Evaluating Generalized Linear Models For Count Data With Application To Pre-Exposure Prophylaxis HIV Sero-Conversion Data

by Jannis Ndungwa Mutisya

Submission date: 23-Nov-2020 05:28PM (UTC+0300)

Submission ID: 1455061013

File name: Jannis_Mutisya_Final_MSC_project.pdf (1.48M)

Word count: 8973

Character count: 48938

University of Nairobi

ISSN: 2410-1397

Master Project in Social Statistics

Evaluating Generalized Linear Models For Count Data With Application To Pre-Exposure Prophylaxis HIV Sero-Conversion Data

Research Report in Mathematics, Number 33, 2020

Jannis Ndungwa Mutisya

November 2020

Submitted to the School of Mathematics in partial fulfilment for a degree in Masters of Science in Social Statistics

School of Mathematics

Master Project in Social Statistics

University of Nairobi
November 2020

Evaluating Generalized Linear Models For Count Data With Application To Pre-Exposure Prophylaxis HIV Sero-Conversion Data

Research Report in Mathematics, Number 33, 2020

Iannis Ndungwa Mutisya

School of Mathematics
College of Biological and Physical sciences
Chiromo, off Riverside Drive
30197-00100 Nairobi, Kenya

Master Thesis

Submitted to the School of Mathematics in partial fulfilment for a degree in Masters of Science in Social Statistics

Submitted to: The Graduate School, University of Nairobi, Kenya

Abstract

Background Generalized Linear Models(GLMs) are a strategy for tackling statistical questions, especially those that involve non-normally distributed data, in such a way that much of the simpleness of the linear model is retained . This study was aimed to evaluate Generalized linear models for count data with application to Pre-Exposure Prophylaxis HIV sero-conversion(PrEP) data.

Methods This study used data that was retrieved from Kenya Health Information System(KHIS) for the period March to April,2019 from 104 health facilities.Poisson Regression Model,Quasi-Poisson Regression model Negative Binomial Regression Model,and Conway-Maxwell Poisson regression models were compared to determine the best model which can be used in modeling HIV sero-conversion among PrEP users in Kenya.The model with the best fit was checked using Akaike information criterion(AIC).

Results and Conclusion From the results, the Conway-Maxwell Poisson regression was considered a better model when analyzing PrEP data in Kenya since its AIC value was the least.

Master Thesis in Mathematics at the University of Nairobi, Kenya.
ISSN 2410-1397: Research Report in Mathematics

DISTRIBUTOR: School of Mathematics, University of Nairobi, Kenya

Declaration and Approval

I the undersigned declare that this dissertation is my original work and to the best of my knowledge, it has not been submitted in support of an award of a degree in any other university or institution of learning.

Signature Date

Jannis Ndungwa Mutisya
Reg No. I56/12046/2018

In my capacity as a supervisor of the candidate's dissertation, I certify that this dissertation has my approval for submission.

Signature Date

Dr Idah Orowe
School of Mathematics
University of Nairobi,
Box 30197, 00100 Nairobi, Kenya.
E-mail: orowe@uonbi.ac.ke

Dedication

I dedicate this project to my husband,Joseph and my sons, David and Jonathan.

Contents

Figures and Tables

Figures

Tables

List of Abbreviations

AIC	Akaike Information Criterion
AIDS	Acquired Immunodeficiency Syndrome
ART	Antiretroviral Therapy
ARVS	Antiretroviral Drug(s)
CDC	Centres for Disease Control
COM	Conway-Maxwell
GLMs	Generalized Linear Models
HPV	Human Papillomavirus
HIV	Human Immunodeficiency Virus
HIM	Human Papillomavirus in Men
KHIS	Kenya Health Information System
MSM	Men who have sex with men
MOH	Ministry of Health
NASCOP	National AIDS and STI Control Programme
PrEP	Pre-Exposure Prophylaxis
STIs	Sexually Transmitted Infections
TG	Transgender
UNAIDS	United Nations Programme on HIV/AIDS
USAID	United States Agency for International Development
WHO	World Health Organization

Acknowledgments

First, I would like to acknowledge the Almighty God, for the gift of life, provision and for enabling me to pursue this course.

I am grateful to my supervisor, Dr Idah Orowe for her continuous guidance, advice and valuable comments throughout all the stages of this project.

I also wish to thank Dr Collins Odhiambo for providing the data used in this project and for invaluable advice and guidance throughout the project.

I also wish to extend my appreciation to all my lecturers and other members of staff of the School of Mathematics, University of Nairobi for their immense support.

To my course mates, who were always available for consultation and discussions,am forever grateful for your time,wise counsel and encouragement, my studies would not have been the same without you.

Special appreciation to my husband Joseph, for his unwavering support and encouragement throughout my studies. To my dear parents for always being there for me, thank you mum and dad. My siblings for their love, prayers and support at every stage my studies.

Jannis Ndungwa Mutisya

Nairobi, 2020.

0.1 Introduction

0.1.1 Background of the study

Generalized Linear Models

Generalized Linear Models (GLMs) are a strategy for tackling statistical questions, especially those that involve non-normally distributed data, in such a way that much of the simpleness of the linear model is retained, McCulloch (2000). GLMs also give a more adequate way of linking the systematic part of a model together with the random part and the fact that a single algorithm is used to fit any model is an indication that quite a small combination of routines can provide a base computing tool to enable researchers to fit models to quite a wider range of data, Nelder & Wedderburn (1972).

GLMs relate a response variable of interest, to other variables or predictors (also called factors, covariates or independent variables) about which you know or have idea about. To be able to accomplish this, it is important to have the distribution of the response first defined, after which the independent variables can then be related to the response thus allowing for the random variation of the data. Therefore, firstly we consider the general form of distribution which is known as exponential family which is used in GLMs. One of the advantages of GLMs is that regression is no more restricted to normal data, but it extends to distributions that are members of the exponential family. This aspect allows for proper modeling of, for example frequency counting or data that is skewed.

Conversely to models suitable for normally distributed data, GLMs allow fitting of skewed distributions while relaxing the assumption of a constant variance(that is,it varies as a function of the mean), which provides a choice of scales, through the link transformation of the mean, which yields an additive (meaning there is non-interaction) linear model. Therefore, many data types;whether continuous or binary and count, can be modelled by models that belong to this rich family of generalized linear models, Lindsey & Jones (1998).

In statistical modeling, if the response variable is count data, the most popularly used regression tool used is the Poisson regression model. However, sometimes the Poisson model fails to provide a sufficient fit when there is existence of the problem of over-dispersion or under-dispersion and zero-inflation since it assumes equi-dispersion (Variance is equal to the mean) which is rarely reflected in real data since in most cases the variance is usually greater than the mean and for this reason the Poisson model has become less ideal for modeling.Therefore, modifications have been done to the Poisson model so as to be able to deal with a wider range of dispersion levels.This study will use Poisson,Negative Binomial, Quasi-Poisson and Conway-Maxwell-Poisson (COM-Poisson) Regression models which all belong to the family of generalized linear models, Nelder & Wedderburn (1972); McCullagh and Nelder 1989).

Departures from poisson model can occur in a variety of ways;the main reasons are: some covariates may be ommitted and/or may not have a uniform effect on all subjects so that population heterogeneity has not been accounted for, and excess number of zero events occured compared to the poisson distribution Lee et al. (2012). The issue of over-dispersion with excess zeroes existed in the Human Papillomavirus Infection in Men (HIM) dataset. The study established a prospective cohort of men in three countries to determine the incidence of genital human papillomavirus (HPV) infections. A HPV incidence rate,along with the exact 95% confidence interval,was estimated based on a Poisson distribution. However, inspection of the data revealed severe over-dispersion,as well as a very large proportion of zero counts for specific HPV-type infections, (Giuliano et al.)

Although GLMs have been widely used in other fields, for example in econometrics, they have not been as widely used in medical statistics as expected. Specifically GLMs have not been extended to Pre-Exposure Prophylaxis (PrEP) data. In this study, we will illustrate modelling Pre-Exposure Prophylaxis (PrEP) count data using the Poisson, Quasi-Poisson, Negative Binomial and Conway-Maxwell-Poisson Regression models. We will fit the generalized linear models for count data on PrEP data and select the model that is most adequate.

Pre-Exposure Prophylaxis

Pre-Exposure Prophylaxis (PrEP) being an anti-HIV medication is given to HIV negative people who are also at a high risk of HIV infection so as to reduce their chances of being infected. Although there have been a number of substantial studies suggesting the high potential efficacy of PrEP, its large-scale implementation has been limited by several factors including cost, adherence, and concern about resistance arising from PrEP, especially in resource constraint settings, in which Antiretroviral treatment options are limited. According to National AIDS & STI Control Programme (NASCOP) (2017), Kenya has made significance steps to contain the HIV pandemic, for example in 2015 the HIV prevalence in Kenya went down by almost 50% from an index of 10.6% in 1995/96 to about 5.9% . This decline has been accomplished through the unwavering implementation of a combination of evidence-based interventions including increased use of antiretroviral therapy, NASCOP(2017).

A routine test for HIV is part of the package of PrEP services so as to prevent development of resistance, and also for timely identification of PrEP users who get infected with HIV. A person's HIV status should be determined and documented at the point of initiating PrEP, in the first month, and every 3 months preceding initiation of PrEP, NASCOP(2017). In 2015, the World Health Organization (WHO) released a number of recommendations which supported the use of drugs such as pre-exposure prophylaxis (PrEP) to prevent the acquisition of HIV. It was recommended that an urgent initiation into antiretroviral treatment (ART) together with adequate provision of PrEP to all those at a high continuous risk of acquiring HIV infection would help to reduce the risk of acquiring HIV.

In 2016, Kenya also adopted the oral PrEP in its ART guidelines for people at substantial risk of HIV infection.Baeten et al. (2016) and in 2017, the PrEP implementation framework was published and has been rolling out PrEP across the public health sector since then. The framework identifies 19 geographical areas with high HIV prevalence where PrEP should be available. It also names the following population groups as priorities for PrEP access: female sex workers, people in discordant couples, pregnant women, fishing communities around Lake Victoria, adolescents and young people, people in the general population with multiple sexual partners, men who have sex with men, and people who inject drugs, NASCOP(2017).

0.1.2 Problem statement

There have been few studies among sero-discordant couples or heterosexual populations, studies examining knowledge and attitudes towards PrEP and related behaviors among priority groups for PrEP have been conducted in a variety of locations, including the United States,Organization et al. (2012). These studies have surveyed a variety of settings, including HIV clinics among others. Focus group participants said that PrEP was acceptable, but potential sexual risk disinhibition, stigma and discrimination associated with PrEP use, and mistrust of healthcare professionals were major concerns, Galea JT.et al.

There is need for proper modeling of the PrEP indicators and since PrEP sero-conversion data is highly skewed and there is no available literature that have similar data features as PrEP that have used GLMs for count data, this study therefore seeks fit PrEP data which was retrieved from Kenya Health Information System (KHIS) for the period covering March 2019 to April 2019, into four GLMs. we will choose the best model that can be used to analyze the current trend of HIV sero-conversion among PreP users and also for prediction of future trends.

Main objective

To compare different count data models of estimating HIV Sero conversion among PrEP users.

Specific objectives

1. To compare the performance of Poisson model, Negative Binomial model, Quasi-Poisson model and COM-Poisson model under different simulated data sets.

2. To evaluate the best fit model for HIV sero conversion among PrEP users by comparing Poisson, Negative Binomial, Quasi-Poisson and COM-Poisson Regression models.

0.1.3 Significance of the study

Despite the high rate of new HIV infections in Kenya, and a growing body of literature on the use of PrEP to reduce the chances of becoming infected, there have been no appropriate modeling techniques employed to estimate the incidence rate of HIV infections and the factors associated with the new infections among PrEP users. This study will therefore aid in identifying the best model for predicting, analyzing trends among PrEP users and for planning purposes.

0.2 Literature review

0.2.1 introduction

More than 34 million people globally are living with HIV on HIV/AIDS et al. (2010). A number of prevention methods are available including; use of condoms, male circumcision, prevention of mother-to-child transmission, use of sterilized needles, however, these approaches have not been sufficient to stop the epidemic. In 2009 alone, an estimated 2.7 million people became newly infected on HIV/AIDS et al. (2010). Therefore the urgent need for additional safe and effective measures to HIV prevention.

Men and transgender women who have sex with men (MSM and TG) have a disproportionate burden of HIV in most countries in the world, even in many countries with generalized HIV epidemics. Worldwide, their odds of being infected with HIV are 19.3 times higher than those for others (WHO). Clearly, existing methods of HIV prevention are not sufficient for MSM and TG. Biomedical prevention has shown promise. Male circumcision has proved effective in protecting heterosexual men who are exposed to HIV during penile-vaginal intercourse, and a vaginal gel has shown some effectiveness in protecting women who are exposed by vaginal intercourse. Pre-exposure prophylaxis (PrEP) is the first biomedical intervention that has proved effective in providing additional protection to men who have unprotected rectal exposure to HIV (WHO).

According to NASCOP (2017), the access to HIV medication by more than one million Kenyans has greatly improved the quality of life of people who are living with HIV. But even with this progress, approximately over 77,000 Kenyans were infected with HIV in the year 2015. About half of this number were young people at the age between 15 to 24 years whereby young women bore a third of all recorded new infections.Fighting the HIV epidemic will not be successful until the index of the new HIV infections goes down. This can be achieved through aggressive input towards HIV prevention programs which will work towards ensuring those who are HIV negative remain uninfected.

As much as substantial progress has been achieved in the reduction of the number of new HIV infections, there are certain populations that are still at a high risk of HIV infection with the HIV prevalence in Kenya remaining high though showing signs of stabilization. Therefore, such interventions as the case of implementing PrEP in a very strategic and calculated approach are key in reducing the HIV infections,NASCOP (2017).

0.2.2 Mathematical modelling

Mathematical modelling is a valid tool in estimating the use of PrEP and the associated new HIV infections. Modeling count variables is a common task in many fields including economics,social sciences and medicine. The statistical approach of count data is different from that of, for example,binary data, whereby observations only take two values, represented by 0 and 1, or ordinal data, which may include integers but where the particular values fall on an arbitrary scale and only the relative ranking is important.For Count data models, the dependent variable has to be count data.Maxwell et al. (2018). The classical Poisson regression model for count data is often of limited use in these disciplines because empirical count data sets typically exhibit over-dispersion and/or an excess number of zeros. This can be resolved by extending the plain Poisson regression model into various dimensions.

Maxwell et al. (2018) in modeling auto crash data, used Poisson Regression Model, Negative Binomial Regression Model, Generalized Poisson Regression Model, and Conway-Maxwell Poisson regression model which were compared to determine a better model used in modeling auto-crashes in Nigeria, the best model for modeling traffic crash data in Nigeria was the Generalized Poisson Regression model based on AIC and BIC values. This study used Generalized Poisson model while in our case the Quasi-Poisson model was included for comparison.

Lee et al. (2012) used five count data models; Poisson regression model, Negative binomial regression model, Zero-inflated Poisson regression model and Zero-inflated Negative binomial model. They illustrated the use of the four models for over-dispersed data (Human Pappillomavirus infection in men study) that may be attributed to excessive zeroes. They recommended Negative Binomial was a better fit, however, Zero-inflated Poisson model showed similar results as the negative binomial although there were computational difficulties with zero-inflated models. They also recommended that zero-inflated models should be used with a lot of caution since cases of small sample size and variable selection of covariates have not yet been well studied in literature. Although our study does not focus on the Zero-inflated models, many researchers have compared them with the models that this study has focused on.

Zeileis et al. (2008) used Poisson model, Quasi-Poisson model, Negative binomial model, Hurdle models and Zero-inflated models to model the demand for medical care as it had been captured by the number of hospital visits. The negative-binomial-based model performed much better in capturing over-dispersion than Poisson models, however, both hurdle and zero-inflated models were able to incorporate both over-dispersion and excess zeroes. Therefore, the hurdle and zero-inflation models led to the best results on the medical care data.

Muoka et al. (2016) used statistical stimulation technique to compare the performance of different count data models that is; Poisson model, Negative Binomial and Hurdle model. They simulated sets of count data with different proportions of zero and Akaike Information Criterion (AIC) was used to compare how good various count data models fit the simulated data sets. from the results, Negative Binomial fitted better to over-dispersed data which has proportion of zeros below 30% and Hurdle model fitted better in data sets with proportion of zeros 30% and above.

Johansson (2014) focused on comparing five regression models OLS, Poisson, Negative Binomial, Hurdle based on Poisson and Hurdle based on Negative Binomial. The study aimed at choosing the best choice model for predicting the number of claims that an insurance company will have in one year from the third -party automobile insurance. since this is count data, the OLS was included to illustrate how much better fit one can get by using an appropriate model. From the results, OLS is a bad choice for modelling claims, which is count data, but the other models fit the data well.
From the result, it was impossible to find the perfect model since the model that is the

best one in one test is worse in another. In the case of insurance, it is not only important to have a large likelihood for the whole data set, but it should also make fair predictions for all groups. One drawback of the thesis was that the data was not current and also lack of available data whereas the claim behavior could have changed since then in such a way that other models would suit better currently.

Jiang & House (2017) undertook a study to examine the performance of six count-data models (Poisson model, Zero-Inflated Poisson model, Hurdle Poisson model, and their negative binomial variations) under different zero-proportion, and skewness levels using simulation studies. They also compared the capabilities of these models on predicting zero-observations, and structural zero-observations, in order to evaluate their capabilities in predicting market structure when applying to the food consumption analysis. From the results, they recommended to the researchers to consider the hurdle models when there is zero-deflation, and the zero-inflated models when there is zero-inflation. If the underlying assumption assumes that there are different types of zero observations, it is recommended to use zero-inflated models.

Miller (2007) compared the fit between the Poisson, ZIP, and Hurdle models together with their negative binomial formulations. Each of the analysis was performed for simulated data with five different proportions of zeros and three different amounts of skew for the nonzero distribution with the intention of clarifying the discrepant results from previous research work. The main aim of the study was to determine superiority of fit for the different models with different proportions of zero-inflation and different levels of skew. The study used deviance statistics and Akaike Information Criterion to examine fit between the models and from their findings, the Negative binomial Poisson model was significantly a better model fit compared to the Poisson model for most of the conditions while the Hurdle model was a better fit of all two-part models. One of the shortcoming in this study is that the research did not present sufficient information, more so on data that is required to calculate skew. Although not all research presented sufficient information, especially data necessary to calculate skew, there was clearly enough variation in results to warrant further research.

Warton (2005) compared 20 data sets with varying sample sizes, zero proportions, and factors/levels. The version of ordinary least squares included the addition of one to all the counts before the logarithm was taken . The other models that do not accommodate zero-inflation were the Poisson and four formulations of the negative binomial Poisson including the quasi-Poisson. The zero-inflated models included the ZIP model and the negative binomial ZIP model. The Akaike Information Criterion (AIC) values were then calculated and averaged. If over-dispersion was present, the negative binomial formulations were a better fit. However, if over-dispersion was absent, the opposite was true for more than 50% of the variables. This implies that the skew level in the model interacts with zero-inflation when we are interested in measuring model adequacy.However, other features of these data set including varying degrees of zero-inflation and overall distributions called for further research toward appropriate model selection.

Lindsey & Jones (1998) compared Normal distribution, Log-normal distribution, Gamma distribution, Inverse Gaussian distribution, Poisson distribution and Negative Binomial distribution in a study comparing generalized linear models to check for difference in T4 cell counts between two groups. Negative Binomial model was identified as the best fitting model, however they concluded that appropriate model selection criteria for any study should be well specified including clinical trials, in order to draw optimum inferences since model selection was not just enough. They also stated that their concentration on GLM family was because of their importance and the ready availability of software to fit them.

Greene (1994) used credit-reporting data to investigate differences between the Poisson, negative binomial Poisson, ZIP, negative binomial ZIP, as well as some of their aforementioned variants and the specification of a probit link rather than the logit link. The data consisted of 1,023 people who had been approved for credit cards. The count variable of concern was the number of major derogatory reports (MDR), which is the number of payment delinquencies in the past 60 days.The negative binomial Poisson resulted in improved fit (based on the Vuong test statistic), increased standard errors and different parameter estimates. The ZIP model resulted in slightly worse fit than the negative binomial Poisson while remaining much better compared to the Poisson model.

In summary, from the literature above, most of the reseachers comparing the Performance of Generalized linear models for count data were inclined to zero-inflated models; some stated computational difficulties and others stated that they should be used with caution as variable selection of the zero models components have not really been studied well in literature. In this study, we focused on the Poisson regression model, Negative binomial model, Quasi-Poisson regression model and the COM-Poisson regression models for comparison of their performance.

0.3 Methodology

In this section, we will be concerned with reviewing models that are mostly used to model the count data, including understanding the model formulation, and parameter estimation.These models include: Poisson model, Quasi-Poisson model, Negative Binomial Regression model and Conway-Maxwell Poisson model.

0.3.1 Framework for Count Data GLMs

The Generalized linear model was initially described by Nelder and Wedderburn (1972) and later developed and explained by Mc Cullagh (1989).
GLMs describe the dependence of a scalar variable $y_i(i = 1, \ldots, n)$ on a vector of regressors xi. All GLMs possess a random component, a systematic component, and a link function. The conditional distribution of $yi|xi$ is a linear exponential family with probability density function;

$$f(y;\lambda,\phi) = exp\left(\frac{y\cdot\lambda - b(\lambda)}{\phi} + c(y,\phi)\right) \tag{1}$$

whereλ is the canonical parameter that depends on the regressor via a linear predictor and ϕ is a dispersion parameter that is often known.The functions $b(\cdot)$ and $c(\cdot)$ are known and they determine which member of the family is used.
Conditional mean and variance of y_i are given by $E[y_i|x_i] = \mu i = b'(\lambda i)$ and $VAR[y_i|x_i] = \phi \cdot b''(\lambda i)$. Therefore, upto a dispersion parameter ϕ, the distribution of y_i is determined by its mean. Its variance is proportional to $v(\mu) = b''(\lambda(\mu))$, also known as variance function.
The dependence of the conditional mean $E(y_i|x_i) = \mu$ on the regressor x_i is specified as;

$$g(\mu) = x_i^T\beta \tag{2}$$

Where $g(\cdot)$ is a known link function and β is the vector of regression coefficients which are usually estimated by Maximum Likelihood (ML).

Poisson models

The Poisson regression model is a special case of Generalized Linear Models (GLM) framework which was derived by Poisson (1837). It is the simplest and the mostly used distribution for modeling count data. Its probability density function is given by;

$$f(y;\mu) = \frac{exp(-\mu)\cdot\mu^{y}}{y!} \tag{3}$$

The canonical link $g(\mu) = \log(\mu)$, resulting in a log-linear relationship between the mean and linear predictor.The mean and the variance of the Poisson distribution are equal $E(y) = var(y) = \mu$. The dispersion parameter of Poisson distribution is $\phi = 1$ while the expected value μ is a linear function of n predictors that take the values $X' = (x_1, ..., x_n)$ for the ith case so that;$\mu = X'\beta$.

Where β is a vector of the parameters to be estimated.

From equation (3), it is clear that the Poisson distribution belongs to the exponential family bacause its probability distribution function can be expressed as;

$$f(y) = exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right) \tag{4}$$

where θ and ϕ are location and scale parameters and $a(\phi)$,$b(\theta)$ and $c(y,\phi)$ are known functions.

The problem with Poisson regression model is that its assumption of mean equal to the variance is very restrictive when it comes to real data where the observed variance exceeds the observed mean.

Parameter Estimation The parameters of Poisson model are estimated by maximum likelihood approach using an iteratively re-weighted least squares algorithm. The log-likelihood for a sample $y_1, ... y_n$ can be expressed as;

$$l = \sum_{i=1}^{n}\left(\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi)\right) \tag{5}$$

The maximum likelihood estimates are therefore obtained by solving the equations;

$$S(\beta_j) = \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n}\left(\frac{y_i - \mu_i}{\phi_i v(\mu_i)} * \frac{x_{ij}}{g'(\mu_i)}\right) = 0 \tag{6}$$

For parameters β_j,where $v(\mu_i)$ is a variance function. The assumption is that $\phi_i = \frac{\phi}{a_i}$ Where ϕ is a single dispersion parameter and *ai* are known prior weights. The estimating equation therefore can be written as;

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \left(\frac{a_i(y_i - \mu_i)}{v(\mu_i)} \cdot \frac{x_{ij}}{g'(\mu_i)} \right) = 0 \tag{7}$$

Equation (7) above is then solved by using Fisher's scoring iterative algorithm such that in the rth iteration, the new estimate $\beta^{(}r+1)$ is obtained from the previous estimate β^r by use of the equation below;

$$\beta^{r+1} = \beta^r + S(\beta^r)E(H(\beta^r))^{-1} \tag{8}$$

Where H is the Hessian matrix,that is,matrix of the second derivatives of the log-likelihood. The parameters are estimated by the equation;

$$\beta^{r+1} = (X^T W^r X)^{-1} X^T W^r Z^r \tag{9}$$

Where $W^r = diag(w_i)$ and the working dependent variable;

$$Z_i = \eta_i^r + (y_i - \mu_i^r)g'(\mu_i^r)$$

and

$$w_i^r = \frac{a_i}{v(u_i^r)(g'(\mu^t))^2}$$

The process is then repeated until successive estimates converge.

Negative-Binomial Regression model

The Negative binomial regression which was first derived by Greenwood and Yule (1920) is a generalization of Poisson regression which is used for modeling over-dispersed count data and eases the restrictive Poisson model assumption that the variance and the mean are equal. Therefore it allows the modeling of Poisson heterogeneity using a gamma distribution. It is thus a Poisson-gamma mixture distributions with probability density function;

$$p(y) = \frac{\Gamma(\theta + y)}{\Gamma(\theta)y!}\left[\frac{\beta}{1+\beta}\right]^{y}\left[\frac{1}{1+\beta}\right]^{\theta}, y_i = 0, 1, 2... \tag{10}$$

with mean $E(y) = \theta\beta$ and variance $var(y) = \theta\beta + \theta\beta^2$. The shape parameter is $\theta, \Gamma(\cdot)$ is the gamma function and the dispersion is $\phi = 1$.

The model can also be expressed in terms of log link function as;

$$\ln(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n \tag{11}$$

For n dependent variables and $\beta_0, \beta_1 ..., \beta_n$ being estimated.

Taking the exponent of equation (11), Negative Binomial distribution can be written as;

$$p(y) = \frac{\Gamma(\theta^{-1+y})}{\Gamma(\theta^{-1})(y+1)}\left[\frac{\theta e^{x_i\beta})}{1+\theta e^{x_i\beta}}\right]^{y}\left[\frac{1}{1+\theta e^{x_i\beta}}\right]^{1/\theta} \tag{12}$$

Where $\mu = \theta\beta, k = 1/\theta, \mu i > 0$, for i=1,2...,n and θ is the Negative binomial over-dispersion parameter.

Parameter Estimation Estimating θ and β using Maximum likelihood approach, the likelihood function is;

$$L(\theta,\beta) = \prod_{i=1}^{n} p(yi) = \prod_{i=1}^{n}\frac{\Gamma(\theta^{-1+y_i})}{\Gamma(\theta^{-1})(y_i+1)}\left[\frac{\theta e^{x_i\beta}}{1+\theta e^{x_i\beta}}\right]^{y_i}\left[\frac{1}{1+\theta e_i^x\beta}\right]^{1/\theta} \tag{13}$$

The log-likelihood function is;

$$\ln L(\theta,\beta) = \prod_{i=1}^{n} p\left(y_i\right) = \sum_{i=1}^{n}\begin{pmatrix} y_i \ln\theta + y_i\left(x_i \cdot \beta\right) - \left(y_i + \frac{1}{\theta}\right)) \\ \ln\left(1+\theta e^{x_i \cdot \beta}\right) + \ln\Gamma\left(y_i + \frac{1}{\theta}\right) \\ -\ln\left(y_i + 1\right) - \ln\Gamma\left(\frac{1}{\theta}\right) \end{pmatrix} \tag{14}$$

The θ and β values that maximize $\ln L(\theta,\beta)$ are the maximum likelihood estimates

Quasi-Poisson model

Quasi-Poisson model is an alternative approach to fit extra-dispersion parameter that takes care of the extra variance(Seyoum et al,2016).It uses the mean regression function and the variance function from Poisson GLM but allows the dispersion parameter ϕ to be unrestricted.

For a random variable y that follows a Quasi-Poisson distribution, the mean and the variance can be expressed as $E(y) = \mu$ and $var(y) = \phi E(Y) = \phi\mu$ in a GLM setting.

And the log-link function can be expressed as;

$$g(\mu_i) = log(\mu_i) = X_i^{T\beta} \tag{15}$$

A quasi-Poisson model function $q(\mu_i, y_i, \phi_i)$ is also the same as the definition of the first order derivative of its log-likelihood function.

The score function for ith observation is;

$$q(\mu_i, y_i, \phi) = \frac{y_i - \mu_i}{\phi V(\mu_i)} = \frac{y_i - \mu_i}{\phi\, \mu_i} \tag{16}$$

Where $V(\mu_i)$ is the variance function.

The quasi-likelihood function for the ith observation can be written as ;

$$Q(\mu_i, y_i, \phi) = \int_{y_i}^{\mu_i} q(s, y, \phi) ds = \int_{y_i}^{\mu i} \frac{y_i - s}{\phi_s} ds \tag{17}$$

and the quasi-likelihood function for the whole sample is the sum of the likelihood function for each observation, that is;

$$Q(\mu, y, \phi) = \sum_{i=1}^{n} Q(\mu_i, y_i, \phi) = \sum_{i=1}^{n} \int_{y_i}^{\mu_i} \frac{y_i - s}{\phi_s} ds \tag{18}$$

The parameters to be estimated will try to maximize $Q(\mu, y, \phi)$ values in the equations;

$$\frac{\partial Q}{\partial \beta} = \sum_{i=1}^{n} q(\mu_i, y_i, \phi) \frac{\partial \mu_i}{\partial \beta} = \sum_{i=1}^{n} \left(\frac{y_i - \mu_i}{\phi\, \mu_i}\right) \frac{\partial \mu_i}{\partial \beta} = 0 \tag{19}$$

Which is equal to;

$$\sum_{i=1}^{n}(y_i-\mu_i)X_i=0 \tag{20}$$

Conway-Maxwell-Poisson (COM-Poisson) Models

This distribution was originally proposed by Conway and Maxwell in 1962[3] as a solution to handling queueing systems with state-dependent service rates.The distribution generalizes the Poisson distribution by adding a parameter to model over-dispersion and under-dispersion and includes the geometric distribution as a special case and the Bernoulli distribution as a limiting case.COM-Poisson is flexible and can therefore handle a wide range of count data. Its probability distribution is;

$$P(X=j)=\frac{1}{Z(\lambda,v)}\cdot\frac{\lambda^j}{(j!)^v}, j\varepsilon Z^+=(0,1,2...) \tag{21}$$

For a random variable Y,where $Z(\lambda,v)$ is a normalizing constant defined by;

$$Z(\lambda,v)=\sum_{i=0}^{\infty}\frac{\lambda^i}{(i!)^v} \tag{22}$$

And v is the dispersion parameter. The domain of admissible parameters is $\lambda, v, > 0$ and $0<\lambda<1$

Parameter Estimation The parameters of the COM-Poisson model will be estimated by maximum likelihood approach.

The log-likelihood function can be expressed as;

$$LogL(y_1,...,y_n|\lambda,v)=log\lambda\sum_{i=1}^{n}y_i-v\sum_{i=1}^{n}log(y_i)-nlogZ(\lambda,v) \tag{23}$$

The maximum likelihood estimates are then obtained by iteratively solving the set of normal equations;

$$E(Y)=\bar{Y} \tag{24}$$

and

$$E(log(Y!)) = lo\bar{g}Y! \tag{25}$$

parameter Estimation The parameters of the GLM models will be estimated by maximum likelihood approach. This will be important to examine the significance of the variables in the models.

0.3.2 Model specification

The models have the number of people who tested HIV positive while on PrEP as the response variables and the four explanatory variables are; Number screened,Number initiated on PrEP,Number tested HIV positive at month1 refill and Defaulters. The models parameterize as follows;

$$\theta_i = \exp(\beta_1 Screened + \beta_2 Initiated + \beta_3 TestedHIVpositiveatmonth1refill + \beta_4 Defaulters)$$

0.3.3 Goodness of fit

One of the most popular used measures for comparing model performance when dealing with several models is AIC. This study used AIC(Akaike Information Criterion) to determine the goodness of fit for model selection.
AIC,which was first developed by Akaike(1973) takes the form;

$$AIC = 2p - 2\log(\hat{\theta})$$

Where p is the number of parameters and $\hat{\theta}$ is the maximum likelihood function for the fitted model.A relatively small value of AIC is mostly preferred for the fitted model.

0.3.4 Simulations

This study used simulation technique in R to generate data that was used for comparing the four count data models.
The random number generation was performed under the following specified conditions:

1. The number of values to generate .
2. The lambda parameter of the response variable which is a count
3. The specification of seed for easy reproduction of the data.

Count data sets were simulated under different conditions so as to get sets of count data with different characteristics by varying sample sizes and varying the lambda parameter of the response variable.The simulated response variable assumed a poisson distribution while the covariates assumed a normal distribution. The simulated sample sizes were 50,200,500 and 1000 while lambda was set at 0.97,5.97,10.97 and 15.97 The average AIC's were then compared to determine goodness-of-fit for the four different count data models.We further investigated how the models performed given different levels of dispersion.

0.3.5 Pre-exposure Prophylaxis

PrEP setting in Kenya The Pre-exposure prophylaxis (PreP) program in the country is embedded within HIV prevention intervention and is closely coordinated by NASCOP through the ministry of health. Kenya's PreP program algorithm is guided by WHO guidelines and involves key stakeholders like donor agencies (CDC and USAID), implementing partners and the government of Kenya. Data of all PreP indicators i.e. number of people screened for the purpose of determining their eligibility for PrEP, number of individuals screened whose sexual partners are HIV Positive, number of those starting PrEP for the first time ever, number of clients who started PrEP prior to current month and either came for refill or they had enough drugs to take them through the whole of this month, number of those who started PrEP prior to current month, stopped and are restarting PrEP this month, number of clients whose main reason for starting PrEP is recurrent use of PEP, among others is captured monthly though Ministry of Health (MOH) HIV prevention tools.

0.4 Data analysis and Results

0.4.1 Introduction

In this section, we will analyze both simulated and the real data. Random data sets with different characteristics will be generated and comparative analysis of the four models will be done to examine their goodness of fit.

0.4.2 Simulations Results

Different sample sizes of count data were simulated with sizes 50,200,500 and 1000 with the response variable assuming a Poisson distribution and the covariates assuming a normal distribution .Poisson, Negative Binomial, Quasi-Poisson and the COM-Poisson regression models were then fit into the data sets and their average AICs compared. The count data was simulated with varying lambda parameter of the response variable, the values were set at 0.97, 5.97,10.97 and 15.97. Regression was performed for each of the randomly generated data sets with the same covariates. The AICs of each of the four models (Poisson, Negative binomial, Quasi-Poisson and COM-Poisson) were then obtained. The mean and the variance of the simulated dependent variables were also noted so as to check for over dispersion.

From table 1, when lambda is 0.97 for a sample size of 50 in the simulated count data set, the average AIC for COM-Poisson model is 131.97 being the lowest while that for negative binomial and COM-Poisson were higher, with both having AIC of 135.42 and 133.42 respectively.The values of the mean(1.06) was higher than that of the variance(0.83), showing a case of under-dispersion, whereby COM-Poisson was a better fit. When lambda is 0.97 and the sample size is 200, the COM-Poisson is also a better fit for a case of over-dispersion. As the sample size increased and also as the value of lambda increased, the AIC values of the Poisson regression model were the least. On average, the Poisson model maintained AIC values that were lower than those of the Negative Binomial and COM-Poisson Regression models. This can be explained by the fact that the mean and variance values of the response variable were almost equal as the sample sizes and the lambda values increased , thus explaining the reason why Poisson model was a better fit. For AIC, the lower the value, the better the model fit. For a small sample size and small mean value of the response variable, the COM-Poisson model was a better fit and most

Table 1. AIC values for simulated data sets

When lambda=0.97

Sample sizes	50	200	500	1000
Poisson model	133.42	567.65	1260.6	2590
Negative Binomial	135.42	567.12	1262.6	2592
COM-Poisson	131.97	566.59	1262.53	2591
Mean(Y)	1.06	1.09	0.916	0.979
Variance(Y)	0.83	1.308	0.915	0.967

When lambda=5.97

Sample sizes	50	200	500	1000
Poisson model	236.7	914.52	2307.1	4632.1
Negative Binomial	238.64	916.52	2309.1	4633.6
COM-Poisson	238.69	916.29	2309.14	4633.92
Mean(Y)	5.8	5.84	5.98	6.029
Variance(Y)	6.94	6.095	6.052	6.268

When lambda=10.97

Sample sizes	50	200	500	1000
Poisson model	280.75	1059.1	2625.1	5230.9
Negative Binomial	282.36	1061.1	2626.8	5232.7
COM-Poisson	281.97	1061.104	2626.866	5232.707
Mean(Y)	11.2	11.45	10.802	10.806
Variance(Y)	13.878	11.41	11.293	11.083

When lambda=15.97

Sample sizes	50	200	500	1000
Poisson model	284.51	1127.6	2782.8	5582.8
Negative Binomial	286.51	1129.6	2784.8	5584.8
COM-Poisson	286.126	1129.547	2784.502	5584.6
Mean(Y)	16.14	16.265	15.662	15.881
Variance(Y)	15.143	15.844	15.378	15.669

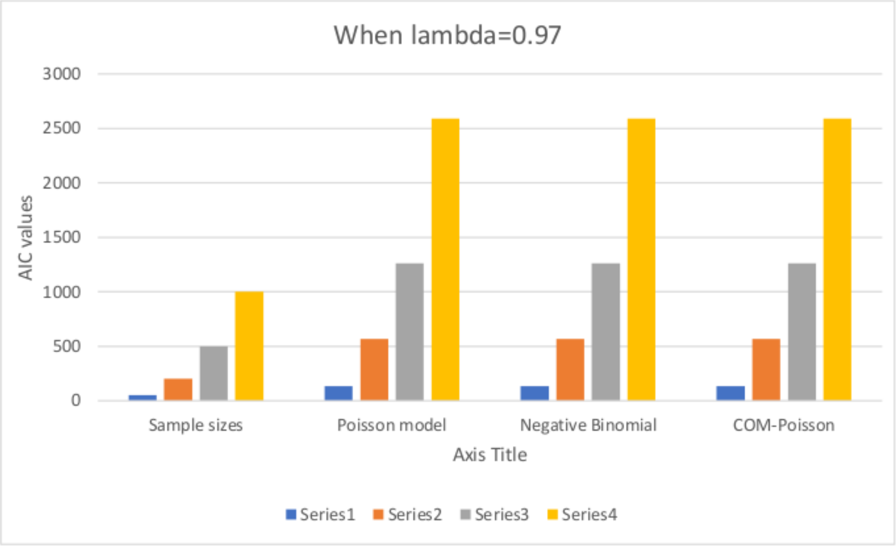

Figure 1. AIC values when lambda is 0.97

preferred compared to Poisson and Negative Binomial regression models while Poisson regression model was preferred for large counts. Therefore from these results,the interpretation is that for small counts, the COM-Poisson was a better fit while Poisson model performed better for large counts. On average the Poisson model was a better fit. These results were explained further by a graph for AIC values against the sample sizes.

It can be seen from the simulations results that, as the sample size increases, the AIC values also increase, implying that a smaller sample size is more preferred. Also, as the value of lambda for the response variable increases, there is notably an increase in the values of AIC, which can be interpreted that, the smaller the lambda, the better the model fit.

0.4.3 Real Data Analysis

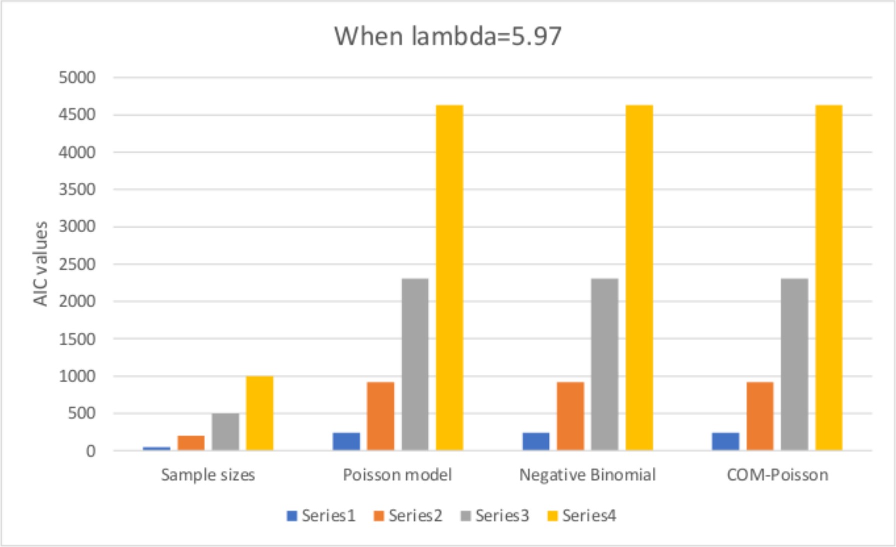

Figure 2. AIC values when lambda is 5.97

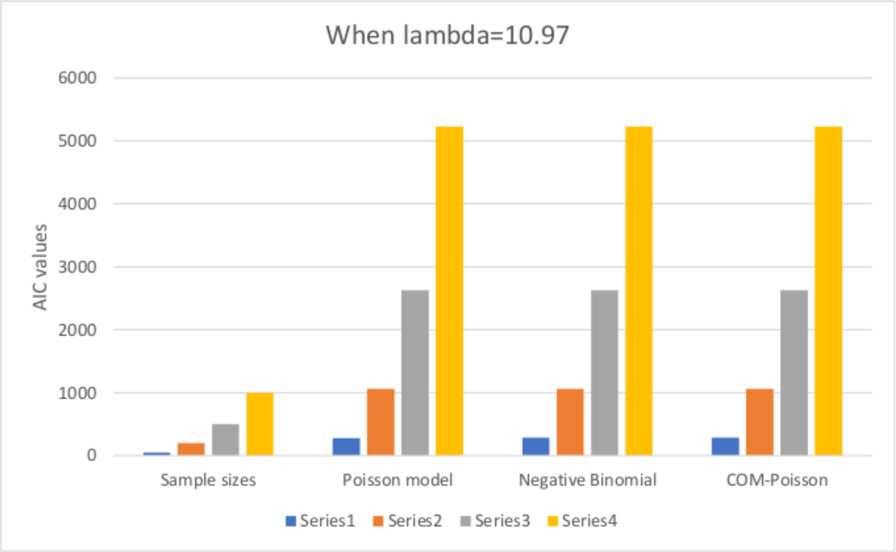

Figure 3. AIC values when lambda is 10.97

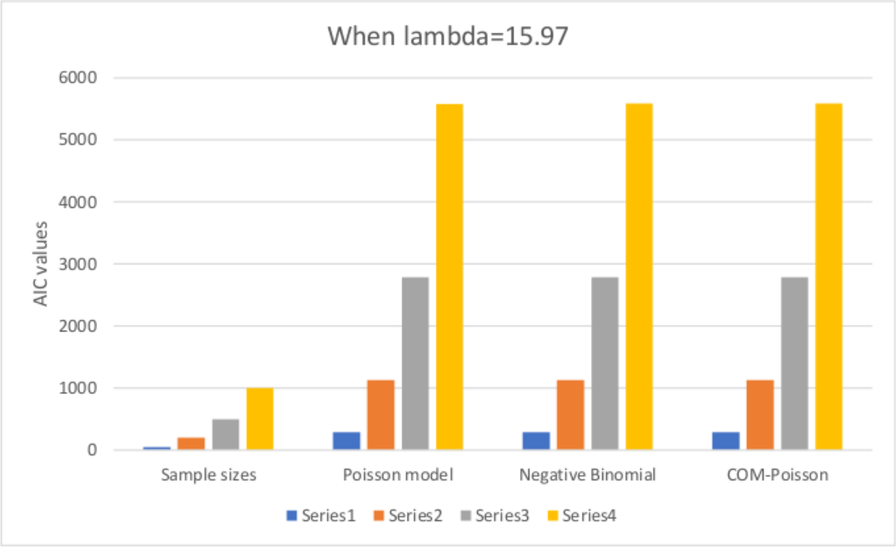

Figure 4. AIC values when lambda is 15.97

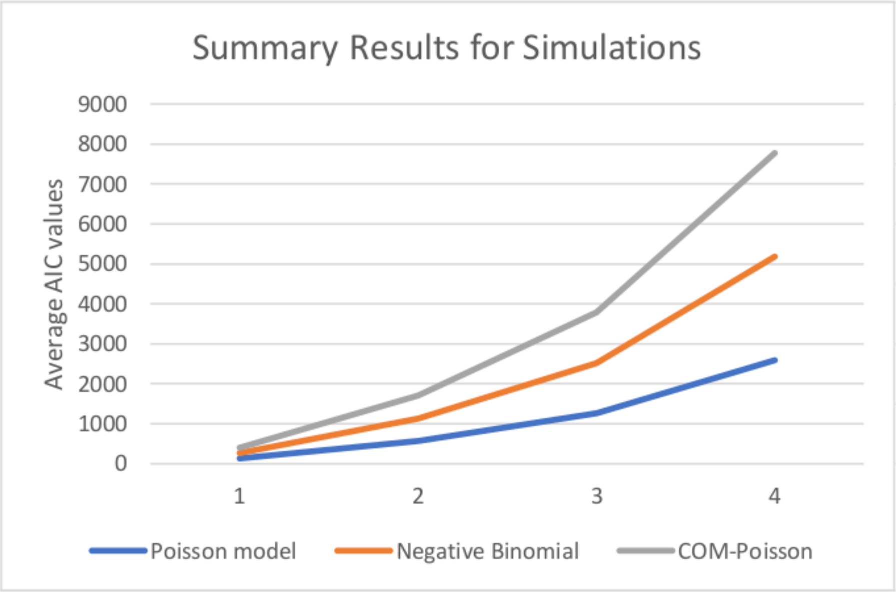

Figure 5. Summary of the average AIC values

Data description PreP data was retrieved from the Kenya Health Information System (KHIS) for a two month period covering March 2019 to April 2019. The number of health facilities were 104. The focus was level IV facilities which are offering PreP services. The 104 tier IV facilities are determined by MOH and some of the features of level IV are; has more comprehensive services and specialized staff than a level 3 hospital. Emergency, general and specialized operations are handled at the facility. Provides a highly active anti-retroviral therapy (HAART), antiretroviral (ARV) prophylaxis for children born of HIV-positive mothers, male circumcision, pelvic inflammatory disease (PID) management, and screening for animal transmitted conditions, the facility should be built on five-acre piece of land or an office space of approximately 2,500 square metres, have a 150- bed capacity for inpatient with 30 beds each for male, female, pediatric, antenatal and postnatal wards.etc.

PrEP data was analyzed using R-software and the four count data models,Poisson, Negative Binomial,Quasi-Poisson and COM-Poisson were used. The results from the four methods were compared using the Akaike Information Criterion (AIC) values. P-values below 0.05 are considered to be statistically significant. The response variable is count data of the number of people who tested HIV positive while on PrEP. The data had sixteen variables and in order to capture the relationship between the number of people who tested HIV positive while on PrEP and all the regressors included in the model, we fitted the basic Poisson Regression model as shown in table 2.

After performing a step wise regression analysis, Akaike Information Criteria was used for model selection and the model with the smallest AIC value was preferred. The covariates that were included in the model were;

1. The number of people screened; This is the number screened in the reporting month for the purpose of determining their eligibility for PrEP.
2. Number initiated; This is the number of those who were starting PrEP for the first time ever.
3. Tested HIV positive at month1 refill; This is a subset of those who tested for HIV at month1 refill.
4. Defaulters; These are clients who were expected to come for refill on the reporting month but they did not turn up.

Table 2. Saturated model

	Estimate	Std. Error	z value	P-value
Intercept	2.02963	12.56904	0.161	0.872
Screened	-0.42	1.57081	-0.269	0.788
eligible for prep	-0.04612	0.4729	-0.098	0.922
initiated	0.16403	0.47185	0.348	0.728
continuing	0.07684	0.48884	0.157	0.875
restarting prep	-0.08075	0.49273	-0.164	0.87
currently on prep	-	-	-	-
refilled at month 1	0.12124	0.3933	0.308	0.758
tested for HIV at month 1 refill	-	-	-	-
HIV+ at month 1 refill	-0.30671	1.11873	-0.274	0.784
refilled at month 3	0.16179	0.44659	0.362	0.717
tested for HIV at month 3 refill	-	-	-	-
HIV+ at month 3 refill	-0.20308	1.12431	-0.181	0.857
diagnoised with STIs while on prep	-0.06733	0.50654	-0.133	0.894
discontinued prep this month	-	-	-	-
recurrent use of prep	0.02938	0.42948	0.068	0.945
defaulters	0.3395	0.41344	0.821	0.412
AIC	236.93			

Table 3. Final model

	Estimate	Std.Error	Z value	P-value
Intercept	-0.7001	0.7369	-0.95	0.342
Screened	-0.1519	0.3073	-0.494	0.621
Initiated	0.1898	0.3765	0.504	0.614
HIV+ at month1	-0.4695	-0.4695	-0.559	0.401
Defaulters	0.3275	0.3565	0.919	0.358
AIC		221.31		

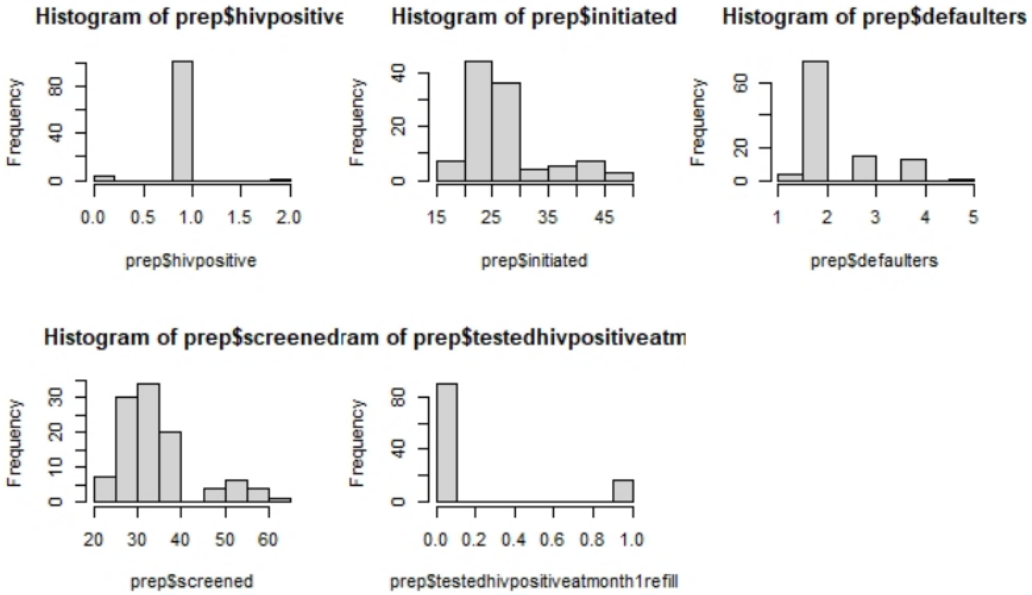

Figure 6. frequency distributions of the response and the explanatory variables

Table 4 gives a summary results of the Poisson, Negative binomial and COM-Poisson regression models. The Quasi-Poisson model, being semi-parametric does not generate AIC values and therefore its results were not included in the summary, it was a control model in the study.

From the Poisson regression model results, given the p-values, there was no significant association between the number of people who tested HIV positive while on PrEP and The number screened, the number initiated, number of people who tested HIV positive at month 1 refill and Defaulters. The Poisson model AIC value was lower than that of the Negative Binomial implying that Poisson model was a better fit compared to Negative binomial regression model.

From the Negative binomial and the COM-Poisson regression models results, the same results were observed with all the covariates included in the model having no significant association with the response variable. The AIC value for the COM-Poisson regression model was the lowest indicating that the best model for analyzing PrEP data in Kenya is the COM-Poisson Regression model.

Table 4. Summary results for the models

	Poisson model			Negative Binomial			COM-Poisson		
	Estimate	S.e	p-value	Estimate	S.e	p-value	Estimate	S.e	p-value
(Intercept)	-0.7	0.737	0.342	-0.7	0.737	0.342	-36.955	8699.82	0.997
Screened	-0.152	0.307	0.621	-0.152	0.307	0.621	-8.976	7565.15	0.999
Initiated	0.19	0.377	0.614	0.19	0.377	0.614	8.021	8875.11	0.999
HIV+at month1	-0.47	0.559	0.401	-0.47	0.559	0.401	8.065	NaN	NaN
Defaulters	0.328	0.357	0.358	0.328	0.357	0.358	75.3914	NaN	NaN
AIC values		221.31			223.31			12	

0.5 Discussion and Conclusion

Simulations

There are very few studies in literature evaluating and comparing performance of models using simulated data. Unlike in this study where our focus was to compare simulated data sets which vary in terms of sample size and the lambda parameter of the response variable, Muoka et al(2016) used simulated data sets with different proportions of zero;from their results Negative binomial model fitted better to over dispersed data with fewer zeroes while the Hurdle model fitted better in data sets with more zeroes.The few studies that used simulation studies majorly focused on zero-inflation which was not our main focus in this study.

From the simulation results in this study, the conclusion is that, the models that were most preferred had small sample sizes and small lambda parameter for the response variable. The COM-Poisson fitted well for under-dispersed data while the Poisson model was a better fit for large counts.

PrEP data

The main focus of this study is to compare the regression models; Poisson Regression Model, Quasi-Poisson Regression model Negative Binomial Regression Model, and Conway-Maxwell Poisson regression model in order to determine a better model which can be used in modeling HIV sero-conversion among PrEP users in Kenya. The criterion for selection of the best model used was Akaike Information Criterion (AIC) whereby the best model is the one that has the smallest AIC value. Based on the result on Table 7 above, the model with the smallest AIC is the COM-Poisson Regression model. Therefore, the best model for analyzing PrEP data in Kenya is the COM-Poisson Regression model.

Limitation of the study

One of the limitations in this study was lack of documented literature of studies that have used Generalized Linear Models for count data to model PrEP data. Also, the interpretation of the results generated from simulated data can only be limited to the conditions set in this study while simulating the data.

Bibliography

Baeten, J. M., Heffron, R., Kidoguchi, L., Mugo, N. R., Katabira, E., Bukusi, E. A., … others (2016). Integrated delivery of antiretroviral treatment and pre-exposure prophylaxis to hiv-1–serodiscordant couples: a prospective implementation study in kenya and uganda. *PLoS medicine*, *13*(8), e1002099.

Greene, W. H. (1994). Accounting for excess zeros and sample selection in poisson and negative binomial regression models.

Jiang, Y., & House, L. A. (2017). Comparison of the performance of count data models under different zero-inflation scenarios using simulation studies.

Johansson, A. (2014). *A comparison of regression models for count data in third party automobile insurance.*

Lee, J.-H., Han, G., Fulp, W., & Giuliano, A. (2012). Analysis of overdispersed count data: application to the human papillomavirus infection in men (him) study. *Epidemiology & Infection*, *140*(6), 1087–1094.

Lindsey, J. K., & Jones, B. (1998). Choosing among generalized linear models applied to medical data. *Statistics in medicine*, *17*(1), 59–68.

Maxwell, O., Mayowa, B. A., Chinedu, I. U., & Peace, A. E. (2018). Modelling count data; a generalized linear model framework. *Americal Journal of Mathematics and Statistics*, *8*(6), 179–183.

McCulloch, C. E. (2000). Generalized linear models. *Journal of the American Statistical Association*, *95*(452), 1320–1324.

Miller, J. M. (2007). *Comparing poisson, hurdle, and zip model fit under varying degrees of skew and zero-inflation* (Unpublished doctoral dissertation). University of Florida.

Muoka, A. K., Waititu, A., & Ngesa, O. O. (2016). Statistical models for count data.

National AIDS & STI Control Programme (NASCOP), M. o. H. (2017). *Framework for the implementation of pre-exposure prophylaxis of hiv in kenya.* NASCOP Nairobi, Kenya.

Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, *135*(3), 370–384.

on HIV/AIDS, J. U. N. P., et al. (2010). *Global report: Unaids report on the global aids epidemic 2010*. Unaids.

Organization, W. H., et al. (2012). *Guidance on pre-exposure oral prophylaxis (prep) for serodiscordant couples, men who have sex with men and transgender women at high risk of hiv in implementation research, annexes* (Tech. Rep.). World Health Organization.

Warton, D. I. (2005). Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data. *Environmetrics: The official journal of the International Environmetrics Society*, *16*(3), 275–289.

Zeileis, A., Kleiber, C., & Jackman, S. (2008). Regression models for count data in r. *Journal of statistical software*, *27*(8), 1–25.

Evaluating Generalized Linear Models For Count Data With Application To Pre-Exposure Prophylaxis HIV Sero-Conversion Data

ORIGINALITY REPORT

13%	%	9%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Cedar Valley College Student Paper	2%
2	Submitted to American Public University System Student Paper	1%
3	Submitted to Indian School of Business Student Paper	1%
4	Jong-Min Kim, Nak-Kyeong Kim, Yoonsung Jung, Sunghae Jun. "Patent data analysis using functional count data model", Soft Computing, 2018 Publication	1%
5	Atikur R. Khan, Christine M. O'Keefe. "Disclosure risk reduction for generalized linear model output in a remote analysis system", Data & Knowledge Engineering, 2017 Publication	1%
6	Submitted to Kwame Nkrumah University of Science and Technology	1%

	Student Paper	
7	Gizem ERKAN, Ozan EVKAYA, Semra TÜRKAN. "Determination of the Affecting Factors of the Number of Babies Born Alive in Multiple Pregnancies with Poisson Models", Turkiye Klinikleri Journal of Biostatistics, 2017 Publication	<1%
8	Imoto, Tomoaki. "A generalized Conway–Maxwell–Poisson distribution which includes the negative binomial distribution", Applied Mathematics and Computation, 2014. Publication	<1%
9	Giuliano Rizzardini, Dean L. Winslow. "Seroconversion on preexposure prophylaxis", AIDS, 2018 Publication	<1%
10	Submitted to uhasselt Student Paper	<1%
11	Submitted to University of Warwick Student Paper	<1%
12	J. K. Lindsey. "Choosing among generalized linear models applied to medical data", Statistics in Medicine, 01/15/1998 Publication	<1%
13	"Methods Based on Extensions of Generalized Linear Models", Springer Texts in Statistics,	<1%

	2003 Publication	
14	Submitted to IIT Delhi Student Paper	<1%
15	Ashenafi A. Yirga, Sileshi F. Melesse, Henry G. Mwambi, Dawit G. Ayele. "Negative binomial mixed models for analyzing longitudinal CD4 count data", Scientific Reports, 2020 Publication	<1%
16	Xiongqing Zhang, Yuancai Lei, Daoxiong Cai, Fengqiang Liu. "Predicting tree recruitment with negative binomial mixture models", Forest Ecology and Management, 2012 Publication	<1%
17	Elizabeth M. Irungu, Kenneth Ngure, Kenneth K. Mugwanya, Merceline Awuor et al. "“Now that PrEP is reducing the risk of transmission of HIV, why then do you still insist that we use condoms?” the condom quandary among PrEP users and health care providers in Kenya", AIDS Care, 2020 Publication	<1%
18	Charalampos Chanialidis, Ludger Evers, Tereza Neocleous, Agostino Nobile. "Efficient Bayesian inference for COM-Poisson regression models", Statistics and Computing, 2017 Publication	<1%

19	Robert E. Gaunt, Satish Iyengar, Adri B. Olde Daalhuis, Burcin Simsek. "An asymptotic expansion for the normalizing constant of the Conway–Maxwell–Poisson distribution", Annals of the Institute of Statistical Mathematics, 2017 Publication	<1%
20	Kayoung Park, Dongmin Jung, Jong-Min Kim. "Control charts based on randomized quantile residuals", Applied Stochastic Models in Business and Industry, 2020 Publication	<1%
21	Submitted to University of Derby Student Paper	<1%
22	Elizabeth M. Irungu, Jared M. Baeten. "PrEP rollout in Africa: status and opportunity", Nature Medicine, 2020 Publication	<1%
23	"Model Selection and Multimodel Inference", Springer Science and Business Media LLC, 2004 Publication	<1%
24	Submitted to essex Student Paper	<1%
25	Till Bärnighausen, Claudia Wallrauch, Alex Welte, Thomas A. McWalter et al. "HIV Incidence in Rural South Africa: Comparison of	<1%

	Estimates from Longitudinal Surveillance and Cross-Sectional cBED Assay Testing", PLoS ONE, 2008 Publication	
26	David N. Burns, Cynthia Grossman, Jim Turpin, Vanessa Elharrar, Fulvia Veronese. "Role of Oral Pre-exposure Prophylaxis (PrEP) in Current and Future HIV Prevention Strategies", Current HIV/AIDS Reports, 2014 Publication	<1%
27	Tony Barnett, Alan Whiteside. "AIDS in the Twenty-First Century", Springer Science and Business Media LLC, 2002 Publication	<1%
28	Submitted to University of Central Florida Student Paper	<1%
29	Submitted to University of York Student Paper	<1%

Exclude quotes	On	Exclude matches	< 20 words
Exclude bibliography	On		