MODELING THE EFFECT OF MEDIATION ON HIV PREVALENCE IN KENYA USING A LOGISTIC REGRESSION MODEL

Ruth Naomi Wanga

A Research Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in Statistics of Masinde Muliro University of Science and Technology.

March, 2023

DECLARATION

The thesis is my original work prepared with no other than indicated sources and support and has not been awarded elsewhere for a degree or any other award.

Signature..... Date .....

Ruth Naomi Wanga

SES/G/01-52824/2018

CERTIFICATION

The undersigned certify that they have read and hereby recommend for acceptance of Masinde Muliro University of Science and Technology a research thesis entitled **"Modeling the Effect of Mediation on HIV Prevalence in Kenya Using a Logistic Regression Model."**

Signature..... **Date** .....

Dr. David Anekeya Alilah

Department of Mathematics

Masinde Muliro University of Science and Technology

Signature..... **Date** .....

Dr. Everlyne Akoth Odero

Department of Mathematics

Masinde Muliro University of Science and Technology

COPYRIGHT

DEDICATION

I dedicate this work to my family, siblings and friends for their love, encouragement and support. May the living God bless them abundantly for their endless contributions towards this work.

ACKNOWLEDGEMENTS

I am thankful and grateful to the Almighty God for the gift of life, good health, care, protection, knowledge and wisdom that He granted me on the journey for the pursuit of this MSc degree. My sincere gratitude and appreciation to my supervisors, Dr. David Alila and Dr. Everlyne Akoth Odero for their continuous support throughout the research work. It is through their intellectual input, resourcefulness, dedication, effort and prudent scientific and moral guidance that am acquainted with a lot of knowledge and skills in various aspects of my research. I thank the entire staff of Mathematics department for their encouragement and support they offered during the entire period of my research. I cannot forget to thank Masinde Muliro University of Science and Technology for permitting my studentship in MSc Statistics. My gratitude goes to my dear family, parents and siblings who encouraged me throughout the entire research work.

ABSTRACT

Over the last decades, major global efforts mounted to address the HIV epidemic has realised notable successes in combating the pandemic. Sub Saharan Africa still remains a global epicenter of the disease, accounting for more than 70% of the global burden of infections. Despite the widespread use of HIV mass media national campaigns as an intervention in HIV prevention due to its numerous advantages since the mid-1980s, HIV prevalence still remains a challenge in especially in some geographic areas and populations. Therefore how HIV mass media interacts with the prevailing HIV risk factors to cause an impact on HIV prevalence remains a question not answered. This study considered Exposure to HIV related media as a mediator variable believed to mediate the relationship between HIV risk factors and HIV prevalence. Two logistic regression models were formulated and used to compare the model with mediation and that without mediation in order to establish the effect of mediation on HIV prevalence. Models were fitted to real data from 2018 Kenya Population-based HIV Impact Assessment survey and estimation of the model parameters was done using Maximum Likelihood Estimation in R. Results of R analysis based on both Akaike's Information Criterion and the McFadden's R^2 value for model with mediation revealed that the model formulated in presence of mediation was better compared to that without mediation since the effects of mediation variable were found to be more significant on HIV prevalence.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS AND ACRONYMS

AIDS	:	Acquired Immunodeficiency Syndrome
AIC	:	Akaike Information Examination
DIC	:	Deviance Information Criterion
GIS	:	Geographic Information System
HIV	:	Human Immunodeficiency Virus
KDHS	:	Kenya Demographic and Health Survey
KENPHIA	:	Kenya Population-based HIV Impact Assessment
KNASP	:	Kenya National AIDS Strategic Plan
MLE	:	Maximum Likelihood Estimation
MSE	:	Mean Squared Error
NACC	:	National AIDs Control Council
NASCOP	:	National AIDS and STI Control Programme
pR^2	:	Pseudo r-squared
pscl	:	Political science computational laboratory
PLWHA	:	People Living With HIV/AIDs
UNAIDS	:	United Nations Programme on HIV/AIDS
WHO	:	World Health Organization

List of Tables

List of Figures

DEFINITION OF TERMS/CONCEPTS

HIV/AIDS	:	An infectious disease caused by te Human Immuno Defficiency Virus that affects and destroys te immune system rendering people more prone to oppotunistic diseases.
Intervention	:	This is any action intended to reduce or avert exposure or the likelihood of exposure to sources which are not part of a controlled practice or which are out of control as a consequence of an accident.
Logistic Regression Model	:	is a statistical analysis method used to predict a binary outcome based on prior observations of a data set. The model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables.
Mass media	:	the means of communication that reach large numbers of people in a short time, such as television, newspapers, magazines, and radio.
Mediator variable	:	Is an intermediate variable between an exposure and the outcome, which is influenced by the exposure on the causal pathway to the outcome.
Mediation	:	is a mechanism by which an independent variable causes change in a dependent variable.
Model formulation	:	is the step where our knowledge of a natural system is translated in mathematical form. It involves the construction of a conceptual model and the formulation of this conceptual model into mathematical equations.
Risk Factors	:	Something that increases the chance of developing a disease. Some examples of risk factors for cancer are age, a family history of certain cancers, use of tobacco products, being exposed to radiation or certain chemicals, infection with certain viruses or bacteria, and certain genetic changes.

CHAPTER ONE

INTRODUCTION

1.1 Background to the Study

Lindsay [21], describes Acquired Immuno Deficiency syndrome (AIDS) as an infectious disease caused by the Human Immuno deficiency Virus (HIV) that affects and destroys the immune system rendering people more prone to opportunistic infections.

According to UNAIDS Global report [11], 39.0 million people worldwide were living with HIV/AIDS, with 1.3 million being the new infections. Over two thirds of all these persons living with HIV/AIDs are found in sub-Saharan Africa. In 2017 Kenya was ranked as one of the country hard hit by the HIV /AIDS epidemic in terms of the estimated number of new HIV infections among adults aged 15 years and older, however in recent years, Kenya has recorded steady progress in HIV/AIDS prevention [45].

The Kenya HIV Estimate report [32], shows that currently Kenya has a HIV prevalence rate of 4.5% among adults between the ages of 15-49 years and 4.9% among adults between the ages of 15-64 years. The variations in HIV prevalence rates cuts across Counties in Kenya ranging from 20% among Counties located around Lake Victoria and 1% among Counties in North Eastern region. The variation further cuts across gender and age. The women having higher HIV prevalence (6.2%) compared to the 3.1% among male , while adults aged 15-64 years having a higher prevalence rate (4.9%) compared to children aged 0-14 years

[37].

Prevention of the HIV epidemic, like other infectious diseases, depends on having a good understanding of the determinants of the spread of the infections and being able to explain its trends in disease magnitude and evaluation of intervention programs, [38]. In sub-Saharan Africa, a number of variables have been linked to the risk of contracting HIV, including individual demographic traits like gender, age, and marital status, socioeconomic status like education and wealth, cultural practices like religion and circumcision, and risk factors related to sexual behaviors [19, 28].

A study by [28] to examine the determinants and cross-national variations in the risk of HIV seropositivity in SSA using multilevel logistic regression models to Demographic and Health Survey data collected during 2003–2008 from 20 countries of SSA revealed that there was gender disparity in socio-economic risk factors, partly explained by sexual behaviour, background socio-economic risk factors were stronger predictors among women than men and there were generally variations in the risk of HIV across countries and regions. Therefore interventions must address specific determinants in order to be effective [38].

According to [20], Mass media interventions are useful in reducing global HIV/AIDS disparities because of their wide reach, standardization and repetition of messages, and the ability to use different content formats, including entertainment, news, and short advertisements or announcements. Intervention strategies applied to prevent HIV infection and to improve the living standards of HIV-infected persons are some of the underlying mechanisms barely estimated in most of the previously

conducted logistic regression studies.

According to [17], Interventions interact with HIV risk factors including the social, economic, legal, political, and built environments that underlie processes and outcomes affecting population health. Prevention efforts are strategically important in meeting the current need in HIV/AIDS prevention by not only identifying the correct and most effective strategies but also targeting the right population for the specific intervention [31].

Exposure to HIV related media campaigns is one of the major interventions employed in Kenya to combat HIV/AIDs infections. This is a powerful tool that educates both the HIV negative persons on avoiding to contract HIV virus and the HIV positive persons on better living standards. However, all these depends on how the information is delivered to the individuals [16]. The biggest challenge with mass media campaign lies in disseminating accurate, objective, balanced and non-judgemental information on HIV/AIDs to individuals, which implies that exposure to media varies across population groups.

While many studies such as [20, 2, 30] have addressed the relationship between the use of mass media and HIV Prevalence in distinct contexts, due to the complexity of this interaction, little is known about how different risk factors interact with Exposure to HIV related media strategies to reduce HIV Prevalence considering that exposure to HIV related mass media is a mediator variable between risk factors and HIV prevalence.

More so, previous studies on HIV/AIDs in Kenya have majorly focused on individual risk factors, ignoring the significance of the social en-

vironment structure in which the individuals live, which has a significant impact on the individuals' health and behavior [24]. However recent developments in statistical models makes it possible to test in studies that seek to examine the additive and interactive effects of individual-level and contextual factors that affect sociological outcomes at the individual level.

The purpose of this study is to establish the effect of exposure to HIV related media intervention in the relationship between various HIV risk factors and HIV prevalence in Kenya using logistic regression modeling. It focuses mainly on the effect of mediation variable such as Exposure to HIV related media on the relationship between various HIV risk factors and HIV prevalence in Kenya.

1.1.1 Risk factors

Risk factors are attributes, characteristics or exposures that increase the chances of a person developing a disease. The HIV epidemic is extremely heterogenic and dynamic and thus a good understanding of county-specific transmission determinants is important in determining the effective HIV/AIDs prevention and control strategies. Most population-based surveys on HIV provide an opportunity to link HIV status with behavioral, social, biological and other risk factors which vary across communities [41].

1.1.2 Exposure to mass media

Mass media has been an important strategy for many health behavior change topics, including heart disease, smoking, family planning, and HIV/AIDS prevention since 1960s. Mass media use in HIV prevention

intervention has been greatly used over other interventions due to its strength in wide reach, standardization and repetition of messages, and the ability to use different content formats, including entertainment, news, and short advertisements or announcements [20]. It is a powerful tool that educates both the HIV negative persons on how to avoid contracting HIV virus and the HIV positive persons on better living standards [16]. The success of all these depends on how the information is delivered, received and utilized by the individuals. This implies that their effect on HIV prevalence varies based on geographical location, population among other factors.

1.1.3 Mediation Analysis

Mediation is described as a mechanism by which an independent variable causes change in a dependent variable. The independent variable causes change in the mediator, which in turn cause change in the dependent variable, hence the effect of an independent variable is at least partially transmitted through a mediator variable to the dependent variable, [13].

The primary objective of mediation analysis is to divide the Total effect (TE) of exposure variable on outcome variable into an indirect effect (IE) via the mediator and a direct effect (DE) whose effects are exclusively attributable to exposure, [7, 15]. Total effect is described as the expected effect of a change in independent variable on outcome variable [14].

According to partial mediation used in this study, there is a direct relationship between the independent and dependent variables as well as a substantial relationship between the mediator and the dependent variable [14]. This suggests that when estimating the impact of expo-

sure variables on outcomes, both a direct and indirect effect must be considered.

Figure 1.1 shows a conceptual diagram of a simple mediation model with a single mediator variable. According to this model, a single mediator variable M can affect the outcome variable Y by at least one exposure variable X. This suggests that there are two possible methods of X affecting Y. According to [14], one route connects X and Y directly and is known as the direct effect of X to Y, whereas the other route connects X and Y via a mediator M and is known as the indirect effect of X to Y.

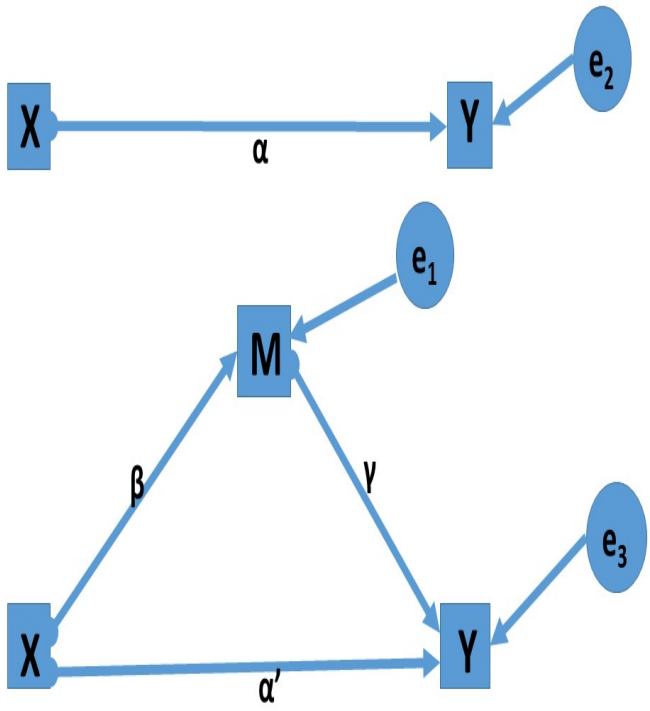

Figure 1.1: **Path diagram: A Simple Mediation Model**

The total effect is represented by α. The direct effect is represented by $\alpha^{'}$. The effect of independent variable on mediator variable is represented by β while that between the mediator and the dependent variable is shown by γ. Mediation is commonly assessed using three standard methods including: causal steps, difference in coefficients and product of coefficients [15]. For a basic single mediator model in Figure 1.1, a series of Ordinary Least Square regression equations below are sufficient to answer the main questions about mediation [14].

$$Y = \kappa_1 + \alpha X + e_1 \tag{1.1}$$

$$M = \kappa_2 + \beta X + e_2 \tag{1.2}$$

$$Y = \kappa_3 + \alpha^{'} X + \gamma M + e_3 \tag{1.3}$$

Equation 1.1 shows the total effect α of the independent variable on dependent variable, Equation 1.2 shows the effect of independent variable, β on the mediator and Equation 1.3 shows the direct effect, $\alpha^{'}$ of independent variable on dependent variable when the effect, γ of mediator on dependent variable, is incorporated in the model. κ_1, κ_2 and κ_3 are intercepts for each of the three equations, while ϵ_1, ϵ_2 and ϵ_3 are respective residuals assumed to follow a normal distribution with mean 0 and variance δ_1^2, δ_2^2 and δ_3^2 respectively.

1.2 Statement of the problem

According to [11] HIV/AIDs claims a life almost every minute in the world and this has culminated to about 630,000 people who have died

of HIV related illness. 39.0M people globally are HIV infected and 1.3M are the new infections. Sub-Saharan Africa though with almost 11% of the worlds population continues to be the global epicenter HIV epidemic, accounting for 70% disease burden.

Kenya has recently shown a steady decline in HIV prevalence which currently stands at 4.9% among population aged 15-64 years. However the UNAIDS/WHO AIDS Epidemic Update shows that the actual number of infected individuals is still rising as a result of new infections and longer life expectancy brought on by the use of anti-retro-viral medications. The Kenya Estimate report [32] further reveals the variations in HIV prevalence across counties and population groups in Kenya.

Prevention of HIV epidemic depend on a good understanding of the determinants of HIV spread, their specific magnitude on the disease burden and intervention program needed to curb the epidemic [38]. A number of factors including the demographic, cultural, behavioral factors have been linked to the risk of contracting HIV in Kenya. A study conducted by [28] using multi-level logistic regression modeling shows that there was generally variations in the risk of HIV across counties and regions, implying need for specific interventions for specific HIV determinants. Mass media is one of the interventions greatly used in HIV prevention due to its numerous advantages over other interventions. It has a wide reach, has standardized repetitive messages and has ability to use different content formats such as entertainment, news, advertisements among others. [20]. These intervention interact with HIV risk factors to built an environment that underlie processess and outcomes affecting population health [17].

Many study including [38, 28, 5] have addressed the relationship between HIV risk factors and HIV prevalence without considering a mediator variable. In addition, a number of studies such as [13, 4] have been done to assess mediation in HIV intervention with various objectives. However, none of the studies explicitly describes the effect of mediator variable in terms of interventions strategies used to control HIV on HIV prevalence in Kenya.

This study models the effect of mediation on HIV prevalence in Kenya using a logistic regression model in order to establish the effect of an intervention, specifically Exposure to HIV related mass media on HIV prevalence in Kenya.

1.3 Objectives of the Study

1.3.1 Main Objective

The main objective of this study is to model the effect of mediation on HIV prevalence using a Logistic regression model.

1.3.2 Specific Objectives

The specific objectives of this study included:

1. To formulate two Logistic regression models; one in the absence of mediation and another in the presence of mediation.

2. To estimate the effect of parameters of the fitted models on HIV/AIDs prevalence using Maximum Likelihood Estimation approach.

3. To compare the performance of the two models and evaluate the adequacy of the model fit.

1.4 Significance of the Study

The study contributes to better understanding of the relationship between an independent variable and a dependent variable when these variables do not have an obvious direct connection. This can inform policy makers in formulation of appropriate intervention strategies aimed at reducing HIV/AIDS prevalence.

1.5 Justification of the Study

Understanding the processes underlying the effect of HIV risk factors on HIV Prevalence through mediation analysis can help improve the effectiveness and help minimize on cost of HIV prevention and treatment interventions.

1.6 Methods of study

1.6.1 Logistic Regression Analysis

Logistic regression analysis studies the association between a categorical dependent variable and a set of independent variables when the dependent variable has only two values, such as 0 and 1 or Yes and No. Logistic regression is suited for analyzing dichotomous outcomes and has been increasingly applied in social science research. It does not require many of the principle assumptions of linear regression models that are based on ordinary least squares method particularly regarding linearity of relationship between the dependent and independent variables, normality of the error distribution, homoscedasticity of the errors, and measurement level of the independent variables. Logistic regression can handle non-linear relationships between the dependent and independent

variables, because it applies a non-linear log transformation of the linear regression [35].

1.6.2 Maximum Likelihood Estimation

This is a method of estimating the parameters of an assumed probability distribution, given some observed data. According to [10] Estimation of parameters of the logistic regression model using the maximum likelihood method involves differentiating the likelihood function, then set this first derivative to 0, and continue to solve the equation to obtain the estimate of parameters. This study employed MLE to estimate the parameters in the models by partially differentiating the log of the likelihood function and equating the results to zero. That is;

$$\frac{\partial ln(L)}{\partial(\theta)} = 0$$

1.6.3 Akaike's Information Criterion (AIC)Approach

AIC is a mathematical method for evaluating how well a model fits the data it was generated from. It is used to compare different possible models and determine which one is the best fit for the data. [6] defined AIC as follows

$$AIC = 2K - 2logL(\beta) \tag{1.4}$$

where K represents the number of parameters in the model and $L(\beta)$ denotes the maximum value of the likelihood function in the model.

Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. This approach was

therefore used in this study to estimate the relative amount of information lost while fitting the model with mediation and the model without mediation. The less information a model looses the higher the qaulity of the model. The model with the least value of AIC is the best model [?].

1.6.4 McFadden's Pseudo-R^2

Among the several Pesudo R^2 measures suggested for assessment of the predictive strength of categorical models, a number of previous studies preferred McFadden's R^2, considering its easy computation, intuitive interpretation, base-rate stability in binary models and possible information theory interpretation [46].

According to Smith, [42] McFadden's R^2 is a metric computed using pR^2 function from the *pscl* package in R. It ranges from 0 to 1, with values close to 0 indicating that the model has no predictive power and values over 0.40 indicating that a model fits the data very well. McFadden's R squared measure is defined as;

$$R^2_{McFadden} = 1 - \frac{LogL(\beta)}{LogL(null)} \tag{1.5}$$

where $L(\beta)$ denotes the (maximized) likelihood value from the fitted model, and $L(null)$ denotes the corresponding value but for the null model – the model with only an intercept and no covariates.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This chapter discusses the literature that exists on HIV risk factors, mediation analysis and the use of Logistic regression in modeling HIV prevalence in relation to the proposed study.

According to Global AIDS report [11], HIV/AIDS has spread at an alarming rate worldwide since first AIDS diagnoses in early 1980s, with the number of new infections rising each year. SSA remains the global epicenter of HIV epidemic accounting for the 2/3 of the new infections in the world. Globally 39 million people are living with HIV/AIDS, 1.3 million people are newly infected and about 630,000 people have died from AIDS related illness [29].

According to Kenya AIDS Response Report [32], Despite the progress made by Kenya in advancing towards the UNAIDS 90-90-90 targets for ending the HIV/AIDs epidemic, there still exist a significant geographic variation in HIV prevalence rates among counties and regions. The Highest prevalence of 20% recorded in counties located around Lake Victoria region and lowest of 1% recorded in counties in North Eastern region.

Johnson *et.al* [18], demonstrated on how HIV is a multidimensional epidemic and various risk factors such as demographic, residential, social, biological, and behavioral factors exerted varying effects on HIV infection for different individuals. Therefore a good understanding of

population-specific transmission risk factors both direct and indirect factors, could be helpful in designing effective mediation for specific population.

Moineddin et al [24] indicated that while individual risk factors have received a lot of attention, the social environment in which people live has been largely ignored despite the fact that it has a significant impact on people's health and behavior. A multilevel models have been identified as highly appropriate in assessing how context affects individual-level health risks and outcomes. Recent developments in statistical models have also made it possible to test sociological theories by enabling researchers to examine the additive and interactive effects of individual-level and contextual factors that affect sociological outcomes at the individual level [28].

According to the study by [44] on Religion and Women's Health in Ghana, one's religious affiliation significantly affects their understanding about AIDS and is linked to changes in some preventative behaviors, most notably the usage of condoms.

Huberman *et al* [15], in their study on Estimating the drivers of species distributions with opportunistic data using mediation analysis, described a spatial estimation method for supplementarily including additional opportunistic data using mediation analysis concepts. The opportunistic data mediate the effect of the covariate on the designed-survey data response, decomposing it into a direct and indirect effect. A component of the indirect effect was then estimated via regressing the mediator on the covariate, while the other components are estimated through a spatial occupancy model. Simulation results suggested that

the mediated method produced an improvement in relative MSE for reasonable quality data. However, the standard, unmediated method is more preferable if the simulated opportunistic data are poorly correlated with the true spatial process.

Magadi and Desta conducted a study on the multilevel analysis of the determinants and cross-national variations of HIV seropositivity in sub-Saharan Africa using data from the Demographic and Health Surveys, collected between 2003 and 2008 from 20 different sub-Saharan African countries, and examined the determinants and cross-national variations in the risk of HIV seropositivity in the region Magadi and Desta, [28]. Using simultaneous confidence intervals of country-level residuals, the risk of being HIV seropositive was compared across nations. The study found that sexual behavior characteristics partly explained some of the gender differences in socioeconomic risk factors.

In a study to estimate HIV prevalence and characterize risk factors among young adults in Asembo, rural western Kenya as conducted by [5], a Community-based cross-sectional survey was designed and potential study participants who included the residents aged 13-34 years for the period October 2003 - April 2004 were randomly selected through stratified sampling by sex and age group using the Demographic Surveillance System (DSS) as a sampling platform and were interviewed on risk behavior and tested for HIV and Herpes Simplex Virus 2 (HSV-2) on voluntary basis. The results of study showed that HIV infection was strongly associated with age, higher number of sex partners, widowhood, and HSV-2 seropositivity in the multivariate models stratified by gender and marital status. The extremely high HIV and HSV-2 preva-

lence, and probable high incidence, were observed among young adults, suggesting that further research in the area. The current study fills this gap by confirming that introduction of a mediator factor, exposure of population especially the young adults to HIV related mass media plays a great role in lowering HIV prevalence.

Despite the many efforts to fight AIDS and the rising awareness of the disease, the epidemic continues to claim lives while imposing heavy costs on the Kenyan economy. A good understanding of the HIV risk factors and their respective magnitude on HIV Prevalence and intervention applied is required for effective HIV control [38].

A multilevel analysis of the determinants and cross-national variations of HIV seropositivity in sub-Saharan Africa using DHS data collected during 2003–2008 from 20 countries of sub-Saharan Africa, indicated that there are generally variations in the risk of HIV across counties and regions [28]. This implies that effective control of HIV requires specific interventions for specific risk factors.

Exposure to mass media through HIV/AIDs mass media campaigns is one of the major interventions employed in Kenya to combat HIV/AIDs infections. According to Irimu and Schwartz [16], Mass media is a powerful tool that educates both the HIV negative persons on how to avoid contracting HIV virus and the HIV positive persons on better living standards. However, all these depend on how the information is delivered to the individuals. The report showed that the biggest challenge with mass media campaign lied in disseminating accurate, objective, balanced and non-judgmental information on HIV/AIDs to individuals. This implied that exposure to media varies across population groups and

its impact on HIV control therefore varies depending on factors affecting the population groups or individuals.

Li,[22] conducted a study to identify the sources of HIV information for general Chinese population and examined how they affected HIV transmission knowledge and level of stigmatization towards People Living With HIV/AIDS (PLWHA). A face-to-face survey on market workers in Fuzhou, China was conducted and multiple regression models were used to describe correlations among respondents' HIV/STD information sources, HIV transmission knowledge, and stigmatizing attitude toward PLWHA. Television programs, newspapers, and magazines, were identified as most frequently used channels for HIV information and exposure to theses multiple sources had potential to improve HIV knowledge and reduce stigmatizing attitude toward PLWHA which in turn plays a great role in lowering HIV prevalence.

According to Myhre and Flora [23], effectiveness of HIV/AIDS Mass media campaigns as a HIV control intervention lies not only on the type of the channel of delivery but also in the level of exposure to HIV media messages by different group populations. Since level of exposure varies across different population groups, there is a probability of varying effects on HIV prevalence across counties as a result of exposure to mass media HIV/AIDs campaigns.

Agha [2], also revealed that the higher level of exposure to campaign media lead to more favourable outcomes such as safer sex, higher perceived self-efficacy in condom use negotiation, and higher perceived condom efficacy.

A lot of research has been done to address the relationship between

HIV risk factors and HIV prevalence. This includes a study by [5] that estimated HIV prevalence and characterized risk factors among adults in Asembo Community in Western Kenya. The study results revealed extremely high HIV and HSV-2 Prevalence among young adults and recommended further research on circumstances around HIV acquisition among the young adults which the current study considered. In addition, several studies on on assessing mediation in HIV Prevalence based on various objectives has been done. This includes study by [4, 13, 25]. However none of these studies explicitly describes the role of mediation on the relationship between HIV risk factors and HIV prevalence.

Mediation analysis is a process of establishing that dependent variable Y is influenced by independent variable X while being able to describe and quantify the mechanism responsible for that effect. It is popular among behavioral researchers as a means of testing hypothetical processes and mechanisms through which an independent variable, X, might affect a dependent variable, Y, indirectly through the mediating variable, M ([27, 26]).

Pirlot and Mackinnon [25], describes the limitations highlighted in previous studies on mediation by measuring confounders of mediation in research studies. He describes the various approaches of improving causal inference from a mediation study. The approaches discussed includ the comprehensive structural equation models, instrumental variable method, Principal stratification and the inverse probability weighting.

Aly, [4] describes the effect of both women's education and empowerment on receiving reproductive health care, using Multiple-Group Path

Analysis method based on the 2014 Egypt's Demographic Health Survey. The study examined the role of poverty and residence on the model where empowerment played a mediating role between education and women's access to reproductive health care. The study recommendation is in line with the current study on creation of a new societal culture through the Ministry of Education and the Ministry of Higher Education by reforming the educational curricula.

The study by [8], aimed at evaluating the effectiveness of five intervention strategies among persons with HIV (PWH) who are out of care using a systematic review of CDC's Prevention Research Synthesis (PRS). A descriptive analysis among the intervention strategies using SPSS was done and effectiveness of intervention was determined using standard meta analysis method. The results showed that the intervention strategies are effective in re-engagement, retention and viral suppression outcomes. This implies that several strategies are effective in improving HIV Care outcomes among PWH who are out of care. However this doesn't bring out the real picture of the underlying factors that interact with the intervention strategies in order cause an effective outcomes. The current study estimated the effect of one of the intervention strategies, Exposure to HIV mass media on HIV prevalence using mediation analysis and logistic regression modeling.

Huberman *et.al* [15], estimated the drivers of species distribution with opportunistic data using mediation analysis where opprotunistic data was used to mediate the effect of covariate on designed survey data response, decomposing it to direct and indirect effect. A component of indirect effect was then estimated via regressing the mediator on the

covariate. The decomposition of the total effect into direct and indirect effects is the key concept leveraged in our current study. The regression equation for capturing the total effect of X on Y is

$$Y = X\alpha + \epsilon_1 \tag{2.1}$$

where $X = (1, x)$; $\alpha = (\alpha_0, \alpha_1)^T$ and ϵ_1 represent a mean-zero normal error. Expectation of Y is therefore given as;

$$E(Y) = X\alpha \tag{2.2}$$

Give that γ is the effect of mediating variable M on Y given the effect of X, The regression Equation showing the decomposition of total effect into direct and indirect becomes;

$$M = X\beta + \epsilon_2 \tag{2.3}$$

and

$$Y/M = X\alpha^{'} + M\gamma + \epsilon_3 \tag{2.4}$$

where ϵ_2 and ϵ_3 are independent mean-zero normal errors.

Using the Equations 2.3 and 2.4, the expectation of Y becomes;

$$E(Y) = E(E(Y/M)) = E(X\alpha^{'} + M\gamma) = X\alpha^{'} + X\beta\gamma = X(\alpha^{'} + \beta\gamma) \tag{2.5}$$

The total effect can be equivalently expressed as

$$\beta = \alpha^{'} + \beta\gamma \tag{2.6}$$

where $\alpha^{'}$ and $\beta\gamma$ represent the direct and indirect effects, respectively [3]. In the current study, the logistic regression model with mediation was formulated based on the derivation by [3], where the total effect

was decomposed into direct and indirect effects. The model was fitted to both simulated data and real data from KENPHIA, 2018 Survey and the parameters of the model were estimated using MLE method.

Previous studies as reviewed by [31] reveal that though intervention strategies are effective in HIV Prevention, they do not reveal the effectiveness of specific interventions, which makes it difficult to tell which intervention the country should largely invest in for effectiveness and in order to minimize on cost of HIV Prevention. The current study closes this gap by estimating the effect of one of the most used intervention strategy, Exposure to HIV media on HIV Prevalence under mediation analysis using Logistic regression analysis.

CHAPTER THREE

MODEL FORMULATION

3.1 Introduction

In this chapter two models were formulated; A Logistic regression model with mediation effects and a Logistic regression model without mediation effects. Logistic regression is used to analyze the relationship between multiple independent variables and a categorical dependent variable in order to estimates the probability of occurrence of an event.

3.2 Model Variables

The study used real data derived from the 2018 Kenya Population-based HIV Impact Assessment (KENPHIA) survey. Based on the KENPHIA data, the response variable used was "HIV final result" (HIV positive-1; HIV Negative-2).

The Mediator variable used was assumed to be "ever heard of HIV" (Yes-1; No-2).

The independent variables; Behavioral variables, Social variables, Demographic variables and Biological variables were assessed using various questions in the survey as follows;

Behavioral variables as "used condom at last sexual encounter in the past 12 months" (Used condom at last sexual intercourse in the past 12 months-1, Did not use condom at last sexual intercourse in the past 12 months -2, No sexual intercourse in the past 12 months-3).

Social variables as "Education level in Kenya" (1 - No primary, 2 - In-

complete Primary, 3 - Complete Primary, 4 - Complete Secondary). Demographic variables as "Urban Area Indicator" (Urban =1 ; Rural = 2) and Biological variables as "Gender" (Male =1; Female =2).

3.3 Formulation of a Logistic Regression Model in the absence of Mediation and Parameter Estimation

The model formulated in the absence of mediation was an update of model used previous studies such as that conducted by [40, 9, 43] while considering binary data. Let N represent total number of populations, n_i representing the number of observations in population i for $i = 1, 2 \cdots N$, where $n = \sum_{i=1}^{N} n_i$ is the total sample size.

The study considered the outcome variable Y, as HIV status of an individual and each element y_i as a random variable taking on values 1 for HIV positive and 0 for HIV negative. The distribution of Y is a Bernoulli and the probability of a given individual y_i sampled from the population i being HIV positive is $\pi_i = P(y_i = 1 \mid i)$ whereas the probability of the sampled individual being HIV negative is $1 - \pi_i = P(y_i = 0 \mid i)$.

According to [43], simple linear regression that has a numerical continuous measurement of both explanatory X and response variable Y, assumes that individual responses vary around the mean based on a normal distribution with variance δ^2, that is $\epsilon \sim N(0, \delta^2)$.

This is contrary for binary data as in the current study, where Y is binary in nature with y_i independent observations taking on two possible values; 1 for individual testing HIV positive and 0 otherwise can be modeled in terms of predictor variable x_i through a linear function given as.

$$E(y_i) \mid x_i) \quad = \quad \alpha_0 + \alpha_1 x_{i2}. \cdots + \alpha_R x_{iR} = \pi_i \tag{3.1}$$

Where $i = 1, 2, \cdots N$; $r = 0, 1, 2, \cdots, R$; x_{ir} are the explanatory variables and α_r are the unknown parameters to be estimated.

The variance of a binary response variable unlike in simple liner model is a function of the probability π_i and the $var(y_i) = \pi_i(1 - \pi_i)$, implying that y_i is also a function of π_i. This makes the assumption of a constant variance δ^2 invalid and since binary response can only take on two possible values 0 and 1, assumption of individual responses varying around the mean according to normal distribution is also violated. To solve this problem, there is a need to model the binary response variable by some curved relationship with predictor variable using logistic regression model [43].

The study therefore used logistic regression model to provide a curved relationship with the predictor variable X. The logistic regression model is more advantageous in a way that it is bounded between two values 0 and 1, it has a hidden linear model that can be revealed when the response variable is transformed and finally the sign associated with α_r shows the direction of the curve.

The mean of response variable is modeled in terms of predictor variable and mediator variable through a linear function [36, 9].

$$\begin{aligned} ln\left(\frac{\pi_i}{1-\pi_i}\right) &= \alpha_0 + \alpha_1 x_{i1} +, \cdots, +\alpha_R x_{iR} \\ &= \sum_{r=0}^{R} \alpha_r x_{ir} \end{aligned} \tag{3.2}$$

Where ; $i = 1, 2, \cdots, n$ and $ln\left(\frac{\pi_i}{1-\pi_i}\right)$ is the odds of an event occurring.

Solving for π_i in Equation 3.2

$$\begin{aligned}
\left(\frac{\pi_i}{1-\pi_i}\right) &= \exp^{\sum_{r=0}^{R}\alpha_r x_{ir}} \\
\pi_i &= \exp^{\sum_{r=0}^{R}\alpha_r x_{ir}}(1-\pi_i) \\
\pi_i(1+\exp^{\sum_{r=0}^{R}\alpha_r x_{ir}}) &= \exp^{\sum_{r=0}^{R}\alpha_r x_{ir}} \\
\pi_i &= \frac{\exp^{\sum_{r=0}^{R}\alpha_r x_{ir}}}{1+\exp^{\sum_{r=0}^{R}\alpha_r x_{ir}}} \qquad (3.3)
\end{aligned}$$

Logistic regression aims at estimating $k+1$ unknown parameters α_r in Equation 3.2 using Maximum Likelihood Estimation which entails finding estimates of the model parameters that are most likely to give us data. Given that the observed responses are independent of each other, the likelihood is a product of π_i and $(1-\pi_i)$. The likelihood function for the binary data in the study expresses the value of α in terms of known fixed values of y_i as follows;

$$\begin{aligned}
L(\alpha \mid Y) &= \prod_{i=1}^{n}\pi_i^{y_i}(1-\pi_i)^{1-y_i} \\
&= \pi_i^{\sum_{i=1}^{n}y_i}(1-\pi_i)^{n-\sum_{i=1}^{n}y_i} \qquad (3.4)
\end{aligned}$$

Taking natural log of the likelihood in Equation 3.4

$$\begin{aligned}
\ln L(\alpha \mid Y) &= \sum_{i=1}^{n} y_i \ln \pi_i + \left(n - \sum_{i=1}^{n} y_i\right)\ln(1-\pi_i) \\
&= \sum_{i=1}^{n} y_i \ln \pi_i - \sum_{i=1}^{n} y_i \ln(1-\pi_i) + n\ln(1-\pi) \\
&= \sum_{i=1}^{n} y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right) + n\ln(1-\pi_i) \qquad (3.5)
\end{aligned}$$

Substituting for $\ln\left(\frac{\pi_i}{1-\pi_i}\right)$ from Equation 3.2 and π_i from Equation

3.3 in Equation 3.5 we have;

$$\begin{aligned}\ln L(\alpha \mid Y) &= \sum_{i=1}^{n} y_i \left(\sum_{r=0}^{R} \alpha_r x_{ir} \right) + n \ln \left(1 - \frac{\exp^{\sum_{r=0}^{R} \alpha_r x_{ir}}}{1 + \exp^{\sum_{r=0}^{R} \alpha_r x_{ir}}} \right) \\ &= \sum_{i=1}^{n} y_i \left(\sum_{r=0}^{R} \alpha_r x_{ir} \right) + n \ln \left(\frac{1}{1 + \exp^{\sum_{r=0}^{R} \alpha_r x_{ir}}} \right) \\ &= \sum_{i=1}^{n} y_i \left(\sum_{r=0}^{R} \alpha_r x_{ir} \right) + n \ln (1 + \exp^{\sum_{r=0}^{R} \alpha_r x_{ir}})^{-1} \end{aligned} \tag{3.6}$$

Recall

$-1 \ln(x) = \ln(x)^{-1}$**, thus we obtain;**

$$\ln L(\alpha \mid Y) = \sum_{i=1}^{n} y_i \left(\sum_{r=0}^{R} \alpha_r x_{ir} \right) - n \ln \left(1 + \exp^{\sum_{r=0}^{R} \alpha_r x_{ir}} \right) \tag{3.7}$$

The log likelihood function in Equation 3.7 represents the formulated logistic regression model in the absence of mediation.

To find the maximum likelihood estimates in the model, we differentiate the log likelihood function in Equation 3.7 with respect to the parameters α_r and set the derivative to zero. In differentiating Eq. 3.7

$$\frac{\partial}{\partial \alpha_r} \ln L(\alpha \mid Y) = \sum_{i=1}^{n} y_i x_{ir} - n_i x_{ir} \frac{\exp^{\sum_{r=0}^{R} \alpha_r x_{ir}}}{1 + \exp^{\sum_{r=0}^{R} \alpha_r x_{ir}}} = 0 \tag{3.8}$$

Setting the Equation 3.8 to zero results to $r + 1$ non-linear equations each having $r + 1$ unknown parameters which when solved gives a solution, which specifies a critical point that is either a maximum or a minimum. This results into an estimated vector of α_r elements that is considered as maximum likelihood estimates of the models,[9, 43].

3.4 Formulation of a Logistic Regression Model in presence of mediation and parameter estimation

Equations 1.1, 1.2 and 1.3 were used to fit a simple mediation model in Figure 1.1. The mediation effect in the model can be estimated using

either product of coefficient method which is a product of β and γ or the difference-between coefficients approach which is the difference between α and $\alpha^{'}$ [39]. This study considered one independent variable, Y_i with multiple covariates, X_i.

According to [14], mediation analysis mainly looks at decomposing total effect (TE) of the exposure variable into Indirect effect through a mediator and direct effect whose impacts solely comes from the exposure variable. The mediation effect is indicated by α and γ paths while the direct effect by $\alpha\prime$ path as shown in Figure 1.1.

In this study both Mediation variable, M and dependent variable, Y were Binary variables and the sample size for estimating the parameters for M-regression and Y-regression equations were the same. We therefore used the product of coefficients (ab) method in this study because of its strength in considering one regression model for the outcome and another regression model for the mediator thus circumventing the model compatibility issue in the difference method Cheng [7]. In this study, our outcome of interest,was Y (HIV Prevalence), exposure variable was X (HIV Risk factors), and a mediator variable was M (Exposure to HIV related mass media). We also observed ϵ, a vector of residual factors in the estimated exposure-outcome association assumed to have equal error variance.

Assuming the conditional mean model of outcome Y_i in Equation 1.3.

$$g(E(Y\backslash X, M, e_3)) = \kappa_3 + \alpha\prime X + \gamma M + e_3 \tag{3.9}$$

where $g(.)$ is the logit link function, since the outcome is Binary in nature while $\alpha\prime$ is the exposure effect on the outcome conditional to the effect

of the mediator and error term. γ represents the relationship between the mediator variable and outcome variable conditional to the effect of the exposure variable and the error term.

In addition, the product method required fitting the mediator model as shown in Equation 1.2

$$h(E(M\backslash X, e_2)) = \kappa_2 + \beta X + e_2 \tag{3.10}$$

Where $h(.)$ is a logit link function given that our mediator variable is Binary and β represents the association between the exposure variable and mediator variable conditional on the effects of the covariates and the error term. ϵ_2 and ϵ_3 are independent mean-zero normal errors. The study assumes that both the outcome and mediator models in Equations 3.9 and 3.10 above shares the same set of covariates and are used to estimate the total effects, Mediation effects and Direct effects.

Huberman *et.al* [15], states that the total effect which is the expectation of X on Y can further be decomposed through the mediator into direct and indirect effects as follows;

The total effect of x_i on y_i can be captured in a regression Equation as;

$Y = X\alpha + e_1$

Where $X = (x_i)$, $\alpha = (\alpha_0, \alpha_1)'$ and e represents a mean-zero normal error.

The expectation of Y was given as;

$$
\begin{aligned}
E(Y) &= E(E(Y/M)) \\
&= E(\alpha\prime X + \gamma M) \\
&= E(\alpha\prime X + \gamma(\beta X)) \\
&= \alpha\prime X + \gamma(\beta X) \\
&= X(\alpha\prime + \gamma\beta) \qquad (3.11)
\end{aligned}
$$

The total effect is represented as

$$\alpha = \alpha\prime + \gamma\beta \qquad (3.12)$$

where $\alpha\prime$ and $\gamma\beta$ represent the direct and indirect effects, respectively [3].

Instead of the fixed effects α shown in the formulated logistic regression Model without mediation, we now have the total effect of independent variable X on dependent variable Y, decomposed into mediation effect, $\gamma\beta$ and direct effect, $\alpha\prime$. As in the case of logistic model in absence of mediation, the response variable y_i is binary and has Bernoulli distribution. The mean of response variable is modeled in terms of predictor variable and mediator variable through a linear function

$$E(Y \mid M) = \alpha_0 + \alpha\prime_r x_{ir} + \beta\gamma \qquad (3.13)$$

where Y is the response variable, X the predictor variable and M is the mediator variable. $\alpha\prime_r$ representing direct effect and $\beta\gamma$ representing indirect effect from mediation estimate are the unknown parameters to be estimated.

$i = 1, 2, \cdots, n$, $r = 1, 2, \cdots, R$ and M is a single mediator variable in the model.

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \sum_{r=0}^{R} \alpha\prime x_{ir} + \beta\gamma \tag{3.14}$$

where $r = 0, 1, 2, \cdots, R$ and $i = 1, 2, \cdots, n$

solving Equation 3.14 by taking anti logarithms on both side

$$\begin{aligned}\left(\frac{\pi_i}{1-\pi_i}\right) &= \exp^{\sum_{r=0}^{R} \alpha\prime x_{ir}+\beta\gamma} \\ \pi_i &= \exp^{\sum_{r=0}^{R} \alpha\prime x_{ir}+\beta\gamma}(1-\pi_i) \\ &= \frac{\exp^{\sum_{r=0}^{R} \alpha\prime x_{ir}+\beta\gamma}}{1+\exp^{\sum_{r=0}^{R} \alpha\prime x_{ir}+\beta\gamma}}\end{aligned} \tag{3.15}$$

where $\alpha\prime$ and $\beta\gamma$ is the decomposed total effect of x_i on y_i and is given as the direct effect and indirect effect respectively, [3].

In order to find the estimates of the model parameters that for which the probability of observed data is greatest, given that the observations are independent, the likelihood for the binary data as in this study is given as

$$\begin{aligned}L((\alpha', \beta\gamma) \mid Y) &= \prod_{i}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i} \\ &= \pi_i^{\sum y_i}(1-y_i)^{n-\sum y_i}\end{aligned} \tag{3.16}$$

Taking the logs of the likelihood

$$\begin{aligned}&\ln L((\alpha', \beta\gamma) \mid Y) \\ &\quad= \sum y_i \ln \pi_i + \left(n - \sum y_i\right)\ln(1-y_i) \\ &\quad= \sum y_i \ln \pi_i - \sum y_i \ln(1-\pi_i) + n\ln(1-\pi_i) \\ &\quad= \sum y_i \left(\ln \frac{\pi_i}{1-\pi_i}\right) + n\ln(1-\pi_i)\end{aligned} \tag{3.17}$$

Substituting Equation 3.14 and 3.15 in Equation 3.17

$$\begin{aligned}
&\ln L((\alpha', \beta\gamma) \mid Y) \qquad (3.18)\\
&= \sum y_i \ln\left(\sum_{r=0}^{R} \alpha\prime x_{ir} + \beta\gamma\right) + n\ln\left(1 - \frac{\exp^{\sum_{r=0}^{R} \alpha\prime x_{ir}+\beta\gamma}}{1+\exp^{\sum_{r=0}^{R} \alpha\prime x_{ir}+\beta\gamma}}\right)\\
&= \sum y_i \ln\left(\sum_{r=0}^{R} \alpha\prime x_{ir} + \beta\gamma\right) + n\ln\left(\frac{1}{1+\exp^{\sum_{r=0}^{R} \alpha\prime x_{ir}+\beta\gamma}}\right)\\
&= \sum y_i \ln\left(\sum_{r=0}^{R} \alpha\prime x_{ir} + \beta\gamma\right) + n\ln(1+\exp^{\sum_{r=0}^{R} (\alpha\prime x_{ir}+\beta\gamma)})^{-1}
\end{aligned}$$

Recall that $-1\ln(x) = \ln(x)^{-1}$, thus we obtain

$$\begin{aligned}
&\ln L((\alpha', \beta\gamma) \mid Y)\\
&= \sum y_i \ln\left(\sum_{r=0}^{R} \alpha\prime x_{ir} + \beta\gamma\right) - n\ln\left(1+\exp^{\sum_{r=0}^{R} \alpha\prime x_{ir}+\beta\gamma}\right) \qquad (3.19)
\end{aligned}$$

Equation 3.19 represents the formulated logistic regression model in the presence of mediation.

Similarly, to find the Maximum likelihood estimates $\alpha\prime$ and $\beta\gamma$, we find the partial derivative of the log likelihood function in Equation 3.19 with respect to the parameters $\alpha\prime$ and $\beta\gamma$, set the derivative to zero and solve.

In differentiating Eq. 3.19

With respect to α'_r

$$\frac{\partial}{\partial(\alpha'_r)} \ln L((\alpha', \beta\gamma) \mid Y) = \sum_{i=1}^{n} y_i x_{ir} - n_i x_{ir} \frac{\exp^{\sum_{r=0}^{R} \alpha_r x_{ir}+\beta\gamma}}{1+\exp^{\sum_{r=0}^{R} \alpha_r x_{ir}+\beta\gamma}} = 0 \quad (3.20)$$

With respect to $\beta\gamma$

$$\begin{aligned}
\frac{\partial \ln L((\alpha', \beta\gamma) \mid Y)}{\partial(\beta\gamma)} &= \frac{1}{\beta\gamma}\sum_{i=1}^{n} y_i - n_i x_{ir} \frac{\exp^{\sum_{r=0}^{R} \alpha_r x_{ir}+\beta\gamma}}{1+\exp^{\sum_{r=0}^{R} \alpha_r x_{ir}+\beta\gamma}} \frac{1}{\beta\gamma}\\
&= \frac{1}{\beta\gamma}\left(\sum_{i=1}^{n} y_i - n_i x_{ir} \frac{\exp^{\sum_{r=0}^{R} \alpha_r x_{ir}+\beta\gamma}}{1+\exp^{\sum_{r=0}^{R} \alpha_r x_{ir}+\beta\gamma}}\right) = 0 \quad (3.21)
\end{aligned}$$

Using the same procedure as mentioned under model in the absence of mediation, $r+1$ non-linear equations each having $r+1$ unknown parameters under this model which when solved gives a solution, which specifies a critical point that is either a maximum or a minimum. This results into an estimated vector of α_r' elements and a vector of $\beta\gamma$that is considered as maximum likelihood estimates of the model representing the direct and indirect effects respectively,[9]& [43].

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1 Introduction

The formulated models were each fit to both KENPHIA data separately, that is the model without mediation and Model with mediation for the purposes of data analysis and estimation of individual parameter effect on HIV/AIDs prevalence based on sign and value associated with the parameter was done.

KENPHIA 2018 survey data was carried out with an aim of building on the previously conducted Kenya AIDS Indicator Survey (KAIS) surveys. The new features in the survey included HIV prevalence of each of the 47 counties and the National HIV prevalence that included for Mandera, Wajir and Garissa counties which were previously excluded from data collected in the KAIS as indicated in the Kenya HIV estimates report [32].

4.1.1 Fitting Logistic Regression Model to KENPHIA data set without mediation and parameter estimates

From the KENPHIA data, our response variable was Final HIV Status; 1-HIV Positive, 0-HIV Negative.

The predictor variables in the model include "Gender" as the Biological factor whose responses were 1-Male and 2-Female, "Education level in Kenya" as the social factor with responses; 1-No primary, 2-Incomplete primary, 3-complete primary and 4-complete secondary), "Urban Area Indicator" as the Demographic variable with responses;

1-Urban, 2-Rural and "Used condom at last sexual intercourse in the past 12 months", as the Behavioral factor with responses 1 - Used condom at last sexual intercourse in the past 12 months 2 - Did not use condom at last sexual intercourse in the past 12 months 3 - No sexual intercourse in the past 12 months.

The logistic regression model without mediation was then fitted as follows;

$$\log\left(\frac{\pi}{1-\pi}\right) = 1.863541 + 0.038917B + 0.014316S + 0.013461D - 0.039834b \tag{4.1}$$

Table 4.1: **Parametric estimates of the fitted regression model to KENPHIA data without Mediation**

Coefficients				
(Intercept)	**B**	**S**	**D**	**b**
1.863541	**0.038917**	**0.014316**	**0.013461**	**-0.039834**
Residual Deviance: 1215.6				
AIC: 525.06				

Table 4.1 shows that there is Positive correlation between Behavioral factors, B and HIV prevalence, that is, a unit increase in Behavioral factors increases the log odds of HIV prevalence by 0.038917. There is positive correlation between Social economic factors, S and HIV prevalence, that is, a unit increase in Social factors increases the log odds of HIV prevalence by 0.014316. There is Positive correlation between Demographic factors, D and HIV prevalence, that is, a unit increase in Demographic factors increases the log odds of HIV prevalence by 0.013461. There is Negative correlation between biological factors, b

and HIV prevalence, that is, a unit increase in biological factors reduces the log odds of HIV prevalence by -0.039834. The value of Akaike Information Criteria (AIC) obtained while fitting the model was 525.06.

Table 4.2: **Estimation of the effect of the parameters of model without mediation on HIV Prevalence using KENPHIA data**

Coefficients		
•	Estimated effect	Std. Error
(Intercept)	1.863541	0.012697
B-used condom	0.038917	0.003255
S-Education level	0.014316	0.002190
D-Urban Rural Indicator	0.013461	0.003647
b-Gender	-0.039834	0.003595

Table 4.2 demonstrates that the Behavioral factor, that is in this case "individuals who used condom in the last sexual intercourse in last 12 months" had a positive effect (0.038917) on the HIV/AIDS prevalence. The Social factor, "education level" had a significant positive effect, (0.014316) on HIV Prevalence. The Demographic factor, "Urban Area Indicator" had a significant positive effect (0.013461) on HIV prevalence while Biological factor, in this case "Gender" had a significant negative effect (-0.039834) on HIV prevalence

4.1.2 Fitting Logistic Regression Model to KENPHIA data set with mediation and parameter estimates

To fit the model with mediation, a mediator variable was introduced. The study assumed that all the individuals tested were exposed to HIV/AIDs. The mediator variable therefore was Ever tested for HIV and responses were; Ever Tested-1, Never tested-2. The logistic regression model with mediation was then fitted as follows;

$$\log\left(\frac{\pi}{1-\pi}\right) = 1.793732 + 0.036647B + 0.018416S$$
$$+0.011259D - 0.031591b + 0.047586M \quad (4.2)$$

Table 4.3: **Parametric estimates of the fitted regression model to KENPHIA data with Mediation**

Coefficients					
(Intercept)	**B**	**S**	**D**	**b**	**M**
1.793732	**0.036647**	**0.018416**	**0.011259**	**-0.031591**	**0.047586**
Residual Deviance: 1210.1					
AIC: 434.01					

Table 4.3 shows that there is a Positive correlation between Behavioral factors and HIV prevalence, that is, a unit increase in Behavioral factors increases the log odds of HIV prevalence by 0.036647. There is positive correlation between Social economic factors and HIV prevalence, that is, a unit increase in Social factors increases the log odds of HIV prevalence by 0.018416. There is Positive correlation between Demographic factors and HIV prevalence, that is, a unit increase in Demographic factors increases the log odds of HIV prevalence by 0.011259. There is Negative correlation between biological factors and HIV prevalence, that is, a unit increase in biological factors reduces the log odds of HIV prevalence by -0.031591. The value of AIC obtained while fitting the model was 434.01.

Table 4.4: **Estimation of the effect of parameters of the model with mediation on HIV Prevalence using KENPHIA data**

Coefficients		
•	Estimated effect	Std. Error
(Intercept)	1.793732	0.014586
B-used condom	0.036647	0.003256
S-Education level	0.018416	0.002226
D-Urban Rural Indicator	0.011259	0.003646
b-Gender	-0.031591	0.003687
M-Ever tested HIV/AIDs	0.047586	0.004928

In Table 4.4 the behavior factor, that is in this case "individuals who used condom in the last sexual intercourse in last 12 months" has a positive value (0.036647) for HIV/AIDS prevalence. The Social factor, "education level in Kenya" had a more positive effect on HIV (0.018416). The Demographic factor, "Urban Area Indicator" had a significant positive effect on HIV prevalence 0.011259, while Biological factor, in this case "Gender" has a more negative effect on HIV (- 0.031591). The mediator variable, in this case "individuals who Ever tested for HIV/AIDs", had a significant positive effect (0.047586) on HIV Prevalence. This implies that a unit increase in the mediator variable increases the odds of HIV prevalence by 0.047586. The study assumed that all who tested for HIV/AIDS had been exposed to information on HIV hence were willing to test [1]. In the study to examine the relationship between exposure to media and Voluntary Counseling and Testing (VCT), it was revealed that media is very crucial in scaling up VCT services and individuals exposed to mass media are more likely to be tested for HIV than those not exposed [34].

4.1.3 Comparison of the performance of the models formulated and fitted with KENPHIA data

The AIC under the model with mediation (434.01) as indicated in Table 4.7 was lower compared to that in the model without mediation (525.06) as indicated in Table 4.5, this clearly indicates that the amount of information lost whe fitting the model with mediation was less compared to amount of information lost whe fitting the model without mediation hence the model with mediation is a better model and well fits the KENPHIA data.

A comparison was also done based on the assessment of models adequacy using McFadden R^2 criterion. A value of 0.4755762 is quite high for McFadden's R^2 for the model with mediation, which indicates that the model fits the data very well and has high predictive power as compared to value of 0.3593836 for the McFadden's R^2 for the model without mediation [42, 46].

In addition the comparison of the two models when fitted with KENPHIA data was done using density curve shown in Figure 4.1;

The distribution reveals that the model without mediation has several peaks that tend to be negatively skewed with the main mode at 9 while the model with mediation effect tends towards a normal distribution with the main peak at 8. This indicates that mediation variable tends to lower the prevalence rate of HIV/AIDs among individuals irrespective of their gender, education level, Urban Rural indicator and Condom use unlike in the model without mediation where HIV/AIDs prevalence varies with the group in which individuals are in terms of the associated risk factor.

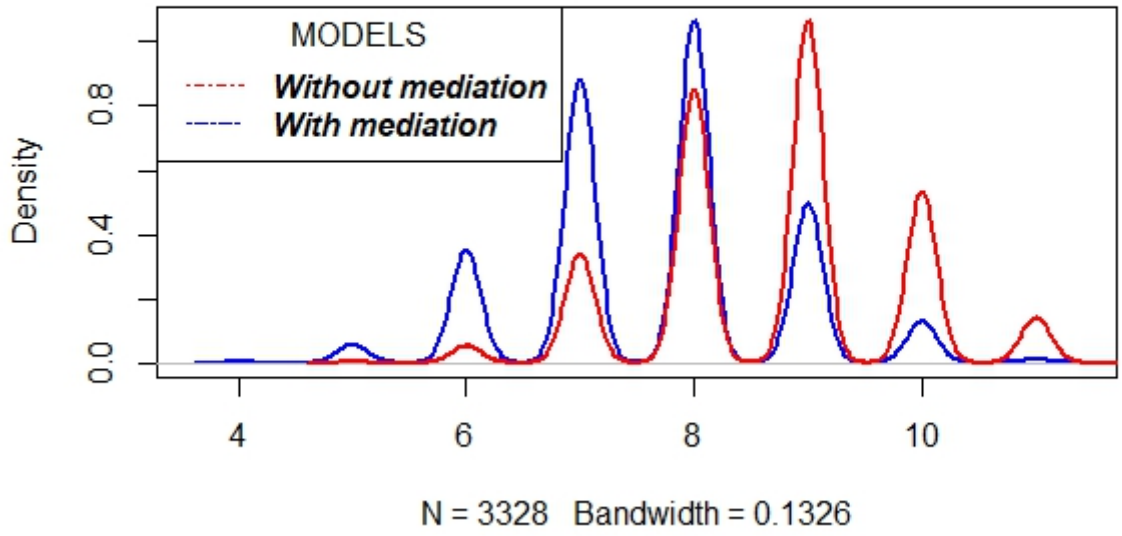

Figure 4.1: **Density Curve of predictor variables for KENPHIA data set in presence and in absence of mediation variable**

Therefore, real data from KENPHIA survey shows that mediation plays a great role in prevention of HIV/AIDS in Kenya.

4.1.4 Evaluation of the Model Adequacy using simulated data

Basing on study by [42, 46], A value of 0.4755762 is quite high for McFadden's R^2 for the model with mediation, which indicates that the model fits the data very well and has high predictive power as compared to value of 0.3593836 for the McFadden's R2 for the model without mediation.

CHAPTER FIVE

CONCLUSIONS AND RECOMMENDATIONS

5.1 Introduction

This chapter discusses the results of the analysis done in chapter Four using KENPHIA data to estimate the effect of mediation on HIV prevalence. It also gives recommendation of some of the areas that can be improved in the study.

5.2 Conclusion

The study looked at modeling the effect of mediation on HIV/AIDS prevalence using the logistic regression model, one in presence of mediation and the other in the absence of mediation. The Binary logistic regression models formulated were fit to data and their parameter estimated using Maximum likelihood estimation. The results of the study revealed a high prevalence of HIV in a model without mediation while a model with mediation showed a low prevalence of HIV. The low AIC in model with mediation revealed that the model was better compared to that without mediation. This indicates that the existence of mass media in Kenya plays a significant role in sensitizing the members of society against HIV, thus reducing the prevalence of the disease. According to [31] its critical to establish the effect of specific intervention strategy used in HIV Prevention for effectiveness and in order to minimize cost of HIV prevention.

5.3 Recommendation for Further Research

There are complexities involved in estimating the effect of specific interventions used to control spread of HIV/AIDS due to prevailing HIV related risk factors such as socio economic, demographic and cultural background of individuals on HIV Prevalence in Kenya. These risk factors will either enhance or hinder the intervention, in this case the mediating factors were used in lowering HIV Prevalence in the country. Therefore, there is need to estimate the effect of other intervention strategies such as male circumcision, PrEP or ART used in HIV/AIDS control assuming they are mediators. This will help the country to channel resources to the specific mediators that are effective and efficient in controlling HIV prevention.

REFERENCES

[1] **Adegboye, O. A., Ezechukwu, H. C., Woodall, H., Brough, M., Robertson-Smith, J., Paba, R., ... & Emeto, T. I. (2022). Media exposure, behavioural risk factors and HIV testing among women of reproductive age in Papua New Guinea: a cross-sectional study. Tropical Medicine and Infectious Disease, 7 (2), 30.**

[2] **Agha, S. (2003). The impact of a mass media campaign on personal risk perception, perceived self-efficacy and on other behavioral predictors.** *AIDS care,* **15(6), 749-762.**

[3] **Alwin, D. F., & Hauser, R. M. (1975). The decomposition of effects in path analysis.** *American sociological review,* **37-47.**

[4] **Aly, R. A. M. (2021). Empowerment as a Mediator between Education and Reproductive Health Care in Egypt: The Impact of Poverty and Residence. Open Journal of Social Sciences, 9(3), 58-76.**

[5] **Amornkul, P. N., Vandenhoudt, H., Nasokho, P., Odhiambo, F., Mwaengo, D., Hightower, A., ... & De Cock, K. M. (2009). HIV prevalence and associated risk factors among individuals aged 13-34 years in Rural Western Kenya. PloS one, 4 (7), e6470.**

[6] **Cavanaugh, J. E., & Neath, A. A. (2019). The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. Wiley Interdisciplinary Reviews: Computational Statistics, 11(3), e1460.**

[7] **Cheng, C., Spiegelman, D., & Li, F. (2021). Estimating the natural indirect effect and the mediation proportion via the product method.** *BMC medical research methodology,* **21 (1), 1-20.**

[8] **Crepaz, N., Mullins, M. M., Adegbite-Johnson, A., Jayleen, K. L., Denard, C., Mizuno, Y., & Project, P. R. S. (2022). Strategies to improve HIV care outcomes for people with HIV who are out of care. AIDS (London, England), 36 (6), 853.**

[9] **Czepiel, S. A. (2002). Maximum likelihood estimation of logistic regression models: theory and implementation. Available at czep. net/stat/mlelr. pdf, 83.**

[10] **Febrianti, R., Widyaningsih, Y., & Soemartojo, S. (2021). The parameter estimation of logistic regression with maximum likelihood method and score function modification. In Journal of Physics: Conference Series (Vol. 1725, No. 1, p. 012014). IOP Publishing.**

[11] **Global, A. I. D. S. (2021). Update. Seizing the moment: tackling entrenched inequalities to end epidemics. Geneva: UNAIDS; 2020.**

[12] **Hamsyiah, N., Nisa, K., & Warsono, W. (2017). Parameter estimation of Bernoulli distribution using maximum likelihood and Bayesian methods. In Prosiding Seminar Nasional METODE KUANTITATIF 2017. Jurusan Matematika FMIPA Unila.**

[13] **Hardnett, F. P., Pals, S. L., Borkowf, C. B., Parsons, J., Gomez, C., & O'Leary, A. (2009). Assessing mediation in HIV intervention studies.** *Public Health Reports,* **124 (2), 288-294.**

[14] **Hayes, A. F. (2009). Beyond Baron and Kenny: Statistical mediation analysis in the new millennium.** *Communication monographs,* **76 (4), 408-420.**

[15] **Huberman, D. B., Reich, B. J., Pacifici, K., & Collazo, J. A. (2020). Estimating the drivers of species distributions with opportunistic data using mediation analysis.** *Ecosphere,* **11 (6), e03165.**

[16] **Irimu, K., & Schwartz, U. (2021). Reporting HIV/AIDS A guide for Kenyan Journalists [Internet]. Friedrich Ebert stiftung Coalition of Media Health Professionals; 2003.**

[17] **Iskarpatyoti, B. S., Lebov, J., Hart, L., Thomas, J., & Mandal, M. (2018). Evaluations of structural interventions for HIV prevention: a review of approaches and methods. AIDS and Behavior, 22, 1253-1264.**

[18] **Johnson, K., & Way, A. (2006). Risk factors for HIV infection in a national adult population: evidence from the 2003 Kenya Demographic and Health Survey.** *JAIDS Journal of Acquired Immune Deficiency Syndromes,* **42 (5), 627-636.**

[19] **Karavasilis, G. J., Kotti, V. K., Tsitsis, D. S., Vassiliadis, V. G., & Rigas, A. G. (2005). Statistical methods and software for risk assessment: applications to a neurophysiological data set. Computational statistics & data analysis, 49 (1), 243-263.**

[20] **LaCroix, J. M., Snyder, L. B., Huedo-Medina, T. B., & Johnson, B. T. (2014). Effectiveness of mass media interventions for HIV**

prevention, 1986–2013: a meta-analysis. JAIDS Journal of Acquired Immune Deficiency Syndromes, 66, S329-S340.

[21] Lindsay E. (2001). Fact sheets of HIV/AIDS for nurses and midwives. *Supplement to Africa Journal of Nursing and Midwifery,* **3 (1), S1-S10.**

[22] Li, L., Rotheram-Borus, M. J., Lu, Y., Wu, Z., Lin, C., & Guan, J. (2009). Mass media and HIV/AIDS in China. Journal of health communication, 14 (5), 424-438.

[23] L. Myhre, June A. Flora, S. (2000). HIV/AIDS communication campaigns: progress and prospects. *Journal of Health Communication,* **5 (sup1), 29-45.**

[24] Moineddin, R., Matheson, F. I., & Glazier, R. H. (2007). A simulation study of sample size for multilevel logistic regression models. *BMC medical research methodology,* **7 (1), 1-10.**

[25] MacKinnon, D. P., & Pirlott, A. G. (2015). Statistical approaches for enhancing causal interpretation of the M to Y relation in mediation analysis. *Personality and Social Psychology Review,* **19 (1), 30-43.**

[26] MacKinnon, D. P., Cheong, J., & Pirlott, A. G. (2012). Statistical mediation analysis.

[27] MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological methods,***7 (1), 83.**

[28] Magadi, M., & Desta, M. (2011). A multilevel analysis of the determinants and cross-national variations of HIV seropositivity in

sub-Saharan Africa: evidence from the DHS. *Health & place,* **17 (5), 1067-1083.**

[29] **Mahy, M. I., Sabin, K. M., Feizzadeh, A., & Wanyeki, I. (2021). Progress towards 2020 global HIV impact and treatment targets. Journal of the International AIDS Society, 24, e25779.**

[30] **Mugoya, G. C. T., Aduloju-Ajijola, N., & Dalmida, S. G. (2016). Relationship between Knowledge of Someone Infected with HIV/AIDS and HIV Stigma: A moderated mediation model of HIV knowledge, gender and hiv test uptake. HIV/AIDS Res Treat Open J.**

[31] **Mwaura, E. (2009).** *HIV/AIDS prevention strategies in Kenya. A critical review* **(Doctoral dissertation, University of Pittsburgh).**

[32] **NACC, N. (2018). Kenya HIV estimates report.** *Nairobi, Kenya: NACC.*

[33] **NASCOP, K. (2020). Preliminary KENPHIA 2018 Report. National AIDS and STI Control Program (NASCOP). Accessed April, 14.**

[34] **Onsomu, E. O., Moore, D., Abuya, B. A., Valentine, P., & Duren-Winfield, V. (2013). Importance of the media in scaling-up HIV testing in Kenya. SAGE Open, 3(3), 2158244013497721.**

[35] **Park, H. A. (2013). An introduction to logistic regression: from basic concepts to interpretation with particular attention to nursing domain.***Journal of Korean Academy of Nursing,* **43(2), 154-164.**

[36] **Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. The journal of educational research, 96(1), 3-14.**

[37] **Plan, M. C. O. (2017). Strategic direction summary.** *US President's Emergency Plan for AIDS Relief (PEPFAR)*

[38] **Prevalence, H. I. V., & Tanzania, M. (2013). DHS WORKING PAPERS**

[39] **Rijnhart, J. J., Twisk, J. W., Eekhout, I., & Heymans, M. W. (2019). Comparison of logistic-regression based methods for simple mediation analysis with a dichotomous outcome variable. BMC medical research methodology, 19, 1-10.**

[40] **Shalizi, C. (2011). Logistic Regression and Newton's Method. CMU Statistics. Pittsburgh, Pennsylvania, 15.**

[41] **Speizer, I. S., Gómez, A. M., Stewart, J., & Voss, P. (2011). Community-level HIV risk behaviors and HIV prevalence among women and men in Zimbabwe.** *AIDS education and prevention,* **23 (5), 437-447.**

[42] **Smith, T. J., & McKenna, C. M. (2013). A comparison of logistic regression pseudo R2 indices. Multiple Linear Regression Viewpoints, 39 (2), 17-26.**

[43] **Stephenson, B., Cook, D., Dixon, P., Duckworth, W., Kaiser, M., Koehler, K., & Meeker, W. (2008). Binary response and logistic regression analysis. available at:¡ a href=" http://www. stat. wisc. edu/mchung/teaching/MIA/reading/GLM.**

logistic. Rpackage. pdf"¿ http://www. stat. wisc. edu/mchung/teaching/MIA/reading/GLM. logistic. Rpackage. pdf¡/a¿(last access: 30 August 2014).

[44] Takyi, B. K. (2003). Religion and women's health in Ghana: Insights into HIV/AIDS preventive and protective behavior. *Social science & medicine,* **56(6), 1221-1234.**

[45] UNAIDS, J. (2017). UNAIDS data 2017. Jt. United Nations Program. HIV/AIDS, 1-248.

[46] Windmeijer, F. A. (1995). Goodness-of-fit measures in binary choice models. Econometric reviews, 14(1), 101-116.

APPENDICES

APPENDIX : Fitting Logistic Regression Model to KENPHIA Data Set Without and With Mediation and Parameter Estimates

$attach(RDataCVS)$

$read.csv(C:Users\ USER\ Documents\ Ruthfolder\ ThesisData\ RDataCVS.csv)$

$model1 = glm(P{-}HIVP \sim b{-}Gender{+}S{-}Educl{+}D{-}URI{+}B{-}Condm, data = RDataCVS)$

$summary(model1)$

$pscl :: pR2(model1)["McFadden"]$

$plot(density(b{-}Gender{+}S{-}Educl{+}D{-}URI{+}B{-}Condm), from = -0.5, to = 0.5, col =' blue', lwd = 2, main =")$

$model2 = glm(P_HIVP \sim b - Gender + S - Educl + D - URI + B - Condm + m - Tested, data = RDataCVS)$

$summary(model2)$

$pscl :: pR2(model2)["McFadden"]$

$plot(density(b{-}Gender{+}S{-}Educl{+}D{-}URI{+}B{-}Condm{+}m{-}Tested, data = RDataCVS), from = -0.5, to = 0.5, col =' red', lwd = 2, main =")$

$plot(density(b - Gender + S - Educl + D - URI + B - Condm), from = -0.5, to = 0.5, col =' blue', lwd = 2, main =")$

$lines(density(b{-}Gender{+}S{-}Educl{+}D{-}URI{+}B{-}Condm{+}m{-}Tested, data = RDataCVS), from = -0.5, to = 0.5, col =' red', lwd = 2, main =")$

plot(density($P{-}HIVP$),from =-0.5, to=0.5,col='red',lwd=2,main=")