inverse Gaussian shared frailty model, mothers and fathers educational level, place of residence and religion of parents were found to be the most significant determinants of age at circumcision of girls. The estimated acceleration factor for the group of mothers who had secondary and higher educational level were highly prolonged age at circumcision of girls by the factor of $\phi = 3.119$ and $\phi = 3.933$ respectively. The log-logistic model with inverse Gaussian shared frailty distribution described age at circumcision of girls better than other models and there was heterogeneity between the regions on age at circumcision. Improving parents access to education would be an important way approach for preventing girls' circumcision.

Keywords Survival data analysis · Time-to-circumcision · Shared frailty · Acceleration factor

Extended author information available on the last page of the article

Abbreviations

FMG	Female genital mutilation
PH	Proportional hazard
AFT	Accelerated failure time
KM	Kaplan–Meier
AIC	Akaike information criterion
BIC	Bayesian information criterion
CSA	Central Statistical Agency
EDHS	Ethiopia Demographic and Health Survey

1 Background

Female Genital Mutilation (FGM), also known as female genital cutting or female circumcision, is one of the deeply rooted traditional practices, in which the external female genital organ is either partially or totally removed for non-medical reasons (WHO 2008a). It is estimated that world-wide between 100 and 140 million women are thought to have undergone FGM and 3 million girls annually are thought to be at risk (WHO 2008b). The practice is primarily performed in Africa where more than 28 countries and more than 3 million girls are at risk of experiencing FGM (UNFPA 2013; Population Reference Bureau 2010). In East Africa; Somalia (98%), Djibouti (93%), Eritrea (89%) and Ethiopia (74%) have the highest FGM prevalence (WHO 2008c). In Ethiopia, FGM is widespread across the majority of regions and ethnic groups, having the highest national prevalence (Macfarlane and Dorkenoo 2014. The national estimated prevalence of FGM among girls and women (age 15–49 years) is 79.9%(Central Statistical Authority and ORC Macro 2001) and 74.3% (Central Statistical Agency Ethiopia and ORC Macro 2006), respectively. The prevalence is estimated to be highest in Afar (91.6%), Somali (97.3%) regions and Dire Dawa city adminstration (92.3%) (Central Statistical Agency Ethiopia and ORC Macro 2006). FGM is carried out on girls at different ages ranging from babies and toddlers to teenagers. It is frequently carried out in unsterile conditions by traditional practitioners. This is both the result of its traditional form and its unlawfulness in many places. Complications can include immediate urinary and genital tract infection, pain and hemorrhage, complications in childbirth and social, psychological and sexual complications (UNFPA 2013). The public health burdens of FGM include both consequences for women mortality and ongoing morbidity concerns through their life span. For the formulation of effective policy to aware the women about the risk of FGM on women health, it is crucial to study the effect of various socio-economic and demographic factors which affect time-to-circumcision of girls. Having these, this study examined factors associated to time-to-circumcision of girls using parametric survival models. Survival analysis is a statistical method for data analysis where the response variable is the time to the occurrence of an event, time-to-circumcision of girls in this study. Survival data is a term used for describing data that measured in a time to the occurrence of a given event of interest. In this study the event of interest is time-to-circumcision of girls. One of the major objectives of this analysis was to model and assess major risk factors responsible for female genital mutilation. Kaplan and Meier have got one important development in non-parametric methods (Kaplan and Meier 1958). The non-parametric methods work well for homogeneous samples; they do not determine whether certain variables are related to the survival times (Klein and Moeschberger 1997). The Cox PH model has the restriction that proportional hazards assumption holds with time-fixed

covariates; and it may not be appropriate in many situations and other modifications such as stratified Cox model or Cox model with time-dependent variables are required (Collett 2003). The Study subjects (circumcised girls) in this study, taken from clustered community and hence clustered circumcised girls survival data may be correlated at the regional level. In this study, Shared Frailty Models were explored assuming that circumcised girls within the same cluster (region) share similar risk factors. Frailty model is common to all individuals in the cluster and responsible for creating dependence between event times (Sastry 1997). The study used Parametric Accelerated Failure time (AFT) models (Exponential, Weibull and log-logistic) with gamma and inverse Gaussian shared frailty distributions in determining the factors which affect the time-to-circumcision of girls.

2 Methods

2.1 Data source

The data for this study was extracted from the published reports of Ethiopian Demographic and Health Survey (EDHS, 2016) which is obtained from Central Statistical Agency (CSA) (Central Statistical Agency 2016). It is the fourth survey conducted in Ethiopia as a part of the worldwide DHS project. The 2016 EDHS was designed to provide estimates for the health and demographic variables of interest for the following domains. Ethiopia as a whole; urban and rural areas (each as a separate domain); and 11 geographic administrative regions (9 regions and 2 city administrations), namely: Tigray, Affar, Amhara, Oromia, Somali, Benishangul-Gumuz, South Nations Nationalities and Peoples (SNNP), Gambela and Harari regional states and two city administrations, Addis Ababa and Dire Dawa. The principal objective of the 2016 EDHS was to provide current and reliable data on fertility and family planning behavior, child mortality, adult and maternal mortality, children's nutritional status, use of maternal and child health services, knowledge of HIV/AIDS, and prevalence of HIV/AIDS and anemia.

2.2 Study population and variables

The total number of sample 15,684 girls below the age of 16 were identified in the households of selected clusters (regions). There were cases in which information on the relevant variables was missing and these cases were excluded from the analysis. Thus, the analysis presented in this study on the risk factors of female circumcision was based on the 2930 girls aged less than 16 years old. The response variable in this study is time-to-circumcision of girls which is measured in years. It is measured as the length of time from birth to age at which the girl has been circumcised. At any point in time, the data include observations in one of the following three categories: (1) those events (FGM) have occurred, (2) those events (FGM) which have not yet occurred but is likely to occur in the future, and (3) those events (FGM) could not occur and may never occur. The events that belongs to second and third categories were taken as right censored events. Risk factors of FGM from birth to the age or period at which girls were exposed to cutting (in years), following reporting by the girl's mother in the EDHS surveys in which regional state of the girls has been considered as a clustering effect in all models with shared frailty distributions were analyzed. Thus, this study attempts to include socio-economic, demographic and environment related factors that are assumed as a potential determinants of FGM adopted

from literature reviews and their theoretical justification. Possible explanatory variables for FGM includes mother education, religion, household socio-economic status, residence, exposure to media, employment status, and the father's education.

2.3 Survival data analysis

Survival analysis is a collection of statistical procedures for data analysis for which the outcome variable of interest is the time until an event occurs. By time, we mean years, months, weeks, or days from the beginning of follow-up of an individual until an event occurs; alternatively, time can refer to the age of an individual when an event occurs. By event, we mean death, disease incidence, relapse from remission, recovery (e.g., return to work) or any designated experience of interest that may happen to an individual. The use of survival analysis, as opposed to the use of other statistical method, is most important when some subjects are lost to follow up or when the period of observation is finite certain patients may not experience the event of interest over the study period. In this latter case one cannot have complete information for such individuals. These incomplete observations are referred to as being censored. Most survival analyses consider a key analytical problem of censoring. In essence, censoring occurs when we have some information about individual survival time, but we do not know the survival time exactly (Aalen et al. 2008).

2.3.1 Survival function

The survival function is defined to be the probability that the survival time of a randomly selected subject is greater than or equal to some specified time. Thus, it gives the probability that an individual surviving beyond a specified time. Let T be a continuous random variable associated with the survival times, t be the specified value of the random variable T and $f(t)$ be the underlying probability density function of the survival time T. The cumulative distribution function $F(t)$, which represents the probability that a subject selected at random will have a survival time less than some stated value t, is given by Cox (1972);

$F(t) = P(T < t) = \int_0^t f(u)du$, where; $t \geq 0$, the survivor function $S(t)$, is given by;

$S(t) = P(T \geq t) = 1 - F(t)$, where; $t \geq 0$, the relationship between $f(t)$ and $S(t)$ is given by;

$$f(t) = \frac{d}{dt}F(t) = \frac{d}{dt}(1 - S(t)) = \frac{-d}{dt}S(t) \geq 0$$

That is, the survival function gives the probability of surviving or being event free beyond time t. Because $S(t)$ is a probability, it is positive and ranges from 0 to 1. It is defined as $S(0) = 1$ that is, at the start of the study, since no one has experienced the event yet, the probability of surviving past time 0 is one and as t approaches positive infinity, $S(t)$ approaches 0 that is, theoretically, if the study period increased without limit, eventually nobody would survive, so the survivor curve must eventually converge to zero.

2.3.2 Hazard function

The hazard function $h(t)$ gives the instantaneous potential for failing at time t, given the individual has survived up to time t. This is the conditional probability of experiencing the event of interest within a very small time interval of size Δt having survived up to

time t. It is a measure of the probability of failure during a very small interval, assuming that the individual has survived at the beginning of the interval. In addition, it is not a probability as it does not lie between 0 and 1. The hazard function, $h(t) \geq 0$ is given as Cox (1972);

$$h(t) = \lim_{\Delta t \to \infty} \frac{p(t \leq T \leq t + \Delta t | t \geq t)}{\Delta t}$$

By applying the theory of conditional probability, the hazard function can be expressed in terms of the underlying probability density function and the survivor function becomes:

$$h(t) = \frac{f(t)}{S(t)} = \frac{-d}{dt} lnS(t)$$

The corresponding cumulative hazard function, $H(t)$, is defined as:

$$H(t) = \int_0^t h(u)du = -lnS(t), \quad \text{then;} \quad S(t) = exp(-H(t)) \quad \text{and} \quad f(t) = h(t)S(t)$$

The survival function is most useful for comparing the survival progress of two or more groups while the hazard function gives a more useful description of the risk of failure at any time point.

2.3.3 The Kaplan–Meier estimator of survival function

The Kaplan–Meier (KM) estimator is the standard non parametric estimator of the survival function, S(t), proposed by Kaplan and Meier (1958) which is not based on the actual observed event and censoring times, but rather on the ordered in which events occur. It is also called the Product-Limit estimator. KM estimator incorporates information from all of the observations available, both censored and uncensored, by considering any point in time as a series of steps defined by the observed survival and censored times. When there is no censoring, the estimator is simply the sample proportion of observations with event times greater than t. The technique becomes a little more complicated but still manageable when censored times are included. Let ordered survival times are given by $0 \leq t1 \leq t2 \leq tj \leq \infty$ then;

$$\hat{S}(t) = \begin{cases} 1, & \text{if} \quad t < t_1 \\ \prod_{j:t_j \leq t} \left[1 - \frac{d_j}{r_j}\right], & \text{if} \quad t \geq t_1 \end{cases}$$

where d_j is the observed number of events at time t_j and r_j is the number of individuals at risk at time t_j. The Kaplan–Meier estimator, $\hat{S}(t)$ is a step function which jumps at the observed event times. The size of the jump at a certain event time t_j depends on the number of events observed at t_j, as well as on the pattern of the censored event times before t_j.

2.4 Accelerated failure time models

Although parametric models are generally applicable to analyze survival data, there are relatively few probability distributions for the survival time that can be used with these models. In these situations, the accelerated failure time model (AFT) is an alternative to the PH model for the analysis of survival time data. Under AFT models we measure the

direct effect of the explanatory variables on the survival time instead of hazard. This characteristic allows for an easier interpretation of the results because the parameters measure the effect of the correspondent covariate on the mean survival time. The members of the AFT model considered in this study are the Exponential AFT, Weibull AFT and log-logistic AFT models. The AFT models are named for the distribution of T rather than the distribution of $log(T)$.

2.4.1 Weibull accelerated failure time model

The Weibull distribution (including the exponential distribution as a special case when the shape parameter is equal to one) can be parameterized as an AFT model, and they are the only family of distributions to have this property. The results of Weibull model can therefore be interpreted in either framework (Klein and Moeschberger 2003). Then the Weibull distribution is very flexible model for time-to-event data. It has a hazard rate which is monotone increasing, decreasing, or constant. The AFT representation of the survival and hazard function of the Weibull model is given by:

$$S_{\varepsilon_i}(t) = exp\left(-exp\left(\frac{log(t) - (\mu + \alpha' x)}{\sigma}\right)\right) = exp\left(-exp\left(\frac{-(\mu + \alpha' x)}{\sigma} t^{\frac{1}{\sigma}}\right)\right)$$

$$h_i(t) = \frac{1}{\sigma} t^{\frac{1}{\sigma}-1} exp\left(\frac{-\mu - \alpha' x}{\sigma}\right)$$

2.4.2 Log-logistic accelerated failure time model

The log-logistic distribution has a fairly flexible functional form, it is one of the parametric survival time models in which the hazard rate may be either decreasing, increasing, or hump-shaped, that is it initially increases and then decreases. In cases where one comes across to censored data, using log-logistic distribution is mathematically more advantageous than other distributions. According to the study of Gupta and Kundu (1999), the log-logistic distribution proved to be suitable in analyzing survival data conducted by Cox (1972), Cox and Oakes (1984), Bennett (1983) and O'Quigley and Stare (2002). The cumulative distribution function can be written in closed form is particularly useful for analysis of survival data with censoring (Bennett 1983). The log-logistic distribution is very similar in shape to the Log-normal distribution, but is more suitable for use in the analysis of survival data. The log-logistic model has two parameters λ and ρ, where λ is the scale parameter and ρ is the shape parameter. Its probability density function is given by O'Quigley and Stare (2002);

$$f(t) = \frac{\lambda \rho t^{\rho-1}}{(1 + \lambda \rho t^{\rho})^2}$$

The corresponding survival and hazard functions are given by;

$$S(t) = \frac{1}{1 + \lambda t^{\rho}}, \quad \text{and} \quad h(t) = \frac{\lambda \rho t^{\rho-1}}{1 + \lambda t^{\rho}}$$

where $\lambda \epsilon R, \quad \rho > 0$, when $\rho \leq 1$, the hazard rate decreases monotonically and when $\rho > 1$, it increases from zero to its maximum point and then decreases to zero. Suppose that the

survival times have log-logistic distribution with parameter λ and ρ, under the AFT model, the hazard function for the ith individual is:

$$h_i(t/x) = h_0(texp(-\alpha' x_i))exp(-\alpha' x_i) = \frac{\rho exp((\lambda)texp(-\alpha' x_i))}{1 + exp(\lambda)\{texp(-\alpha' x_i)\}^{\rho}}$$

The log-logistic AFT model with a covariate x is given by;$Y = log(T) = \mu + \alpha x_i + \sigma\varepsilon$, where; $\alpha' = (\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_p)$; ε has standard logistic distribution. The survival and hazard with covariate x is given as follows:

$$S_T(t/x) = \frac{1}{1 + \lambda exp(\beta' x)t^{\rho}} = \frac{1}{1 + exp(log\lambda + \beta' x)}$$
$$h_t(t/x) = \frac{\rho t^{\rho-1}\lambda exp(\alpha' x)}{1 + \lambda exp(\alpha' x)t^{\rho}} = \frac{\rho t^{\rho-1}\lambda exp(\alpha' x)}{1 + exp(log\lambda + \alpha' x)}$$

To interpret the factor $exp(\beta' x)$ for log-logistic model, one can notice that the odds of survival beyond time t for log-logistic model is given by: $\frac{S_T(t)}{1-S_T(t)}$

We can see that the log-logistic distribution has the proportional odds property. So this model is also a proportional odds model, in which the odds of an individual surviving beyond time t are expressed as:

$$\frac{S_T(t)}{1 - S_T(t)} = exp(-\alpha' x)\frac{S_0(t)}{1 - S_0(t)}$$

The factor $exp(-\alpha' x)$ is an estimate of how much the baseline odds of survival at any time changes when individual has covariate x. And $exp(\alpha' x)$ is the relative odds of experiencing the event for an individual with covariate x relative to an individual with the baseline characteristics. As this representation of log-logistic regression is as accelerated failure time model with a log-logistic baseline survival function, then the log-logistic model is the only parametric model with both a proportional odds and an accelerated failure-time representation. If T_i has a log-logistic distribution, then ε_i has a logistic distribution. The survival function of log-logistic distribution is given by Collett (2003):

$$S_{\varepsilon_i}(\varepsilon) = \frac{1}{1 + exp(\varepsilon)}$$

Then, the AFT representation of log-logistic survival function is given by:

$$S_T(t) = \left[1 + t^{\frac{1}{\sigma}} exp\left(\frac{-\mu - \alpha' x}{\sigma}\right)\right]^{-1}$$

And the associated hazard function for the ith individual is given by;

$$h_t(t) = \frac{1}{\sigma t}\left[1 + t^{\frac{-1}{\sigma}} exp\left(\frac{-\mu - \alpha' x}{\sigma}\right)\right]^{-1}$$

If the plot of $log[\frac{1-s(t)}{S(t)}]$ against $log(t)$ is linear,the log-logistic distribution is appropriate for the given data set.

2.4.3 Parameter estimation

Parameters of AFT models can be estimated by maximum likelihood method. The likelihood of n observed survival times, $t_1, t_2, t_3, \ldots, t_n$, the likelihood function for right censored data is given by:

$$L(\alpha, \mu, \sigma) = \Pi_{j=1}^{n} f_i(t_i)^{\delta_i} * S_i(t_i)^{1-\delta_i}$$

where $f_j(t_j)$ the density function of the ith individual at time t_i , $S_i(t_i)$ is the survival function of the ith individual at time t_i, δ_i is indicator variable. The logarithm of the above equation yields;

$$log(\alpha, \mu, \sigma) = \sum_{j=1}^{n} = \{-\delta_i log(\delta t_i + \delta_i log f_i(x_i) + (1 - \delta_i) log S_i(W_i))\}$$

where $W_i = \{logt_i - \frac{\mu+\alpha_{1i}x_{1i}+\cdots+\alpha_{pi}x_{pi}}{\delta}\}$, $Z = \{z_{ji}\}$ is vector of covariates for the jth subject. The maximum likelihood parameters estimates are found by using Newton–Raphson procedure.

2.5 Shared frailty model

The frailty approach is a statistical modeling concept which aims to account for heterogeneity, caused by unmeasured covariates. In statistical terms, a frailty model is a random effect model for time-to-event data, where the random effect (the frailty) has a multiplicative effect on the baseline hazard function (Wienke et al. 2003). Vaupel et al. (1979) used the frailty approach to derive the individual hazard function based on the population hazard function obtained from life tables.

The shared frailty approach assumes that all failure times in a cluster are conditionally independent given the frailties. The value of the frailty term is constant over time and common to all individuals in the cluster, and thus it is responsible for creating dependence between event times in a cluster. This dependence is always positive in shared frailty models. Conditional on the random effect, called the frailty denoted by u_i , the survival times in cluster i $(1 \leq i \leq n)$ are assumed to be independent and the proportional hazard frailty model assumes:

$$h_{ij}(t/x_{ij}, u_i) = h_0(t) exp(\beta' x_{ij} + u_i)$$

where i indicates the ith cluster and j indicates the jth individual for the ith cluster, $h_0(.)$ is the baseline hazard function, u_i the random term of all the subjects in cluster i, x_{ij} the vector of covariates for subject j in cluster i, and β the vector of regression coefficients. If the proportional hazards assumption does not hold, the accelerated failure time frailty model which assumes:

$$h_{ij}(t/x_{ij}, u_i) = h_0(exp(\beta' x_{ij} + u_i)) exp(\beta' x_{ij} + u_i)$$

If the number of subjects n_i is 1 for all groups, the univariate frailty model is obtained (Wienke et al. 2010); otherwise the model is called the shared frailty model because all subjects in the same cluster share the same frailty value (Hougaard 2012; Duchateau and Janssen 2008). Let us assume $Z = exp(u_i)$ and assume Z has the gamma or the inverse

Gaussian distribution, so that the hazard function depends upon this frailty that acts multiplicatively on it. Shared frailty models are very important in analyzing multivariate or clustered survival data. Shared frailty model assumes that all individuals in a subgroup or pair share the same frailty $Z_i (i = 1, 2, \ldots, n)$, and because of this it is called shared frailty model, but frailty from group to group may differ. Shared frailty model is similar to the individual frailty model except the only difference is that frailty is now shared among the n_i observations in the ith group. Proportional hazard shared frailty model and accelerated failure time shared frailty model assumes:-

$$h_{ij}(t) = Z_i h_0(t) exp(\beta' x_{ij}), \quad \text{and} \quad h_{ij}(t) = Z_i h_0(t) exp(\beta' x_{ij})(Z_i(\beta' x_{ij} t)) \quad \text{respectively.}$$

2.5.1 Gamma shared frailty distribution

The gamma distribution has been widely applied as a mixture distribution by Greenwood and Yule (1920) and Hougaard (2012). From a computational point of view, it fits very well into survival models, because it is easy to derive the formulas for any number of events. The gamma frailty distribution has been widely used in parametric modeling of intra-cluster dependency because of its simple interpretation, flexibility and mathematical tractability (Vaupel et al. 1979; Clayton and Cuzick 1985). To make the model identifiable, we restrict that expectation of the frailty equals one and variance be finite, so that only one parameter needs to be estimated. Thus, the distribution of frailty Z is the one parameter gamma distribution. Under the restriction, the corresponding density function and Laplace transformation of gamma distribution is given by Gutierrez (2002):

$$f_z(Z) = \frac{Z_i^{\left(\frac{1}{\theta}\right)-1}}{\theta^{\frac{1}{\theta}}\Gamma\left(\frac{1}{\theta}\right)} exp\left(\frac{-Z_i}{\theta}\right), \theta > 0$$

where $\Gamma(.)$ is the gamma function, it corresponds to a Gamma distribution $Gam(\mu, \theta)$ with μ fixed to 1 for identifiability and its variance is θ. The associated Laplace transform is:

$$L(u) = \left(1 + \frac{u}{\theta}\right)^{-\theta}, \theta > 0$$

Note that if $\theta > 0$, there is heterogeneity. So the large values of θ reflect a greater degree of heterogeneity among groups and a stronger association within groups. The conditional survival and hazard function of the gamma frailty distribution is given by Gutierrez (2002):

$$S_\theta(t) = [1 - \theta ln(S(t))]^{-\frac{1}{\theta}}, \quad \text{and} \quad h_\theta(t) = h(t)[1 - \theta ln(S(t))]^{-1}$$

where $S(t)$ and $h(t)$ are the survival and the hazard functions of the baseline distributions. For the Gamma distribution, the Kendall's Tau (Hougaard 2012), which measures the association between any two event times from the same cluster in the multivariate case. It is an overall measure of dependence and independent of transformations on the time scale and the frailty model used. The associations within group members are measured by Kendall's, which is given by:

$$\tau = \frac{\theta}{\theta + 2} \epsilon(0, 1)$$

2.5.2 Inverse Gaussian shared frailty distribution

Similar to the gamma frailty model, simple closed-form expressions exist for the unconditional survival and hazard functions, this makes the model attractive. The probability density function of an inverse Gaussian shared distributed random variable with parameter $\theta > 0$ is given by:

$$f_Z(Z_i) = \left(\frac{1}{2\pi\theta}\right)^{\frac{1}{2}} Z_i^{-3/2} exp\left(\frac{-(Z_i-1)^2}{2\theta Z_i}\right), \theta > 0, z > 0$$

For identifiability, we assume Z has expected value equal to one and variance θ. The Laplace transformation of the inverse Gaussian distribution is:-

$$L(s) = exp\left[\frac{1-(1+2\theta s)^{\frac{1}{2}}}{\theta}\right], \theta > 0, s > 0$$

For the inverse Gaussian frailty distribution the conditional survival function is given by Gutierrez (2002):

$$S_\theta(t) = exp\left\{\frac{1}{\theta}\left(1-[1-2\theta ln\{S(t)\}]^{\frac{1}{2}}\right)\right\} \theta > 0$$

For the inverse Gaussian frailty distribution the conditional hazard function is given by Gutierrez (2002):

$$h_\theta(t) = h(t)[1-2\theta ln\{S(t)\}]^{\frac{-1}{2}}, \theta > 0$$

where $S(t)$ and $h(t)$ are the survival and the hazard functions of the baseline distributions. With multivariate data, an Inverse Gaussian distributed frailty yields a Kendall's Tau given by:

$$\tau = \frac{1}{2} - \frac{1}{\theta} + 2\frac{exp(2/\theta)}{\theta^2}\int_{2/\theta}^{\infty}\frac{exp(-u)}{u}du, \quad \text{where} \quad \tau\epsilon(0,1/2).$$

2.5.3 Parameter estimation

Under assumptions of non-informative right -censoring and of independence between the censoring time and the survival time random variables, given the covariate information, the marginal log-likelihood of the observed data is given by Gutierrez (2002):

$$l_{marg}(\Psi,\beta,\theta;Z,X) = \prod_{i=1}^{s}\left[\left(\prod_{j=1}^{n_i}(h_0(y_{ij})exp(X_{ij}^T\beta))^{\delta_{ij}}\right)X \int_0^{\infty} Z_i^{d_i} exp\left(-Z_i\sum_{j=1}^{n_i}H_0(y_{ij})exp(X_{ij}^T\beta)\right)f(Z_i)dz_i\right]$$
$$= \prod_{i=1}^{s}\left[\left(\prod_{j=1}^{n_i}(h_0(y_{ij})exp(X_{ij}^T\beta))^{\delta_{ij}}\right)X(-1)^{d_i}L^{(d_i)} \left(\sum_{j=1}^{n_i}H_0(y_{ij})exp(X_{ij}^T\beta)\right)\right]$$

Taking the logarithm, the marginal likelihood is:

$$l_{marg}(\Psi,\beta,\theta;Z,X) = \Sigma_{i=1}^{s}\left\{\left[\sum_{j=1}^{n_i}\delta_{ij}(log(h_0(y_{ij})) + X_{ij}^T\beta)\right] + log\left[(-1)^{d_i}L^{d}\left(\left[\sum_{j=1}^{n_i}H_0(y_{ij})exp(X_{ij}^T\beta)\right]\right)\right]\right\}$$

where $d_i = \sum_{j=1}^{n_i}\delta_{ij}$ is the number of events in the ith cluster, and $L^{(q)}(.)$ is the qth derivative of the Laplace transform of the frailty distribution defined as:

$$L(s) = E[exp(-Zs)] = \int_0^{\infty} exp(-Z_i s)f(Z_i)dz_i, s \geq 0$$

where Ψ represents a vector of parameters of the baseline hazard function, β the vector of regression coefficients and θ the variance of the random effect. Estimates of Ψ, β, θ are obtained by maximizing the marginal log-likelihood of the above; this can be done if one is able to compute higher order derivatives $L^{(q)}(.)$ of the Laplace transform up to $q = max\{d_1, \ldots, d_s\}$.

2.5.4 Model selection

For comparing models that are not nested, the Akaike's information criterion (AIC) and Bayesian information criterion (BIC) are used which are respectively defined as:

$$AIC = -2log(L) + 2(k + c + 1) \quad \text{and} \quad BIC = -2log(L) + log(nk)$$

where k is the number of covariates,n is sample size and c the number of model specific distributional parameters. This study used the AIC and BIC to compare various candidates of non- nested parametric models. The preferred model is the one with the lowest values of the AIC and BIC (Wit et al. 2012).

2.6 Model diagnosis

2.6.1 Checking the adequacy of parametric baselines

The graphical methods can be used to check if a parametric distribution fits the observed data. Model with the Weibull baseline has a property that the $log(-log(S(t)))$ is linear with the log of time, where $S(t) = exp(-\lambda t^{\rho)})$. Hence, $log(-log(S(t))) = log(\lambda) + \rho log(t)$. This property allows a graphical evaluation of the appropriateness of a Weibull model by plotting $loq(-log(\hat{S}(T)))$ versus $log(t)$ where $\hat{S}(t)$ is Kaplan–Meier survival estimate (Datwyler and Stucki 2011). The log-failure odd versus log time of the log-logistic model is linear. Where the failure odds of log-logistic survival model can be computed as:

$$\frac{1 - s(t)}{s(t)} = \frac{\frac{\lambda t^{\rho}}{1+\lambda t^{\rho}}}{\frac{1}{1+\lambda t^{\rho}}} = \lambda t^{\rho}$$

Therefore, the log-failure odds is given by: $log(\frac{1-S(t)}{S(t)}) = log(\lambda t^{\rho}) = log(\lambda) + \rho log(t)$. Therefore, the appropriateness of model with the log-logistic baseline can graphically be evaluated by plotting $log(\frac{\hat{S}(t)}{1-\hat{S}(t)})$ versus $log(time)$ where $\hat{S}(t)$ is Kaplan–Meier survival estimate (Datwyler and Stucki 2011). If the plot is straight line, log-logistic distribution will fit the given dataset well. If the plot $\frac{\hat{S}(t)}{1-\hat{S}(t)}$ against t is linear, the Exponential distribution will be appropriate for the given data set.

2.6.2 Using residual plots

For the parametric regression problem, analogs of the semi parametric residual plots can be made with a redefinition of the various residuals to incorporate the parametric form of the baseline hazard rates (Klein and Moeschberger 2003). The first such residual is the Cox–Snell residual that provides a check of the overall fit of the model. The Cox–Snell residual, r_j, is defined by:

$$r_j = \hat{H}(T_j|X_j)$$

where $\hat{H}$ is the cumulative hazard function of the fitted model. If the model fits the data, then the r_j's should have a standard ($\lambda = 1$) exponential distribution, so that a hazard plot of r_j versus the Nelson-Aalen estimator of the cumulative hazard of the r_j's should be a straight line with slope 1.

3 Results

3.1 Summary statistics

A total of 2930 girls were included in the study from nine regional states and two city administrations. The time interval between the girls' date of birth and the time circumcision took place was an interest of this study. Of all 2930 girls considered 655 (22.4%) were circumcised and the rest of 2275 (77.6%) did not experience circumcision between the age of 1 year and 15 years. The median age at circumcision was 3 years while the minimum

and maximum observed event time was 1 year and 15 years, respectively. Furthermore, among 23.7% girls were circumcised in the 1 year age,which indicates that circumcision of girls in Ethiopia takes palace at the early ages of the daughters. From Table 1, out of 2930 girls' mothers, 2312 (78.9%) lives in rural while 618 (21.1%) of them were residing in urban. About 1999 (68.2%) of the girls' mothers had no work and 931 (31.8%) had a work. The proportion of mothers who have access to media was 2063 (70.4%) while 867 (29.6%) of mothers had no access to media. Majority of the girls' mothers 1375 (46.9%) were in the ages of 25–34, while 1105 (37.7%) and 450 (15.4%) were in the ages of 35–49 and 15–24 respectively.

Out of 2930 total number of mothers of the girls, 1466 (50.0%) were Christian, 1441 (49.2%) Muslim, and 23 (0.8%) of them were from other religious group. About 1376 (47.0%) of the mothers wealth indexes were classified as poor while 435 (14.8%) had medium income and 1119 (38.2%) were rich. The study revealed that educational attainments of girls' mothers; about 1901 (64.9%) had no education while 739 (25.2%) had primary education and the remaining 290 (9.9%)had attended secondary and higher education. From the total number of fathers who were included in the study, 1509 (51.5%) of them were illiterate (no education), 903 (30.8%) of the fathers had attended primary education and the remaining 518 (17.7%) were secondary and higher education level. The bar chart of Fig. 1 reveals that circumcision (event) of girls in Ethiopia is highly prevalent in Afar region followed by Amahara and Somali regions compared to the other regions in the country.

Plots of the KM curves to the survival and hazard experience of time- to- circumcision of girl is shown in Fig. 2, the survival plot decreases throughout the given time. This implies that the more girls approach to the age 15, they are more likely to get circumcised.

The survival plots for time-to-circumcision of girls by place of residence, mothers' age,wealth index and religion are shown in Fig. 3. The plot indicates that the risk that the girls being circumcised is similar for those girls whose mothers lived either in rural or urban at birth. However, the difference becomes visible at the middle till the end of the curves. From the middle point to the end of the curves, the survival plot of circumcision of girls whose mothers lived in rural is below that of girls whose mothers lived in urban. This implied that the risk that the girls whose mothers lived rural being circumcised is much higher than those girls whom their mothers are from urban. The plot also depicts that as there is highest risks of circumcision for those girls whose mothers age group is 35–49 compared to those girls whose mothers age group belongs to other categories. The plot demonstrates that girls whose mothers wealth index is poor have highest risk towards circumcision compared to girls whose mothers are from middle and rich wealth index. The plot also reveals that the risk of circumcision is high for those girls whose mothers are muslim and other religion compared to those whose mothers are Christian.

The survival plot of time- to- circumcision of girls by mothers education, fathers education,exposure to media and employment status are shown in Fig. 4. From this plot we can observe that the risk that girls being circumcised after birth is similar for all groups at the beginning. But the difference becomes visible at the middle of the curves. At the middle point of the curves, the survival plot of girl circumcision for those girls whose mothers have no education is below that of those girls whose mothers are in primary,secondary and higher education. This implied that the risk of girls circumcision for whom their mothers have education is much higher than those girls for whom their mothers belong to primary education and above. Also the plot shows that girls whose fathers have no education have highest risk of circumcision compared to those whose fathers education level belong to primary school and above. This plot suggested that the risk of girl circumcision is high

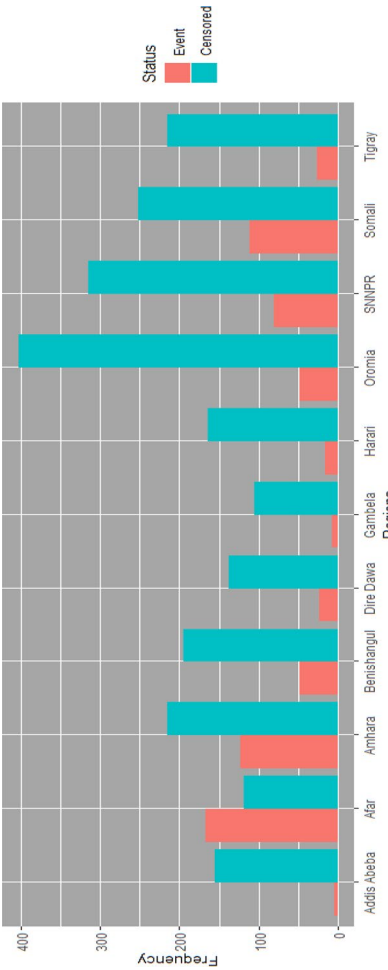

Fig. 1 Bar chart of circumcision status of girls from birth to date of interview by region

Table 1 Descriptive summary of FMG data

Covariates	Categories	Status		Total (%)
		Censored (%)	Event (%)	
Residence	Rural	1711 (75.2)	601 (91.8)	2312 (78.9)
	Urban	564 (24.8)	54 (8.2)	618 (21.1)
Employment status	No	1513 (66.5)	486 (74.2)	1999 (68.2)
	Yes	762 (33.5)	169 (25.8)	931 (31.8)
Exposer to media	No	1557 (68.4)	506 (77.3)	2063 (70.4)
	Yes	718 (31.6)	149 (22.7)	867 (29.6)
Age of mothers	15–24	394 (17.3)	56 (8.5)	450 (15.4)
	25–34	1128 (49.6)	247 (37.7)	1375 (46.9)
	35–49	753 (33.1)	352 (53.7)	1105 (37.7)
Religion	Christian	1228 (54.0)	238 (36.3)	1466 (50.0)
	Muslim	1031 (45.3)	410 (62.6)	1441 (49.2)
	Other	16 (0.7)	7 (1.1)	23 (0.8)
Wealth index	Poor	990 (43.5)	386 (58.9)	1376 (47.0)
	Middle	342 (15.0)	93 (14.2)	435 (14.8)
	Rich	943 (41.5)	176 (26.9)	1119 (38.2)
Mothers education	No education	1352 (59.4)	549 (83.8)	1901 (64.9)
	Primary	639 (28.1)	100 (15.3)	739 (25.2)
	Secondary	177 (7.8)	4 (0.6)	181 (6.2)
	Higher	107 (4.7)	2 (0.3)	109 (3.7)
Fathers education	No education	1030 (45.3)	479 (73.1)	1509 (51.5)
	Primary	765 (33.6)	138 (21.1)	903 (30.8)
	Secondary	272 (12.0)	23 (3.5)	295 (10.1)
	Higher	208 (9.1)	15 (2.3)	223 (7.6)
Status		2275 (77.6)	655 (22.4)	2930 (100)
Time		Minimum	Median	Maximum
		1	3	15

for those girls whose mothers are not exposed to media. Also the curves reveals that girls whose mothers are not employed have higher risk of circumcision compared to those girls whose mothers are employed.

3.1.1 Simulation study

In this study, survival data which accounted with five covariates with random values of β was simulated from Weibull distribution. The main goal of doing this was to see the performance of different survival data models for different parameter values. Three popular survival baseline models, Exponential, Weibull and log-logistic models with various frailty distributions were considered. Four binary covariates which were generated from Bernoulli distribution with success probability of 0.4, 0.25, 0.6 and 0.37 and one covariate from uniform distribution between the values of 1.5 and 4.5 were included. Random values for the

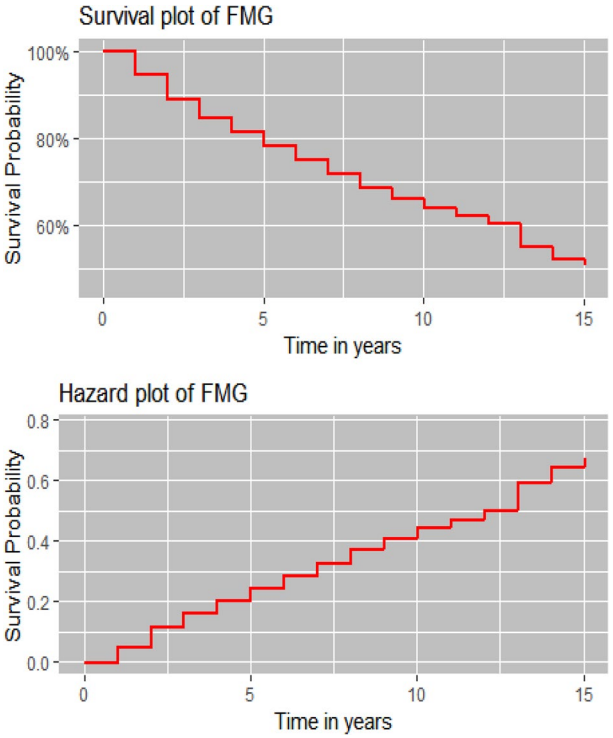

Fig. 2 The K–M plots of survival and hazard functions of FMG after marriage

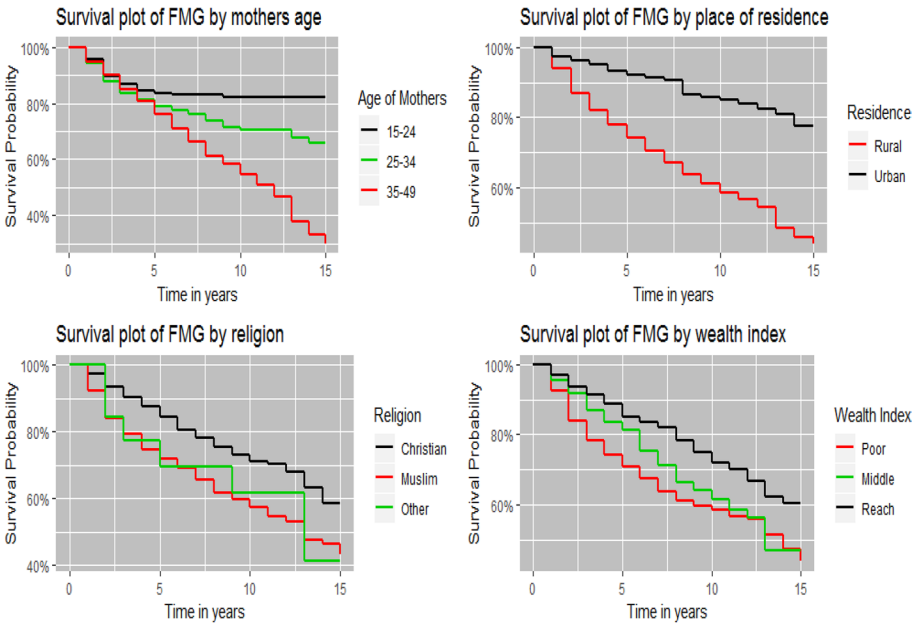

Fig. 3 Survival of time-to-circumcision of girls by place of residence, mothers age, wealth index and religion

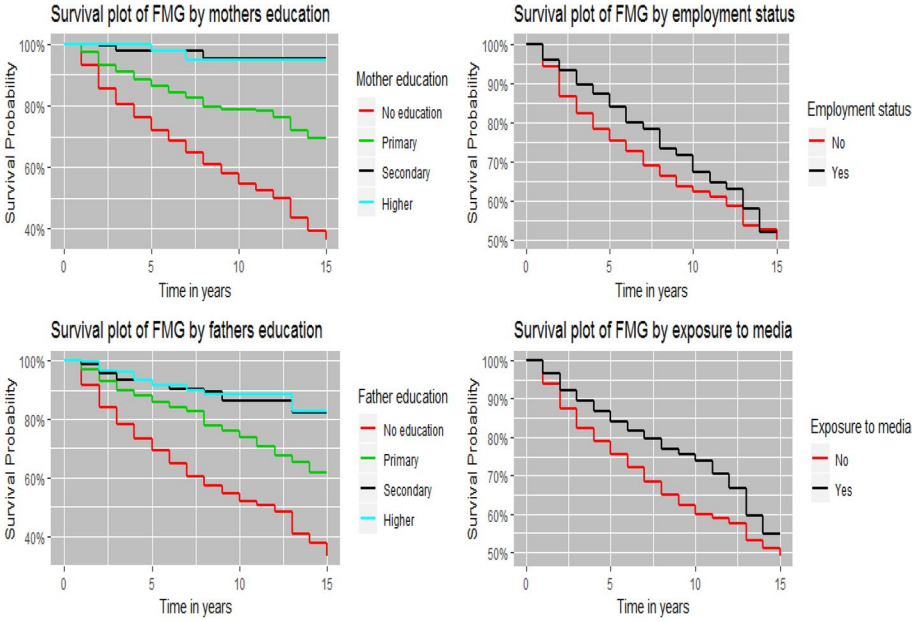

Fig. 4 Survival of time-to-circumcision of girls by mothers education, fathers education, exposure to media and employment status

parameters of covariates and ancillary parameters for the time to event distribution and time to censoring distribution in the simulations were considered. List of vectors indicating the effect of the corresponding covariate $(\beta) = (-0.423, 0.250, -1.702, 1.038, 0.902)$, ancillary parameter for the time to event distribution $(\beta_1) = 0.5$, β_0 parameter for the time to event distribution $= 0.268$, ancillary parameter for the time to censoring distribution $(\beta_1) = 0.5$ and β_0 parameter for the time to censoring distribution $= 1.368$ were considered. The trial was repeated for 20 times for sample size of 1000 with maximum time of follow-up 1825 days (5 years). In this study we were interested to explore how the available frailty distributions in the survival data models behave with different sets of parameters. AIC statistic was considered to compare the efficiency of each models. Figure 5 shows that log-logistic with inverse Gaussian shared frailty has superior performance than the others survival models, that is, Exponential and Weibull models.

As we can see from Fig. 5 of simulation experiment, the AIC and BIC values for all models (Exponential,Weibull and log-logistic) with no shared frailties are high compared to those models with gamma and inverse Gaussian shared frailty distributions. The AIC and BIC values of log-logistic gamma and inverse Gaussian shared frailty model are substantially small compared with Exponential and Weibull models with shared frailties which depicts the appropriateness of log-logistic inverse Gaussian model for survival data with shared frailty distribution (Fig. 6).

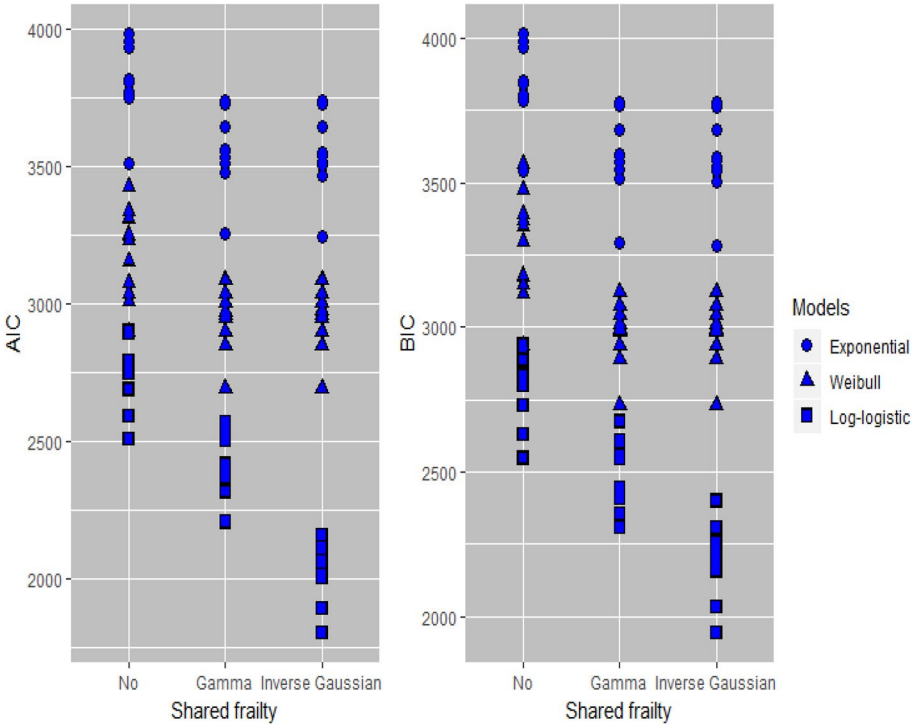

Fig. 5 AIC and BIC statistics of the three models when diferent shared frailties are considered (survival data simulated from Weibull distribution)

3.2 Accelerated failure time model results

3.2.1 Univariate analysis

This study used univariate analysis in order to see the effects of each covariate on time-to-circumcision of girls before proceeding to the multivariate analysis. The univariate analyses was fitted for every covariate by AFT models using different baseline distributions i.e., exponential, weibull and log-logistic. In all univariate analysis of baseline AFT models, age of mothers, place of residence, religion of mothers and education level of both mothers and fathers were significantly associated with girls circumcision while access to media, employment status and wealth index were not significant at 5% level of significance. Based on the univariate analysis, except those insignificant covariates, all explanatory variables are candidate predictors for further analysis.

3.3 Multivariate AFT analysis

The summary of multivariate analysis is given in Tables 2 and 3 and model comparisons were presented in Table 4 and Fig. 5. Accordingly, they suggested that log-logistic inverse Gaussian shared frailty AFT model with a minimum 3456.258 AIC value appears to be

Table 2 Results of survival AFT models based on different shared frailty types

Covariates	Estimates	No frailty				Estimates	Gamma shared frailty				Estimates	Inverse Gaussian shared frailty			
		SE	ϕ	95% CI for ϕ	p value		SE	ϕ	95% CI for ϕ	p value		SE	ϕ	95% CI for ϕ	p value
Exponential—accelerated failure-time form															
Residence															
Urban	0.582	0.165	1.789	[1.293, 2.477]	0.000	0.551	0.177	1.736	[1.226, 2.459]	0.002	0.553	0.176	1.740	[1.230, 2.461	0.002
Employment status															
Yes	0.077	0.091	1.080	[0.903, 1.292]	0.395	− 0.083	0.095	0.919	[0.762, 1.109]	0.384	− 0.085	0.095	0.918	[0.761, 1.107]	0.372
Exposure to media															
Yes	0.129	0.099	1.138	[0.937, 1.382]	0.190	0.103	0.100	1.109	[0.911, 1.350]	0.300	0.103	0.100	1.108	[0.910, 1.349]	0.303
Age															
25–34	− 0.389	0.150	0.677	[0.504, 0.909]	0.010	− 0.459	0.150	0.631	[0.470, 0.848]	0.002	− 0.461	0.150	0.630	[0.469, 0.847]	0.002
35–49	− 0.657	0.149	0.518	[0.387, 0.694]	0.000	− 0.768	0.149	0.463	[0.345, 0.621]	0.000	− 0.770	0.149	0.462	[0.345, 0.620]	0.000
Religion															
Muslim	− 0.352	0.086	0.703	[0.593, 0.833]	0.000	− 0.195	0.121	0.822	[0.648, 1.043]	0.107	− 0.189	0.121	0.827	[0.651, 1.050]	0.120
Other	− 0.229	0.386	1.257	[0.590, 2.681]	0.552	− 0.743	.393	0.475	[0.219, 1.027]	0.059	− 0.743	0.393	0.475	[0.219, 1.027]	0.059
Wealth index															
Middle	0.082	0.119	1.085	[0.858, 1.372]	0.494	− 0.166	0.130	0.846	[0.655, 1.093]	0.202	− 0.173	0.130	0.841	[0.651, 1.086]	0.185
Rich	− 0.100	0.111	0.904	[0.726, 1.126]	0.370	− 0.285	0.124	0.751	[0.588, 0.958]	0.022	− 0.290	0.124	0.747	[0.585, 0.955]	0.020
Mothers education															
Primary	0.318	0.121	1.374	[1.082, 1.745]	0.009	0.222	0.122	1.248	[0.982, 1.586]	0.069	0.221	0.122	1.247	[0.981, 1.585]	0.071
Secondary	0.318	0.121	5.881	[2.114, 16.360]	0.001	1.586	0.524	4.885	[1.748, 13.654]	0.002	1.587	0.524	4.893	[1.751, 13.670]	0.002
Higher	1.877	.742	6.535	[1.525, 28.010]	0.011	1.950	0.745	7.032	[1.630, 30.335]	0.009	1.956	0.745	7.071	1.639, 30.500]	0.009
Fathers education															
Primary	0.521	0.103	1.683	[1.373, 2.063]	0.000	0.365	0.108	1.441	[1.165, 1.781]	0.001	0.364	0.108	1.439	[1.164, 1.779]	0.001
Secondary	0.627	0.228	1.872	[1.195, 2.931]	0.006	0.550	0.229	1.733	[1.104, 2.719]	0.017	0.550	0.229	1.733	[1.104, 2.720]	0.017
Higher	0.403	0.289	1.496	[0.848, 2.638]	0.163	0.244	0.293	1.277	[0.718, 2.271]	0.405	0.246	0.293	1.278	[0.719, 2.274]	0.402
Theta (θ)						0.299	0.138				0.339	0.182			

Table 2 (continued)

Covariates	Estimates	No frailty				Estimates	Gamma shared frailty				Estimates	Inverse Gaussian shared frailty			
		SE	ϕ	95% CI for ϕ	p value		SE	ϕ	95% CI for ϕ	p value		SE	ϕ	95% CI for ϕ	p value
Weibull—accelerated failure-time form															
Residence															
Urban	0.486	0.140	1.625	[1.233, 2.143]	0.001	0.459	0.150	1.582	[1.178, 2.126]	0.002	0.460	0.149	1.585	[1.181, 2.127]	0.002
Employment status															
Yes	0.065	0.077	1.067	[0.918, 1.242]	0.393	− 0.071	0.081	0.931	[0.794, 1.091]	0.381	− 0.072	0.081	0.929	[0.793, 1.090]	0.370
Exposure to media															
Yes	0.115	0.083	1.122	[0.953, 1.321]	0.167	0.095	0.085	1.100	[0.931, 1.299]	0.263	0.094	0.085	1.099	[0.930, 1.298]	0.266
Age of mothers															
25–34	− 0.349	0.126	0.704	[0.549, 0.904]	0.006	− 0.405	0.127	0.666	[0.518, 0.855]	0.001	− 0.407	0.127	0.665	[0.518, 0.854]	0.001
35–49	− 0.553	0.126	0.574	[0.448, 0.736]	0.000	− 0.648	0.127	0.522	[0.407, 0.670]	0.000	− 0.650	0.127	0.521	[0.406, 0.669]	0.000
Religion															
Muslim	− 0.285	0.074	0.751	[0.649, 0.868]	0.000	− 0.151	0.102	0.859	[0.703, 1.05]0	0.140	− 0.146	0.102	0.863	[0.706, 1.057]	0.155
Other	− 0.162	0.325	0.850	[0.449, 1.609]	0.619	− 0.602	0.333	0.547	[0.284, 1.052]	0.071	− 0.601	0.333	0.547	[0.285, 1.052]	0.071
Wealth index															
Middle	0.062	0.101	1.064	[0.872, 1.297]	0.539	− 0.147	0.110	0.863	[0.694, 1.072]	0.184	− 0.153	0.111	0.858	[0.690, 1.066]	0.168
Rich	− 0.076	0.094	0.926	[0.769, 1.114]	0.418	− 0.235	0.105	0.790	[0.642, 0.971]	0.026	− 0.239	0.105	0.786	[0.639, 0.968]	0.024
Mothers education															
Primary	0.278	0.102	1.321	[1.079, 1.616]	0.007	0.201	0.103	1.223	[0.998, 1.499]	0.052	0.200	0.103	1.222	[0.997, 1.498]	0.053
Secondary	1.507	0.441	4.515	[1.899, 10.738]	0.001	1.359	0.444	3.893	[1.629, 9.301]	0.002	1.360	0.444	3.898	[1.632, 9.310]	0.002
Higher	1.599	0.627	4.948	[1.446, 16.929]	0.011	1.674	0.632	5.336	[1.545, 18.427]	0.008	1.679	0.632	5.361	[1.552, 18.513]	0.008
Fathers education															
Primary	0.446	0.088	1.563	[1.314, 1.858]	0.000	0.324	0.091	1.383	[1.155, 1.655]	0.000	0.323	0.091	1.382	[1.154, 1.654]	0.000
Secondary	0.532	0.193	1.702	[1.165, 2.487]	0.006	0.477	0.194	1.612	[1.100, 2.361]	0.014	0.477	0.194	1.612	[1.100, 2.362]	0.014

Table 2 (continued)

Covariates	Estimates	No frailty				Estimates	Gamma shared frailty				Estimates	Inverse Gaussian shared frailty			
		SE	ϕ	95% CI for ϕ	p value		SE	ϕ	95% CI for ϕ	p value		SE	ϕ	95% CI for ϕ	p value
Higher	0.333	0.244	1.396	[0.864, 2.255]	0.173	0.202	0.249	1.224	[0.750, 1.997]	0.418	0.203	0.249	1.225	[0.751, 2.000]	0.415
Theta (θ)						0.297	0.138				0.338	0.182			

ϕ Acceleration factor, *SE* standard error

Table 3 Results of survival AFT models based on different shared frailty types

Covariates	Estimates	No frailty				Estimates	Gamma shared frailty				Estimates	Inverse Gaussian shared frailty			
		SE	ϕ	95% CI for ϕ	p value		SE	ϕ	95% CI for ϕ	p value		SE	ϕ	95% CI for ϕ	p value
Log-logistic—accelerated failure-time form															
Residence															
Urban	0.506	0.137	1.659	[1.267, 2.174]	0.000	0.433	0.145	1.541	[1.159, 2.050]	0.003	0.421	0.144	1.524	[1.149, 2.023]	0.003
Employment status															
Yes	0.110	0.080	1.116	[0.953, 1.308]	0.172	0.060	0.086	1.062	[0.897, 1.258]	0.483	0.062	0.085	1.064	[0.900, 1.259]	0.464
Exposure to media															
Yes	0.136	0.087	1.146	[0.966, 1.360]	0.117	0.102	0.088	1.108	[0.932, 1.317]	0.243	0.105	0.087	1.111	[0.937, 1.318]	0.224
Age of mothers															
25–34	− 0.347	0.129	0.706	[0.547, 0.911]	0.007	− 0.222	0.121	0.800	[0.631, 1.016]	0.067	− 0.200	0.117	0.818	[0.650, 1.030]	0.088
35–49	− 0.446	0.129	0.640	[0.496, 0.825]	0.001	− 0.125	0.128	0.882	[0.685, 1.135]	0.330	− 0.095	0.122	0.908	[0.714, 1.156]	0.437
Religion															
Muslim	− 0.340	0.076	0.711	[0.612, 0.826]	0.000	− 0.429	0.092	0.651	[0.542, 0.781]	0.000	− 0.424	0.091	0.654	[0.546, 0.782]	0.000
Other	− 0.203	0.365	0.816	[0.398, 1.671]	0.579	− 0.675	0.416	0.508	[0.224, 1.151]	0.105	− 0.660	0.419	0.516	[0.226, 1.175]	0.115
Wealth index															
Middle	0.097	0.105	1.102	[0.897, 1.355]	0.352	0.030	0.110	1.031	[0.830, 1.280]	0.781	0.029	0.109	1.030	[0.830, 1.277]	0.786
Rich	− 0.071	0.101	0.931	0.763, 1.135]	0.482	− 0.116	0.110	0.890	[0.716, 1.105]	0.292	− 0.119	0.109	0.887	[0.715, 1.099]	0.275
Mothers education															
Primary	0.297	0.102	1.346	[1.102, 1.644]	0.004	0.188	0.100	1.207	[0.990, 1.470]	0.062	0.181	0.099	1.199	[0.985, 1.458]	0.069
Secondary	1.393	0.390	4.029	[1.875, 8.659]	0.000	1.150	0.310	3.159	[1.719, 5.805]	0.000	1.137	0.304	3.119	[1.716, 5.670]	0.000
Higher	1.454	0.550	4.283	[1.457, 12.590]	0.008	1.372	0.414	3.946	[1.752, 8.886]	0.001	1.369	0.403	3.933	[1.784, 8.669]	0.001
Fathers education															
Primary	0.471	0.089	1.601	[1.342, 1.910]	0.000	0.289	0.095	1.336	[1.107, 1.611]	0.002	0.278	0.094	1.321	[1.097, 1.591]	0.003
Secondary	0.543	0.185	1.721	[1.196, 2.479]	0.003	0.393	0.174	1.481	[1.052, 2.087]	0.024	0.382	0.171	1.465	[1.046, 2.051]	0.026
Higher	0.363	0.232	1.437	[0.911, 2.267]	0.118	0.294	0.210	1.342	[0.888, 2.026]	0.162	0.305	0.206	1.357	[0.905, 2.036]	0.139

Table 3 (continued)

Covariates	Estimates	No frailty				Estimates	Gamma shared frailty				Estimates	Inverse Gaussian shared frailty			
		SE	ϕ	95% CI for ϕ	p value		SE	ϕ	95% CI for ϕ	p value		SE	ϕ	95% CI for ϕ	p value
Gamma (γ)	1.720	0.022	0.720	[0.677, 0.765]		2.429	0.045	0.429	[0.348, 0.527]		4.399	0.042	0.399	[0.324, 0.490]	
Theta (θ)						2.007	0.758	2.007	[0.957, 4.209]		9.708	4.976	9.708	[3.555, 26.512]	
Likelihood-ratio test of theta (θ) = 0: $\chi^2_1 = 136.35$, Prob $\geq \chi^2_1 = 0.000$						Kendall's tau $\tau = 0.473$									

ϕ Acceleration factor, *SE* standard error

Table 4 Summary of AIC and BIC values for different survival AFT models with different shared frailties

Information criterions	Models	Frailty distributions		
		No frailty	Gamma shared frailty	Inverse Gaussian shared frailty
AIC	Exponential	3646.219	3527.396	3526.225
	Weibull	3619.913	3503.102	3501.976
	Log-logistic	3590.611	3471.312	3456.258
BIC	Exponential	3741.943	3620.103	3627.932
	Weibull	3721.620	3610.792	3609.666
	Log-logistic	3692.318	3579.001	3563.948

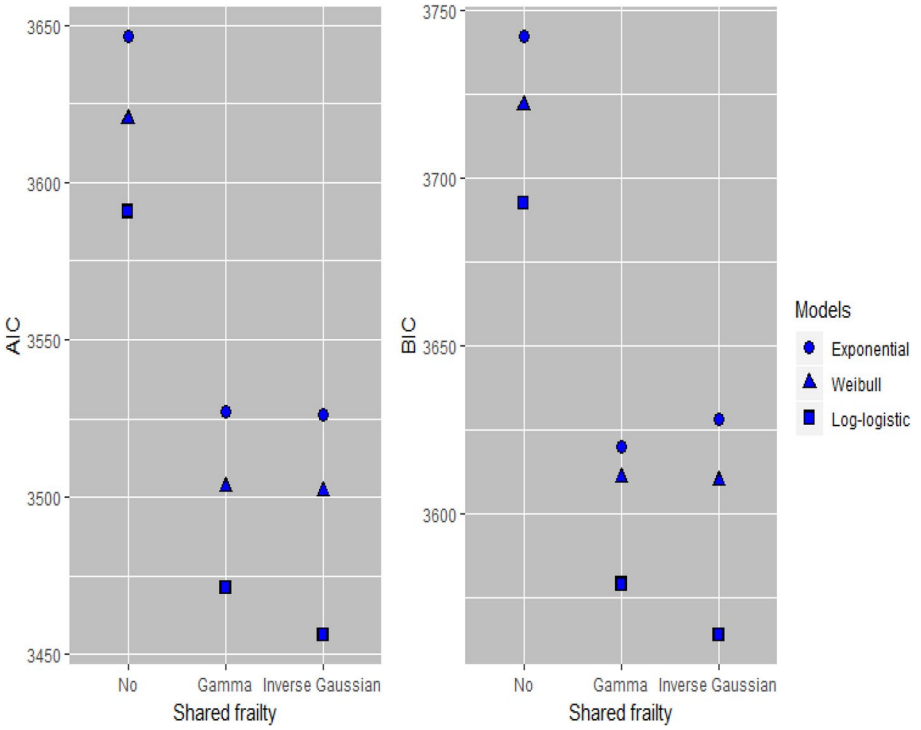

Fig. 6 AIC and BIC statistics of the three models when different shared frailties are considered (survival data obtained from time-to-circumcision dataset)

appropriate model compared with other models. This indicates it is more efficient model to describe determinant factors of time-to-circumcision of girls. The frailty in this model is assumed to follow a inverse gaussian distribution with mean 1 and variance equal to theta (θ). The heterogeneity in the population of the region which is used as a clusters are estimated by the selected model is $\theta = 9.708$ and the dependence within the clusters (region) is measured by Kendall's tau is $\tau = 0.473$. A variance of zero ($\theta = 0$) would indicate that the frailty component does not contribute to the model. A likelihood ratio test for the hypothesis

$\theta = 0$ is shown at the bottom of Table 3 and indicates a chi-square χ^2 value of 136.35 with one degree of freedom resulted a highly significant p value of 0.000. This implied that the frailty component had significant contribution to the model. The estimate of shape parameter in the log-logistic inverse Gaussian shared frailty model is $\gamma = 4.399$. This value shows the shape of hazard function is unimodal because the value is greater than unity i.e., it increases up to some time and then decreases. In this model all categorical variables were significant except wealth index, exposure to media and employment status. From Table 3 the confidence intervals of the acceleration factor for all significant categorical covariates do not include one at 5% level of significance. This shows that they were significant factors for determining the survival time-to-circumcision of girls in Ethiopia. However, from the variable of religion category, for those mothers who were following others religion, it was not significant when using Christian as the reference category with (p value = 0.115, $\phi = 0.516$, 95% CI = [0.226, 1.175]). Also, those mothers with primary education was insignificant by using no education as a reference category with (p value = 0.069, $\phi = 1.199$, 95% CI = [0.985, 1.458]). From the categories of fathers education, higher education was not significant when no education was used as reference category with (p value = 0.139, $\phi = 1.357$, 95% CI = [0.905, 2.036]). The estimated coefficient of the parameter for girls mothers who were residing in urban was 0.421. "The sign of the coefficient was positive which implies increase in log of survival time and hence, elongated expected duration of time-to-circumcision of girls whose mothers had lived in urban areas. The 95% confidence interval for acceleration factor of mothers educational levels was [0.985, 1.458], [1.716, 5.670] and [1.784, 8.669] for the group of primary, secondary and higher education's respectively. These confidence interval do not include one for secondary and higher education level. Indicating secondary and higher education were significantly relevant factors for the age at circumcision of girls by using uneducated mothers as a reference category. Accordingly, the age at circumcision of girls prolonged by a factor of ($\phi = 3.119$ and $\phi = 3.933$) for secondary and higher education respectively at 5% level of significance. Religion of mothers was statistically determining age at circumcision of girls. The time rate and 95% confidence interval of acceleration factors for religious group of mothers for the category of muslim was 0.654, [0.546, 0.782] when compared to the category of Christian religion as reference. The estimated coefficient of the parameters for girls whom their mothers were following muslim was − 0.424. The sign of the coefficient is negative which implies that decreasing log of survival time and hence, shorter expected duration of age at circumcision of girls. The time rate and 95% confidence interval for acceleration factor of girls fathers with education category of primary and secondary was 1.321, [1.097, 1.591] and 1.465, [1.046, 2.051] respectively by using uneducated fathers as a reference categories.

3.4 Model diagnostics

3.4.1 Checking model adequacy of parametric baselines using graphical methods

After the model has been fitted, it is desirable to determine whether a fitted parametric model adequately describes the data or not. Therefore, the appropriateness of model with Exponential baseline can be graphically evaluated by plotting cumulative hazard function versus time, the Weibull baseline by plotting log-cumulative hazard function versus log(time) and log-logistic baseline by plotting log-failure odd versus log(time). If the plot is linear, the given baseline distribution is appropriate for the given dataset. Accordingly, the

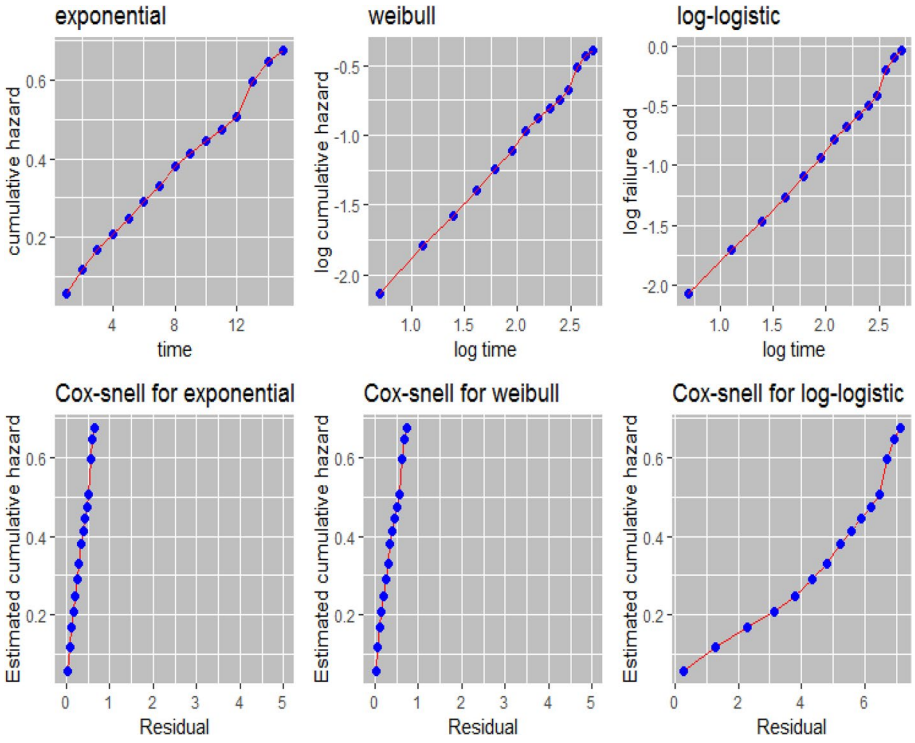

Fig. 7 Diagnosis plots of exponential, Weibull and log-logistic models

models plots are given in Fig. 7 and the plot for the log-logistic baseline distribution make a bit straight line better than Exponential and Weibull baseline distributions. This evidence also strengthens the decision made by AIC and BIC values of simulation experiment and actual dataset, that log-logistic baseline distribution is an appropriate for the given dataset.

3.4.2 Cox–Snell residuals plots

The Cox–Snell residuals are one way to investigate how well the model fits the data. The plots for fitted models of residuals for Exponential,Weibull and log-logistic to our data via maximum likelihood estimation with cumulative hazard functions are given in Fig. 7. If the model fits the data, the plot of cumulative hazard function of residuals against Cox–Snell residuals should be approximately a straight line with slope 1. The plot makes straight lines through the origin for log-logistic with inverse Gaussian share frailty distribution suggesting that it is appropriate for time-to-circumcision of girls dataset.

4 Discussion

The main goal of the study was modeling the determinants of age at circumcision of girls in Ethiopia using AFT parametric shared frailty models by considering three baseline distributions: Exponential, Weibull and log-logistic distributions and the frailty distributions:

Gamma share frailty and inverse Gaussian shared frailty. In this study, region was used as a clustering (frailty) effect on modeling the determinants of time-to-circumcision of girls in Ethiopian using 2016 EDHS data, in which the comparison of models was performed using AIC criteria where the model with minimum AIC value was accepted to be the best model for the given data set. From simulation study and actual dataset,we have found a small AIC value for log-logistic inverse Gaussian shared frailty model than its counter Exponential and Weibull models. Accordingly, log-logistic inverse Gaussian shared frailty model was selected as a best model. The clustering effect were also significant (p value < 0.000) in log-logistic inverse Gaussian shared frailty model. This showed that there is heterogeneity between regions by assuming girls within the same region share similar risk factors towards circumcision. That is, the correlation within regions cannot be ignored and clustering effect was important in modeling the hazard function. In this study the adequacy of baseline distributions and distributions with shared frailty were checked by using graphs in Fig. 5. From the plot of baseline Exponential, Weibull and log-logistic distributions; the plot of baseline log-logistic was more straight line compared with Exponential and Weibull for age at circumcision dataset. The Cox–snell plot of log-logistic inverse Gaussian shared frailty showed straight line through the origin with approximately slope 1. Suggesting that the model is appropriate for time-to-circumcision of girls dataset. These findings were consistent with Cox (1972), O'Quigley and Stare (2002) and Bennett (1983) for log-logistic model. The results of this study suggested that place of residences was significant predictive factor for age at circumcision of girls in Ethiopia. This shows that girls whose mothers lived in urban areas had longer survival with respect to age at circumcision than girls whose mothers resided in rural areas. It showed that age at circumcision for girls whose mothers lived in urban was prolonged by the factor of $\phi = 1.524$ when rural is used as reference. A similar study that has been conducted in Ethiopia by Setegn et al. (2016) revealed as the rural girls were more vulnerable towards circumcision than urban girls. The findings of this study also exposed that the educational level of girls mothers had a significant effect on the survival of age at circumcision with 5% level of significance and it prolonged age at circumcision by the factor of $\phi = 3.119$ and 3.933 for secondary and higher education respectively when illiterate mothers was used as the reference category. The result of the study shows that girls whose mothers had secondary and higher education were more survived than those uneducated and primary education. This is consistent with the study conducted in Burkina Faso by Karmaker et al. (2011) they revealed that the prevalence of girls circumcision is high for those daughters whom their mothers are illiterate. Mothers religion was found to be one of the significant factors for determining age at circumcision in this study. It showed that age at circumcision for those girls whom their mothers religion were muslim was shortened by the factor of $\phi = 0.654$ when mothers with Christian religion is used as reference. Similar study conducted by Abdisa et al. (2017) found that girls whom their mothers are muslims are more vulnerable towards circumcision. The result of this study also demonstrated that fathers educations are determining factors for age at circumcision of girls in Ethiopia which agrees with the study conducted in Liberia by Adetunji (2018).

5 Conclusions

This study was based on a dataset of time-to-circumcision of girls in Ethiopia which was obtained from Central Statistics Agency with an aim of modeling the determinants of time-to-circumcision of girls by using different parametric shared frailty models. Out of

the total of 2930 girls, about 22.4% were experienced an event (circumcised) and 77.6% were censored (uncircumcised) between the age of 1 and 15 years. The estimated median age of girls at circumcision was 3 years. To model the determinants of time-to-circumcision of girls, various parametric AFT shared frailty models by using different baseline distributions were employed. Among these using AIC, the log-logistic AFT inverse gaussian shared frailty model is best fitted to circumcision dataset than other parametric AFT shared frailty models. There is a frailty (clustering) effect on age at circumcision of girls dataset that arises due to differences in distribution of time-to-circumcision of girls among regions of Ethiopia. Goodness of the fit of baseline distributions and distribution with shared frailty was checked by means of graphical method and Cox–Snell residuals plots in Fig. 5 and revealed that baseline log-logistic and log-logistic inverse gaussian shared frailty model were better fitted the age at circumcision of girls dataset compared to other models. The result of log-logistic inverse gaussian shared frailty model showed that the factors that determine the timing of age at circumcision are place of residence, religion of mothers, mothers educational level and fathers educational level of the respondents are statistically significant. As educational level of mothers and fathers of girls increases, age at circumcision of girls in Ethiopia highly prolonged. This indicates that parents of the girls have to be given an opportunity of getting access to adult education to get awareness about harmfulness of circumcision of girls. Place of residence and religion of mothers prolong and shorten age at circumcision of girls respectively. Awareness has to be given on risk of girls circumcision specially for the society residing in rural area. The education can plays a crucial role in this regard. Since parent education is the most determinant factor of age at circumcision, parents of the girls have to be given special attention in giving them education opportunity and aware them on the complications that FGM would bring after circumcising girls. Further cross-sectional studies should be conducted in each region of Ethiopia and identify other factors of age at circumcision of girls that are not identified in this study.

Acknowledgements I acknowledge the Ethiopian central statistical agency for providing me the data.

Compliance with ethical standards

Conflict of interest The authors declare that they have no competing interests.

Availability of data and materials I can provide the dataset that has been used during the current study on reasonable request.

References

Aalen, O., Borgan, O., Gjessing, H.: Survival and Event History Analysis: A Process Point of View. Springer, Berlin (2008)

Abdisa, B., Desalegn, M., Tesew, A.: Assessment of the prevalence of FGM and associated factors among women's of reproductive age group in Kebirbeyah Town, Somali Region Eastern Ethiopia, 2017. Health Sci. J. **11**(4), 517–526 (2017)

Adetunji, S.: The impact of parental education level, wealth status, and location on female genital mutilation prevalence in northwestern Liberia (2018)
Bennett, S.: Analysis of survival data by the proportional odds model. Stat. Med. **2**(2), 273–277 (1983)
Central Statistical Agency: Ethiopian Demographic and Health Survey Addis Ababa Ethiopia (2016)
Central Statistical Authority and ORC Macro: Ethiopia Demographic and Health Survey 2000. Addis Ababa, Ethiopia, and Calverton, Maryland (2001)
Central Statistical Agency Ethiopia and ORC Macro: Ethiopia Demographic and Health Survey 2005. Addis Ababa, Ethiopia, and Calverton, Maryland, USA (2006)
Clayton, D., Cuzick, J.: Multivariate generalizations of the proportional hazards model. J. R. Stat. Soc. Ser. A (General) **148**, 82–117 (1985)
Collett, D.: Modeling Survival Data in Medical Research. Chapman & Hall, London (2003)
Cox, D.R.: Regression models and life tables (with discussion). J. R. Stat. Soc. Ser. B **34**(2), 187 (1972)
Cox, D.R., Oakes, D.: Analysis of Survival Data. CRC Press, Boca Raton (1984)
Datwyler, C., Stucki,, T.: Parametric Survival Model. Handout (2011)
Duchateau, L., Janssen, P.: The Frailty Model. Springer, New York (2008)
Greenwood, M., Yule, G.U.: An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. J. R. Stat. Soc. **83**, 255–279 (1920)
Gupta, R.D., Kundu, D.: Generalized exponential distribution: Existing results and some recent developments. J. Stat. Plan. Inference **137**(11), 3537–3547 (2007)
Gutierrez, R.G.: Parametric frailty and shared frailty survival models. Stata J. **2**(1), 22–44 (2002)
Hougaard, P.: Analysis of Multivariate Survival Data. Springer, Berlin (2012)
Kaplan, E.L., Meier, P.: nonparametric estimation from incomplete observations. J. Am. Stat. Assoc. **1958**(53), 457–81 (1958)
Karmaker, B., Kandala, N.B., Chung, D., Clarke, A.: Factors associated with female genital mutilation in Burkina Faso and its policy implications. Int. J. Equity Health **10**(1), 20 (2011)
Klein, J.P., Moeschberger, M.L.: Survival Analysis: Techniques for Censored and Truncated Data. Springer, New York (1997)
Klein, J.P., Moeschberger, M.L.: Survival Analysis: Techniques for Censored and Truncated Data. Springer, Berlin (2003)
Macfarlane, A., Dorkenoo, E.: Female genital mutilation in England and Wales: updated statistical estimates of the numbers of affected women living in England and Wales and girls at risk. City University London, Northampton Square, London (2014)
O'Quigley, J., Stare, J.: Proportional hazards models with frailties and random effects. Stat. Med. **21**, 3219–3233 (2002)
Population Reference Bureau (PRB). Female genital mutilation/cutting: data and trends. Washington (2010)
Sastry, N.: A nested frailty model for survival data, with an application to the study of child survival in Northeast Brazil. J. Am. Stat. Assoc. **92**, 426–43 (1997)
Setegn, T., Lakew, Y., Deribe, K.: Geographic variation and factors associated with female genital mutilation among reproductive age women in Ethiopia: a national population based survey. PLoS ONE **11**(1), e0145329 (2016)
UNFPA. Global consultation on female genital mutilation/cutting. Technical report 2009 (2013)
Vaupel, J.W., Manton, K.G., Stallard, E.: The impact of heterogeneity in individual frailty on the dynamics of mortality. Demography **16**(3), 439–454 (1979)
Wienke, A., Lichtenstein, P., Yashin, A.I.: A bivariate frailty model with a cure fraction for modeling familial correlations in diseases. Biometrics **59**, 1178–83 (2003)
Wienke, A., Ripatti, S., Palmgren, J., Yashin, A.I.: A bivariate survival model with compound Poisson frailty. Stat. Med. **29**, 27583 (2010)
Wit, E., van den Heuvel, E., Romeyn, J.-W.: 'All models are wrong...': an introduction to model uncertainty. Stat. Neerl. **66**(3), 217236 (2012)
World Health Organization (WHO). Eliminating Female Genital Mutilation: An Interagency Statement. Geneva (2008a)
World Health Organization fact sheet on Female genital mutilation (2008b)
World Health Organization (WHO). Eliminating Female Genital Mutilation: An Interagency Statement, OHCHR, UNAIDS, UNDP, UNECA, UNESCO, UNFPA, UNHCR, UNICEF, and UNIFEM. WHO, Geneva (2008c)

Affiliations

Daniel Biftu Bekalo[1]

✉ Daniel Biftu Bekalo
danibiftu@gmail.com

[1] Department of Statistics, Haramaya University, Dire Dawa, Ethiopia