International Journal of Statistical Distributions and Applications
2023; 9(3): 68-80
http://www.sciencepublishinggroup.com/j/ijsda
doi: 10.11648/j.ijsd.20230903.11
ISSN: 2472-3487 (Print); ISSN: 2472-3509 (Online)

Modelling Count Data for HIV-Positive Patients on Antiretroviral Treatment in Kenya

Anna Nanjala Muricho[*], Thomas Mageto, Samuel Mwalili

Department of Statistics and Actuarial Science, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya

Email address:

annamuricho@gmail.com (Anna Nanjala Muricho), tttmageto@jkuat.ac.ke (Thomas Mageto),
samuel.mwalili@gmail.com (Samuel Mwalili)

[*]Corresponding author

To cite this article:

Anna Nanjala Muricho, Thomas Mageto, Samuel Mwalili. Modelling Count Data for HIV-Positive Patients on Antiretroviral Treatment in Kenya. *International Journal of Statistical Distributions and Applications*. Vol. 9, No. 3, 2023, pp. 68-80. doi: 10.11648/j.ijsd.20230903.11

Received: October 3, 2023; **Accepted**: October 20, 2023; **Published**: October 31, 2023

Abstract: The Acquired Immunodeficiency Syndrome (AIDS), caused by the Human Immunodeficiency Virus (HIV), is a lentivirus that weakens a person's resistance to infection. The National AIDS Control Programme (NASCOP) guidelines advise patients to begin antiretroviral therapy (ART) when an individual`s CD4+ cell count is below 350 cells/ml or when they begin to exhibit symptoms of HIV infection, as defined by WHO stages I through IV. To achieve HIV viral suppression, antiretroviral drug adherence is essential. Measurements on a variable are gathered for each individual at several points in longitudinal research. Although variables with repeated measurements within an individual are correlated, the between individuals are typically presumed to pose independence, and this is a major characteristic of such longitudinal data. A Retrospective Longitudinal study of HIV-Positive patients enrolled on ART from 2018 to 2021 those above 9 years when they sign up for ART. In total, 1489 individuals were involved during research. Data was examined by descriptive statistics. A generalized linear mixed effect model was fitted which took into account the within and between variations due to its flexibility. The number of patients enrolled on ART increases by Age and Gender over the four years. In 2018, 2019, 2020, and 2021 ART coverage was 22.4%, 24.2%, 26.1%, and 27.3% respectively. The variables age, gender and year were found to be the significant predictors. The GLMM with negative binomial distribution was used to analyze the data due to overdispersion in the data and the fact that there wpas a random factor. The AIC was used as the model selection approach. A model considered as the baseline was built with all possible interactions and major effects, and the best fitting model was defined as the one with the lowest AIC.

Keywords: Antiretroviral Treatment (ART), Generalized Linear Models (GLM), Generalized Linear Mixed Models (GLMM)

1. Introduction

The Joint United Nations Programme on HIV/AIDS (UNAIDS) estimates that 34 million persons worldwide are HIV positive. In sub-Saharan Africa, a total of 23.5 million, an estimate of 14 million patients are qualified for HIV therapy, and 8 million people worldwide are using ART. [1]. A patient's unique traits should be taken into consideration while starting antiretroviral therapy (ART). Before beginning the therapy, sociological variables including the presence of medics with constant provision of medications should be taken into account. The HIV replication is inhibited by HAART, resulting in demonstrable decreases in morbidity and mortality. At the moment, the infected individuals in Kenya should visit the medical centers frequently and without missing out on ART program. The approach is one-size-fits-all. This strategy has not been particularly successful in preventing HIV/AIDS to date: First of all, most patients don't reliably commit to attending their visits because of their lifestyle, and with time, they stop responding to the recommended treatments. These lifestyles include erratic schedules at employment and drug use. The lack of resources in hospitals to carry out these programs is the second major barrier. The issue of stigmatization is the last one. Because they worry about discrimination, some infected individuals are unwilling to publicize their HIV status. In the end, these difficulties result in ART non-response, which increases the risk to the patients. There are currently HIV/AIDS prediction systems in place to manage HIV/AIDS in accordance with the aforementioned issues. These forecasting techniques have not proven to be successful since they only specify the types and dosages of ARVs to be used, supporting a one-size-fits-all strategy. As a result, medical institutions are overworked and the patients who are unstable might not get the care they need. It's possible that the desires and expectations of stable HIV patients won't be met. When repeated counts are measured on the same individual overtime, the assumption of independence is no longer reasonable, instead they are correlated. Mixed effects models include the between-source variation and the within source variation, by addition of random effects in the models. Mixed effects models have the significant the advantage of allowing for specific-individual inference as well as the more frequent population-average inference. The mixed models incorporated additional effects that are random to account for data correlation, are an extension of the general linear models [2]. Construction of a statistical model with fewer limitations is possible through the process of moving to complicated models. The negative binomial mixed effect model is suggested to take into account the time-dependent over dispersion [3]. Simply maximizing the likelihood function will yield the maximum likelihood estimator (MLE) for GLMs. Statisticians have suggested a number of methods for estimating the parameters of GLMM, either by defining likelihood substitutes or approximating the likelihood function. When the models are properly stated, the mixed effects models are effective and frugal. Additionally, V. Culshaw demonstrated that several of the prior prediction models employed regression approaches, which are incompatible with non-normal, missing, or correlated data, leading to erroneous statistical judgments [4]. To handle non-normal data, they did not apply the most adaptable and potent longitudinal models.

Y. Liang and L. Zeger demonstrated that the two most commonly used models are generalized estimating equation (GEE) [5] and the models with mixed effects [6]. Therefore, the most adaptable and effective longitudinal model for handling non-normal, incomplete, or correlated data was used in this study to create a model that predicts the association between ART uptake and HIV repeated measurements. This model is called the GLMM.

2. Literature Review

2.1. Antiretroviral Treatment Status Globally

From a peak of 2.1 million deaths worldwide each year in 2004 to an expected 0.7 million in 2019, the AIDS-related mortality has consistently reduced. The decrease is due to the greater accessibility of ART and supporting initiatives for those infected HIV, especially in regions with little resources. According to the 2011 KDHS report, 59% of the 1.6 million are women living with HIV, an estimate of 1 million children have been left bereft, and there are further infections are being detected in people between the ages of 15 and 35. In sub-Saharan Africa, where in 2009, there were 20% fewer deaths from HIV/AIDS-related causes than there were in 2004, the impacts are particularly noticeable drastically altered because in areas with limited resources, ART accessibility and availability began to drastically increase, and the regimen has also been streamlined and the majority of clients citing negligible side effects and adverse events. The WHO/UNAIDS 2010 report's main results were as follows: In the low- and middle-income countries, 36 percent of the 14.5 million people who needed ART at the end of 2009 were receiving it, with the number of AIDS-related fatalities had decreasing. According to the most recent HIV progress indicators report from Kenya, 78.9% of all people who knew their HIV status had started taking ARVs. Males had a greater Anti-Retroviral Therapy (ART) coverage rate of 85.3% compared to females' 69.7%. According to the research, by 2020, 70 children had access to ARVs for every 100 children living with HIV in Kenya, up from the 18 children who did so in 2008, according to the scale-up of ART. As a result, a chronic disease that was once fatal has become more manageable thanks to its use. HIV/AIDS patients now have significantly different quality and length of lives. The regimen has also been streamlined, with the majority of clients reporting minor unpleasant responses and side effects [7].

2.2. Antiretroviral Treatment in Kenya

The first HIV case was noted in 1984 in Kenya. In combating the HIV/AIDS epidemic actively by encouraging safer sexual habits, Kenya adopted this slowly compared to other African countries like Uganda and Botswana. S. McClelland suggested a Kenya National HIV/AIDS Strategic Plan [8] was unveiled in by the National Syndemic Disease Control Council (NSDCC), an urgency that directly reports to the president's office. For five key areas: management and coordination of HIV treatment, mitigation of social and economic impact, treatment, continuum of care, and support, and prevention and advocacy. This strategy established a multi-sectoral approach. In 2001, the Kenyan government began supplying ART to the public health sectors, opening five centers. According to NACC estimates, there were 1.8 million orphans and 1.45 million persons infected with HIV in 2003. The HIV patients dropped to 1.35 million in 2005. It was estimated that 1.5 million Kenyans in 2011 were infected with the disease. Although young adults are the group most likely to contract HIV/AIDS, a sizable percentage of seniors are living with the disease. 60% of new HIV infections occurred in people between the ages of 15 and 35 in 2009, with one out of every 11 PLWHIV being between the ages of 50 and 64. (KNBS, 2010). According to estimates, there would be 1.4 million infected with HIV in 2020, comprising of an adult prevalence of 4.2% and 86% of 33100 were PLWHIV were on antiretroviral therapy.

Antiretroviral therapy is one of the methods used by Kenya and other poor nations in treatment HIV infections. This is a commendable endeavor since HIV/AIDS-related cases and fatalities have significantly decreased. HIV infections have become a chronic, treatable condition because of the use of ART. All governmental health facilities, as well as private hospitals and faith-based organizations (FBO), offer the treatment free of charge. The following ART objectives were listed in the WHO, Kenya National ART (2010) guideline: the quality of life [9] of the HIV positive people is improved, reducing viral load and thus hinder disease progression, reducing morbidity and mortality, restoring immunological function, and reducing HIV cases related to transmission from the mother to the child transmission. However, it is advised that there should be a major rise in HIV illness knowledge so that these objectives are achieved and for the clients to completely benefit from ART, it is important to increase its accessibility to all social classes (MOH, 2010). Even though ART is now more widely accessible, the fight against HIV in this country is far from ended. As improved therapies assist maintain excellent health and lengthen lifespans, the number of PLWHIV keeps rising. However, a tiny number of patients, particularly those from distant locations, are still unable to take use of the HIV prevention and treatment facilities NASCOP [10] reports that access to treatment has improved in metropolitan areas, with 72% of people who are eligible for treatment having been able to take advantage of the services, especially those in rural locations. According to NASCOP, more adults who are qualified for treatment are able to get therapy in metropolitan regions, with 72% of them able to do so. Between 30 and 40 percent of babies are born to mothers who are HIV positive. In the future, HIV/AIDS might cause more pediatric deaths than measles and malaria combined (KDHS, 2010).

2.3. Empirical Studies

Wu and Zhang investigated censored factors effects of models that are mixed with relevance to research studies on HIV [11]. The conclusion of model were the following restrictions: (i) a lack of availability of such models in many applications and (ii) computational challenges. This is due to the fact that these models are frequently non-linear, making computing a significant obstacle to likelihood inference.

Yu and Wu modelled the linkage of viral loads and CD4 counts for HIV/AIDS data that is complex [12]. The study proved that viral load and CD4 are crucial variables in HIV/AIDS research. The analysis found that many previous studies did not target errors of measurement and the outliers that are common characteristics of HIV related data. This study was the number one to their knowledge to tackle each of these data issues separately. The findings in the research showed that viral load and CD4 cell count are adversely linked with time, regardless of how CD4 is conceptualized-as continuous, binary, count, response, or covariate. However, errors of measurement and outliers are not taken into account during data analysis, the degree of the association's strength may be grossly underestimated. T. Wendler and S. Grottrup suggested resilience of two step techniques in combining Linear Mixed Effect/Generalized Linear Mixed Model [13] and joint Negative Linear Mixed Effect with LME that models performed well, according to simulation findings. By incorporating a model for missing data (for example binary mixed effects model for the missing data indicators), which also generates a joint model, the suggested methods can be used with missing data.

X. Lu demonstrated statistical studies and modeling have made significant contributions to our knowledge on the HIV-1 infection development [14]. Additionally, they provide recommendations to care for infected individuals and the assessment of Antiretroviral Treatment. The CD4 counts and viral loads were modeled using a variety of statistical methods, in particular, nonlinear mixed-effects models. These methods frequently make the assumption that people infected with HIV come from a homogeneous population and have same trajectories. This presumption hides significant differences in patient subgroups' treatment responses and illness progression due to biologically distinct disease trajectories. This could result in skewed inference. They used Markov chain Monte Carlo (MCMC) to create a mixed dynamic model and related Bayesian inferences in their study. Latent class models which are also well-known as finite mixture models which employ to capture population diversity that is not predetermined. To do this, the population is divided into latent classes with a finite number, and the population is modelled using a mixture distribution. This crucial component might aid medical professionals in better comprehending the course of a patient's illness and adjusting their ART plan in advance. Using longitudinally recorded CD4 counts and their potential predictors, S. E Holte et al. [15] examined the relationship between HIV infection progressions and longitudinal analysis approaches. For the statistical analysis of ART data, two modeling approaches (GEE and GLMM) were evaluated. It was discovered that GLMM provided a better match with less disturbance for this data than GEE. The study also discovered that after patients started an ART protocol, their CD4 counts typically increased over time in a quadratic pattern (i.e., due to the medication, the immune system strengthens while the disease's progression declines).

Generalized linear mixed models are preferred over generalized estimating equations for correlated data, even though the decision between the two must be made based on the subject matter at hand.

2.4. Generalized Linear Mixed Model

In generalized linear mixed models (GLMM), correlations within each cluster are taken into account by the random effects incorporated into the linear predictor. Analyzing correlated non-normal data and describe over-dispersion, Mixed Models further removes the independence restriction of the observations. A fixed variable is one that is thought to be accurately measured. Additionally, it is expected that a fixed variable's values will remain constant from one study to the next and that they will match up between investigations.

Categorical outcomes, normally distributed into outcomes that are continuous, together with outcomes that are non-normally distributed like counts can all be analyzed using the Mixed-effects Models (GLMMs). GLMMs make allows the possibility to account for within-subject associations. For binary and count data the simple random-effects models serve as the basis for GLMMs. The estimates in GLMMs for the random effects, is employed with addition of the marginal probability to be a foundation for inferences for the parameters which are effects that are fixed. In GLMMs, both measure variables and unobserved random effects are dependable of the mean response in the model; the latter's inclusion, when the distribution of random effects is [16] averaged, slightly induces correlation among the repeated answers. The independence assumption in linear models makes the general linear regression model not suitable for longitudinal data. This is because the response y is observed repeatedly from the same individuals, and so it is much more likely to assume that errors within an individual are correlated to some degree.

3. Methodology

Counts: Count Data comes from counting events of interest in an experimental unit. Counts are non-negative integers, often right skewed, with a Poisson or Negative Binomial distribution. Count data is unbounded, i.e., there are no predetermined limits imposed on the range of values.

3.1. Generalized Linear Models (GLMs)

Generalized Linear Models (GLM) were first developed by McCullagh and Nelder in 1989. If there are n participants in the sample, then let Y_i (X_i) represent a response that is continuous. The conventional linear model is provided by:

$$Y_i = X^T\beta_i + \varepsilon_j,\ 1 \leq \mathrm{i} \leq \mathrm{n}, \tag{1}$$

When the models are extended to non-continuous response such as binary they are expressed as: -

$$Y_i|X_i \sim \mathrm{N}\ (\mu_i, \delta^2),\ \mu_i = E(Y_i|X_i) = X_i^T\beta, 1\ \leq \mathrm{i} \leq \mathrm{n} \tag{2}$$

that is $Y_i|X_i$ denoting the distribution that is conditional of Y_i where X_i and $E(Y_i|X_i)$ represents the mean that is conditional. The class of GLM is obtained by substituting alternative distributions suited for the type of response for the normal. This is given by:

$$Y_i|X_i \sim f(\mu_i), g(\mu_i){=}X_i^T\beta, 1\ \leq \mathrm{i} \leq \mathrm{n} \tag{3}$$

Since $g(\mu)$ is a linkage to the explanatory variables and the mean the pdf of Y, is the component that is random in the GLM and $X_i^T\beta$ is the component that is systematic.

3.2. The GLMM

The GLMM for the measurement y of individual i given below,

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \epsilon_{ij} \tag{4}$$

Considering a research study that is longitudinal with times points T and let Y_{it} with X_{it} represent the outcome and predictors.

The GLMM is specified by:

$$Y_{it}|X_{it}, Z_{it}, b_i \sim f(\mu\hat{}_{it}) \tag{5}$$

$$g(\mu_{it}) = X_{it}^T\beta\hat{} + Z_{it}^T b_i, b_i, 1 \leq i \leq n, 1 \leq t \leq T \tag{6}$$

Z_{it} is a vector of predictors and $g(\mu)$ is the link function. The random effects are represented by the vector of latent variables b_i which represents individual variations from the mean of the population, b_i which is the fixed effects. Correlated responses are supported by the GLMM by directly simulating their joint distribution. The correlated responses are modelled using latent variables (b_i). Thus, even if Y_{it} for each time point t is still modelled, Y_{it} it is given a conditional specification that includes the random effect b_i.

The GLMM takes the form:

$$g\big(E\big[Y_{ij}|b_i\big]\big) = X_{ij}\beta + Z_{ij}b_i \tag{7}$$

GLMMs is an extension of GLMs to incorporate random effects. B. M Bolker et al. [17] suggested that the Generalized linear mixed models are an extension of generalized linear models that allow for random effects, demonstrating how ecologists and evolutionary biologists should assess non-normal data with random effects.

$$f(y_i|b_i) = exp\left\{\frac{y_i\theta_i - d(\theta_i)}{a_i(\emptyset)} + c(y_i, \emptyset)\right\} \tag{8}$$

Given count data where the covariate of interest is time the GLMM is of the form:

$$log\big(E\big[Y_{ij}|b_{0i}, b_{1i}\big]\big) = b_{0i} + \beta_0 + b_{1i}t_{ij} + \beta_1 t_{ij} \tag{9}$$

The variance and the mean that are condition in the distribution are: -

$$Var\big(Y_{ij}|b_i\big) = E\big(Y_{ij}|b_i\big) \tag{10}$$

In GLMM both the fixed and random effects are estimated $Y_{ij}|\beta \sim EF(\theta)$ *and* $b_i \sim N(0, G)$.

The first approach is to estimate full likelihood using the equation and conditional likelihood. The second method involves estimating the parameters from the complete likelihood using Bayesian inference via Bayesian Hierachial models. The final step is to estimate the parameter from the complete likelihood using the EM algorithm technique.

The Marginal quasi-likelihood is of the form:

$$L(\beta, \alpha) = exp\{l(\beta, \alpha)\} = \int exp\{\textstyle\sum_{i=1}^n l_i\,(\beta, \alpha|b)\,\}dF(b, \alpha) \tag{11}$$

Empirical Bayes for random effect:

$$l_i(\beta, \alpha|b)\alpha \int_{y_i}^{\mu_i^b} \frac{a_i(y_i - \mu)}{\phi Var(\mu)} \tag{12}$$

$$l(\beta, \alpha) = logL(\beta, \alpha) = log \int exp\{\textstyle\sum_{i=1}^n l_i\,(\beta, \alpha|b)\,\}dF(b, \alpha) \tag{13}$$

3.3. Model Formulation

The distribution of the observations, this is distribution of the observation's conditional on the random model effects. The linear predictor if there are random model effects, include them in the linear predictor. The link function species which link function is being used.

3.3.1. The Distribution of the Observations

Observed count of a specific individual in a given year as a Poisson distribution $Y_{ij} \sim Poisson(\lambda_i)$

The expected count proportion is $\lambda_{ij} = \frac{Y_{ij}}{n_{ij}}$

3.3.2. The Linear Predictor

In the GLMM approach, link function $\log(\lambda)$ is the natural parameter. The linear predictor

$$\eta_{ij} = \eta + \tau_i + b_j \quad (14)$$

For the simple Poisson GLMM

$$Y_{ijk}|r_k, w_{ik} \sim Poisson(\lambda_{ijk})$$

3.3.3. Conditional Model

A random unit level known as within subject effect is added to the linear predictor by conditional model.

$$\omega_{ik} \sim N(0, \sigma)$$

where serial correlation structure is specified by the covariance matrix Σ. It follows that a fully conditional generalized linear mixed model of repeated measures is presented as:

$$\eta_{ijk} = E(Y_{ijk}|r_k, \omega_{ik}) = \beta_0 + \alpha_i + \beta_j + \alpha\beta_{ij} + r_k + \omega_{ik} + S_{ijk} \quad (15)$$

3.4. Model Evaluation

Goodness of fit: adequacy with which a model accounts for variability in the data. With non-Gaussian data inadequacy occurs when: - Assuming the wrong distribution for the observed data, terms have been left out of the linear predictor and over dispersion occurs when the model fails to adequately account for all the sources of variability in the data. In a conditional GLMM using the Laplace approximation method, a good fit is indicated by the random scatter of points of residuals and assessment of over dispersion is detected by deviance.

Model:

The Generalized Linear Mixed-effect Model will be presented as;

$$Y_{ij=}\beta_{oj} + \beta_{1j}X_1 + \beta_{2j}X_2 + \beta_{3j}X_3 + \cdots + \beta_{pi}X_p + b_i, 0 + bi, 1X_1 + bi, 2X_2 + bi, 3X_3 + bi, PX_p + Error_{ij} \quad (16)$$

3.4.1. Estimation and Statistical Inference Procedure for GLMM

The parameters are estimated using the Laplace approximation method and maximum likelihood (ML). Assuming that the estimations for the fixed effects are accurate, the random effects standard deviation are calculated by the ML. The data's likelihood of being a function of unknown parameters is:

$$L(\beta, \alpha, Y) = \pi_{it}^{m} \int \pi_{it}^{ni} f(Y_{ij}|\beta, b_i) f(b_i|\alpha) d\, b_i \quad (17)$$

Above is the random effects' joint distribution with the data integral over the unobserved random effects.

3.4.2. GLMM Model Checking Method

The effects that are random are presumed to be distributed normally and the error term are uncorrelated. To visually assess the effects normalcy and to spot any unusual effect categories, residual plots were employed. G. Fitzmaurice et al. [18] suggested a better impression can be gained by looking at the fitted values plotted by any relevant covariates against the standardized residuals.

3.4.3. The Likelihood of Negative Binomial

In the analysis, Y, is the dependent variable and it is a count variable which represents the number of patients on ART in the 47 counties from 2018 to 2021. The assumption made is that the dependent variable follows a negative binomial distribution. Therefore, Y being a count variable, the distribution that could be used is a Poisson. Preliminary analysis of the data was done and it was demonstrated that the requirement of the Poisson process was not met. The outcome is assumed to be *i.i.d.* Additionally, the assumption of the model belonging to a family that is Gaussian is number two. Every distribution represented as Gaussian distribution is validated by this principle.

The distribution of Negative Binomial is: -

$$f(y) = \frac{\Gamma(y+n)}{\Gamma(n)\Gamma(y+1)} p^n (1-p)^y \quad (18)$$

4. Results and Discussion

The number of HIV patients on ART increased over the years: 2018 (1,018,899), 2019 (1,104,140), 2019 (1,188,663), 2020 (1,245,819). Over the years, the total number of female patients on ART is 3,078,202 while male patients 1,479,319. Additionally, the distribution of ART by age was 10-14 (149,146), 15-19 (146,561), 20-24 (216,831), 25+(4,044,983). Nairobi County recorded the highest number of patients on ART (618,609) while Mandera county the lowest (2,529) cumulatively over the four years. Additionally, the observations in counties ranged between 10,000 and 180,000**.** The number of HIV-patients on ART in Kenya during the study period continued to increase from year to year, with an average of 1.1 million patients (25%). The greatest number of HIV patients receiving ART during the study period over the years was Nairobi County with an average of 1.5 thousand female (25%) patients aged 25+ years.

4.1. Analysis of Exploratory Data

The diagonal alignment of the variables shown in Figure 5 the scatter plot indicates independence structure which is the structure of the correlation. This structure assumes that the correlations between subsequent measurements are zero, also within people the measurements are assumed to be independent of one another.

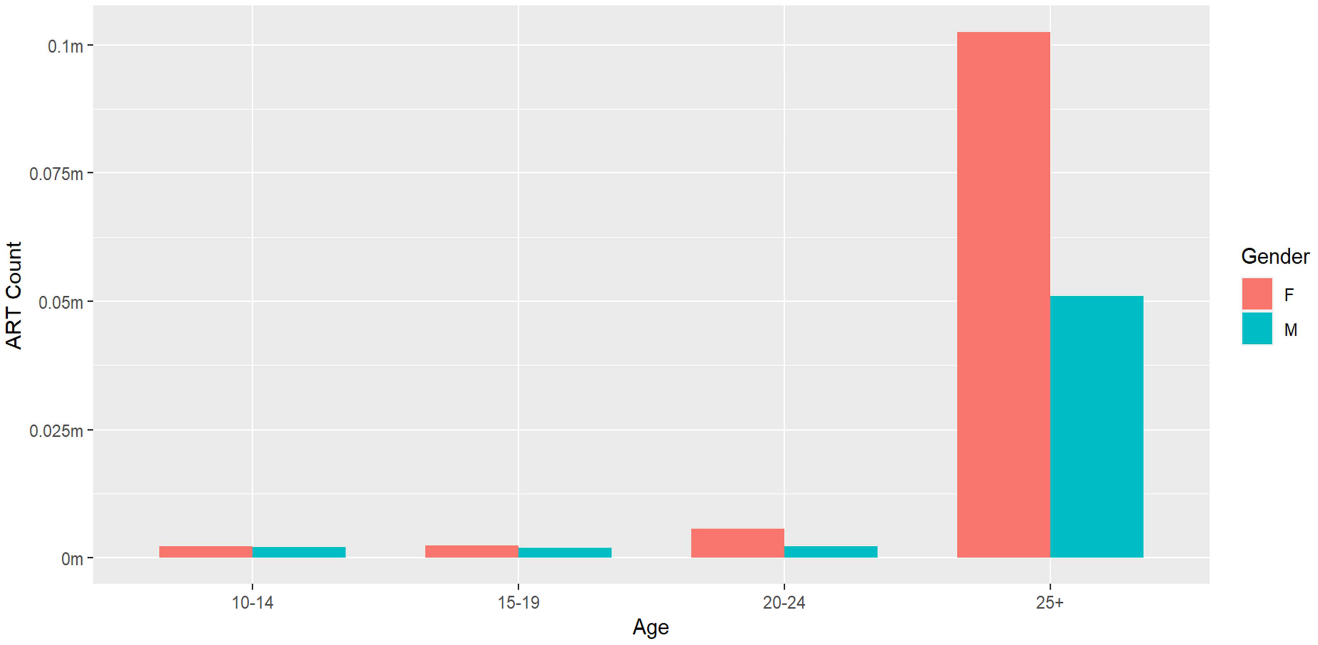

Figure 1. *Age and Gender characteristics bar graph.*

The graph depicts the age and gender distribution of ART patients aged 25 years and above.

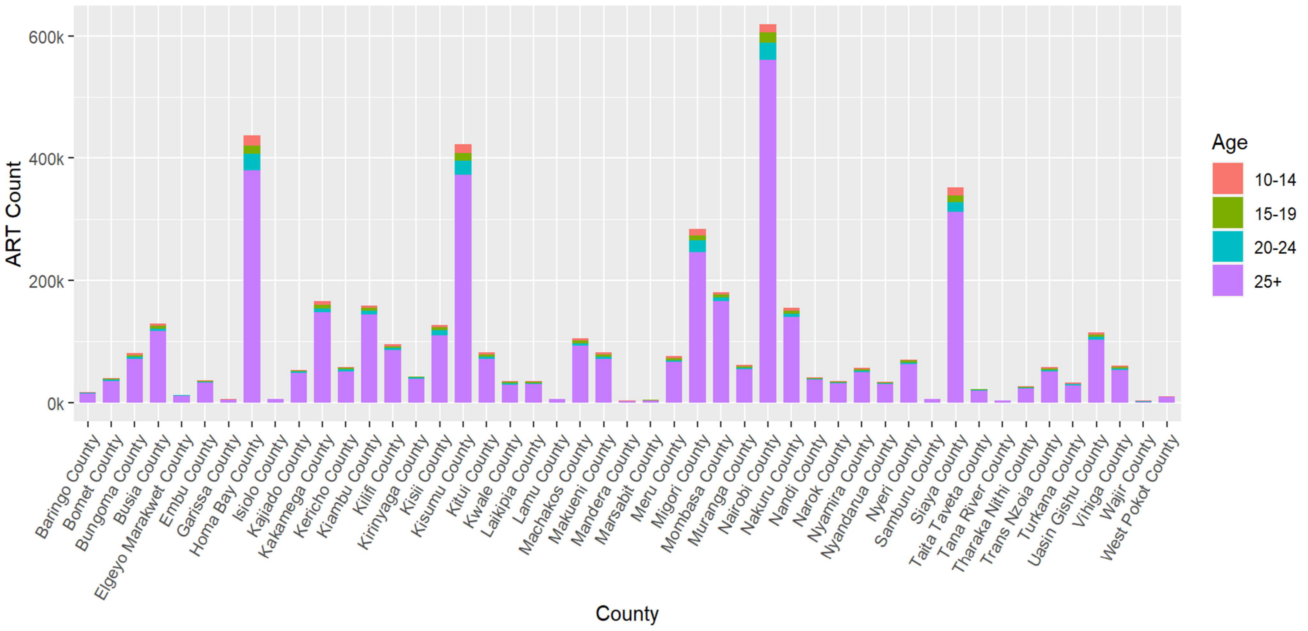

Figure 2. *Distribution of ART by county.*

The graph depicts counties with the highest ART frequencies.

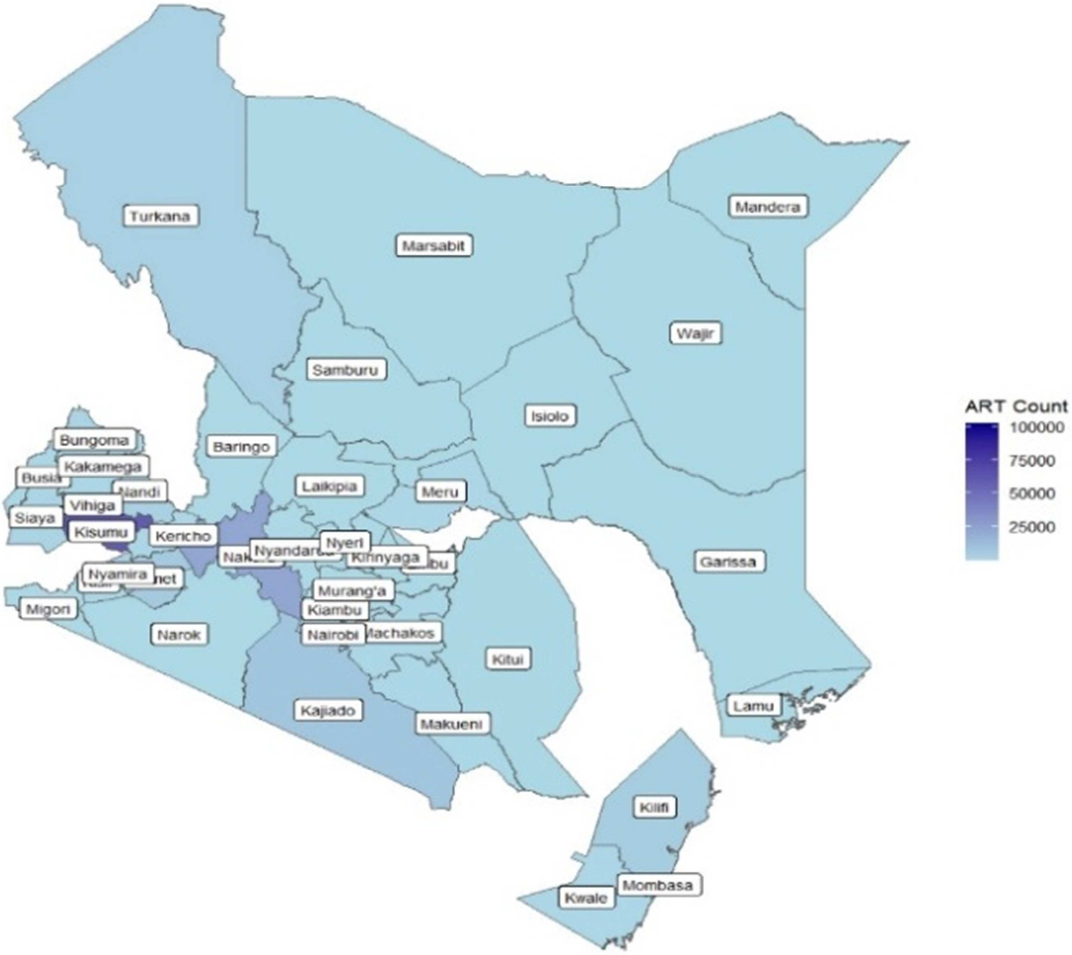

Figure 3. *Distribution of ART map.*

The counties with the highest ART frequencies were Nairobi, Kisumu, Homabay and Siaya.

4.2. Modelling

The ART data was examined in this section using Poisson and Negative Binomial Generalized Linear Mixed Models. Tables 1 and 2 illustrate the results:

Checking for overdispersion:

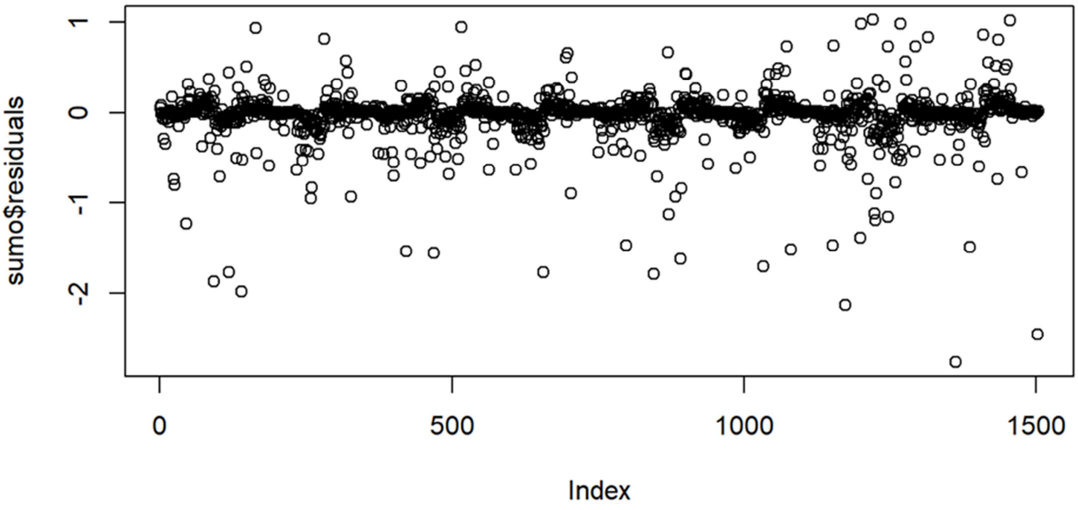

Figure 4. *Residual plot.*

The p-value of the model was 4.232e-09<0.05 hence there was evidence of over dispersion.

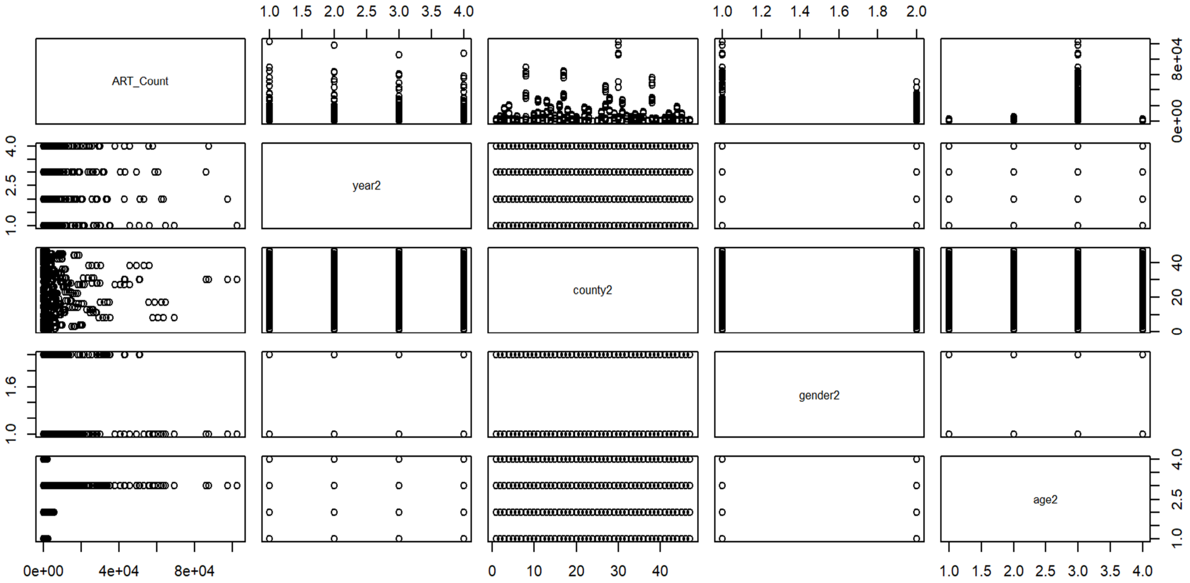

Figure 5. *Scatter plot matrix of the variables.*

A correlation that is negative (-0.12) was exhibited among ART count with gender, this implied that the model's predictive and analytical capabilities would be enhanced by the inclusion of this predictors. There exists a positive correlation (0.14) between ART count and age groups. Additionally, there exists a negative correlation (-0.023) between ART count and years. Consequently, there exist a negative correlation (-0.025) between ART count and county regions.

4.2.1. The Data Management Process

Cleaning of Data: A preliminary investigation of the data before running the model revealed that there were outliers in the data, which violated a fundamental assumption of the model. To overcame this the data was further cleaned. Identifying Outliers: The data was visualized through summary statistics, histograms and scatter plot in order to identify potential outliers. Definition and handling Criteria for Outliers: The threshold/criteria for the outlier constitution were noted based on standard deviations/ percentiles and datapoints that met outlier criteria deleted. Exclusion of patients aged <1 and 0-9 years. Data Validation and Correction: the data values recorded wrongly and data entry errors. Missing data was treated as not initiated on ART. Rechecking Assumptions: After data cleaning model assumptions were reassessed to meet assumptions necessary for the chosen modeling technique.

4.2.2. Regression Coefficients Interpretation

The male coefficient was -0. 2811 thus ($exp^{-0.2811}$) is equal to 0.75 which is the odds ratio. Therefore 100 (1-0.75)% = 25%, hence a male is 25% less likely to enroll on ART as compared to a female patient. For patients aged 15-19 the coefficient is -0.0539 giving 0.95 as the odds ratio since ($exp^{-0.0539}$). Thus 100 (1-0.95) % = 5%, hence a patient aged 15-19 is 5% less likely to enroll on ART as compared to a patient aged 10-14.

The patients aged 20-24, the coefficient was 0.0986 that is 0.91 as the odds ratio because ($exp^{-0.0986}$). Hence 100 (1-0.91)% = 9%, therefore an individual aged 20-24 is 9% less likely to enroll on ART as compared to a patient aged 10-14. The coefficient for 2019 was -0.0559 therefore ($exp^{-0.0559}$) equals to 0.94 which is the odds ratio. Thus 100 (1-0.94)% = 6%, hence in 2019 patients were 6% less likely to enroll on ART as compared to the year 2018. In the year 2021 the coefficient was -0.1799 that is 0.84 as the odds ratio of because ($exp^{-0.1799}$). Therefore 100 (1-0.84)% = 16%, hence in 2021 patients were 16% less likely to enroll on ART as compared to the year 2018.

Table 1. *GLMM-Poisson Results.*

AIC	BIC	LogLik	Deviance	df.resid
11409	11449.7	-5695.4	11390.9	683
Scaled residuals				
Min	IQ	Median	3Q	Max
8.6042	-2.0205	-0.4625	1.6711	23.5215
Random Effects				
Groups	Name	Variance	Std. Dev	
county	Intercept	0.9906	0.9953	
Number of obs: 692	Groups:	County, 47		

AIC	BIC	LogLik	Deviance	df.resid		
Fixed Effects						
Coefficients	estimate	san.se	z value	p		
Intercept	4.9868	0.1597	31.228	<2e-16***		
male	-0.2942	0.0073	-40.227	<2e-16***		
15-19	-0.0505	0.0085	-5.922	<3.18e-09***		
20-24	0.0299	0.0092	3.267	0.109		
25+	2.6176	0.0240	8.860	<2e-16***		
2019	-0.0625	0.0097	-6.508	<2e-16***		
2020	-0.0906	0.0095	-9.540	<2e-16***		
2021	-0.1695	0.0096	-12.582	<2e-16***		
Correlation of fixed effects						
	(Intr)	male	15-19	20-24	25+	2019
male	-0.023					
15-19	-0.025	-0.042				
20-24	-0.027	-0.169	0.462			
25+	-0.016	-0.173	0.188	0.208		
2019	-0.029	-0.005	0.004	-0.006	-0.030	
2020	-0.031	0.032	-0.013	-0.013	-0.017	0.495
2021	-0.031	-0.054	-0.024	-0.013	-0.062	0.490

Comparing each of the coefficient's p-values in table 1 to the level of significance of 0.05, age groups (15-19, 25+), gender and year were found to be significant predictors of ART count at alpha<0 05. As shown in table, the non-significant predictor was age (20-24). The column of standard deviation under random effects, the standard deviation of 0.9953 due to county.

Table 2. *Output from GLMM-Negative Binomial.*

AIC	BIC	LogLik	Deviance	df.resid
6833	6878	-3406.4	6812.7	682
Random Effects				
Groups	Name	Variance	Std. Dev	
county	Intercept	0.9626	0.9811	
Number of obs: 692	Groups:	County, 47		
Fixed Effects				
Coefficients	estimate	san.se	z value	p
Intercept	4.9925	0.1602	31.16	<2e-16***
male	-0.2811	0.0238	-11.80	<2e-16***
15-19	-0.0539	0.0272	-1.980	0.04760
20-24	0.0986	0.0301	0.030	0.9739
25+	2.5458	0.0721	835.33	<2e-16***
2019	-0.0559	0.0305	-1.830	0.06711
2020	-0.0828	0.0315	-2.750	0.00603***
2021	-0.1799	0.0311	-5.790	6.94e-09***

The random effect considered was County as it constitutes a random sample. Age, Gender, Year were considered constant factors. In table 2, under random effects the standard deviation of 0.9626 is a measure of the amount of variation in the dependent measure due to county Additionally, the variance estimate of random effect county is 0.9811. Due to this difference from zero, it implies that ART count varies for the different patients based on gender, age groups and year.

Table 3. *Output from GLM-Poisson.*

Residuals Deviance				
Min	**1Q**	**Median**	**3Q**	**Max**
-18.01	-9.2750	-0.0985	5.3874	15.1790
Coefficient:				
	Estimate	Std. Error	z value	Pr (>|z|)
(Intercept)	4.915088	0.008882	553.356	<2e-16***
15-19	-0.024404	0.008440	-2.891	0.00383***
20-24	0.003175	0.008428	0.377	0.70737
25+	0.492931	0.015246	32.331	<2e-16***
Male	-0.054262	0.006829	-7.946	<2e-16***
2019	0.053041	0.009539	-5.561	<3e-9***
2020	0.040905	0.009401	-4.351	<2e-6***
2021	0.060606	0.009424	-6.431	<2e-11***
Dispersion parameter for Poisson family taken to be 1				

Residuals Deviance				
Min	**1Q**	**Median**	**3Q**	**Max**
Null deviance	50517	On 691 degrees of freedom		
Residual deviance	49403	On 684 degrees of freedom		
AIC: 53745				
Number of Fisher Scoring iteration: 5				

Table 4. *Output from GLM-NB.*

Deviance Residuals				
Min	**1Q**	**Median**	**3Q**	**Max**
-2.9720	-1.0939	-0.0092	0.4968	1.3367
Coefficient:				
	Estimate	Std. Error	z value	Pr (>|z|)
(Intercept)	4.9201	0.09016	54.574	<2e-16***
15-19	-0.0273	0.08365	-0.326	0.74445
20-24	0.02672	0.08395	1.593	0.9975
25+	0.4938	0.01866	2.646	0.00813***
Male	-0.5832	0.06857	-0.850	0.39505
2019	-0.05343	0.09618	-0.556	0.57852
2020	-0.04134	0.09509	-0.454	0.65005
2021	-0.06187	0.09502	-0.651	0.51492
Dispersion parameter for Negative Binomial (1.2963) family taken to be 1				
Null deviance	786.41	On 691 degrees of freedom	Theta: 1.2963	
Residual deviance	776.58	On 691 degrees of freedom	Std. Err: 0.0646	
AIC: 8103			2 x log-likelihood	-8084.9640
Number of Fisher Scoring iteration: 1				

4.2.3. Model Fitting

The test employed in testing the significance of the model was likelihood ratio test. Compared the Generalized Linear Mixed Model incorporated with effects that are random for county to the negative binomial regression model to the one with only the fixed factors. At the 0.05 significance, the null hypothesis that the Generalized Linear Mixed Model-NB model does not provide a better fit that is more significant than the GLM negative binomial model is ruled out. In addition, the negative binomial mixed model was compared to the negative binomial regression GLM model. When the output from Tables 2 and Table 4 are compared, the results demonstrated the AIC for fit 1 Mixed Model-NB is 6833 while the AIC in the GLM-NB is 8103. The model with fixed and random effects has a lower AIC, making it the better model.

4.3. Residual Diagnostics for the Regression Models

A simulation-based approach DHARMa (diagnostics for hierarchical regression models) was used to provide easily interpretable quantile residuals for models with random effects inclusion.

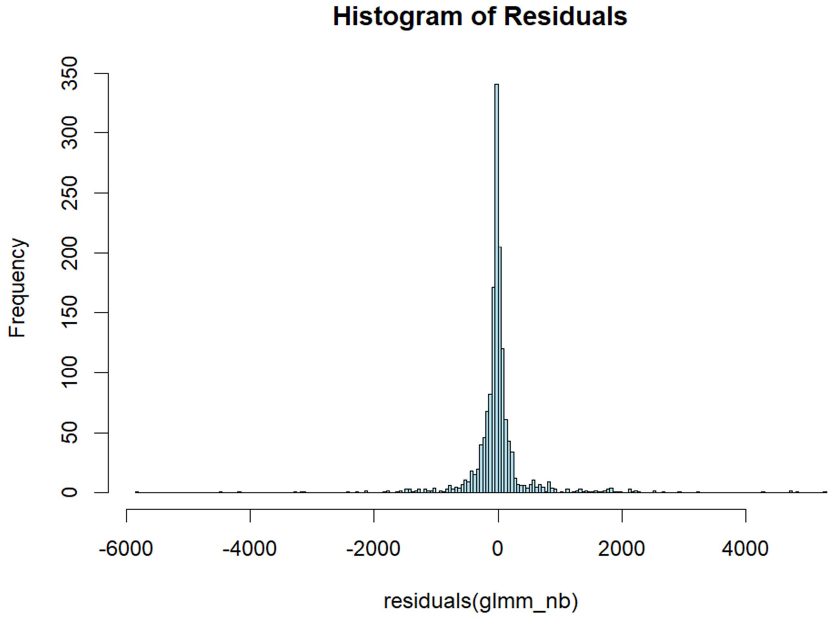

Figure 6. *Residual histogram.*

Figure 6 shows the residuals are not normally distributed, as there is positive skewness. In figure 7 and 8, the QQ-plot created in the left panel detects overall deviations from the expected distribution, by default with added tests for correct distribution (KS test), dispersion and outliers. Note that outliers are values that are by default defined as values outside the simulation envelope, not in terms of a particular quantile. Thus, which values will appear as outliers will depend on the number of simulations. Visualization of the residuals against the expected value exhibited in the panel that is on the right.

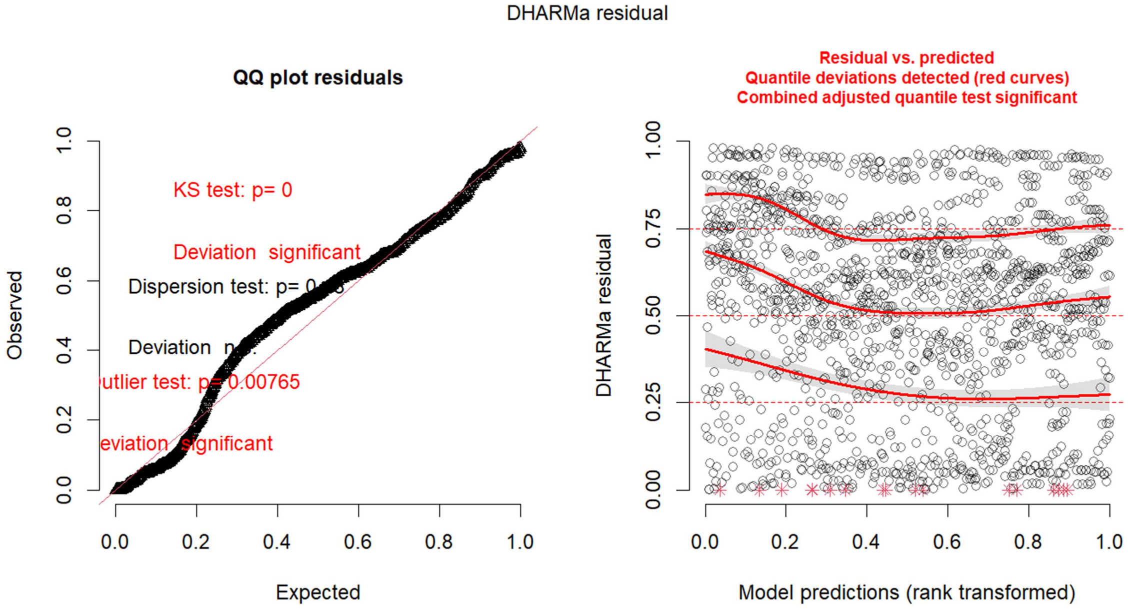

Figure 7. DHARMa residual Poisson GLMM plot.

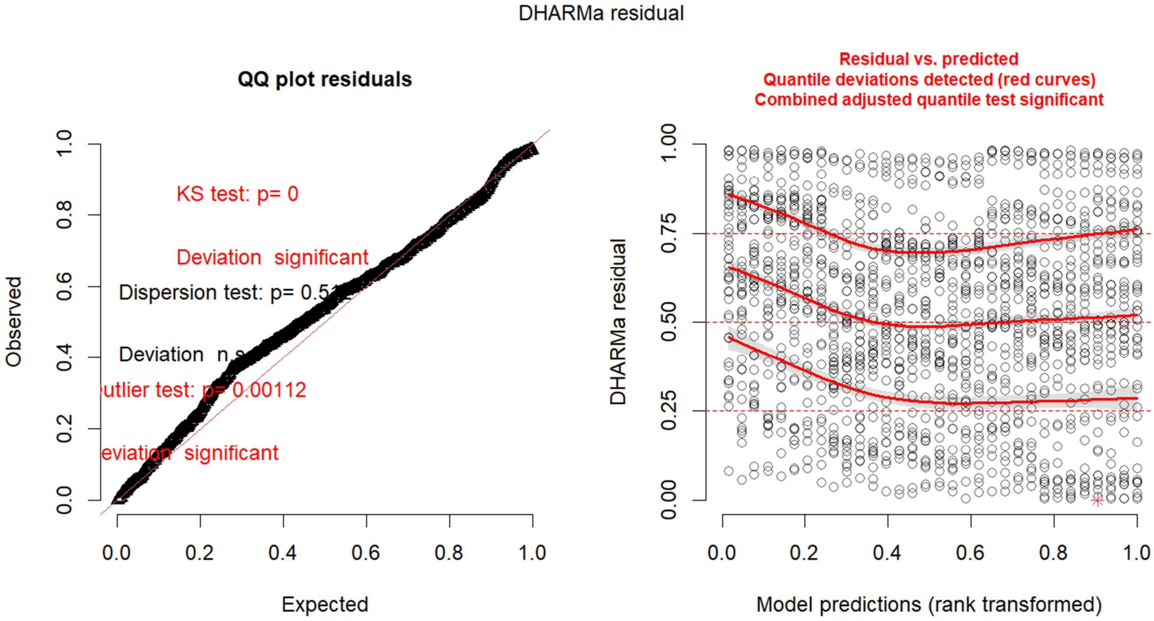

Figure 8. DHARMa residual Negative binomial GLMM plot.

Basing on the above residual plots: -

QQ: Distribution test (KS) significant in both model plots

QQ: Dispersion test is not significant in both model plots

QQ: Outlier test significant in both model plots

Res ~ predicted: Quantile fits are spread out too evenly

The research focused on modelling ART count repeated measures for HIV-positive patients in Kenya. According to the findings of the study, more women than males were initiated on ART. The study findings exhibited almost identical outcomes to the results of Kenya Aids Indicator Survey 2012 that demonstrated, 59 percent were women infected at a rate prevalence of 8 percent, compared to 4 percent for men [19]. The findings revealed that the highest proportion of Kenyans have HIV/AIDS enrolled on ART were patients aged 25+. Similarly, NASCOP 2010 indicated an increase in the number of orphans as a result of HIV infection in people aged 15 to 45, with 60 percent of new infections occurring between the ages of 35 and above. In both the models it was seen that some of the fixed effect coefficients (β_2, β_4) showed opposing signs for the two models, owing to the average of the population and specific subject interpretation for the two models. This is in support with the findings of D. Renard on modelling multilevel and longitudinal data [20].

5. Conclusion and Recommendation

In the research study, longitudinal models were used to assess the relationship between ART count and potential predictors. Statistically, the modeling technique; GLMM, an extension of the GLM, was utilized to analyze ART data from patients on ART. The emphasis was on model parameter interpretation and calculation. The distinctions between the two GLMMs and GLM when modeling a count response were examined for parameter interpretation.

The study found Age and Gender to be significant predictors of ART count under both models. The Generalized Linear Mixed-effects Model (GLMM) allowed for regression analysis using correlated data and components with specified variance that represented within-subject and between-subject variation in outcomes. For the known covariance structures, AIC and the BIC comparison was performed, the model that best suited was the GLMM-negative binomial model. HIV being vital, modeling the data assists in identifying the elements that impact the efficacy of ART in order to postpone the rapid advancement of HIV. Thus, new studies in HIV research should be conducted using these flexible statistical approaches, including additional covariates, to enhance the predication of the model. It aided the monitoring of patients and follow-up ensuring provision of proper care. Despite the construction of GLMM models being complex, the models flexible in that they can be formulated for any design structure. Repeated measures (a.k.a. longitudinal) design structure can be used effectively to account for sources of variations in the long-term experiments.

Furthermore, whereas the choice of GLMM for longitudinal data is limited to subject matter, using GLMM for correlated data has lot of emphasis since the model can handle both the within subject variations and by integrating random effects, within measurement variance and between individual variability can be accounted for. As a result, the model with mixed effects fits a given data set with a minor disturbance.

This research can help to inform public education, particularly for patients, as well as policy and therapeutic management. Existing demand for ART implementers and those planning to build up HIV and AIDS programs for stigma reduction in which all HIV-positive patients can come together and discuss their experiences.

The education on health that is extensive regarding ART and HIV in general should be delivered to all groups regardless of education levels. To obtain thorough investigation on ART adherence influencing factors in the same study group, qualitative research is required with inclusion of participants from a wider range. More research is needed to broaden understanding and knowledge of longitudinal data analysis, and also to include more covariates.

References

[1] Global report, UNAIDS Report on the global AIDS epidemic 2012, pp. 6.

[2] E. Ziegal, (2000). COMPSTAT: Proceedings in Computational Statistics. Technometrics, vol. 44 no. 1, pp. 96.

[3] F. Ye, C. Yue and Y. Yang (2013). Modeling time-independent overdispersion in longitudinal count data. Computational Statistics and Data Analysis, vol. 58 no. pp. 257-264.

[4] Rebecca V. Culshaw (2006), Mathematical Modeling of AIDS Progression: Limitations, Expectations, and Future Directions. Journal of American Physicians and Surgeons Vol 11 no. 4.

[5] Y. Liang and L. Zeger (1986), Longitudinal data analysis using generalized linear models. Biometrika, vol. 73 no. 1, pp. 13-22.

[6] M. Laird and H. Ware, (1982). Random-effects models for longitudinal data. Biometrics, vol. 38 no. 4, pp. 963-974.

[7] UNAIDS report on the Global Aids Epidemic 2010, pp. 107.

[8] S. McClelland, (2009). Public Health Aspects of HIV/AIDS in Low- and Middle-Income Countries: Epidemiology, Prevention and Care. JAMA, vol. 302 no. 5, pp. 573-577.

[9] F. J. Palella, K. M. Delaney, A. C. Moorman, M. O. Loveless, J. Fuhrer, G. A. Satten, S. D. Holmberg, (1998). Declining Morbidity and Mortality among Patients with Advanced Human Immunodeficiency Virus Infection. New England Journal of Medicine, vol. 338 no. 13, pp. 853–860.

[10] Kenya Hiv Prevention Revolution Road Map, NASCOP 2014, pp. 7-8.

[11] H. Zhang, H. Wong, and L. Wu, (2018). A mechanistic nonlinear model for censored and mis-measured covariates in longitudinal models, with application in AIDS studies. Statistics in Medicine, Vol. 37 no. 1, pp. 167-178.

[12] T. Yu and L. Wu, (2018). Robust modelling of the relationship between CD4 and viral load for complex AIDS data. Journal of Applied Statistics, vol. 45 no. 2, pp. 367-383.

[13] T. Wendler and S. Grottrup, (2016). Data Mining with SPSS Modeler, Theory, Exercises and Solutions.

[14] X. Lu, (2014). Statistical Modeling and Prediction of HIV/AIDS Prognosis: Bayesian Analyses of Nonlinear Dynamic Mixtures. USF Tampa Graduate Theses and Dissertations.

[15] S. E. Holte, T. W. Randolph, J. Ding, J. Tien, R. S. McClelland, J. M. Baeten, and J. Overbaugh, (2012). Efficient use of longitudinal CD4 counts and viral load measures in survival analysis. Stat. Med, vol. 31 no. 2 and no. 19, pp. 2086–2097.

[16] H. Donald and G. Robert D. Longitudinal Data Analysis. *Wiley*, 2006.

[17] B. M. Bolker, M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. H. Stevens, and J. S. S. White, (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution*, vol. 24 no. 3, pp. 127-135.

[18] G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs, (Eds.). (2008). *Longitudinal data analysis*. CRC press. pp. 19-20.

[19] Kenya Aids Indicator Survey, 2012. *Acquir Immune Defic Syndr*. Volume 66, Supplement 1, May 1, 2014.

[20] D. Renard, (2002). Topics in modeling multilevel and longitudinal data. PhD thesis.