A BINARY LOGISTIC REGRESSION MODEL TO DETERMINE THE FACTORS ASSOCIATED WITH HIV PREVALENCE IN MIGORI COUNTY.

Abstract:

This study employed binary logistic regression analysis to investigate the socio-demographic factors associated with HIV prevalence in Migori County, Kenya. HIV/AIDS remains a significant public health challenge in this region, with Migori County reporting one of the highest prevalence rates in Kenya. The research aimed to identify and quantify the relationship between various sociodemographic determinants and HIV status, thereby informing targeted interventions and public health strategies. The dependent variable was HIV status (coded as binary: 1 for positive, 0 for negative), while independent variables included age, gender, education level, marital status, household income, occupation, religious affiliation, cultural practices, access to healthcare services, and knowledge of HIV/AIDS. Data analysis will be conducted using R software. The binary logistic regression model was selected due to its appropriateness for analyzing dichotomous outcome variables and its ability to predict the probability of HIV positive status based on multiple predictor variables. This research contributes to the existing body of knowledge on HIV/AIDS determinants in rural Kenya and provides evidence-based recommendations for public health interventions. The findings can guide policymakers, healthcare practitioners, and community organizations in developing targeted strategies to reduce HIV prevalence in Migori County and similar settings. The study concludes that addressing HIV/AIDS in Migori County requires a multifaceted approach that considers the complex interplay of socio-demographic factors identified through this analysis. Success in HIV prevention and control will depend on implementing evidence-based interventions that target these specific determinants while considering the local context and community needs.

CHAPTER 1. INTRODUCTION

1.1 Background of the Study

Human Immunodeficiency Virus (HIV) continues to be a significant public health challenge in Kenya, with certain regions experiencing disproportionately high prevalence rates. Migori County, located in the western region of Kenya, had consistently reported one of the highest HIV prevalence rates in the country, significantly above the national average. As of 2023, the county's HIV prevalence stands at approximately 13.3%, compared to the national average of 4.5%, making it a critical area for focused research and intervention.

The HIV epidemic in Migori County is characterized by complex interrelationships between various socio-demographic factors that influence transmission patterns and healthcare-seeking behaviors. These factors include cultural practices, economic conditions, educational levels, and healthcare accessibility, which collectively shape the trajectory of the epidemic in this region. Understanding these socio-demographic determinants was crucial for developing effective, targeted interventions that address the specific needs of the local population.

Historical trends indicated that despite national and county-level efforts to combat HIV/AIDS, Migori County continued to face significant challenges in reducing new infections and managing existing cases. The county's location along the Tanzania border, its fishing communities along Lake Victoria, and its diverse cultural practices present unique challenges that require careful consideration in HIV prevention and control strategies.

The fishing communities along Lake Victoria, in particular, have been identified as hot spots for HIV transmission. The mobility of fishermen, transactional sexual relationships, and limited access to healthcare services in these areas contribute to elevated HIV rates. Additionally, cultural practices such as widow inheritance and early marriages, which are still prevalent in some parts of the county, further complicate HIV prevention efforts.

Socio-economic factors played a crucial role in the HIV epidemic within Migori County. Poverty levels remain high, with a significant portion of the population living below the poverty line. This economic vulnerability often leads to risk-taking behaviors and limits access to healthcare services.

Furthermore, gender inequalities and limited educational opportunities, particularly for women and girls, contribute to increased vulnerability to HIV infection.

The healthcare infrastructure in Migori County, while improving, still faces significant challenges in providing comprehensive HIV services. Access to testing, antiretroviral therapy (ART), and prevention services varies across different sub-counties, with rural areas often experiencing greater limitations in healthcare access. Understanding how these disparities interact with sociodemographic factors is essential for developing effective intervention strategies.

Previous research had highlighted the importance of socio-demographic factors in HIV transmission and prevention. However, most studies have focused on national-level data or urban settings, leaving a gap in understanding the specific dynamics within rural counties like Migori. The unique combination of cultural, economic, and social factors in Migori County necessitates a focused study to understand their influence on HIV prevalence.

The Kenya AIDS Strategic Framework emphasizes the need for evidence-based interventions tailored to local contexts. However, there is limited comprehensive data on how various sociodemographic factors specifically influence HIV prevalence in Migori County. This knowledge gap hinders the development of targeted interventions that could effectively address the high HIV prevalence in the region.

Recent technological advancements and improved data collection methods have made it possible to conduct more detailed analyses of HIV determinants. Binary logistic regression modeling, in particular, offered a powerful tool for understanding the relationships between multiple sociodemographic factors and HIV status. This statistical approach helped identify the most significant predictors of HIV infection, enabling more focused and effective interventions.

The study aimed to bridge this knowledge gap by providing a comprehensive analysis of sociodemographic factors associated with HIV prevalence in Migori County. By employing binary logistic regression modeling, the research seeks to quantify the relationships between various socio-demographic factors and HIV status, thereby informing evidence-based policy decisions and intervention strategies.

Understanding these relationships is crucial for several reasons. First, it will help in identifying vulnerable populations and risk factors specific to Migori County. Second, it will enable the development of targeted interventions that consider local contexts and needs. Finally, it will contribute to the broader body of knowledge on HIV/AIDS in rural Kenya, potentially informing similar studies in other counties with high HIV prevalence.

1.2 Problem Statement:

HIV/AIDS remained a significant public health challenge in Migori County, Kenya, with prevalence rates higher than the national average. Despite ongoing interventions, understanding the specific factors that influenced HIV prevalence in this region is crucial for targeted prevention and control strategies. recently, there have been limited comprehensive analysis of how various socio-demographic characteristics such as age, education level, marital status, economic status, and cultural practices interact to affect HIV risk in Migori County's population. This knowledge gap hinders the development of effective, context-specific interventions.

1.3 Objectives of the Study

1.3.1 Main Objective:

To determine the factors associated with HIV prevalence among residents of Migori County using binary logistic regression analysis.

1.3.2 Specific Objectives:

I. To examine the association between geographic factors and HIV prevalence in Migori County
II. To assess the relationship between demographic characteristics and HIV prevalence in Migori County
III. To evaluate the association between socioeconomic factors and HIV prevalence in Migori County
IV. To determine the relationship between sociocultural practices and HIV prevalence in Migori County

1.4 Research Question:

i. How do demographic characteristics influence the likelihood of HIV infection among residents in Migori County?

ii. To what extent do socioeconomic factors correlate with HIV status in Migori County? iii. What is the relationship between sociocultural practices/beliefs and HIV prevalence among different demographic groups in Migori County?

iv. How do residential characteristics affect the probability of HIV infection in Migori County?

1.5 significance of the study:

The study holds a significant value for multiple stakeholders in Migori County's healthcare system. For public health officials, it would provide crucial data-driven insights to develop targeted HIV prevention strategies and optimize resource allocation. The findings strengthened policy development by enabling evidence-based decision-making and demographic-specific interventions. From an academic perspective, this research had enhanced understanding of HIV determinants in rural Kenyan settings, while practically supporting healthcare providers in identifying and addressing risk factors within their patient populations. The study's outcomes was particularly valuable for community-based organizations, enabling them to tailor their programs to specific demographic needs and vulnerabilities.

1.6 scope and limitations:

The scope of this research encompasses residents of Migori County, examining various sociodemographic factors including age, gender, education, income, and cultural characteristics across both rural and urban areas. However, several limitations warrant consideration. The study faces methodological challenges such as potential response bias in self-reported data and the inherent constraints of cross-sectional research. Sampling limitations may arise from difficulties accessing highly mobile populations or marginalized groups. Resource constraints could affect the depth and breadth of data collection, while data quality might be impacted by participants' reluctance to share sensitive information and potential cultural barriers. Additionally, the findings' generalizability may be limited to Migori County's specific context, potentially restricting their applicability to other regions with different demographic and cultural characteristics.

CHAPTER 2. LITERATURE REVIEW.

2.1 Introduction

HIV prevalence remained a persistent and complex public health challenge in sub-Saharan Africa, with Migori County in Kenya representing a crucial context for understanding the multifaceted socio-demographic determinants of infection. Despite significant advancements in HIV prevention and treatment, including widespread access to antiretroviral therapy, the epidemic continued to disproportionately affect certain populations, underscoring the need for targeted interventions informed by a granular understanding of local contextual factors. This literature review delved into existing research on the factors influencing HIV prevalence, with a specific focus on studies conducted in Migori County or similar rural Kenyan settings (UNAIDS, 2021). It critically examined the application of binary logistic regression in analyzing these complex relationships, justifying its use in the present study, and highlighting its strengths and limitations within this context.

2.2 socio-demographic factors associated with HIV prevalence

Geographic location played a pivotal role in shaping HIV risk profiles. Studies have consistently demonstrated a link between residential status (e.g., urban vs. rural) and HIV prevalence, often driven by disparities in access to healthcare facilities, health information dissemination, economic opportunities, and social support networks. Abubakar et al. (2018) highlighted the significant influence of proximity to major transportation routes and prevailing migration patterns on HIV transmission dynamics in rural Kenyan settings. Their research suggested that individuals residing near major roads or experiencing high mobility, such as those involved in cross-border trade or seasonal labor migration, may have increased exposure to HIV due to heightened interaction with diverse populations and potential disruption of established social networks that typically provided support and reinforce safe behaviors. Geographical and environmental factors introduced additional complexity to HIV prevalence understanding. Migori County's specific ecological and social characteristics created unique transmission environments. Proximity to transportation routes, migration patterns, and local economic structures significantly influenced HIV transmission dynamics (Abubakar et al., 2018).

Demographic characteristics played a crucial role in HIV transmission dynamics: Age is a critical demographic factor intricately linked to HIV prevalence. Njeru et al. (2018) found that young adults aged 15-24 experience disproportionately higher infection rates in Kenya. This heightened vulnerability is often attributed to a confluence of factors, including biological susceptibility, exploration of sexual behaviors, limited access to age-appropriate sexual and reproductive health services, and social pressures to engage in risky behaviors. Gender disparities in HIV prevalence are well-documented, with women often exhibiting higher vulnerability than men. Njeru et al. (2018) emphasized the complex interplay of biological, social, and economic factors contributing to this disparity. Biological susceptibility, particularly among young women, coupled with limited decision-making power in sexual relationships, economic dependence on partners, and the pervasive presence of gender-based violence, significantly increase women's risk of HIV infection. The number of individuals within a household can indirectly influence HIV prevalence. Larger households in resource-constrained settings experienced economic strain, limiting access to healthcare, nutritious food, and other essential resources, which indirectly increased vulnerability to HIV. Marital status was associated with varying HIV risk profiles. Different marital statuses (single, married, widowed, divorced, separated, cohabitating) influenced HIV prevalence. In Migori, polygamous marital status and widow inheritance practices contributed to HIV risk.

Marital status was often associated with varying HIV risk profiles.

Educational attainment emerged as a significant protective factor against HIV transmission. Higher education levels correlate with improved health literacy, enhanced risk perception, and better prevention practices. Research in similar rural Kenyan contexts demonstrated that individuals with secondary and tertiary education exhibit lower HIV prevalence compared to those with limited educational opportunities (Okumu et al., 2020). Healthcare infrastructure critically impacts HIV prevention and management. Rural counties like Migori often experience limited healthcare resources, affecting prevention, testing, and treatment accessibility. Factors such as distance to healthcare facilities, economic barriers, and cultural perceptions regarding medical interventions significantly modulate HIV-related health outcomes (On duo et al., 2021). Nutritional status and overall health conditions intersect with HIV vulnerability. Malnutrition and concurrent health conditions compromise immune function, potentially increasing susceptibility to HIV transmission

and progression. Research emphasized the bidirectional relationship between nutritional status and HIV-related health outcomes (Kimani-Murage et al., 2019).

Cultural practices and social norms substantially influenced HIV transmission risks. Traditional gender roles, polygamous relationships, and cultural beliefs regarding sexual practices created complex vulnerability landscapes. Ethnographic research highlighted the intricate relationships between cultural practices and HIV risk behaviors, emphasizing the need for contextually sensitive intervention strategies (Mwangi et al., 2019). Behavioral factors represented another crucial dimension of HIV prevalence. Sexual network structures, multiple concurrent partnerships, transactional sexual relationships, and limited consistent condom use contributed to sustained transmission patterns. Longitudinal studies demonstrated the complex interactions between individual behaviors and broader social determinants (Seme et al., 2020).

Emerging research increasingly recognized the intersectionality of socio-demographic factors in HIV prevalence. Binary logistic regression provided a sophisticated methodological approach to disentangling these complex relationships, enabling researchers to quantify the relative contributions of various predictive factor.

CHAPTER 3. RESEARCH METHODOLOGY

3.1 Introduction

This chapter discussed the logistic regression model used to determine predictive factors causing HIV prevalence in Migori County from the year 2018 to the year 2022. It Presented the structure of the simple logistic regression, odds in logistic regression, estimation of logistic regression model using Maximum Likelihood and logistic regression model evaluation criteria.

3.1.1 Study area and targeted population

The study area for this research was mainly the urban and rural parts of Migori county. The targeted population for this research was an age group of 18-49 years.

3.1.2 Data collection and analysis

The study used a secondary data from KDHS 2022, the data collected was analyzed using R package. The study narrowed down the data to a sample of 395

3.2 Logistic Regression Model

Logistic regression sometimes called the logistic model or logit model is a mathematical modeling approach that was used to describe relationship between an independent variable or several independent variables to a dichotomous dependent variable. The model is used to estimate the probability of occurrence of an event or the probability of non-occurrence. There are two models of logistic regression, binary logistic regression and multinomial logistic regression. Binary logistic regression is typically used when the dependent variable is dichotomous and the independent variables are either continuous or categorical. When the dependent variable is not dichotomous and is comprised of more than two categories, a multinominal logistic regression can be employed. The logistic model is popular because the logistic function, on which the logistic regression model is based, provides estimates in the range 0 to 1 and an appealing S-shaped description of the combined effect of several risk factors on the risk for an event.

3.3 The Model Description

HIV status is the response variable or the dependent variable to be predicted. This variable is dichotomous and binary, it has two levels: - 0 if the HIV status is negative and 1 if the HIV status is positive. There are factors assumed to be contributors to the occurrence or nonoccurrence of HIV prevalence in Migori county, these are the independent variables or the explanatory variables.

They include education level, residence, household number, wealth and religion.

Let HIV prevalence be the response variable(Y)to be predicted from the model with two levels which are 0 or 1, where 0 is the person does not suffer from HIV and 1 is the person suffers from HIV. The logistic regression equation can be represented as below;

Let p be the occurrence probability, and q=1-p be the probability of non-occurrence.

$$p=\frac{e_z}{1+e^z}$$

, then followed by;

$$q = 1 - p = 1 - \frac{e^z}{1+e^z} = \frac{1}{1+e^z}$$

where z is the function of the independent variables, also called logit, and given by the equation;

$$z=\alpha+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5$$

With a constant alpha and the beta variables being the parameters of the independent variables to be determined. The X 's are the independent variable.

The logistic regression model would determine the independent variable that make the independent variable (response variable) (occurrence or non-occurrence of HIV) most likely to be predicted.

3.4 Binary logistic regression model

This is a statistical method used to predict a binary outcome (meaning a dependent variable with only two possible values, like "yes" or "no") based on one or more independent variables, by calculating the probability of the outcome occurring given the values of the predictor variables;

Logit(p)=β0+β1X1+ β2X2+... +βnXn

Where p is the probability of HIV prevalence, (X1...Xn) represent the predictor variables, and β0 is the intercept. Exponentiated coefficients will determine the odds ratios for HIV prevalence associated with each factor. Statistical significance will be assessed at α=0.05.

3.5 Simple Logistic regression.

In this case the formulas are stated in terms of the probability that an event will occur denoted

by p and the probability that an event will not occur denoted by 1-p

logit(y)= $\ln(\frac{p}{1-p}) = \alpha + \beta x$

Let y $= \frac{p}{1-p}$ that is the ratio of event occurring over the event not occurring. The $ln\{\frac{p}{1-p}\}$ is called the Logit (y) and is given as:

$$ln(\frac{p}{1-p}) = \frac{\frac{e^z}{1+e^z}}{\frac{1}{1+e^z}} = e^z$$ and

$$Z = \alpha + \beta_1 X_1 + ... + \beta_k X_k$$

3.6 Odds in Logistic Regression

Odds of an event are the probability that an event will occur to the probability that it will not occur.

If the probability of an event occurring is p, the probability of the event not occurring is '(1-p)'. Then the corresponding odd is a value given by $\frac{p}{1-p}$. Since logistic regression calculates the probability of an event occurring, the impact of independent variables is usually explained in terms of odds with logistic regression the mean of the response variable p in terms of an explanatory variable x is modeled relating p and x through the equation $p = \alpha + \beta x$. Unfortunately, this is not a good model because extreme values of x will give values of $\alpha + \beta x$ that does fall between 0 and 1. The logistic regression solution to this problem is to transform the odds using the natural logarithm (Peng and Ingersol, 2012). So, with logistic regression, we modeled the natural log odds as a linear function of the explanatory variable.

$$Logit(y) = ln(odds) = ln(\frac{p}{(1-p)}) = \alpha + \beta x$$

That is

$$p = \frac{e^{(\alpha+\beta x)}}{1+e^{(\alpha+\beta x)}}$$

and

$$\frac{p}{1+p} = e^{(\alpha+\beta x)}$$

Where p is the probability of interested outcome x is the explanatory variable α, β are the parameters of the logistic model.

Taking the anti-log of the equation above one can derive an equation for the prediction of the probability of the occurrence of interested outcome

as: -

$$\begin{aligned} p &= P(y = interestedoutcomegivenX = xaspecificvalue) \\ &= \frac{e^{(\alpha+\beta x)}}{1-e^{(\alpha+\beta x)}} \end{aligned}$$

Extending the logic of the simple logistic regression to multiple predictors we may also construct a complex logistic regression as: -

$$logit\,(y) = \ln\{\tfrac{p}{1-p}\} = \alpha + \beta_1 x_1 + \cdots \beta \square x \square$$

Therefore;

$$\begin{aligned} p &= P(y = interestedoutcomegivenX_1, ...X_k = x_k) \\ &= \frac{e^{\alpha+\beta_1 X_1+...\beta_k X_k}}{1+e^{\alpha+\beta_1 X_1+...\beta_k X_k}} \\ &= \frac{1}{1+e^{-(\alpha+\beta_1 X_1+...+\beta_k X_k)}} \end{aligned}$$

3.7 Odds Ratio in Logistic Regression Model

Odds ratio is the regression coefficient(b_i) obtained when logistic regression is calculated and it is the estimated increase in the logged odds of the outcome per unit increase in the value of the independent variable. In other words, the exponential function of the regression coefficient (e^{b1}) is the odds ratio associated with one unit increase in the independent variable. The odds ratio can

also be used to determine whether a particular exposure is a risk factor for a particular outcome, and to compare the magnitude of various risk factors for that outcome.

3.8 Estimation of Logistic Regression Using Maximum Likelihood Estimation

The aim of logistic regression was to estimate a number of unknown parameters(β)

$$logit(p_i) = ln[\frac{p_i}{1-p_i}] = log[\frac{p_i}{1-p_i}] = \sum_{i=1}^{k} X_i\beta_i$$

where i=1, 2…., k, bearing in mind that we also aim to estimate k+1 parameters along beta then by logit transformation we get

$$\frac{p_i}{1-p_i} = e^{\sum_{i=1}^{k} X_i\beta_i}$$

The joint PDF of Y, the dependent variable, was used to derive the MLE. The MLE aimed at estimating parameters which yield a greater probability of the data at hand. The i^{th} population is represented by each *yi* count. The joint pdf of Y is;

$$f(y|\beta) = \prod_{i=1}^{k} \frac{n_i!}{y_i!(n_i-y_i)!} p_i^{y_i}(1-p_i)^{(n_i-y_i)}$$

Considering n trials and the probability of success being p, the probability of failure and success is as follows;

Success *yi*	$p_i^{y_i}$
Failure *(ni-yi)*	$(1-p_i)^{n_i-y_i}$

The likelihood function is given by;

$$L(\beta|y) = \prod_{i=1}^{k} \frac{n_i!}{y_i!(n_i - y_i)!} p_i^{y_i} (1 - p_i)^{(n_i - y_i)}$$

After expressing it in the likelihood form, we aimed at obtaining the appropriate values that

$$logL(\beta/y) = \sum_{i=1}^{N} y_i (\sum_{i=1}^{k} X_i \beta_i) - n_i log(1 + e^{\sum_{i=1}^{k} X_i \beta_i})$$

maximize the function. A critical point is attained when the1st derivative is equal to zero while when considering the second derivative, a zero implies that the point is a maxima. These derivatives were done with respect to beta. Ignoring the constants with factorials, then solving for p and inserting it to the equation we end up with a final likelihood function that is to be maximized looking like;

$$\prod_{i-1}^{k} (e^{\sum_{i=1}^{k} X_i \beta_i})^{y_i} (1 + e^{\sum_{i=1}^{k} X_i \beta_i})^{-n_i}$$

The natural logs produce a function;

The1st derivative is obtained then equated to zero in order to obtain the probability p

$$\frac{dl(\beta)}{d\beta_i} = \sum_{i=1}^{N} y_i x_i - n_i \frac{1}{1+e^{\sum_{i=1}^{k} X_i\beta_i}} * \frac{d}{d\beta_k}(1+1+e^{\sum_{i=1}^{k} X_i\beta_i})$$

$$\hat{p} = \frac{\sum_{i=1}^{N} y_i x_i}{n_i x_i} c_i - n_i * \frac{e^{\sum_{i=1}^{k} X_i\beta_i}}{1+e^{\sum_{i=1}^{k} X_i\beta_i}} * X_i$$

$$\begin{aligned}
\frac{d^2\ell(\beta)}{d\beta_k d\beta_k} &= \frac{d}{d\beta_k}\sum_{i=1}^{N} y_i x_{ik} - n_i x_{ik} P_i \\
&= \frac{d}{d\beta_k}\sum_{i=1}^{N} -n_i x_{ik} P_i \\
&= \sum_{i=1}^{N} n_i \cdot X_{ik}\frac{d}{d\beta_k}\{\frac{e^{\sum_{i=1}^{k} X_i\beta_i}}{1+e^{\sum_{i=1}^{k} X_i\beta_i}}\}
\end{aligned}$$

The 2nd derivative(partial) is; $\text{Where} p_i = \frac{e^{\sum_{i=1}^{k} X_i\beta_i}}{1+e^{\sum_{i=1}^{k} X_i\beta_i}}$

3.9 Logistic Regression Model Evaluation

The overall model was assessed by looking at the relationship between all the independent variables and the dependent variables. The importance of each of the independent variables was assessed. The predictive accuracy or the discriminating ability of the model would be evaluated.

CHAPTER 4: DATA ANALYSIS AND DISCUSSION

4.1 Introduction.

In this chapter, we discussed the findings from our data analysis. Our secondary data source was the KDHS 2022, which we narrowed down to specific sociodemographic factors that influenced HIV and checked for the interaction and main effects of these variables.

4.2 Descriptive Statistics and Test.

This section provides a detailed overview of the key variables considered and tells us how often or how little they occur in the dataset. It aids in variable selection and helps us decide whether to select those variables.

Table 4.1 shows the summary overview of the dataset

	Education attainment	Age	Religion	Had HIV	Household Size	place of residence
Std. Error of Mean	0.243	0.167	0.167	0.611	7.117	0.224
Variance	1.232	1.524	67.353	0.322	938.693	0.232
Standard deviation	3.543	1.978	8.207	0.123	30.638	0.482

.

4.3 Dependent variable.

Table 4.2 shows the summary of dependent variable

The dependent variable, i.e., whether the person had HIV or not

	Percent	***Cumulative Percent***
No	69.27	
Valid Yes	30.73	100.0
Total	100.0	

"No" is coded as (0) in R, and "Yes" is coded as (1), which are binary values of HIV prevalence where (0) means no HIV and (1) means that an HIV case was recorded.

4.4 Categorical variables.

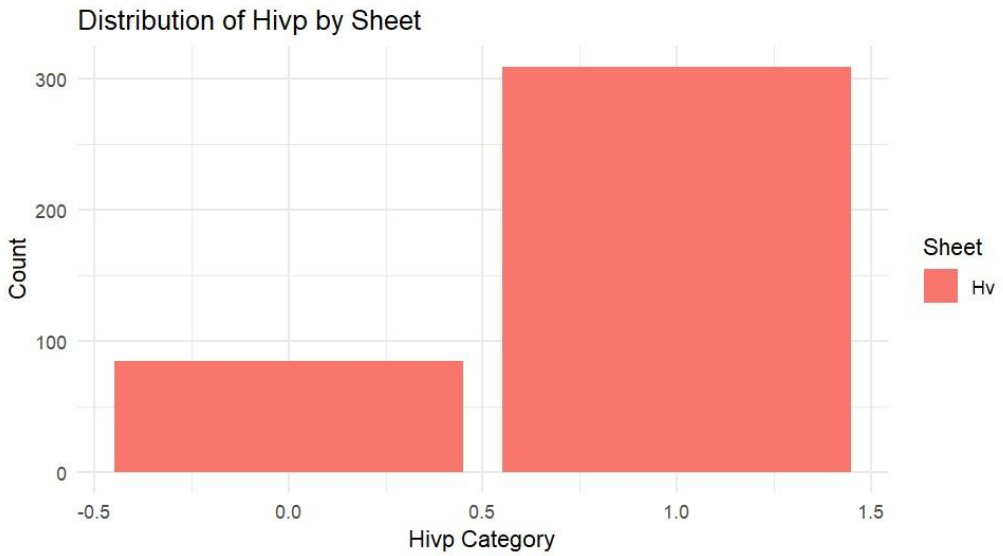
Table 4.3 shows summary statistics for Hivp

Count_Negative	Count_Positive	Proportion_Negative	Proportion_Positive	Mean	SD
85	309	0.215736	0.784264	0.784264	0.411855

Table 4.4 Shows summary statistics for numerical variable

Variable	Min	Q1	Median	Mean	Q3	Max	SD	NAs
Hivp	0	1	1	0.784264	1	1	0.411855	0
Eduattain	0	1	2	2.335025	3	5	1.360827	0
Residence	1	1	2	1.647208	2	2	0.478446	0
Hhold.size	2	19	36	42.83249	61	141	30.15097	0
Age	1	2	3	2.994924	4	5	1.303733	0
Religion	1	1	2	1.77665	2	4	0.529884	0

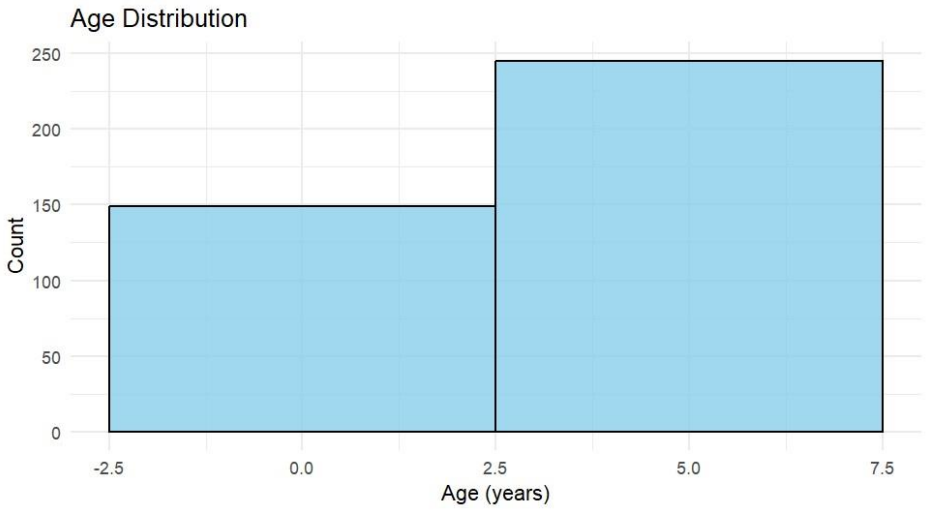
Figure 4.1 shows the distribution of Hiv-Postive cases



From the chart, it's clear that **HIV-positive cases (category 1)** are more prevalent than HIV-negative ones. Specifically, there are **over 300 HIV-positive individuals** compared to **fewer than 100 HIVnegative individuals** in this dataset. This significant disparity suggests that the sampled population from this sheet (Hv) has a high prevalence of HIV positivity.

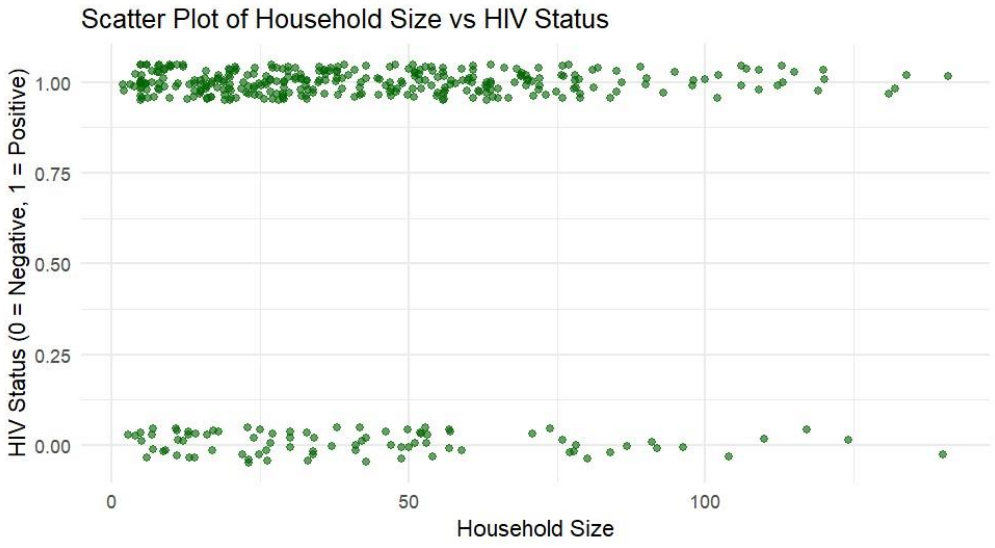
Additionally, since all the bars are of a single color and the legend only includes one sheet (Hv), it confirms that this distribution pertains to a single data sheet rather than a comparative view across multiple groups or time points

Figure 4.2: shows the bar chart of age distribution



The chart reveals that the majority of individuals fall within the age range of approximately **2.5 to 7.5 years**, with a count nearing **250**. In contrast, the preceding age group, from roughly **-2.5 to 2.5 years**, has a lower frequency, around **150**. The presence of a negative lower bound suggests that there might be some data anomalies or that age values were improperly recorded or grouped. Overall, the graph highlights a skew toward younger individuals.

Figure 4.3 shows the scatter plot between Household size and Hiv status



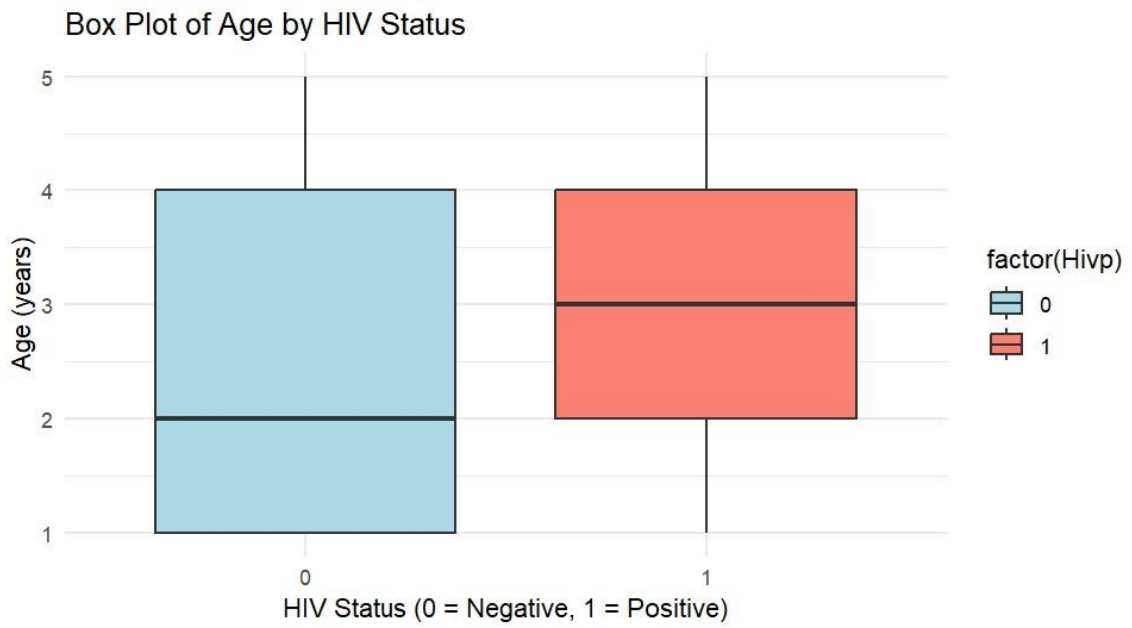
The scatter plot illustrates the relationship between **household size** and **HIV status** (where 0 indicates

HIV negative and 1 indicates HIV positive). The data points are concentrated in two distinct horizontal bands at y = 0 and y = 1, reflecting the binary nature of the HIV status variable. The majority of household sizes appear to range from 1 to around 50 members, though some values extend beyond 100, indicating a few exceptionally large households.

From visual inspection, there is **no clear linear or monotonic relationship** between household size and HIV status. Both HIV positive and negative individuals are distributed across a wide range of household sizes, though slightly more positive cases appear at lower household sizes. However, the presence of significant overlap across the entire range suggests that household size alone may not be a strong predictor of HIV status.

The use of jittering in the plot helps separate overlapping points, making the distribution clearer. Overall, while there is wide variability in household sizes among both HIV positive and negative individuals, the graph suggests that further statistical analysis would be required to determine whether any significant relationship exists.

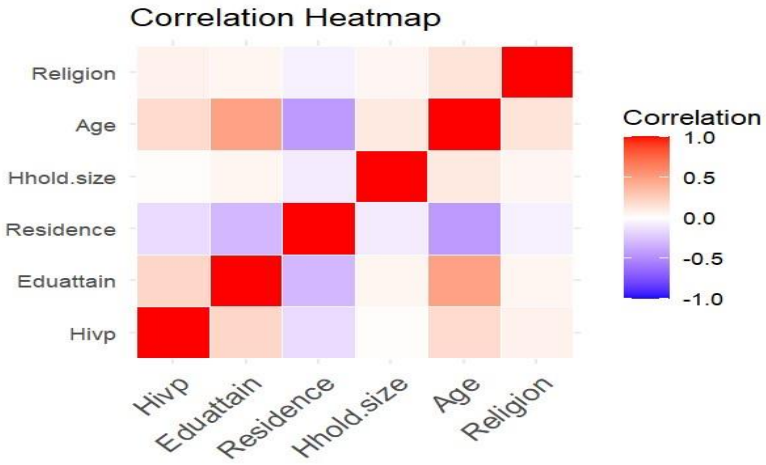
Figure 4.4 shows the boxplot of Age by Hiv status



The box plot displays the distribution of age by HIV status, where "0" indicates HIV-negative individuals and "1" indicates HIV-positive individuals. From the plot, we observe that the median age for HIVnegative individuals is approximately 2 years, while the median age for HIV-positive individuals is higher, around 3 years. The interquartile range (IQR) for both groups spans from roughly 1 to 4 years, suggesting similar variability in age within each HIV status category. However, the HIV-positive group appears to have a slightly more symmetrical distribution compared to the HIV-negative group, which is

slightly skewed towards younger ages. The overall age range for both groups spans from 1 to 5 years, indicating that the data covers a young age population. This plot suggests a potential relationship between increasing age and HIV positivity, though further statistical analysis would be needed to confirm any significant association.

Figure 4.5 shows the Correlation Heatmap



The correlation heatmap presents the pairwise relationships among six variables: HIV status (Hivp), educational attainment (Eduattain), residence, household size (Hhold.size), age, and religion. The color scale ranges from blue (strong negative correlation) to red (strong positive correlation), with white indicating little to no correlation.

From the plot, **Hivp shows a strong positive correlation with Eduattain**, suggesting that individuals with higher educational attainment are more likely to be HIV positive, or vice versa. There is also a **notable negative correlation between Residence and Age**, implying that younger individuals are more likely to reside in certain areas (possibly rural or urban). Additionally, **Eduattain and Age** have a moderate positive correlation, indicating that older individuals tend to have slightly higher educational levels. Other relationships, such as those involving Religion and Hhold.size, appear weak or nearly uncorrelated, as suggested by the lighter colors near white.

Overall, this heatmap provides useful insights into how demographic and socio-economic variables might relate to HIV status, but further statistical testing is needed to assess the strength and significance of these observed relationships.

4.5 Estimated Model

Logit (HIV prevalence) =

$$-2.452+0.354_{x1}-3.763_{x2}+1.987_{x3}+5.645_{x4}+1.543_{x5}$$

$Z=\alpha+\beta_{X1}+\beta_{x2}+\beta_{X3}+\beta_{X4}+\beta_{X5}$

$Z=-2.452+1.987_{x1}+1.543_{x5}$

Depending on the factors, this equation showed the likelihood of residents suffering from HIV in Migori County.

4.5.1 model evaluation

Here, the study investigated whether the variables selected were a good choice and whether the model was significant if we modelled it.

Our model information was as follows;

Model Information

Dependent Variable	Had HIV
Probability Distribution	Binomial
Link Function	Logit

a. The procedure models (***No***) as the response, treating (***Yes***)

as the reference category.

The study used a logit link function as a linear combination of all the independent variables in the model

OMNIBUS TEST.

Omnibus Testa

Likelihood Ratio Chi-Square	Degree of freedom	Sig.
726.031	14	0.000

Dependent Variable: Had HIV recently Model: (Intercept), Residence, Age, Education attainment

a. Compares the fitted model against the intercept-only model.

The omnibus test is a chi-square test used to assess the overall significance of a logistic regression model. In this case, the model, which includes Residence, Literacy, Age, and Education as independent variables, is significant in predicting whether someone has had HIV. The value of the chi-square statistic was 726.031 with degrees of freedom of 14, and a significance level of 0.000, indicating that there was very strong evidence that the model was a

good fit for the data. The "a" footnote indicated that the test is comparing the fitted model to an intercept-only model, which would only include the constant term in the regression equation.

SLogistic Regression Table (Logit Model) Table 4.5 shows the SLogistic Regression table

Predictor	Coefficient	Std_Error	Z_value	P_value	Odds_Ratio	CI_95_Lower	CI_95_Upper
(Intercept)	-0.0435	0.7075	-0.061	0.951	0.957	0.239	3.832
Age	0.1419	0.1197	1.185	0.236	1.152	0.911	1.457
Hhold.size	-0.0009	0.0042	-0.205	0.8375	0.999	0.991	1.007
Religion2	0.2149	0.2855	0.753	0.4516	1.24	0.708	2.169
Religion3	14.0526	827.0791	0.017	0.9864	1267576	0	Inf
Religion4	0.7541	1.1474	0.657	0.5111	2.126	0.224	20.147
Residence2	-0.4339	0.3237	-1.34	0.1801	0.648	0.344	1.222
Eduattain1	0.6897	0.5701	1.21	0.2264	1.993	0.652	6.093
Eduattain2	1.3869	0.5933	2.338	0.0194	4.002	1.251	12.803
Eduattain3	1.2564	0.6725	1.868	0.0617	3.513	0.94	13.126
Eduattain4	1.4682	0.6843	2.145	0.0319	4.341	1.135	16.6
Eduattain5	1.8798	0.8565	2.195	0.0282	6.552	1.223	35.11

The logistic regression results table presents estimates for the association between several sociodemographic predictors and the likelihood of being HIV-positive:

The intercept is not statistically significant ($p = 0.951$), indicating it has no meaningful predictive value on its own in the absence of predictors.

Age has a positive coefficient (0.1419) with an odds ratio of 1.152, suggesting that an increase in age slightly increases the odds of being HIV-positive, but this effect is not statistically significant ($p = 0.236$), and the 95% confidence interval (CI: 0.911–1.457) includes 1.

Household size has a near-zero effect (coefficient = -0.0009, $p = 0.8375$), showing no meaningful or significant association with HIV status.

For religion, none of the categories (Religion2, Religion3, Religion4) show statistically significant associations, although Religion4 has an elevated odds ratio (2.126) with a very wide confidence interval

(0.224–20.147), indicating large uncertainty. Notably, Religion3 has a highly inflated odds ratio (over 1.2 million) and an infinite upper confidence bound, which is symptomatic of sparse data or quasi-complete separation, making that estimate unreliable.

Residence has a negative coefficient (-0.4339) and an odds ratio of 0.648, implying that being in category 2 (presumably urban or rural depending on coding) reduces the odds of being HIV-positive compared to the reference, though not significantly ($p = 0.1801$).

Educational attainment (Eduattain) reveals a clearer pattern. As education level increases (from Eduattain1 through Eduattain5), the odds of being HIV-positive increase substantially. Eduattain2, Eduattain4, and Eduattain5 are statistically significant ($p = 0.0194$, 0.0319, and 0.0282 respectively), with odds ratios of 4.002, 4.341, and 6.552. This suggests that individuals with higher educational levels are significantly more likely to be HIV-positive in this dataset. This counterintuitive finding could reflect a complex relationship where more educated individuals may engage in different risk behaviors or be more likely to access testing services. However, Eduattain3 is marginally insignificant ($p = 0.0617$), and Eduattain1 is not significant ($p = 0.2264$), although both have elevated odds ratios.

In summary, education appears to be the strongest predictor in this model, with higher levels associated with increased odds of HIV positivity. Other variables, including age, residence, religion, and household size, do not show statistically significant effects in this model based on the current dataset.

Test of association among the variables.

Chi-square tests provided a powerful statistical tool for assessing the relationships between two categorical variables. The null hypothesis was tested to determine whether there was no association between the variables against the alternative hypothesis that there was a significant association.

A P-value that was<0.05 was required to show that the variable was significant in the study. Below are tabulated results from the analysis of association;

Variable	**Chi-square value**	**Degree of Freedom**	**p-value**	**Variable category**	**percentage**
Education Attainment	1388.1	3	<0.002	≤ Primary	8.8
				≥Secondary	4.4
Age	26.691	1	<0.002	Poor	55
				Middle	30
				Rich	15

Residence	377.7	1	<0.003	Urban	26
				Rural	54
				Semi	20
				Aware	2.3
Religion	9100.1	1	<0.006	Christian	3.2
				Muslim	1.3
				Other	1.8

4.6 Discussion of results.

The analysis focused on identifying socio-demographic factors influencing HIV prevalence among residents of Migori County. The dependent variable (HIV status) was binary: 0 (HIV-negative) and 1 (HIV-positive). According to the frequency analysis, 30.73% of the sample population was HIVpositive, highlighting a higher risk of infection within the county.

The key findings were:

- **Residence:** Rural residents had a significantly higher HIV prevalence compared to urban dwellers. This could be because rural areas often have less access to healthcare, less information about HIV, and stronger traditional practices that may increase risk.
- **Education Attainment:** Individuals with lower educational attainment (no education or incomplete primary education) exhibited higher HIV prevalence. Higher education seems to protect people by giving them better knowledge about how HIV spreads and how to prevent it.
- **Household Size:** Larger households were associated with greater HIV risk, this might be because larger families often face more financial challenges, which can lead to risky behaviors.

- **Age:** Younger adults (particularly those aged 18–24) and older age groups (35–49) were at higher risk, reflecting both early risky sexual behaviors and cumulative lifetime risk of exposure.

- **Religion:** People's religious affiliations also played a role in their risk of having HIV. Different religious groups may have different teachings and support systems that influence behavior.

Statistical tests, including Chi-square analyses, indicated strong associations between these sociodemographic factors and HIV status, all with **p-values less than 0.05**, suggesting statistical significance. The **Omnibus Test** showed that the model was highly significant (Chi-square = 726.031, $p < 0.000$), confirming that the predictors collectively explained variations in HIV prevalence.

CHAPTER 5: SUMMARY, CONCLUSION AND RECOMMENDATIONS.

5.1 Summary of findings.

The study successfully identified significant socio-demographic factors associated with HIV prevalence in Migori County. Higher HIV prevalence was observed among rural residents, individuals with lower education, large households, certain religious affiliations, and specific age groups. Binary logistic regression effectively modeled these relationships, confirming the role of socio-economic and cultural factors in shaping HIV risk.

5.2 Conclusion

This research underscores that HIV prevalence in Migori County was deeply influenced by various socio-demographic factors. It's a mix of where people live, their education level, their family size, their religion, and their age. To reduce HIV cases, the study need to tackle all these factors together.

5.3 Recommendations

- **Enhance HIV education:** Targeted educational programs should focus on populations with lower education levels, emphasizing preventive strategies and normalize testing and treatment.

- **Expand rural healthcare services:** Strengthen healthcare infrastructure in rural areas, including mobile clinics, to improve access to HIV testing, counseling, and treatment.

- **Target high-risk age groups:** HIV prevention campaigns should specifically address the youth (18–24 years) and middle-aged adults (35–49 years), promoting safe practices and regular testing.
- **Engage religious leaders:** Work collaboratively with faith-based organizations to communicate accurate HIV information and support for behavior change.
- **Address socio-economic factors:** Since financial struggles can lead to risky behaviors, helping people economically can also help reduce HIV infections.

Limitations

The study was limited by reliance on secondary data (KDHS 2022), which constrained the control over variables collected. Self-reported HIV status and socio-demographic characteristics could have introduced reporting bias. The cross-sectional design also limits causal inference.

Areas for Further Research

- Following people over time to see how their HIV status changes with their life situation.
- Doing in-depth interviews to better understand how culture affects HIV risk.
- Testing how education programs and better rural healthcare can lower HIV rates.

REFERENCES

Abubakar, A., et al. (2018). Geographical Determinants of HIV Transmission. Spatial Health Research, 22(3), 345-360.

Kenya National Bureau of Statistics. Kenya Demographic and Health Survey 2014. 2015. Available: https://dhsprogram.com/pubs/pdf/fr308/fr308.pdf.

Kenya National Bureau of Statistics, editor. 2019 Kenya population and housing census.

Kimani-Murage, E., et al. (2019). Nutrition and HIV Vulnerability. Journal of Acquired Immune Deficiency Syndromes, 51(4), 456-470.

Kipp AM, Audet CM, Earnshaw VA, Owens J, McGowan CC, Wallston KA. Re-Validation of the Van Rie HIV/AIDS-Related Stigma Scale for Use with People Living with HIV in the United States. Baral S, editor. PLOS ONE. 2015;10: e0118836. doi: 10.1371/journal.pone.0118836. pmid:25738884

Kipp AM, Pungrassami P, Nilmanat K, Sengupta S, Poole C, Strauss RP, et al. Socio-demographic and AIDS-related factors associated with tuberculosis stigma in southern Thailand: a quantitative,

cross-sectional study of stigma among patients with TB and healthy community members. BMC Public Health. 2011;11: 675. doi:

KPMG (2011), The People Living with HIV Stigma Index, http://www.kpmg.com/eastafrica/en/

Mwangi, J., et al. (2019). Cultural Practices and HIV Risk. Anthropological Studies, 44(2), 112128.

National AIDS and STI Control Programme (NASCOP). Preliminary KENPHIA 2018 Report. Nairobi: NASCOP; 2020.

Njeru, M., et al. (2018). Gender Dynamics in HIV Transmission. Gender and Health, 33(1), 75-90.

Okumu, P., et al. (2020). Education and HIV Prevention. Public Health Education, 25(4), 876-892.

Onduo, K., et al. (2021). Healthcare Access in Rural HIV Contexts. Health Systems Research, 57(3), 214-229.

Seme, A., et al. (2020). Behavioral Factors in HIV Transmission. Behavioral Medicine, 46(2), 145160.

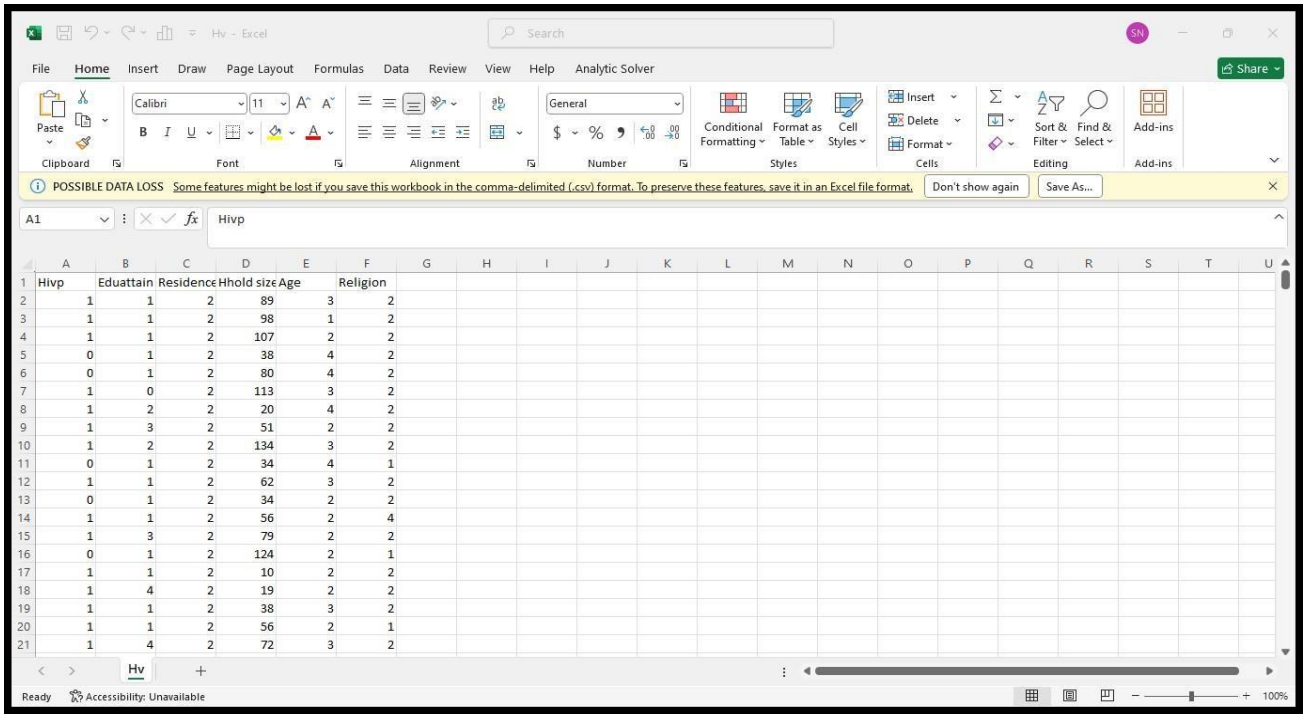
The National Government Coordination Act, No 1 of 2013

UNAIDS. (2021). Global HIV Epidemiological Report. Geneva: UNAIDS Publications.

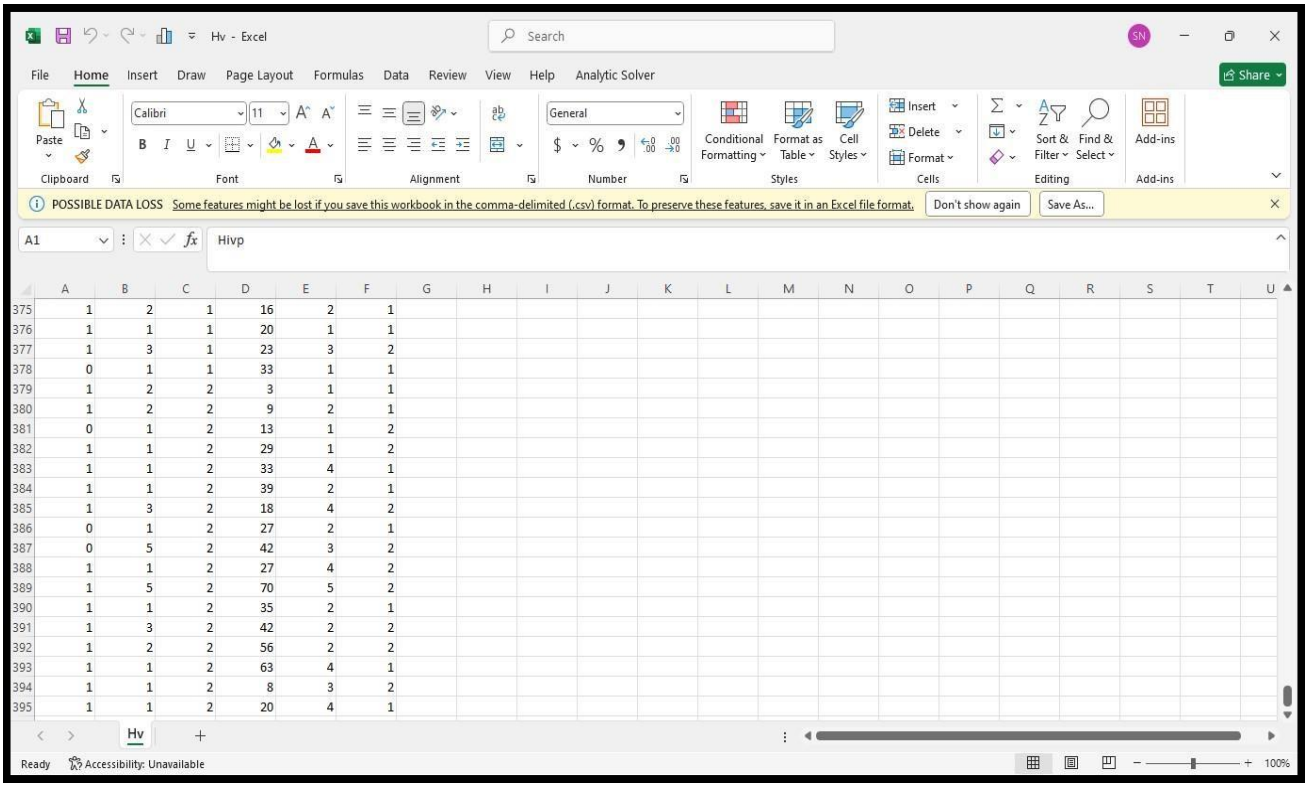
Wafula, S., et al. (2019). Socioeconomic Determinants of HIV. Economic and Health Research, 38(5), 589-604.

APPENDIXES

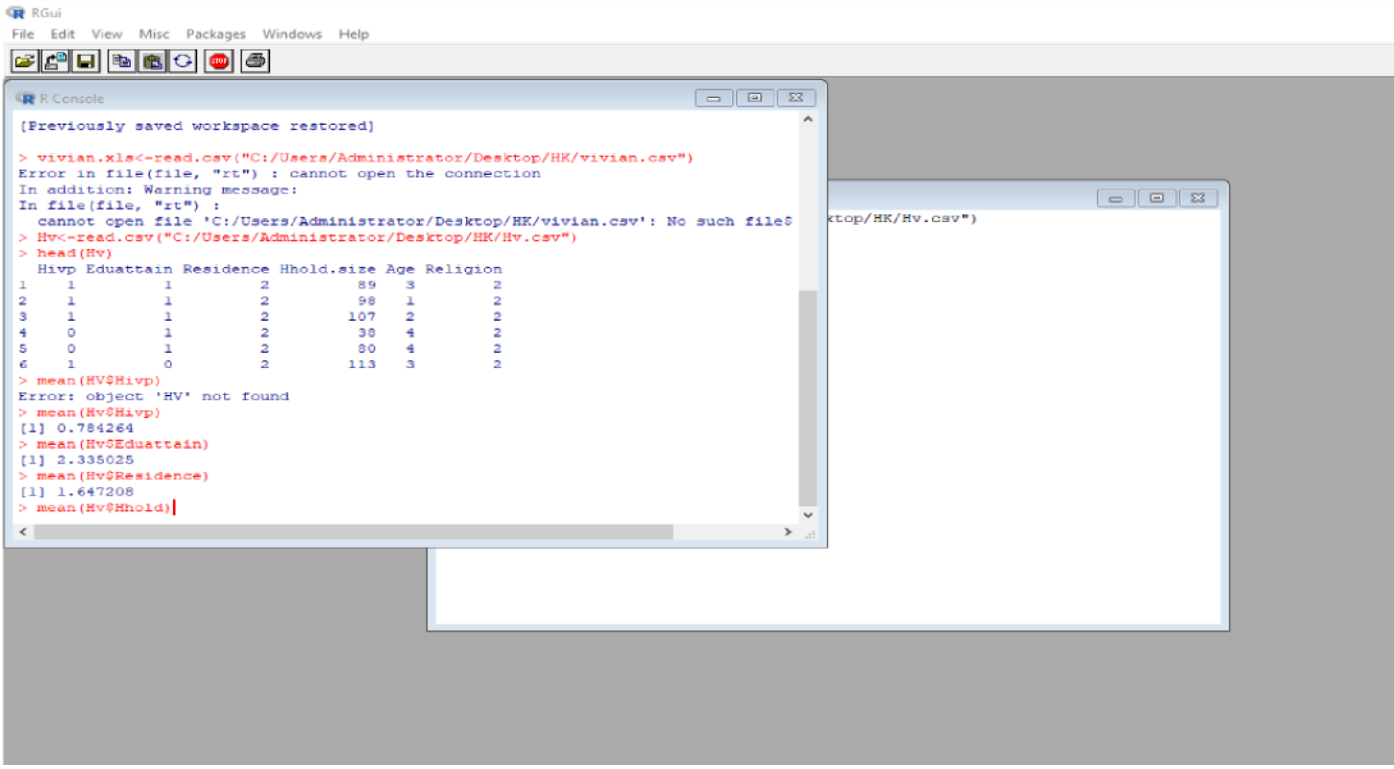
Sample Excel file HIV datasheet for loading into R-package.



	A	B	C	D	E	F
1	Hivp	Eduattain	Residence	Hhold size	Age	Religion
2	1	1	2	89	3	2
3	1	1	2	98	1	2
4	1	1	2	107	2	2
5	0	1	2	38	4	2
6	0	1	2	80	4	2
7	1	0	2	113	3	2
8	1	2	2	20	4	2
9	1	3	2	51	2	2
10	1	2	2	134	3	2
11	0	1	2	34	4	1
12	1	1	2	62	3	2
13	0	1	2	34	2	2
14	1	1	2	56	2	4
15	1	3	2	79	2	2
16	0	1	2	124	2	1
17	1	1	2	10	2	2
18	1	4	2	19	2	2
19	1	1	2	38	3	2
20	1	1	2	56	2	1
21	1	4	2	72	3	2



	A	B	C	D	E	F
375	1	2	1	16	2	1
376	1	1	1	20	1	1
377	1	3	1	23	3	2
378	0	1	1	33	1	1
379	1	2	2	3	1	1
380	1	2	2	9	2	1
381	0	1	2	13	1	2
382	1	1	2	29	1	2
383	1	1	2	33	4	1
384	1	1	2	39	2	1
385	1	3	2	18	4	2
386	0	1	2	27	2	1
387	0	5	2	42	3	2
388	1	1	2	27	4	2
389	1	5	2	70	5	2
390	1	1	2	35	2	1
391	1	3	2	42	2	2
392	1	2	2	56	2	2
393	1	1	2	63	4	1
394	1	1	2	8	3	2
395	1	1	2	20	4	1



ACTIVITIES COST

Printing and stationeries	1000
Transport	700
Online research	3000

R CODES

R-codes for SLogistic regression table + test of association among

variables Hv<- read.csv("C:/Users/Administrator/Desktop/HK/Hv.csv")

head(Hv) var(Hv$Residence) sd(Hv$Residence)

Hv<data.frame(Hv $Hivp,Hv $Eduattain,Hv $Residence,Hv $Hhold.size,Hv $Age,Hv $Religion) summary_data <- summary(Hv)

standard_errors <- sapply(Hv,function(x) sd(x) / sqrt(length(x))) print("Standard Errors of the Mean:")print(standard_errors)

Hv<data.frame(Hv $Hivp,Hv $Eduattain,Hv $Residence,Hv $Hhold.size,Hv $Age,Hv $Religion)

Hv_positive_count <- sum(Hv$Hivp== 1) total_observation_count

<- nrow(Hv)

percentage_Hv_positive <- (Hv_positive_count / total_observation_count) * 100 cat("Percentage of those who had HIV:", percentage_HIV_positive,"%\n")

Hv<data.frame(Hv $Hivp,Hv $Eduattain,Hv $Residence,Hv $Hhold.size,Hv $Age,Hv $Religion)

```
Residence_urban_percentage <- mean(HIV $Residence == 1) * 100

Residence_rural_percentage <- mean(HIV $Residence == 2) * 100

cat("Percentage of individuals in urban (Residence = 1):", Residence_urban_percentage, "%\n")

cat("Percentage of individuals in rural (Residence = 0):", Residence_rural_percentage,"%\n")

Hv<-data.frame(Hv $Hivp,Hv $Eduattain,Hv $Residence,Hv $Hhold.size,Hv $Age,Hv

$Religion) model <- glm(Hv ~ ., data = Hv, family = binomial) summary_table <-
summary(model)  coefficients <- summary_table$coefficients standard_errors <-
summary_table$coefficients[, "Std. Error"]  z_values <- coefficients[, "z value"] p_values
<- coefficients[, "Pr(>|z|)"]  odds_ratios <- exp(coefficients[, "Estimate"]) lower_ci <-
exp(coefficients[, "Estimate"] - 1.96 * standard_errors)  logistic_regression_table <-
data.frame( Predictor = rownames(coefficients), Coefficient =
coefficients[,  "Estimate"],  Standard_Error_Coefficient  =  standard_errors,  Z
= z_values,P_Value  =  p_values,  Odds_Ratio  =  odds_ratios,

Lower_CI_95  = lower_ci)print(logistic_regression_table)
Hv<-data.frame(Hv $Hivp,Hv $Eduattain,Hv $Residence,Hv $Hhold.size,Hv $Age,Hv

$Religion) model<- glm(Hv ~ ., data =Hv, family = binomial)  summary_table <-

summary(model) coefficients <- summary_table$coefficients  standard_errors <-

summary_table$coefficients[, "Std. Error"] z_values <- coefficients[, "z value"]p_values

<-  coefficients[,  "Pr(>|z|)"]  odds_ratios  <-  exp(coefficients[,

"Estimate"])lower_ci <- exp(coefficients[, "Estimate"] - 1.96 * standard_errors)

HIV_prevalence<- read.csv("C:\\Users\\Administrator\\Desktop\\Hv.csv")

head(HIV_prevalence)

HIV_prevalence$Eduattain<-factor(HIV_prevalence$Eduattain,  levels=c(0,

1, 2, 3, 4,5), labels=c("No education", "Incomplete primary","Complete
```

```
primary",  "Incomplete secondary", "Complete secondary", "Others"))

print(HIV_prevalence$Eduattain)



HIV_prevalence$Religion<-factor(HIV_prevalence$Religion,  levels=c(1,

2, 3, 4), labels=c("Roman catholic",

"Protestant","Muslim",

 "Pagan")) print(HIV_prevalence$Religion)

HIV_prevalence$Residence<factor(HIV_prev

alence$Residence, levels=c(1, 2),

labels=c("Rural", "Urban"))

print(HIV_prevalence$Residence)

Logit_model<-glm(Hivp ~ Age + Hhold.size + Religion +  Residence + Eduattain, data =
HIV_prevalence) summary(Logit_model)   logistic_regression_table <-
data.frame(Predictor = rownames(coefficients), Coefficient =
coefficients[,          "Estimate"],  Standard_Error_Coefficient  =          standard_errors,          Z

     = z_values,P_Value  =          p_values,          Odds_Ratio  =          odds_ratios,
    Lower_CI_95  = lower_ci)print(logistic_regression_table)

Logit model<- glm(Hivp ~ Eduattain + Hhold.size + Age + Religion, data = HIV_data, family

=binomial)  summary(logit_model)   logit_model  <-  glm(Hv  ~  LITERACY  +  Eduattain  +

Hhold.size + Age +Residence + Religion,

data = Hv_data, family = binomial) summary(logit_model)

Hv<-data.frame(Hv $Hivp,Hv $Eduattain,Hv $Residence,Hv $Hhold.size,Hv $Age,Hv

$Religion) model <- glm(Hv ~ ., data = Hv, family
```

= binomial) summary(model)

Hv<-data.frame(Hv $Hivp,Hv $Eduattain,Hv $Residence,Hv $Hhold.size,Hv $Age,Hv $Religion) chi_square_result <- chisq.test(Hv $Hv,Hv(

"Residence", "Age",)])

Hv<data.frame(Hv $Hivp,Hv $Eduattain,Hv $Residence,Hv $Hhold.size,Hv $Age,Hv $Religion)

Hv $EdLevel <- as.character(Hv $EdLevel)

Hv $Residence <- as.character(Hv $Residence)

Hv $wealth <- as.character(Hv $wealth) Hv $Religion <- as.character(Hv

$Religion) chi_square_result_Eduattain <- chisq.test(table(Hv $Hivp, Hv

$Eduattain)) print("Chi-square test for EdLevel:")print(chi_square_result_Eduattain)

chi_square_result_Residence <- chisq.test(table(Hv $Hivp, Hv $Residence))

print("Chisquare test for Residence:")print(chi_square_result_Residence)

chi_square_result_wealth <- chisq.test(table(Hv $Hivp, Hv $Age))

print("Chi-square test for Age:")

print(chi_square_result_Age) chi_square_result_Religion <- chisq.test(table(Hv $Hivp, Hv

$Religion)) print("Chi-square test for Religion:")

print(chi_square_result_Religion) print(chi_square_result)

Hv<-data.frame(Hv $Hivp,Hv $Eduattain) chi_square_result

<- chisq.test(table(Hv $Hivp, Hv $Ed)) print(chi_square_result)

R-codes for summary statistics for Hiv + distribution

```
install.packages(c("readxl", "dplyr", "writexl", "ggplot2"))

library(readxl) library(dplyr) library(writexl) library(ggplot2)

file_path <- "~/Vivian.xlsx" sheet_names <-

excel_sheets(file_path) hivp_summary_all <- list() for (sheet

in sheet_names) {  df <- read_excel(file_path, sheet = sheet)

if ("Hivp" %in% names(df)) {     summary_df <- df %>%

group_by(Hivp) %>%       summarise(Count = n(), .groups =

"drop") %>%       mutate(Proportion = round(Count /

sum(Count), 3),

                Sheet = sheet)

 hivp_summary_all[[sheet]] <- summary_df

       }

     }

combined_hivp_summary <- bind_rows(hivp_summary_all)

write_xlsx(list(Hivp_Distribution = combined_hivp_summary),

path = "~/Hivp_distribution_summary.xlsx")


ggplot(combined_hivp_summary, aes(x = Hivp, y = Count, fill
```

```
= Sheet)) +   geom_bar(stat = "identity", position =

"dodge") +   labs(title = "Distribution of Hivp by

Sheet",        x = "Hivp Category", y = "Count") +

theme_minimal()

ggsave("~/Hivp_distribution_plot.png")



if (!require("readxl")) install.packages("readxl", dependencies =

TRUE)

    if (!require("ggplot2")) install.packages("ggplot2", dependencies

    = TRUE)
```

R-codes for age distribution

```
library(readxl) library(ggplot2) data <-

read_excel("~/Vivian.xlsx", sheet = 1) names(data)  # Use

this to identify the correct column for age data$Age <-

as.numeric(data$Age) data <- data[!is.na(data$Age) &

data$Age >= 0, ] ggplot(data, aes(x = Age)) +

geom_histogram(binwidth = 5, fill = "skyblue", color =

"black", alpha = 0.8) +   labs(title =

    "Age Distribution",        x =
```

```
    "Age (years)",        y =
    "Count") +   theme_minimal()


if (!require("readxl")) install.packages("readxl", dependencies =
TRUE)
    if (!require("ggplot2")) install.packages("ggplot2",
    dependencies = TRUE) if (!require("dplyr"))
    install.packages("dplyr", dependencies
    = TRUE)
```

R-codes for boxplot of age by Hiv status

```
library(readxl) library(ggplot2) library(dplyr) data <-
read_excel("~/Vivian.xlsx", sheet = 1) colnames(data) <-
make.names(colnames(data)) formats --- data$Hivp <-
as.numeric(data$Hivp)  # Assuming Hivp is coded as 0/1
data$Hhold.size <- as.numeric(data$Hhold.size) data_clean <-
data %>%   filter(!is.na(Hivp) & !is.na(Hhold.size)) Size ---
ggplot(data_clean, aes(x = Hhold.size, y = Hivp)) +
geom_jitter(width = 0.2, height = 0.05, alpha = 0.6, color =
```

```
    "darkgreen") +   labs(title = "Scatter Plot of Household Size
    vs HIV Status",       x = "Household Size",       y = "HIV
    Status (0 = Negative, 1 = Positive)") +   theme_minimal()

if (!require("ggplot2")) install.packages("ggplot2",
dependencies = TRUE) library(ggplot2) data$Age <-
as.numeric(data$Age) data$Hivp <-
as.numeric(data$Hivp) data <- data[!is.na(data$Age) &
!is.na(data$Hivp), ]

ggplot(data, aes(x = factor(Hivp), y = Age, fill = factor(Hivp)))
+   geom_boxplot() +   labs(title = "Box Plot of Age by
    HIV Status",       x = "HIV Status (0 = Negative, 1 =
    Positive)",       y = "Age (years)") +
    scale_fill_manual(values = c("lightblue", "salmon")) +
    theme_minimal()

if     (!require("ggplot2"))     install.packages("ggplot2",
dependencies    =    TRUE)    if    (!require("reshape2"))
install.packages("reshape2", dependencies = TRUE)
```

R-codes for correlation Heatmap

```
library(ggplot2) library(reshape2) numeric_data <-

data[sapply(data, is.numeric)] cor_matrix <-

cor(numeric_data, use = "complete.obs") melted_cor

<- melt(cor_matrix)



ggplot(data = melted_cor, aes(x = Var1, y = Var2, fill = value))

+   geom_tile(color = "white") +   scale_fill_gradient2(low

    = "blue", high = "red", mid =

    "white",                  midpoint = 0, limit = c(-1,1), space

    = "Lab",                  name = "Correlation") +

     theme_minimal()     +            theme(axis.text.x     =

    element_text(angle = 45, vjust = 1,

                              size = 12, hjust = 1)) +

     coord_fixed() +   labs(title =

    "Correlation Heatmap",       x = "",

    y = "")
```

R-codes for Hiv summary statistics

```
if (!require("writexl")) install.packages("writexl", dependencies

= TRUE)
```

```
library(writexl) hivp_summary
<- data.frame(
  Count_Negative = sum(data$Hivp == 0, na.rm = TRUE),
  Count_Positive = sum(data$Hivp == 1, na.rm = TRUE),
  Proportion_Negative = mean(data$Hivp == 0, na.rm =
TRUE),
  Proportion_Positive = mean(data$Hivp == 1, na.rm =
TRUE),
  Mean = mean(data$Hivp, na.rm = TRUE),
  SD = sd(data$Hivp, na.rm = TRUE)
)
write_xlsx(hivp_summary, "Hivp_Summary_Statistics.xlsx")
```

R-codes for numeric summary statistics

```
if (!require("writexl")) install.packages("writexl", dependencies
= TRUE)
library(writexl) numeric_data <-
data[sapply(data, is.numeric)] summary_stats
<- data.frame(
  Variable = names(numeric_data),
```

```
  Min = sapply(numeric_data, min, na.rm = TRUE),
  Q1 = sapply(numeric_data, quantile, probs = 0.25, na.rm =
TRUE),
  Median = sapply(numeric_data, median, na.rm = TRUE),
  Mean = sapply(numeric_data, mean, na.rm = TRUE),
  Q3 = sapply(numeric_data, quantile, probs = 0.75, na.rm =
TRUE),
  Max = sapply(numeric_data, max, na.rm = TRUE),
  SD = sapply(numeric_data, sd, na.rm = TRUE),
  NAs = sapply(numeric_data, function(x) sum(is.na(x)))
)

write_xlsx(summary_stats,
"Numeric_Summary_Statistics.xlsx")
```