Scientific Research Publishing

Open Journal of Statistics, 2019, 9, 664-685
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Comparative Assessment of Zero-Inflated Models with Application to HIV Exposed Infants Data

Faith Nekesa, Collins Odhiambo, Linda Chaba

Strathmore Institute of Mathematical Sciences, Strathmore University, Nairobi, Kenya
Email: codhiambo@strathmore.edu

How to cite this paper: Nekesa, F., Odhiambo, C. and Chaba, L. (2019) Comparative Assessment of Zero-Inflated Models with Application to HIV Exposed Infants Data. *Open Journal of Statistics*, **9**, 664-685.
https://doi.org/10.4236/ojs.2019.96043

Received: October 18, 2019
Accepted: December 10, 2019
Published: December 13, 2019

Abstract

In a typical Kenyan HIV clinical setting, there is a likelihood of registering many zeros during the routine monthly data collection of new HIV infections among HIV exposed infants (HEI). This is attributed to the implementation of the prevention of mother to child transmission (PMTCT) policies. However, even though the PMTCT policy is implemented uniformly across all public health facilities, implementation naturally differs from every facility due to differential health systems and infrastructure. This leads to structured zero among reported positive HEI (where PMTCT implementation is optimum) and non-structured zero among reported positive HEI (where PMTCT implementation is not optimum). Hence the classical zero-inflated and hurdle models that do not account for the abundance of structured and non-structured zeros in the data can give misleading results. The purpose of this study is to systematically compare performance of the various zero-inflated models with an application to HIV Exposed Infants (HEI) in the context of structured and unstructured zeros. We revisit zero-inflated, hurdle models, Poisson and negative binomial count models and conduct the simulations by varying sample size and levels of abundance zeros. Results from simulation study and real data analysis of exposed infant diagnosis show the negative binomial emerging as the best performing model when fitting data with both structured and non-structured zeros under various settings.

Keywords

Zero-Inflated Models, HIV Exposed Infants, Structured Zeroes, Mother-to-Child Transmission, Count Data

1. Introduction

Kenya has over the years implemented the World Health Organization (WHO) policy guidelines, in particular, prevention of mother-to-child transmission (PMTCT) policy, in an effort to mitigate sero-conversion among HIV exposed infants (HEI). The PMTCT policy includes averting transmission of HIV from mothers who live with HIV to their infants [1] [2] [3] [4] [5]. Research has shown that, majority of HIV sero-conversion among HEI occurred in the course of delivery, pregnancy or breastfeeding hence making the PMTCT policy a priority in the public health sector [6] [7] [8]. HEI sero-status can be HIV negative if deterrence of PMTCT is adhered effectively. The deterrence proportion with absence of PMTCT intervention is roughly between 15% and 45%; however with interventions, this has abridged to as low as 2% [3] [5]. Due to effective PMTCT intervention at different facilities, sero-conversion among HEI has reduced considerably [9] [10] [11] [12]; hence data collected is zero-inflated (ZI) and is therefore difficult to predict.

In Kenyan, HIV clinic setting, there is a likelihood of registering many zeros during routine monthly data collection of new HIV infections among HEI. This is attributed to implementation of PMTCT policies. Even though the PMTCT policy is affected uniformly across all public health facilities, implementation naturally differs at different facilities due to differential health systems and infrastructure. This leads to structured zero among reported positive HEI (where PMTCT implementation is optimum) and non-structured zero among reported positive HEI (where PMTCT implementation is sub-optimum). Failure and inadvertence of accommodating structured and non-structured zero-inflation may result in false inference [13]. Hence the classical ZI and hurdle models that do not account for the abundance of structured and non-structured zeros in data can give misleading results. Several rigorous and non-rigorous count data analysis approaches with zero inflation have been proposed by different researchers. These ZI models [13] and ZA models [14], also known as hurdle models are implemented to model extra zeros using logistics regression and count using count regression but they do not account for both structured and non-structured zeros [15]. Javali *et al.* [16] carried out a study whose aim was to determine factors associated with experience of dental caries. The dataset contained abundant zeros and was analyzed using ZI models. Results showed, the ZIP model performed well over conventional Poisson model. The ZINB also did well compare to the NB model. In conclusion, the ZINB model had performed well than the ZIP model when analyzing DMF count data. Akbarzadeh *et al.* [17] employed ZIP mixed models in evaluating hepatitis C's prognostic factors. Results showed, the mixed ZIP model was the best fit and was able to depict over dispersion, serial dependence, and zero-inflation in longitudinal setting. Also Francois *et al.* [18] compared the performances of Poisson, ZIP, NB, and ZINB models by fitting lesion count data. Results showed the NB and ZINB models are superior to the Poisson and ZIP models. The main objective of this study is to determine the best models to use when dealing with both structured and non-structured zeros

in a zero-inflation setting. To assess the different techniques when dealing with ZI and zero-altered (ZA) count data, observations from both the simulated data and empirical data are systematically analyzed. The ZI and ZA, NB and Poisson models are applied to HEI data and the performance assessed, to determine the most effective model. The next section looks at materials and method, followed by analysis and results section for both simulated and HEI data. We will then do discussions and draw a conclusion.

2. Materials and Methods

2.1. Summary of Zero-Inflated (ZI) and Zero-Altered (ZA) Count Data Models

2.1.1. Zero-Inflated Poisson (ZIP) Model

The ZIP model was introduced by Lambert [13] with reference to defects in a manufacturing process. The outcomes $\Phi = \left(\Phi_1, \Phi_2, \cdots, \Phi_n\right)^t$ are independent. A postulation of ZIP model is that observations 0 are given with probability p, and probability $(1-p)$, for Poisson (λ) variable is examined in Φ. The mathematical expectation and variance of ZIP is given as:

$$E\left(\Phi_i\right) = \left(1-p\right)\lambda_i. \tag{1}$$

$$Var\left(\Phi_i\right) = \left(1-p_i\right)\left(\lambda_1 + \lambda_2\right) - \left(\left(1-P_i\right)\lambda_i\right). \tag{2}$$

The Poisson mean vector $\lambda = \lambda_i, \cdots, \lambda_n$ has a Poisson canonical link $\log\left(\lambda\right)$ function and satisfies:

$$\log\left(\lambda\right) = \Theta_\beta. \tag{3}$$

$$\lambda = \exp\left\{\Theta_\beta\right\}.$$

For the parameter vector $p = \left(p_1, \cdots, p_n\right)$, the canonical logit link function is given by:

$$\text{logit} = \log\left(\frac{p}{1-p}\right) = \Lambda_\gamma. \tag{4}$$

Note that this distribution approaches to Poisson. The ZIP model has two components, one component is to model the probability of being the ordered/structured zeros p using the logistic regression and to model the Poisson mean μ as the other component. Thus, the presence of ordered/structural zeros gives rise not only to a more complex distribution, but also creates an additional link function for modeling the effect of explanatory variables for the occurrence of such zeros. In other words, the ZIP model enables us to better understand the effect of covariates by distinguishing the effects of each specific covariate on structural zeros and on the non-structural zeros.

2.1.2. Zero-Inflated Negative Binomial (ZINB) Model

The ZINB model [19] [20] [21] is defined as largely defined as the mixture distribution, with probability, p assigned to zero-inflation and probability $(1-p)$ assigned to the counts that follow NB distribution. The NB distribution is usual-

ly given in the form:

$$Pr(\Phi=\phi)=\frac{\gamma(\phi-\tau)}{\phi!\phi(\tau)}\left[\frac{\tau}{\tau+\lambda}\right]^{\tau}\left[\frac{\lambda}{\tau+\lambda}\right]^{\lambda},\phi=1,2,\cdots;\lambda,\tau>0. \quad (5)$$

where $\lambda=E(\Phi)$, τ is a shape parameter and Φ is the response variable. The variance of Φ is given as $\lambda+\frac{\lambda^2}{\tau}$. The distribution gets closer to the ZIP distribution and NB distribution as τ approaches inf, and p approaches 0, respectively.

2.1.3. Zero-Altered Poisson (ZAP) Regression

Also called the Poisson hurdle (PH). ZAP model has the hurdle part which models non-zero against the zero counts, and another part (Poisson count) that is utilised for the non-zero counts:

$$p(Y_i=0)=\rho_i \quad (6)$$

$$p(Y_i=k)=(1-\rho_i),k=0,1,\cdots. \quad (7)$$

where ρ_i models all zeros. The ZAP model does not categorize the zeros in the data as structured zeros or unstructured zeros. It overlooks on that concept which may bring about false interpretations of results and the study findings.

2.1.4. Zero-Altered Negative Binomial (ZANB) Regression

It is also known as the negative binomial logit hurdle (NBLH) [?]. Similarly, ZANB can be used in case of over-dispersion instead of applying the Poisson distribution. The ZANB model which is an extension of ZAP model also assumes the existence of the structured zeros and unstructured zeros. It overlooks on that concept hence may bring about false interpretations of results and the study findings.

2.2. Study Design and Setting

The PMTCT program in Kenya is coordinated by the National AIDS & STI Control Programme (NASCOP) through the government of Kenya. Kenya has been conducting exposed infant diagnosis (EID) among HEI since 2007. There has been increase of resources in particular, the year 2008-2009 when the guidelines were modified to include testing of all HEI. The EID testing for HEI algorithm from 2012 has been implemented as follows: maternal and EID-PCR testing is conducted during the first visit for all HEI with unknown HIV status aged-after stopping of breastfeeding. The NASCOP database covers all infants receiving EID-PCR testing in Kenya. Data from the NASCOP database is publicly available and can be viewed on a national dashboard (http://eid.nascop.org/).

2.2.1. Data Description

Data was extracted from the DHIS data for a period covering January 2016 to December 2017. The study focused on a representative of facilities reporting to DHIS. Data was collected from facility level and focused mainly on the three cities in Kenya; Nairobi, Kisumu, and Mombasa with selected health institutions

from the same. The specific items of measurements include; the health facility attended by the mother, the number of EID Positive, EID Testing Point, PCR Type, testing Point, HEI prophylaxis and the maternal Prophylaxis.

2.2.2. Study Population

From study sampling frame, a total of 413 samples were collected from HEI visiting 60 health facilities across the three cities in Kenya (Mombasa, Kisumu and Nairobi) between January 2016 and January 2017 and obtained PCR testing together with the results. HEI with missing age or greater than 2 years old were excluded from analysis. HEI with other missing predictor variables were also excluded in the study.

2.2.3. Ethical Approval

Data collected from this study is secondary and readily available from National AIDS and STI Control Programme (NASCOP) website. No patient identification information is included in NASCOP database. Furthermore, we also obtained ethical approval from Stathmore University Institutional Ethics review committee (*SUIERC*-0446/19).

2.2.4. Statistical Analysis

We conducted simulations to compare different models. To get the model which best fits the data, and is also a model with lower prediction error, stepwise regression for model selection was utilized. The models were also fitted to HEI data and comparison of the performance using AIC was used. Demographic information was summarized using descriptive statistics. The main outcome was the number of infants who turned HIV positive. We then examined the health facility attended by the HIV positive mothers paired to the HEI, the number of EID Positive, EID Testing Point, polymerase chain reaction (PCR) type, testing points, HEI prophylaxis and the maternal prophylaxis. From **Figure 1**, actual

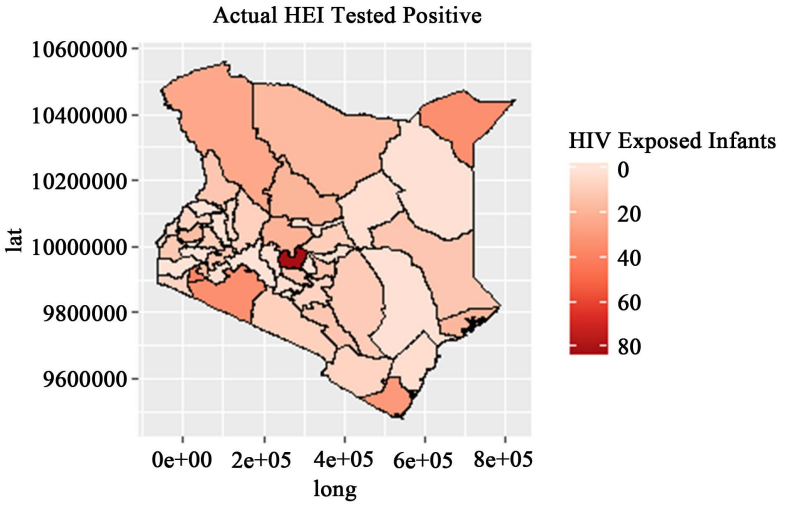


Figure 1. Spatial spread of HIV Exposed infants who were born by HIV positive mothers and turned positive across Kenya.

spatial data, Nairobi, Mombasa and Kisumu shows high numbers of HIV sero-conversion among HEI. Real HEI data was fitted to different ZI models. The most appropriate model based on AIC was used to determine covariates that were associated with the outcome of interest (EID positive). Analysis was conducted using R Studio version 3.5.3.

2.2.5. Simulations

Simulated data was created with unpredictable percentages of zeros and a fixed sample size of 500. A condition which has no zero-inflation ($\omega = 0.00$) will be tested and used as a standard comparison point. The effect of over-dispersion was observed in the non-zero part. The dispersion parameter k will be used with the following values: 1, 10, 50, and 100 which were pre-stipulated. These values represent a range of dispersion which is practical to aid in the assessment of the value of different models under study with varying distributions. The larger the value of k, the less dispersed the variable is and it approaches a Poisson distribution when $k > 10$. Negative binomial distribution was used to generate the response variable with different proportion of zeros added. Two covariates, X_1 and X_2, were also simulated. They were both assumed to come from a binomial distribution with $\mu = 4$ and 1 trial for X_1 and 10 trials for X_2.

2.2.6. Model Selection Criteria

To determine the best model, Akaike information criterion (AIC) was used. The model with minimum AIC was considered as the best model to fit the data [22]. AIC is given by:

$$\text{AIC} = -2\log L(\theta) + 2c, \tag{8}$$

where $L(\theta)$ is the maximized likelihood function for the estimated model and $-L(\theta)$ offers summary information on how much discrepancy exists between the model and the data, where c is the number of free parameters in the model.

3. Results

3.1. Simulation Results

The model with the lowest value of Akaike Information Criteria (AIC) depicted a more preferable model. Under the condition of non-zero inflation, ($\omega = 0.00$), the Poisson model was preferable under the dispersion parameter k = 10 since it had the lowest AIC value with a low dispersion (see **Table 1**). When k = 1, 50 and 100 under the same condition of no zero inflation, the negative binomial is the most preferred model since it had a lower AIC compared to the other models. When data exhibited 20% of zero inflation, ZIP model was most preferred at k = 10. When data exhibited 40% of zero inflation, the most preferred model was a negative binomial with a low dispersion of k = 1. When data exhibited 60% of zero inflation, the model with the lowest AIC was 173 ZIP with k = 100. With 80% of zeros, the best preferred model was ZAP when k = 1, 174 Poisson had the highest AIC value hence the least preferred among the models. Generally, ZAP

Table 1. Simulation design for the different Zero-Inflated Models adjusted for factors A and B.

Factor A:	Factor B:	Factor C:
ω	k	**Models tested on each condition**
0.00	1	Poisson regression model (Poisson)
0.20	10	Negative binomial regression model (NB)
0.40	50	Zero-inflated Poisson model (ZIP)
0.60	100	Zero -inflated negative binomial model (ZINB)
0.80		Zero -altered Poisson model (ZAP)
		Zero -altered negative binomial model (ZANB

had the lowest AIC value of 467.95 at 80% of zeros. This showed clearly; the most appropriate model when using simulation data from different setting. See graphical results at the appendix.

3.2. Results from Empirical Data Analysis

Descriptive Statistics for Variables

Descriptive statistics which include means, frequencies, and percentages for the variables of EID Positive, County, EID Testing Point, HEI prophylaxis and Maternal Prophylaxis is shown in Table 2. The median number of HEI positive recorded from the facilities was 0 (IQR = 0.13). 8.2% of the facilities sampled were from Kisumu county, 47.5% from Mombasa county and 44.3% from Nairobi county. Testing of HIV for exposed infants were mainly done when they were less than 2 months (33.2%) since early detection of HIV infection to the child could assist in early treatment and special care be given to the child. The HEI Prophylaxis mostly prescribed at the facilities for the infants was NVP + AZT (31.2%) and the least prescribed was NVP for 12 weeks (3.9%). For the case of maternal prophylaxis, the most prescribed ARV dose for the mothers was AZT + 3TC + ATV/r (15.3%) and the least prescribed as TDF + 3TC + DTG (0.2%).

3.3. Model Comparison Based on HEI Data

The HIV exposed infants data is fitted with the zero-inflated models which are; ZIP, ZAP, ZINB and ZANB. The performance of the inflated models will be compared using the AIC values. The results are presented below.

Four models described in methods section were used to fit the data which had a mixture of structured and non-structured zeros. The AIC values for the different models are presented in Table 3. The ZAP model had the lowest AIC value (490.81) indicating the best fit to the data and also works well when we have a mixture of both structured and non-structured zeros. ZINB model had the highest AIC value (492.11) indicating a poor fit for the model. Estimates of the regression coefficients and standard errors are presented separately for all the 4 models in Tables 4-7. To determine the significant covariates, the ZAP

Table 2. Descriptive statistics for HEI data.

Variables	Freg (%) or median (IQR)
No. of EID Positive	Median = 0, IQR = 0.13
County	
Kisumu	34 (8.2%)
Mombasa	196 (47.5%)
Nairobi	183 (44.3%)
EID Testing Point	
12 - 24 months	69 (16.7%)
2 - 9 months	69 (16.7%)
9 - 12 months	70 (16.9%)
Above 24 months	68 (16.5%)
less 2 months	137 (33.2%)
HEI prophylaxis	
AZT for 6 weeks + NVP for over 12 weeks	123(29.8%)
AZT for 6 weeks + NVP for 12 weeks	16 (3.9%)
NVP during BF	25 (6.1%)
NVP for 12 Wks	16 (3.9%)
NVP for 6 weeks (Mother on HAART or not BF)	30 (7.3%)
NVP + AZT	129 (31.2%)
Others	18 (4.4%)
Sd NVP + AZT + 3TC	39 (9.4%)
Sd NVP Only	17 (4.1%)
Maternal Prophylaxis	
AZT (From 14 wks or later) + sdNVP + 3TC + AZT + 3TC for 7 days	17 (4.1%)
AZT + 3TC + ATV/r	63 (15.3%)
AZT + 3TC + EFV	33 (8.0%)
AZT + 3TC + LPV/r	35 (8.5%)
AZT + 3TC + NVP	33 (8.0%)
Interrupted HAART (HAART until end of BF)	3 (0.7%)
SdNVP Only	5 (1.2%)
TDF + 3TC + ATV/r	33 (8.08%)
TDF + 3TC + DTG	1 (0.2%)
TDF + 3TC + EFV	124 (30.0%)
TDF + 3TC + LPV/r	33 (8.0%)
TDF + 3TC + NVP	33 (8.0%)

Table 3. Model fit comparison for HEI data. The best model fit for the HEI data is Hurdle Poisson and the worst model fit is ZINB.

Model	AIC value for HEI Data
Hurdle Poisson	490.81
Zero-Inflated Poisson	491.18
Hurdle Binomial	491.73
Zero-Inflated Negative Binomial	492.11

Table 4. AIC values for different variables in the poisson model.

	Df	Deviance	AIC
none		271.47	474.69
Testing Point	4	281.81	477.02
EID Testing Point	4	283.48	478.70
PCR Type	2	285.03	484.25
HEI prophylaxis	8	368.24	555.45
Maternal Prophylaxis	11	448.51	629.73

Table 5. Zero-inflated poisson model results.

Count model coefficients (poisson with log link):	Estimate	Std. Error	z value	p-value
Intercept	0.7160	0.5738	1.248	0.21210
EID Testing Point				
2 - 9 months	0.6228	0.2085	2.986	0.00282**
9 - 12 months	0.1734	0.2831	0.612	0.54027
12 - 24 months	−0.0601	0.3569	−0.168	0.86628
Above 24 months	0.2580	0.2530	1.019	0.30798
PCR Type				
Confirmatory PCR	−2.8241	1.1495	−2.457	0.01402*
2nd/3rd PCR	0.3652	0.5440	0.671	0.50204
Zero-inflation model coefficients (binomial with logit link):				
	Estimate	Std. Error	z value	p-value
(Intercept)	3.9251	0.7719	5.085	3.68e−07***
EID Testing Point				
2 - 9 months	−0.7319	0.4359	−1.679	0.09311 .
9 - 12 months	0.1948	0.5275	0.369	0.71198
12 - 24 months	0.2892	0.5635	0.513	0.60772
Above 24 months	-0.2998	0.4730	−0.634	0.52621
PCR Type				
Confirmatory PCR	−10.5797	82.6014	−0.128	0.89808
2nd/3rd PCR	−2.0356	0.7455	−2.731	0.00632**

Table 6. Zero-inflated Negative Binomial outcomes.

	Estimate	Std. Error	z value	p-value
(Intercept)	0.68653	0.63718	1.077	0.28128
EID Testing Point				
2 - 9 months	0.63452	0.23899	2.655	0.00793**
9 - 12 months	0.17165	0.31957	0.537	0.59118
12 - 24 months	−0.06536	0.39827	−0.164	0.86965
Above 24 months	0.26338	0.28738	0.916	0.35941
PCR Type				
Confirmatory PCR	−2.79575	1.18103	−2.367	0.01792*
2nd/3rd PCR	0.37511	0.60213	0.623	0.53330
Log(theta)	2.59197	1.19069	2.177	0.02949*
Zero-inflation model coefficients (binomial with logit link):				
	Estimate	Std. Error	z value	p-value
(Intercept)	3.8921	0.7788	4.998	5.8e−07***
EID Testing Point				
2 - 9 months	−0.7150	0.4382	−1.632	0.10273
9 - 12 months	0.1998	0.5301	0.377	0.70630
12 - 24 months	0.2866	0.5680	0.505	0.61384
Above 24 months	−0.2927	0.4757	−0.615	0.53840
PCR Type				
Confirmatory PCR	−19.2306	6342.1830	−0.003	0.99758
2nd/3rd PCR	−2.0276	0.7507	−2.701	0.00691**

Table 7. Negative binomial hurdle outcome.

Count model coefficients (truncated negbin with log link):				
	Estimate	Std. Error	z value	p-value
(Intercept)	0.63862	0.63717	1.002	0.31621
EID Testing Point				
2 - 9 months	0.65677	0.24183	2.716	0.00661**
9 - 12 months	0.20880	0.32211	0.648	0.51685
12 - 24 months	−0.05442	0.39480	−0.138	0.89036
Above 24 months	0.28004	0.28758	0.974	0.33017
PCR Type				
Confirmatory PCR	−9.80622	133.47361	−0.073	0.94143
2nd/3rd PCR	0.40607	0.60548	0.671	0.50244
Log(theta)	2.58135	1.18422	2.180	0.02927*

Continued

Zero hurdle model coefficients (binomial with logit link):				
	Estimate	Std. Error	z value	p-value
(Intercept)	−4.0023	0.7556	-5.297	1.18e−07***
EID Testing Point				
2 - 9 months	0.7273	0.4259	1.708	0.08768 .
9 - 12 months	−0.2469	0.5167	−0.478	0.63274
12 - 24 months	−0.3400	0.5508	−0.617	0.53702
Above 24 months	0.2962	0.4643	0.638	0.52350
PCR Type				
Confirmatory PCR	2.1431	1.3091	1.637	0.10161
2nd/3rd PCR	2.0897	0.7344	2.845	0.00444**

model is used since its the best model based on the AIC value. The hurdle poisson (ZAP model), using the step wise model selection dropped the insignificant variables and was left with EID Testing Point and PCR Type. The baseline odds of having a positive count verses zero is 2.12 (exp(0.7556)). This odds is increased by 3.7 (exp(2.0897)) times if 2nd/3rd PCR test is done as compared to the initial test. EID Testing point does not have significant effect. Given the response is positive, the average count is 1.97 (exp(0.67656)). This is increased by 1.9 (exp(0.63982)) times if testing is done at 2 - 9 months compared to less than 2 months. PCR Type does not have significant effect. Similar interpretation can be made for the rest of the models.

Significant covariates in this model are; PCR type 2nd/3rd with p-value 0.003685, HEI Prophylaxis of NVP during BF with p-value 0.0000000211, NVP for 6 weeks (Mother on HAART or not BF) with p-value 7.74e−11, others with p-value 0.001331 and Sd NVP Only with p-value 3.33e−06. Maternal prophylaxis-AZT, 3TC, ATV/r with p-value 1.46e−08; AZT, 3TC, EFV with p-value 0.000259; TDF, 3TC, ATV/r with p-value 0.000345; TDF, 3TC, EFV with p-value 8.49e−15; TDF, 3TC, LPV/r with p-value 0.00000766; TDF, 3TC, NVP with p-value 0.000763.

For the Poisson model (referred to as model 1 in the analysis), using the stepwise model selection criteria dropped the variables that were not significant (county) and the variables that remained included EID Testing Point, PCR Type, Testing Point, HEI prophylaxis and Maternal Prophylaxis. In the EID Testing Point, the significant testing point is between 2 - 9 months which shows that the risk is 2.39 times higher for HEI between 2 - 9 months compared to testing between 0 - 2 months. For the PCR Type, the chain reaction which was significant was that of 2nd/3rd PCR hence it indicates that a HEI is 2.9 times more likely to detect the HIV virus compared to the initial PCR. In the HEI prophylaxis with comparison to using AZT for 6 weeks + NVP for over 12 weeks; the risk of using

Nevirapine during breastfeeding on HEI is 5 times higher, the risk of using nevirapine for 6 weeks (mother not breastfeeding) is on the HEI is 6.5 times higher, the risk of using other drugs is 3.2 times high, the risk of using a combination of Sd NVP + AZT + 3TC is 4.6 times high and lastly the risk of using Sd NVP only is 3.4 times high. Under the Maternal Prophylaxis, in comparison to the use of AZT (From 14 wks or later) + Sd NVP + 3TC + AZT + 3TC for 7 days the risk of using a combination of AZT + 3T + EFV (Efavirenz, which is a capsule and taken by mouth with plenty of water) by the mother to the infant is 5.6 times less, then the risk of using combination of AZT + 3TC + LPV/r (Lopinavir/Ritonavir, which come in tablet forms) is 3.6 times less, the risk of using a combination of TDF + 3TC + ATV/r is 3.5 times less, the risk of using a combination of TDF + 3TC + LPV/r is 3.9 times less and lastly the risk of using a combination of TDF + 3TC + NVP is 3.3 times lesser. The AIC value after fitting the Poisson model is 474.69, which is the second best fitting model for the EID data.

Negative binomial model (referred to as model 2 in analysis), had the AIC value of 429.19 which had the lowest AIC value hence it was considered as the best model. Using the step wise model selection, the following variables which were considered significant and had an effect on the final AIC value were retained; EID Testing Point, PCR Type, Testing Point, HEI prophylaxis and Maternal Prophylaxis. The PCR type that was significant was that of 2nd/3rd PCR type, which indicates that a HEI is 2 times more likely to detect the HIV virus in comparison to the initial PCR. In the HEI prophylaxis with comparison to using AZT for 6 weeks + NVP for over 12 weeks; the risk of using NVP during breastfeeding on HEI is 4.2 times higher, then the risk of using nevirapine for 6 weeks (mother not breastfeeding) is on the HEI is 5.4 times higher, the risk of using other drugs is 2.3 times high, the risk of using a combination of Sd NVP + AZT + 3TC is 3.3 times high and lastly the risk of using Sd NVP only is 2.4 times high according to the results above. Under the Maternal Prophylaxis, in comparison to the use of AZT (From 14 wks or later) + Sd NVP + 3TC + AZT + 3TC for 7 days, the risk of using a combination of AZT + 3T + ATV/r by the mother to the infant is 2.4 times higher, then the risk of using combination of AZT + 3TC + LPV/r (Lopinavir/Ritonavir, which come in tablet forms) is 3.1 times less, the risk of using a combination of TDF + 3TC + ATV/r is 3.4 times less, the risk of using a combination of TDF + 3TC + EFV is 5.9 times lesser, then the risk of using a combination of TDF + 3TC + LPV/r is 3.4 times less and lastly the risk of using a combination of TDF + 3TC + NVP is 3 times lesser.

In the ZIP model, fitting the data using all the variables and using stepwise model selection dropped most of the models and retained EID Testing Point and PCR Type which were the significant variables. Under the EID Testing Point, the risk of testing the infant between 2 - 9 months is 2.9 times higher to testing between 0 - 2 months. For the PCR Type in comparison to the initial PCR, it indicates that the HEI is 2 times less likely to detect the HIV virus during the Confirmatory PCR. Analyzing the model with the 2 variables gave an AIC value of 491.18.

The ZINB model using the stepwise regression, and the direction as backward dropped most of the variables that were not significant and was left with 2 variables which were EID Testing Point and PCR Type. Under the EID Testing Point, the risk of testing the infant between 2 - 9 months is 2.6 times higher to testing between 0 - 2 months. For the PCR Type in comparison to the initial PCR, it indicates that the HEI is 2.7 times less likely to detect the HIV virus during the Confirmatory PCR. The AIC value for the ZINB model is 492.11, hence regarded as the worst model fit for the data.

In the hurdle binomial (ZANB), using the stepwise regression also dropped down the insignificant variables and was left with EID Testing Point and PCR Type. In the EID Testing Point, the risk of testing the infant between 2 - 9 months is 2.65 times higher to testing between 0 - 2 months. For the PCR Type in comparison to the initial PCR, it indicates that the HEI is 2.7 times less likely to detect the HIV virus during the Confirmatory PCR. The AIC value using the 2 variables was 491.73, which is the 2nd worst model fit for the data hence not preferred.

Count data with high number of zeros are commonly registered in medical research and public health particularly, monthly number of HEI. Yip [23] and Lambert [13] proposed ZI Poisson distribution and Heilbron [24] utilised ZAP and NB distributions to model ZI data. Li *et al.* [25] derived a multivariate version of ZIP model and used it to analyse equipment problems in processing of electronics. Although different authors have widely used zero-inflated distributions, there is no practical study that systematically compares zero-inflated outcomes in HIV exposed infant settings. Because the ZI model involves state parameter k and parameter ρ, we extensively conducted simulations by varying percentages of zeros and these parameters. The results of simulation show that ZAP generally had the lowest AIC value, when the percentage of zero was high. This is consistent with the results from the application data because the percentage of zeros in the HEI dependent variable is 88% (see **Table 8**). The simulation procedure selected limited important model terms to maximize the ZI likelihood functions. In all these ZI models, EID testing point and PCR type were statistically significant. Based on HEI data analysis, the proportion of HIV sero-conversion was high for EID tested between 2 - 9 months compared to those tested earlier. The patient outcomes of studies done recently showed sero-status was not different between boys and girls [3] [7] [9]. This was however, not verifiable in our data because we did not collect gender covariate. There are several studies that have attempted to implement ZI model extensions in order to accommodate unstructured effects *i.e.* [15] [24] but not in a public health setting where government policies are not implemented uniformly. In the ZIP model, fitting the data using all the variables and using stepwise model selection dropped most of the models and retained EID Testing Point and PCR type which were the significant variables. Under the EID Testing Point, the risk of testing the infant between 2 - 9 months is 2.9 times higher to testing between 0 - 2 months. For the PCR type in comparison to the initial PCR, it indicates that

Table 8. Hurdle (ZAP) model outcomes. Significant covariates in this model are; EID Testing Point of between 2 - 9 months with p-value 0.00235 and PCR Type 2nd/3rd with p-value 0.00444.

Count model coefficients (Poisson with log link):	Estimate	Std. Error	z value	p-value
(Intercept)	0.67656	0.57662	1.173	0.24066
EID Testing Point				
2 - 9 months	0.63982	0.21034	3.042	0.00235**
9 - 12 months	0.20265	0.28420	0.713	0.47582
12 - 24 months	−0.05207	0.35466	−0.147	0.88328
Above 24 months	0.27111	0.25325	1.071	0.28438
PCR Type				
Confirmatory PCR	−9.26276	103.50896	−0.089	0.92869
2nd/3rd PCR	0.39130	0.54995	0.712	0.47677
Zero hurdle model coefficients (binomial with logit link)				:
	Estimate	Std. Error	z value	p-value
(Intercept)	−4.0023	0.7556	-5.297	1.18e−07***
EID Testing Point				
2 - 9 months	0.7273	0.4259	1.708	0.08768
9 - 12 months	−0.2469	0.5167	−0.478	0.63274
12 - 24 months	−0.3400	0.5508	−0.617	0.53702
Above 24 months	0.2962	0.4643	0.638	0.52350
PCR Type				
Confirmatory PCR	2.1431	1.3091	1.637	0.10161
2nd/3rd PCR	2.0897	0.7344	2.845	0.00444**

the HEI is 2 times less likely to detect the HIV virus during the Confirmatory PCR. Analyzing the model with the 2 variables gave an AIC value of 491.18. This was similar to the mixed ZIP model introduced by Miller [26]. For Poisson model (referred to as model 1 in the analysis), using the step-wise model selection criteria dropped the variables that were not significant (county) and the variables that remained included in the EID Testing Point, PCR Type, Testing Point, HEI prophylaxis and Maternal Prophylaxis. In the EID Testing Point, the significant testing point is between 2 - 9 months which shows that the risk is 2.39 times higher for HEI between 2 - 9 months compared to testing between 0 - 2 months. For the PCR type, the chain reaction which was significant was that of 2nd/3rd PCR hence it indicates that a HEI is 2.9 times likely to detect HIV positive result when compared to initial PCR. In the HEI prophylaxis, in comparison to using AZT for the first 6 weeks plus NVP for over 12 weeks. The risk of using nevirapine during breastfeeding on HEI is 5 times higher. The risk of using nevirapine for 6 weeks (mother not breastfeeding) on the HEI is 6.5 times higher

while the risk of using other drugs is 3.2 times high. The risk of using a combination of April 17, 2019 10/17 Sd NVP + AZT + 3TC is 4.6 times high and lastly the risk of using Sd NVP only is 3.4 times high. Under the Maternal Prophylaxis, in comparison to the use of AZT (From 14wks or later) + Sd NVP + 3TC + AZT + 3TC for 7 days, the risk of using a combination of AZT + 3T + EFV (Efavirenz, which is a capsule and taken by mouth with plenty of water) by the mother to the infant is 5.6 times less, than the risk of using a combination of AZT + 3TC + LPV/r (Lopinavir/Ritonavir, which come in tablet forms) is 3.6 times less, the risk of using a combination of TDF + 3TC + ATV/r is 3.5 times less; the risk of using a combination of TDF + 3TC + LPV/r is 3.9 times less and lastly the risk of using a combination of TDF + 3TC + NVP is 3.3 times less. The AIC value after fitting the Poisson model is 474.69, which yield the second appropriate model for fitting EID data. The NB model (referred to as model 2 in analysis), had the AIC value of 429.19 which had the lowest AIC value hence it was considered the best model. Using the stepwise model selection, the following variables which were considered significant and had an effect on the final AIC value were retained; EID Testing Point, PCR Type, Testing Point, HEI prophylaxis and Maternal Prophylaxis. The PCR type that was significant was that of 2nd/3rd PCR type, which indicates that a HEI is 2 times more likely to detect the HIV virus in comparison to the initial PCR. In the HEI prophylaxis with comparison to using AZT for first 6 weeks plus NVP for over 12 weeks. The risk of using NVP during breastfeeding on HEI is 4.2 times higher, then the risk of using NVP for 6 weeks (mother not breastfeeding) on the HEI is 5.4 times higher. The risk of using other drugs is 2.3 times high, the risk of using a combination of Sd NVP + AZT + 3TC is 3.3 times high and lastly the risk of using Sd NVP only is 2.4 times high according to the results above. Under the Maternal Prophylaxis, in comparison to the use of AZT (From 4 weeks or later) + Sd NVP + 3TC + AZT + 3TC for 7 days, the risk of using a combination of AZT + 3T + ATV/r by the mother to the infant is 2.4 times higher, then the risk of using combination of AZT + 3TC + LPV/r (Lopinavir/Ritonavir, which come in tablet form) is 3.1 times less, the risk of using a combination of TDF + 3TC + ATV/r is 3.4 times less, the risk of using a combination of TDF + 3TC + EFV is 5.9 times less, then the risk of using a combination of TDF + 3TC + LPV/r is 3.4 times less and lastly the risk of using a combination of TDF + 3TC + NVP is 3 times less. The ZINB model using the stepwise regression, and the direction as backward dropped most of the variables that were not significant and was left with 2 variables which were EID Testing Points and PCR Type. Under the EID Testing Point, the risk of testing the infant between 2 - 9 months is 2.6 times higher to testing between 0 - 2 months. For the PCR Type, in comparison to the initial PCR, it indicates that the HEI is 2.7 times less likely to detect the HIV virus during the Confirmatory PCR. The AIC value for the ZINB model is 492.11, hence regarded as the worst model fit for the data. In the hurdle binomial (ZANB), using the stepwise regression also dropped down the insignificant variables and was left with EID Testing Point and PCR Type. In the EID Testing Point, the risk of testing the infant between 2 - 9

months is 2.65 times higher in comparison to testing between 0 - 2 months. For the PCR Type in comparison to the initial PCR, it indicates that the HEI is 2.7 times less likely to detect the HIV virus during the Confirmatory PCR. The AIC value using the 2 variables was 491.73, which is the 2nd worst model fit for the data hence not preferred. ZI models and zero-altered models give almost similar results as shown from both simulated data and the HEI data. The decision when choosing between these two according to the study, heavily relied on the AIC value found after the analysis of the work. Failure to account for the zero-inflation while analyzing such data may result in false inferences. After the simulation study and analysis of EID data, the negative binomial emerges as the gold-standard for us fitting the data with both structured and non-structured zeros.

4. Conclusion

Simulation results offer a general idea as to which model is most appropriate; however, more conditions will need to be examined to get a more accurate relationship between the model selection and different levels whether structured or unstructured zero-inflation. One of the limitations of the study is that, predictive variables for both zero and other count data models were considered the same. One area for further research is the issue of imbalanced covariates with missing data.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Mahy, M., Stover, J., Kiragu, K., Hayashi, C., Akwara, P., Luo, C., Stanecki, K., Ekpini, K. and Shaffer, N. (2010) What Will It Take to Achieve Virtual Elimination of Mother to Child Transmission of HIV? An Assessment of Current Progress and Future Needs. *Journal of Sexually Transmission Infection*, **86**, ii48-ii55. https://doi.org/10.1136/sti.2010.045989

[2] Ministry of Health, Nairobi, Kenya (2012) Kenya AIDs Indicator Survey: National AIDs and STI Control Programme.

[3] UNAIDS (2012) Joint United Nations Programme on HIV/AIDs, Global Report UNAIDS Report on Global AIDs Epidemic. Geneva.

[4] UNICEF New York (2015) United Nation Children's Fund: Towards an AIDs-Free Generation-Children and AIDs, Stock Taking Report.

[5] WHO (2012) World Health Organization: Programmatic Update Use of Antiretroviral Drugs for Treating Pregnant Women and Preventing HIV Infection in Infants.

[6] Essomo, M., Meye, J.F., Belembaogo, E., Engoghan, E. and Ondo, A. (2008) Prevention of Mother-to-Child Transmission of HIV in Gabon: The Problem of Children Lost to Follow-Up.

[7] Asefa, A. and Mitike, G. (2014) Prevention of Mother-to-Child Transmission (PMTCT) of HIV Services in Adama Town: BMC Pregnancy and Childbirth.

[8] Fewtrell, M.S., Kennedy, K., Singhal, A., Martin, R.M., Ness, A. and Hadders-Algra, M. (2008) How Much Loss to Follow-Up Is Acceptable in Long-Term Randomized Trials and Prospective Studies? *Archives of Disease in Childhood*, **93**, 458-461. https://doi.org/10.1136/adc.2007.127316

[9] Cook, R., Ciampa, P., Sidat, M., Blevin, M., Burlison, J., Davidson, M.A., Arroz, J.A., Vergara, A.E., Vermund, S.H. and Moon, T.D. (2011) Predictors of Successful Early Infant Diagnosis of HIV in Rural District Hospital in Zambezia Mozambique. *Journal Acquired Immune Deficiency Syndromes*, **56**, e104-e109. https://doi.org/10.1097/QAI.0b013e318207a535

[10] Gourlay, A.A., Birdthistle, I., Mburu, G., Iorpenda, K. and Wringe, A. (2013) Barriers and Facilitating Factors to the Uptake of Antiretroviral Drugs for Prevention of Mother-to-Child Transmission of HIV in Sub-Saharan Africa: A Systematic Review. *Journal of the International AIDS Society*, **16**, 18588. https://doi.org/10.7448/IAS.16.1.18588

[11] Le Coeur, S., Kanshana, S. and Jourdain, G. (2003) HIV-1 Transmission from Mother to Child and Its Prevention. Médecine Tropicale. Revue du Corps de Santé Colonial.

[12] Le May, A. and Holmes, S. (2012) Introduction to Nursing Research. Hodder Arnold, London. https://doi.org/10.1201/b13215

[13] Lambert, D. (1992) Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. *Technometrics*, **34**, 1-14. https://doi.org/10.2307/1269547

[14] Mullahy, J. (1986) Specification and Testing of Some Modified Count Data Models. *Journal of Econometrics*, **33**, 341-365. https://doi.org/10.1016/0304-4076(86)90002-3

[15] Hu, M., Pavlicova, M. and Nunes, E. (2011) Zero-Inflated and Hurdle Models of Count Data with Extra Zeros: Examples from an HIV-Risk Reduction Intervention Trial. *The American Journal of Drug and Alcohol Abuse*, **37**, 367-375. https://doi.org/10.3109/00952990.2011.597280

[16] Moses, J., Rangeeth, B.N. and Gurunathan, D. (2011) Prevalence of Dental Caries, Socio-Economic Status and Treatment Needs among 5 to 15 Year Old School Going Children of Chidambaram. *Journal of Clinical and Diagnostic Research*, **5**, 146-151.

[17] Akbarzadeh Baghban, A., Pourhoseingholi, A., Zayeri, F., Jafari, A.A. and Alavian, S.M. (2013) Application of Zero-Inflated Poisson Mixed Models in Prognostic Factors of Hepatitis C. *BioMed Research International*, **2013**, Article ID: 403151. https://doi.org/10.1155/2013/403151

[18] Francois, M., Peter, C. and Gordon, F. (2012) Dealing with Excess of Zeros in the Statistical Analysis of Magnetic Resonance Imaging Lesion Count in Multiple Sclerosis. *Pharmaceutical Statistics*, **11**, 417-424. https://doi.org/10.1002/pst.1529

[19] Min, Y. and Agresti, A. (2005) Random Effect Models for Repeated Measures of Zero-Inflated Count Data. *Statistical Modelling*, **5**, 1-19. https://doi.org/10.1191/1471082X05st084oa

[20] Loeys, T., Moerkerke, B., Smet, O.D., *et al*. (2012) The Analysis of Zero-Inflated Count Data: Beyond Zero Inflated Poisson Regression. *British Journal of Mathematical and Statistical Psychology*, **65**, 163-180. https://doi.org/10.1111/j.2044-8317.2011.02031.x

[21] Sileshi, G., Hailu, G. and Nyadzi, G.I. (2009) Traditional Occupancy-Abundance Models Are Inadequate for Zero-Inflated Ecological Count Data. *Ecological Modelling*, **220**, 1764-1775. https://doi.org/10.1016/j.ecolmodel.2009.03.024

[22] Akaike, H. (1973) Information Theory and an Extension of the Maximum Likelih-

ood Principle. In: Petrov, B.N. and Csaki, F., Eds., *Proceedings of the* 2*nd International Symposium on Information Theory*, Akademiai Kiado, Budapest, 267-281.

[23] Yip, P. (1988) Inference about the Mean of a Poisson Distribution in the Presence of a Nuisance Parameter. *Australian Journal of Statistics*, **30**, 299-306. https://doi.org/10.1111/j.1467-842X.1988.tb00624.x

[24] Heilbron, D.C. (1994) Zero-Altered and Other Regression Models for Count Data with Added Zeros. *Journal of Mathematical Methods in Biosciences*, **36**, 531-547. https://doi.org/10.1002/bimj.4710360505

[25] Li, C.S., Lu, J.C., Park, J., Kim, K.M., Brinkley, P.A. and Peterson, J. (1999) A Multivariate Zero-Inflated Poisson Distribution and Its Inferences. *Technometrics*, **41**, 29-38. https://doi.org/10.1080/00401706.1999.10485593

[26] Miller, J.M. (2007) Comparing Poisson, Hurdle and ZIP Model Fit under Varying Degrees of Skew and Zero-Inflation. PhD Thesis, University of Florida, Gainesville.

Appendix

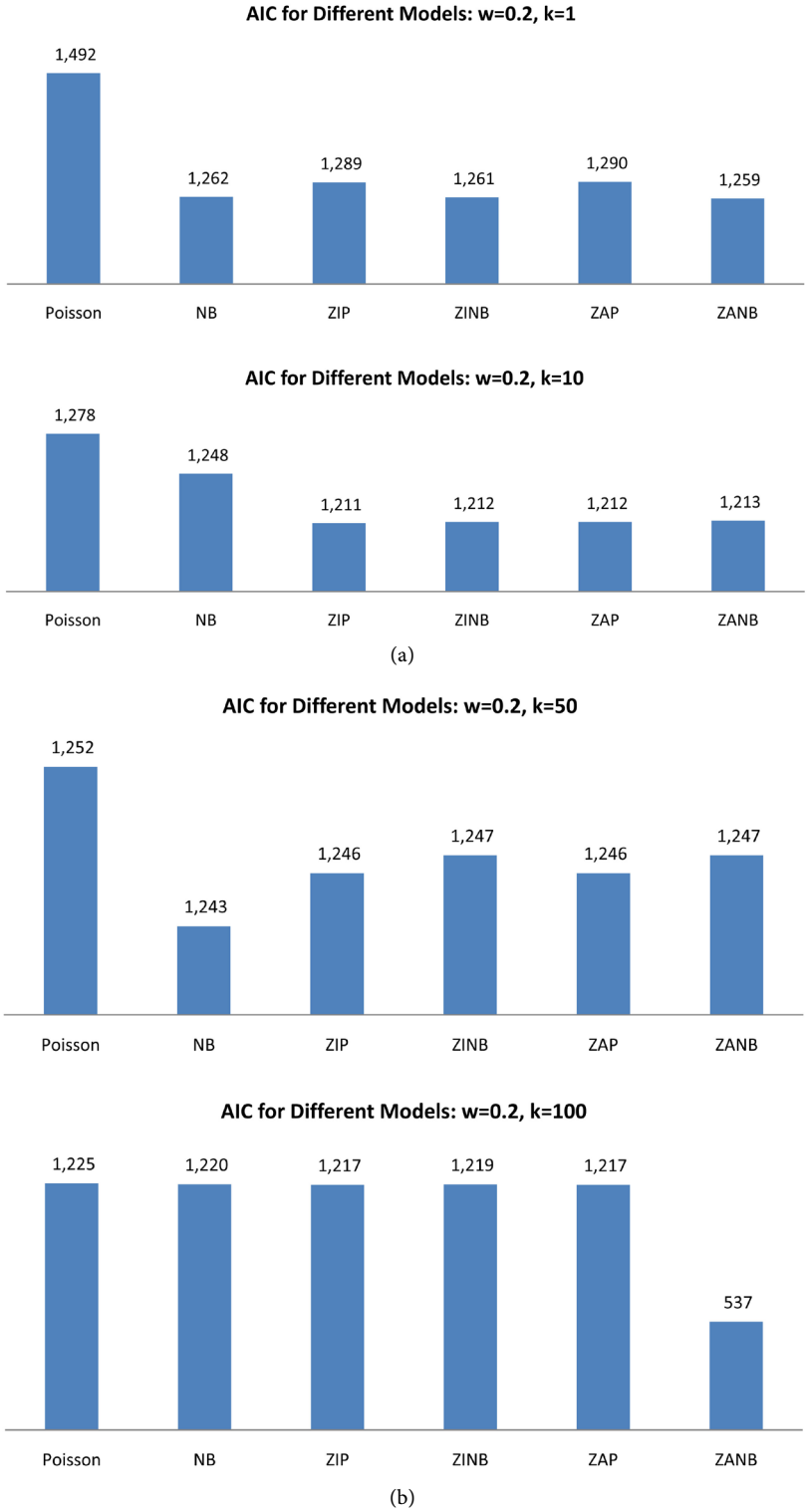
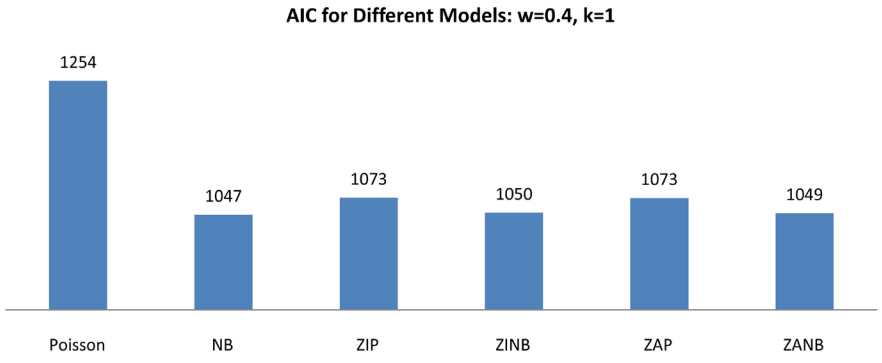
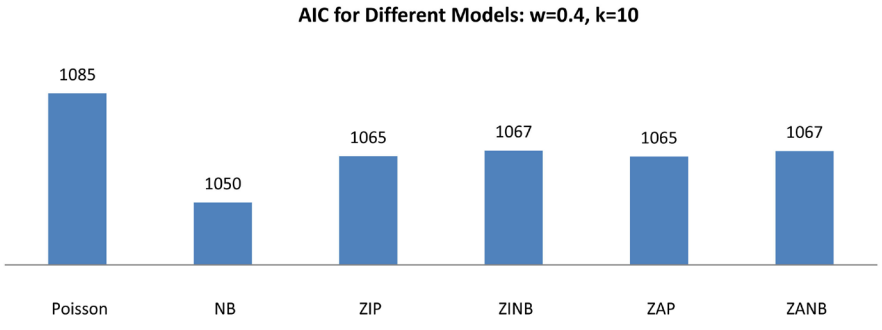
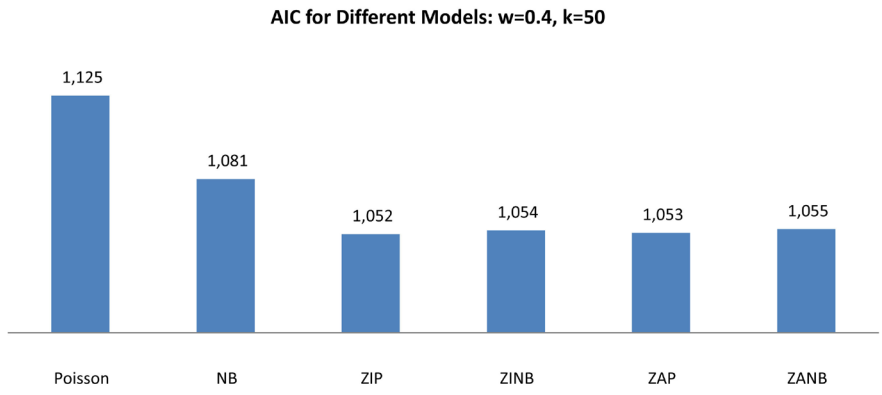


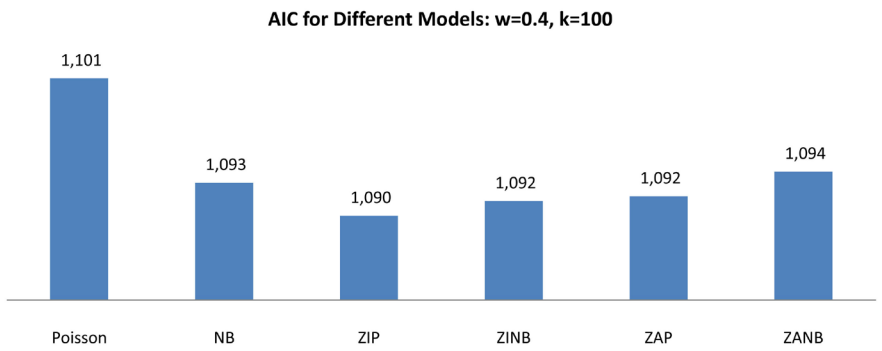
Figure A1. Model comparison based on simulated datesets. Simulations were generated by varying $k \in (1,10,50,100)$ while $\omega = 0.0$ remained constant. (a) $k \in (1,10)$; (b) $k \in (50,100)$.





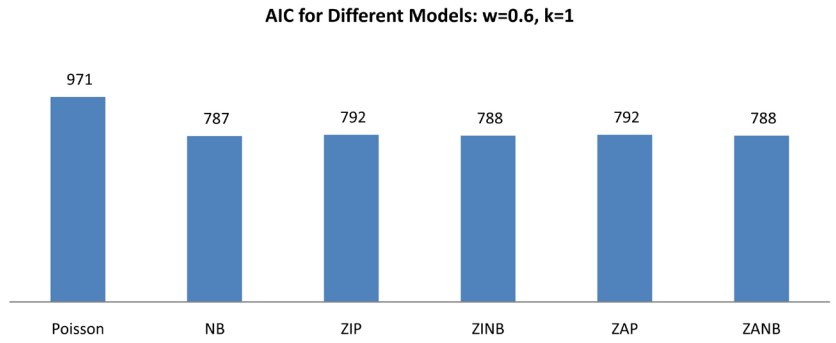
(a)

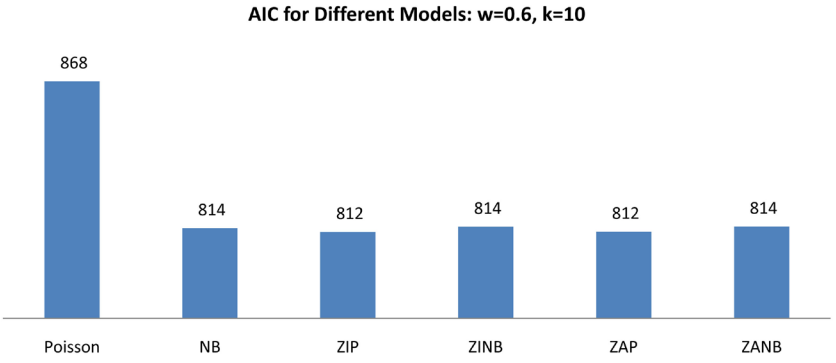




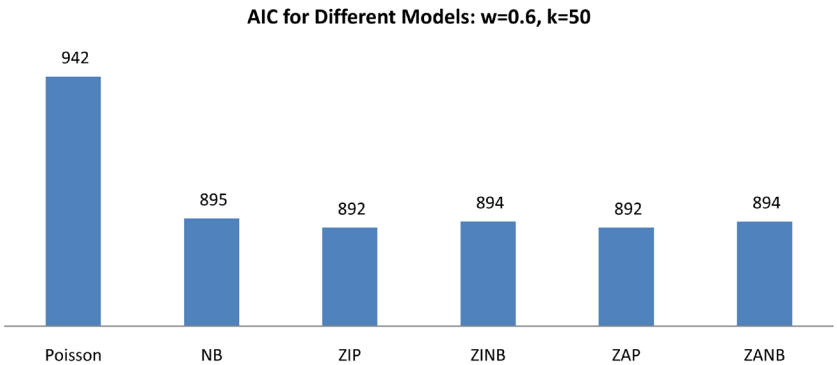
(b)

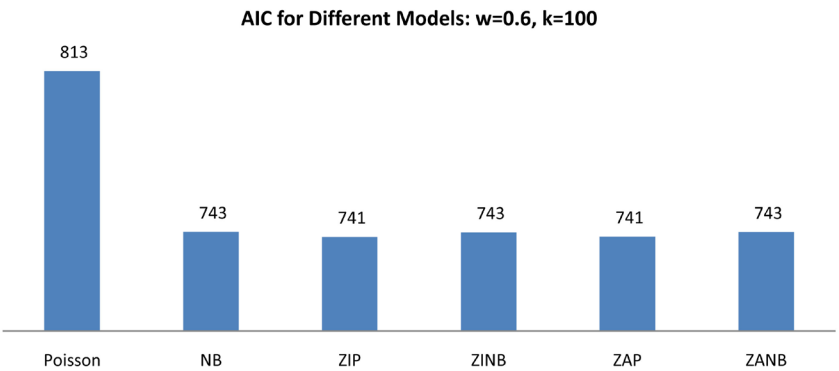
Figure A2. Model comparison based on simulated datesets. Simulations were generated by varying $k \in (1,10,50,100)$ while $\omega = 0.2$ remained constant. (a) $k \in (1,10)$; (b) $k \in (50,100)$.





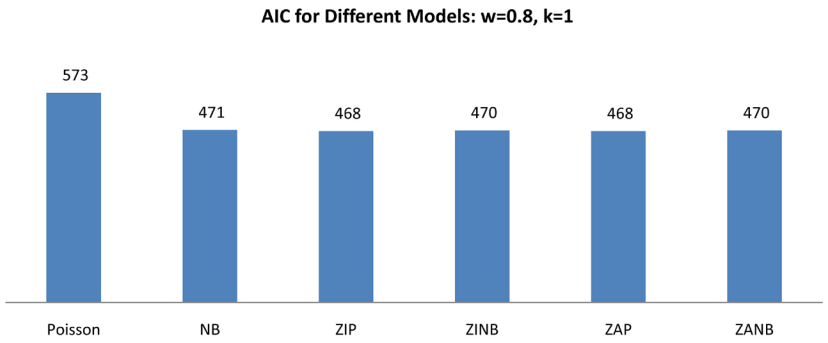
(a)

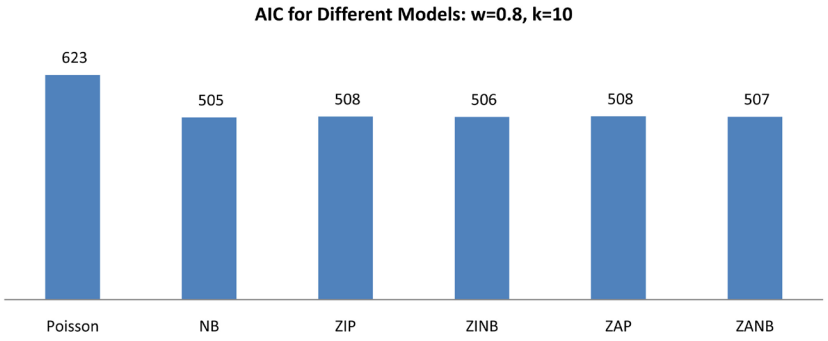




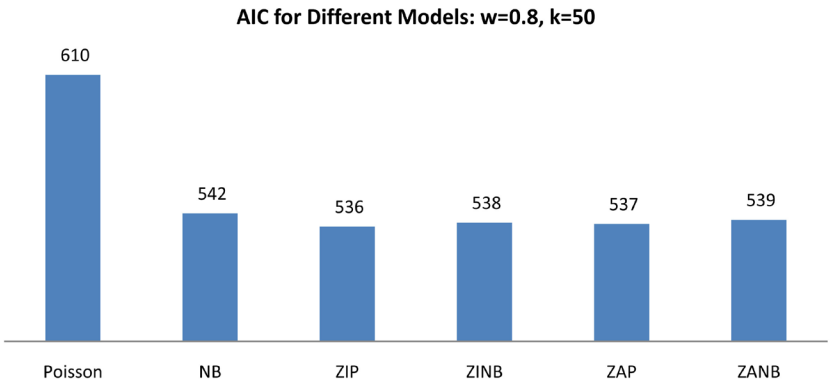
(b)

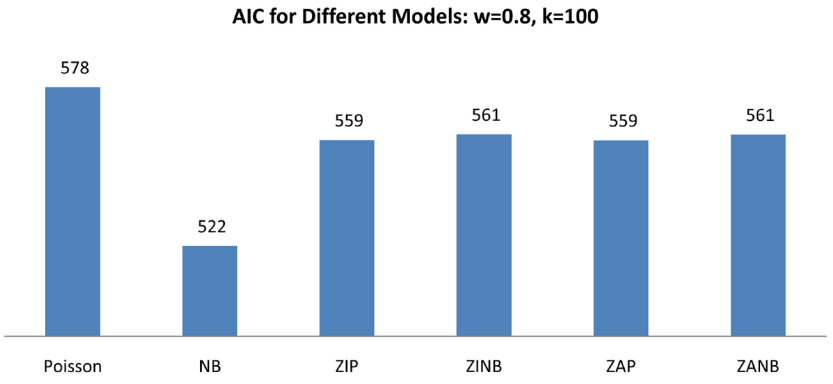
Figure A3. Model comparison based on simulated datesets. Simulations were generated by varying $k \in (1,10,50,100)$ while $\omega = 0.5$ remained constant. (a) $k \in (1,10)$; (b) $k \in (50,100)$.





(a)





(b)

Figure A4. Model comparison based on simulated datesets. Simulations were generated by varying $k \in (1,10,50,100)$ while $\omega = 0.8$ remained constant. (a) $k \in (1,10)$; (b) $k \in (50,100)$.