Data processing and analysis

Data cleaning was conducted to check for the consistency with the EDHS-2016 descriptive report. Recoding, variable generation, labeling and analysis were done by using STATA/SE version 14.0. Descriptive statistics were done to describe the study participants in relation to socio-demographic characteristics which were presented in tables and text. Sample weight (gen wt=v005/1000, 000) was used to compensate the unequal probability of selection between the strata that were geographically defined and for non-responses. Multilevel analysis was conducted after checking the data was eligible to multilevel analysis (by using intra-cluster correction coefficient. When the ICC is greater than 10% (ICC= 22.5%) the community level factors affects the dependent variable. There for it is better to identify community level factors to develop and take different interventions. Since EDHS data are hierarchical (individual “level 1”were nested with in community “level 2”), a two level mixed effects logistic regression model was fitted to estimate both independent (fixed) effects of the explanatory variables and community –level random effects on early sexual initiation among 15-24 years old female. The log of the probability of early sexual initiation was modeled using a two level multilevel model as follows (34): $\mathrm{Log}\ [\frac{\Pi ij}{1-\Pi ij}] = \beta_0 + \beta_1 X_{ij} + B_2\ Z_{ij} + \mu_j + e_{ij}$

Where I and j are individual level and community level (2) unites respectively; X and Z refers to individual and community level variables respectively; πij is the probability of early sexual initiation for the ith youth in the jth community; β’s indicates the fixed coefficients. (B_0) is the intercept, the effect on the probability of early sexual initiation in the absence of influencing factors; and μj showed the random effect (the effect of the community on early sexual initiation of the jth community) and eij showed random errors at individual level. By assuming each community had different intercept (B_0) and fixed coefficient (β), the clustered data nature and intra and inter community variations were taken into account.

During analysis first, bi-variable multilevel logistic regression was fitted and variables with p value less than 0.2 at model I and model II were selected to develop the 3rd model (the final model). The analysis was done in four models. The first model was, model-0 (empty model or null model/ without explanatory variable; to secure the need to multilevel analysis). The second model was, model-I (analyzing only individual level variable), the 3rd model was, model-II (analyzing only community level variable), the last model, model-III (analyzing both community level and individual level variables based on the cutoff point).

The measure of association (fixed effects) estimate the association between the likelihood of early sexual initiation among female youth and different explanatory factors were expressed by Adjusted Odds Ratio (AOR) with respective 95% confidence level. Variables with p- value less than 0.05 at model-III were significantly associated with early sexual initiation. The random-effects (variations) were measured by using ICC (model-0), Median Odds Ratio (MOR) in (model-I and II) and Proportional Change in Variance (PCV) was measured to show variation between clusters.

ICC shows the variation in early sexual initiation among female youth due to community characteristics. The higher the ICC, the community characteristics are more relevant to understand individual variation for early sexual initiation. It is calculated as: $\mathrm{ICC}=(\frac{\delta^2}{\delta^2+\frac{\pi^2}{3}})$, where$\boldsymbol{\delta}^2$ indicates estimated variance of clusters.

MOR is the median value of the odds ratio between the area at highest risk and the area the lowest risk when randomly picking out two areas and it was calculated as: MOR= exp.

($\sqrt{2 \times \delta^2 + .6745}$)≈ exp$^{(0.95\boldsymbol{\delta})}$. In this study, MOR shows the extent to which the individual probability of early sexual initiation for female youth determined by place of residence. PCV measures the total variation attributed by individual level variables and area (35) level variables in the final model (model-III).

It is calculated as PCV= [$(\delta^2 of\ null\ model - \delta^2\ of\ \ each\ model)/\delta^2 of\ null\ model$]. $\boldsymbol{\delta}^2$ of the null model is used as reference.

Multicollinearity was checked among explanatory variables by using standard error at cutoff point ±2. There is no Multicollinearity that is the standard errors were between ±2. The log likelihood test was used to estimate the goodness of fit of the adjusted final model (model-III) in comparison to the preceding models (model-I and model-II) individual and community model adjustments respectively.