Published in final edited form as:
Stat Med. 2014 December 20; 33(29): 5151–5165. doi:10.1002/sim.6293.

A Marginalized Zero-Inflated Poisson Regression Model with Overall Exposure Effects

D. Leann Long[a,*], **John S. Preisser**[b], **Amy H. Herring**[b,c], and **Carol E. Golin**[d]

[a]Department of Biostatistics, West Virginia University

[b]Department of Biostatistics, University of North Carolina

[c]Carolina Population Center, University of North Carolina

[d]Department of Health Behavior, University of North Carolina

Abstract

The zero-inflated Poisson (ZIP) regression model is often employed in public health research to examine the relationships between exposures of interest and a count outcome exhibiting many zeros, in excess of the amount expected under sampling from a Poisson distribution. The regression coefficients of the ZIP model have latent class interpretations, which correspond to a susceptible subpopulation at risk for the condition with counts generated from a Poisson distribution and a non-susceptible subpopulation that provide the extra or excess zeros. The ZIP model parameters, however, are not well suited for inference targeted at marginal means, specifically, in quantifying the effect of an explanatory variable in the overall mixture population. We develop a marginalized ZIP model approach for independent responses to model the population mean count directly, allowing straightforward inference for overall exposure effects and empirical robust variance estimation for overall log incidence density ratios. Through simulation studies, the performance of maximum likelihood estimation of the marginalized ZIP model is assessed and compared to other methods of estimating overall exposure effects. The marginalized ZIP model is applied to a recent study of a motivational interviewing-based safer sex counseling intervention, designed to reduce unprotected sexual act counts.

Keywords

Incidence; Marginalized Models; Unprotected intercourse; Zero-inflation

1. Introduction

Zero-inflated count data exist in many areas of medical and public health research. Because Poisson regression is often inadequate in describing count data with many zeros [1], Mullahy [2] proposed the zero-inflated Poisson (ZIP) regression model, based on a mixture of a Poisson distribution and a degenerate distribution at zero. The ZIP model has two sets

*Correspondence to: Department of Biostatistics, West Virginia University, One Medical Center Drive, P.O. Box 9190, Morgantown, WV 26506-9190. dllong@hsc.wvu.edu.

of regression parameters that have latent class interpretations, one for the Poisson mean and the other for the probability of being an excess zero. These latent classes are often thought to classify some *at-risk* and *not-at-risk* populations, indicating a difference in susceptibility between the two groups. In a manufacturing example as described by Lambert [3], a subpopulation of imperfect machines produce defective output at some rate and all other machines produce perfect output; thus, zero observations are a mixture of perfect machines and imperfect machines with no defective output. In contrast, hurdle models consider all zero observations separately from positive realizations, with the notion that all zeros arise from the same data-generating mechanism, distinct from the process producing positive observations [2].

While the ZIP model parameters have latent class interpretations on these two subpopulations, researchers sometimes seek to make inference on the marginal mean of the sampled population. Examples include population-based sample surveys aimed at describing an entire population, intervention studies that target populations where all members are considered to have some risk for the outcome of interest or where interest is in the global effect in the population as a whole. Albert *et al*. [4] argue that insufficient emphasis has been given to the effects of risk factors on the overall population from which the study sample was drawn and propose estimators of overall exposure effects using a causal inference perspective under the zero-inflated modeling framework. Although such marginal effects of predictors are commonly sought, some analysts may find estimating them to be difficult in the traditional ZIP model framework. While transformation techniques, such as those employing the delta method for variance estimation, may be employed to estimate marginal effects of an exposure of interest, these can prove tedious, and the treatment of covariates is not straightforward [5].

The search for easily implementable overall exposure effect estimation in the ZIP model leads to the consideration of the marginalized models literature. Heagerty [6] proposed marginalized multilevel models, which directly model the marginal means by linking marginal and conditional models with a function of covariates, marginal parameters and random effects specification. Lee *et al*. [7] explore hurdle models in the context of marginalized models to analyze clustered data with excess zeros, marginalizing over the random effects. Combining overdispersion, random effects and marginalized models methods, Iddi and Molenberghs [8] obtain population-averaged interpretations for discrete outcomes. These methods for regression of correlated outcomes combine the desire for population average interpretations with the convenience of estimation with a likelihood function constructed with random effects. In a comparatively simple implementation of the principle of marginalization, the marginalized models approach can be adapted in the ZIP model in order to achieve population-wide parameter interpretations for independent count responses with many zeros. Instead of integrating (averaging) over mixtures of distributions defined by random effects, our approach marginalizes over the Poisson and degenerate components of the two-part ZIP model to obtain overall effects.

In studies of risky sexual behavior among HIV-positive individuals, one zero-inflated count variable often studied is the Unprotected Anal and Vaginal Intercourse count (UAVI), the number of unprotected anal or vaginal intercourse acts with any partner over a specified

time period. Golin *et al*. [9] developed the SafeTalk program, a multicomponent, motivational interviewing-based, safer sex intervention for this at-risk population to reduce the number of unprotected sexual acts. In several populations, sexual behavior count data have displayed a distribution with excess zeros [10, 11], and population-averaged effects of covariates on sexual behavior are often desired.

To obtain inference across the marginalized means of the ZIP model, this article proposes a new method for zero-inflated counts in which overall exposure effect estimates are easily obtained via a model for the marginal mean count. Section 2 reviews the traditional ZIP model and outlines issues with overall exposure effect estimation in the ZIP model. Section 3 introduces the marginalized ZIP model that includes parameters with log-incidence density ratio (IDR) interpretations which are estimated by a maximum likelihood procedure. Section 4 presents a simulation study, which examines the properties of the marginalized ZIP and compares it to existing methods for estimating marginal effects. Section 5 presents analysis of the SafeTalk sexual behavior data, using the marginalized ZIP model. A discussion follows in Section 6.

2. Traditional ZIP Model

The ZIP regression model allows the count variable of interest, say Y_i, $i = 1, \ldots, n$ to take the value of zero from a Bernoulli distribution, with probability ψ_i, or be drawn from a Poisson distribution, with mean μ_i, with probability $1 - \psi_i$. Thus,

$$\begin{aligned} Y_i=0 \quad & \text{with} \quad \text{probability} \psi_i+(1-\psi_i)e^{-\mu_i} \\ =k \quad & \text{with} \quad \text{probability} (1-\psi_i)e^{-\mu_i}\mu_i^k/k!, k \in \mathscr{Z}^+ \end{aligned}$$

The likelihood for this ZIP model is

$$L(\boldsymbol{\psi}, \boldsymbol{\mu}|\boldsymbol{y})=\prod_{y_i=0}[(\frac{\psi_i}{1-\psi_i}+e^{-\mu_i})(1-\psi_i)]\prod_{y_i>0}[(1-\psi_i)e^{-\mu_i}\mu_i^{y_i}/(y_i!)]. \quad (1)$$

Lambert [3] proposed models for the parameters μ_i and ψ_i: $\text{logit}(\psi_i)=\boldsymbol{Z}_i'\boldsymbol{\gamma}$ and $\log(\mu_i)=\boldsymbol{X}_i'\boldsymbol{\beta}$, where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{p1})'$ is a $(p_1 \times 1)$ column vector of parameters associated with the excess zeros, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{p2})'$ is a $(p_2 \times 1)$ vector of parameters associated with the Poisson process, and $\boldsymbol{Z}'_{i_{(1\times p_1)}}$ and $\boldsymbol{X}'_{i_{(1\times p_2)}}$ are the vectors of covariates for the i^{th} individual for excess zero and Poisson processes, respectively.

Importantly, the parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ have latent class interpretations; that is, γ_j is the log-odds ratio of a one-unit increase in the j^{th} element of $\boldsymbol{Z}$ on the probability of being an *excess* zero and β_j is the log-incidence density ratio of a one-unit increase in the j^{th} element of $\boldsymbol{X}$ on the mean of the *susceptible* subpopulation. In general, no simple summary of the exposure effect on the overall population mean of the outcome is directly available. Specifically, consider the marginal mean of Y_i, say $\nu_i \equiv E[Y_i]$, often the primary interest of investigators. The relationship between ν_i and the parameters from the ZIP model is

$$\nu_i=(1-\psi_i)\mu_i=\frac{e^{\boldsymbol{X_i'}\beta}}{1+e^{\boldsymbol{Z_i'}\gamma}}. \quad (2)$$

In (2), the population mean is a function of *all* covariates and parameters from both model parts. For the j^{th} covariate in a ZIP model where $Z_i = X_i$ as is commonly specified, the ratio of means for a one-unit increase in x_{ij} is

$$\frac{E(Y_i|x_{ij}=j+1,\tilde{x}_i=\tilde{x}_i)}{E(Y_i|x_{ij}=j,\tilde{x}_i=\tilde{x}_i)}=\exp(\beta_j)\frac{1+\exp(j\gamma_j+\tilde{x}_i'\tilde{\gamma})}{1+\exp[(j+1)\gamma_j+\tilde{x}_i'\tilde{\gamma}]}$$

where x_i ˜indicates all covariates except x_{ij} and γ ˜is created by removing γ_j from $\boldsymbol{\gamma}$. Thus, unless $\gamma_j = 0$, the incidence density ratio (IDR) is not constant across various levels of the extraneous covariates included in the logistic portion of the ZIP model. Additionally, in order to make statements regarding the variability of any IDR estimates at fixed levels of the non-exposure covariates, formal statistical techniques, such as the delta method or bootstrap resampling methods, are required [4]. The computational tools needed for these transformations are typically not readily available in standard software packages, meaning that these calculations can be arduous for many applied analysts.

3. Marginalized ZIP Model

Because population-wide parameter interpretations are desired, the overall mean ν_i can be modeled directly to give overall exposure effect estimates. The marginalized ZIP model specifies

$$\begin{aligned}\text{logit}(\psi_i) &= \boldsymbol{Z_i'}\boldsymbol{\gamma}\\ \log(\nu_i) &= \boldsymbol{X_i'}\boldsymbol{\alpha}.\end{aligned} \quad (3)$$

Then,

$$\nu_i=\exp(\boldsymbol{X_i'}\boldsymbol{\alpha}) \quad (4)$$

allows log-IDR interpretations of the elements of $\boldsymbol{\alpha}$. Thus, $\exp(\alpha_j)$ is the amount by which the mean ν_i is multiplied per unit change in x_j, providing the same interpretation as in Poisson regression. In order to utilize the ZIP model likelihood framework, we redefine $\mu_i = \exp(\delta_i)$, where δ_i is not necessarily a linear function of model parameters. Rather, solving $\nu_i = (1 - \psi_i)\mu_i$, with substitution for (3), provides

$$\delta_i=\boldsymbol{X_i'}\boldsymbol{\alpha}+\log[1+\exp(\boldsymbol{Z_i'}\boldsymbol{\gamma})].$$

Substituting $\psi_i=\exp(\boldsymbol{Z_i'}\boldsymbol{\gamma})/(1+\exp(\boldsymbol{Z_i'}\boldsymbol{\gamma}))$ and $\mu_i = \exp(\delta_i)$ into (1), the likelihood of the marginalized ZIP model for $(\boldsymbol{\gamma},\boldsymbol{\alpha})$ is

$$L(\boldsymbol{\gamma},\boldsymbol{\alpha}|\boldsymbol{y})=\prod_{y_i}(1+e^{\boldsymbol{Z}_i'\boldsymbol{\gamma}})^{-1}\prod_{y_i=0}(e^{\boldsymbol{Z}_i'\boldsymbol{\gamma}}+e^{-(1+\exp(\boldsymbol{Z}_i'\boldsymbol{\gamma}))\exp(\boldsymbol{X}_i'\boldsymbol{\alpha})})\times\prod_{y_i>0}[e^{-(1+\exp(\boldsymbol{Z}_i'\boldsymbol{\gamma}))\exp(\boldsymbol{X}_i'\boldsymbol{\alpha})}(1+e^{\boldsymbol{Z}_i'\boldsymbol{\gamma}})^{y_i}e^{\boldsymbol{X}_i'\boldsymbol{\alpha}y_i}/(y_i!)] \quad (5)$$

with score equations $\boldsymbol{U}_i=\left[\begin{array}{cc}\frac{\partial l(\gamma,\alpha)}{\partial\boldsymbol{\gamma}} & \frac{\partial l(\gamma,\alpha)}{\partial\boldsymbol{\alpha}}\end{array}\right]'$ where

$$\frac{\partial l(\boldsymbol{\gamma},\boldsymbol{\alpha})}{\partial\boldsymbol{\gamma}}=\sum_i\left[\frac{I(y_i=0)\psi_i(1-\psi_i)^{-1}(e^{\nu_i(1-\psi_i)^{-1}}-\nu_i)}{\psi_i(1-\psi_i)^{-1}e^{\nu_i(1-\psi_i)^{-1}}+1}+\psi_i(y_i-1)-I(y_i>0)\psi_i(1-\psi_i)^{-1}\nu_i\right]\boldsymbol{Z}_i'$$

$$\frac{\partial l(\boldsymbol{\gamma},\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}}=\sum_i\left[(y_i-\nu_i(1-\psi_i)^{-1})I(y_i>0)-\frac{\nu_i(1-\psi_i)^{-1}I(y_i=0)}{\psi_i(1-\psi_i)^{-1}e^{\nu_i(1-\psi_i)^{-1}}+1}\right]\boldsymbol{X}_i'$$

and $\nu_i = \nu_i(\boldsymbol{\alpha})$ and $\psi_i = \psi_i(\boldsymbol{\gamma})$. Given the Fisher information $I(\boldsymbol{\gamma},\boldsymbol{\alpha})$, the model-based standard errors of the parameter estimates are

$$\mathrm{se}_M(\hat{\gamma},\hat{\alpha})=\sqrt{\mathrm{diag}(I(\boldsymbol{\gamma},\boldsymbol{\alpha})^{-1})}.$$

To address possibly overdispersed counts relative to the ZIP model, the robust (empirical) estimates of the standard errors are

$$\mathrm{se}_R(\hat{\gamma},\hat{\alpha})=\{\mathrm{diag}[I(\boldsymbol{\gamma},\boldsymbol{\alpha})^{-1}\left(\sum_{i=1}^{n}\boldsymbol{U}_i\boldsymbol{U}_i'\right)I(\boldsymbol{\gamma},\boldsymbol{\alpha})]^{-1}\}^{1/2},$$

with substitution of the MLE's $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\alpha}}$ for $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$, respectively [12].

While parameter estimation can be implemented using various techniques, such as MCMC methods or the EM algorithm, all results herein are obtained through nonlinear optimization by the quasi-Newton method, implemented in SAS 9.3 IML (SAS Institute, Cary, NC). SAS NLMIXED can also be utilized to estimate parameters, and sample code has been provided in the Appendix. As SAS NLMIXED does not readily provide robust variance estimates, the SAS IML code to calculate the robust estimates of standard error for our motivating example has been provided in the online supplementary material. Additionally, the likelihood derivations, as well as those used to obtain the Fisher information, are provided in the Appendix.

4. Simulation Study

Simulation studies were performed to examine the properties of the new marginalized ZIP model under different scenarios, implemented in SAS 9.3 IML. Let Y_i be the zero-inflated Poisson outcome of interest for the i^{th} participant. Also, let x_{i1} be the exposure variable of interest and let x_{i2} be an additional covariate desired in a regression model. In the SafeTalk example, Y_i is the UAVI count, x_{i1} is an indicator of randomization to SafeTalk intervention, and the additional covariate x_{i2} is the baseline UAVI count. Thus the simulated marginalized ZIP regression model is

$$\begin{aligned}\text{logit}(\psi_i) &= \gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2} \\ \log(\nu_i) &= \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2}\end{aligned}$$

To examine the finite sample performance of the marginalized ZIP in estimating specific parameter estimates, we simulated data using the above model. Specifically, x_{i1} ~ Bernoulli(0.5) and x_{i2} follows the standard lognormal distribution, where x_{i1}, x_{i2} are generated independently for a fixed sample size, creating an essentially balanced covariate x_{i2} across the binary exposure of interest x_{i1}. Together with fixed vectors of $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$, these x_{i1} and x_{i2} were used to define ψ_i and μ_i, which were employed to randomly generate excess zeros and Poisson counts, the latter through $\mu_i = \nu_i = (1 - \psi_i)$. Then the marginalized ZIP model was fit to these simulated data and all parameter estimates retained for examination; the simulation was performed 10,000 times and summary measures were calculated. Specifically, for sample sizes of 100, 200 and 1000, Table 1 presents the relative median bias, simulation standard deviation, median model-based and robust standard errors and their corresponding coverage probabilities for each parameter in the model; 95% Wald-type confidence intervals are used. In Table 1, the true parameter values are $\{\gamma_0 = 0.60, \alpha_0 = 0.25, \gamma_1 = -2, \alpha_1 = \log(1.5), \gamma_2 = \alpha_2 = 0.25\}$.

From Table 1, we note that the marginalized ZIP has low bias for $\boldsymbol{\alpha}$ and the bias generally decreases with increasing sample size. For sample size of 1000, the model-based standard errors are similar to the standard deviation of the simulated parameter estimates, implying adequate estimation of the standard error of the parameter estimates; otherwise, standard errors are slightly underestimated for smaller sample sizes and more so for $\hat{\boldsymbol{\gamma}}$ than $\hat{\boldsymbol{\alpha}}$. For all sample sizes, Wald-type confidence intervals of most the marginalized ZIP parameters have model-based coverage probabilities near the expected 0.95, and coverage probabilities created using the robust standard error have fractionally less coverage. For the sample size of 100, the marginalized ZIP Wald-type confidence intervals have slightly less than desirable coverage for the skewed extraneous covariate parameters γ_2 and α_2, but the coverage nears the expected 0.95 as the sample size increases.

Additionally, a simulation study was performed to compare the new marginalized ZIP model to several existing methods for estimating overall exposure effects, namely Poisson regression, both with and without Pearson scaling for overdispersion. Using data generation as described above, the marginalized ZIP and Poisson models were both applied, examining estimates of the log-IDR and standard error of log-IDR. Additionally, 95% Wald-type confidence intervals for the log-IDR were created using the point estimate and respective standard error. For all methods described, Table 2 presents the relative median bias in estimating the IDR and log-IDR, Table 3 presents coverage probabilities and Table 4 displays power. For the marginalized ZIP and Poisson regression models, robust estimators of the covariance matrix were also employed to calculate the 95% Wald-type confidence intervals, as well as their corresponding coverage probabilities and power. Results are presented for varying levels of the true incidence density ratio e^{α_1}, where $\{\gamma_0 = 0.60, \alpha_0 = 0.25, \gamma_1 = -2, \alpha_1 = \{\log(1), \log(1.25), \log(1.5), \log(2)\}, \gamma_2 = \alpha_2 = 0.25\}$.

With regards to bias, Table 2 shows that the marginalized ZIP model has low relative bias in estimating both the log-IDR and IDR while the Poisson model consistently overestimates the overall exposure effect. Although the bias of the Poisson model generally decreases with increasing sample size, the marginalized ZIP relative median bias is notably smaller for each effect and sample size.

Table 3 displays the coverage probabilities for the 95% confidence intervals for each method described. While the marginalized ZIP model has appropriate coverage across effect size and sample size, each of the various Poisson methods have less than desirable coverage, which is not surprising given the relative bias results in Table 2. Examining the power from each method, the marginalized ZIP model has increasing power with increasing effect size and sample size (Table 4). Table 4 also provides observed Type I error rates for each model, where $\alpha_1 = 0$. Note the inflated Type I error for each of the Poisson models, which increases with sample size. Tables 2, 3 and 4 collectively show the inability of the Poisson model to consistently and efficiently estimate overall exposure effects in the presence of a highly skewed independent covariate, which can arise in practice as seen in our motivating example.

In addition to the simulation scenario with a lognormal covariate presented here, the marginalized ZIP model performance was also assessed in the presence of a binary covariate. Under this binary covariate scenario, both the Poisson model with robust variance estimates and Poisson model scaled for overdispersion had comparable bias, coverage probabilities and power to the marginalized ZIP model (results presented in the online supplementary material).

Finally, simulation results which compare the marginalized ZIP model performance to the traditional ZIP model with delta method transformations at fixed levels of the covariate, as well as the average predicted value ratio from Albert *et al*. [4] are presented in the online supplementary material. In the presence of the lognormal extraneous covariate, the traditional ZIP model has increased relative median bias compared to the marginalized ZIP, less than desired coverage probabilities and inflated Type I error. However, in the presence of a binary covariate, the traditional ZIP model with transformations at the mean covariate value yielded similar bias, coverage and power performance to the marginalized ZIP model. Unlike the marginal inference of the marginalized ZIP and Poisson regression models presented, the traditional ZIP model method does not estimate the marginal mean, but the ‘overall’ mean at fixed covariate levels.

5. Motivational Interviewing Intervention Example

Reducing risky sexual behavior among people living with HIV/AIDS is one area of focus among infectious disease researchers, and one measure of risky behavior is the UAVI count, the number of Unprotected Anal or Vaginal sexual Intercourse acts within a given time period. The SafeTalk program was developed as a motivational interviewing-based intervention to reduce sexual behavior, particularly UAVI [9,13]. To assess SafeTalk’s efficacy at reducing unprotected sex acts in this population, a randomized clinical trial was performed with subjects recruited at three sites being randomized to receive either SafeTalk

or a nutritional intervention as control. The participants were then surveyed every four months for one year to measure their self-reported sexual acts in the previous three-month period. The primary research question for this study is whether those in the SafeTalk intervention have lower UAVI than those in the control at the eight-month follow-up visit, indicating cross-sectional methods are appropriate despite the longitudinal nature of the data collection.

For this analysis, there are 357 participants with non-missing UAVI counts at the 8-month visit, excluding eight participants with UAVI counts greater than 18 for the purposes of this illustration. Figure 1 shows the distribution of UAVI counts, which contains 300 (84%) zeros and 8 '10+' counts (2.2%). Since the randomization scheme stratified by site, the marginalized ZIP model to be fit is

$$\begin{aligned} \text{logit}(\psi_i) &= \gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \gamma_3 x_{i3} + \gamma_4 x_{i4} \\ \log(\nu_i) &= \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \alpha_3 x_{i3} + \alpha_4 x_{i4} \end{aligned}$$

where x_{i1} is an indicator of whether the i^{th} participant received the SafeTalk intervention and x_{i2} and x_{i3} are indicators of whether the i^{th} participant was randomized at the second and third study sites, respectively. Additionally, the analysis controls for baseline UAVI count x_{i4}.

In order to calculate the 'overall' effect of the SafeTalk treatment for the traditional ZIP model, the proportions observed at Site 2 (0.3221) and Site 3 (0.0588) and mean baseline UAVI count (0.9748) are used for the delta method calculations. For the traditional ZIP with delta method, the log-IDR for the intervention is 0.2133 (0.2872), which yields an IDR estimate of 1.2378 and 95% confidence interval (0.705, 2.173). For Sites 1, 2 and 3, the IDR (and corresponding 95% confidence intervals) from the transformed ZIP with fixed mean baseline UAVI count are 1.2360 (0.706, 2.165), 1.2399 (0.704, 2.184), and 1.2429 (0.701, 2.203), respectively. Examining the range of IDR across baseline UAVI counts, the IDR and corresponding 95% confidence intervals for zero and 18 baseline UAVI counts are 1.2487 (0.702, 2.222) and 0.9598 (0.727, 1.267). For this particular example, there does not appear to be much difference in the IDR of treatment across sites, but note the moderate change in IDR estimates for the different baseline UAVI counts. Although none of these estimates are statistically significant, the estimates for different combinations of covariates demonstrate the lack of a single IDR measure when using traditional ZIP with the delta method. In fact, particular transformed ZIP analyses may yield very different IDR estimates for various combinations of covariate values. Also, notice the transformed ZIP methods require more effort and expertise in deriving and programming than the direct estimation of the log-IDR through the marginalized ZIP model.

Table 5 presents the results of the marginalized ZIP analysis on the SafeTalk example. By exponentiating α_1, the estimate of the IDR for treatment is exp(−0.0666) = 0.9355; thus, the marginalized ZIP model reveals those on SafeTalk intervention have 6% fewer unprotected sexual acts at the eight-month followup visit than those participants randomized to control. The 95% model-based Wald-type confidence interval for the treatment IDR is (0.559, 1.567), implying there is no significant difference between the two treatment groups.

However, this illustrative analysis is not considered definitive due to the deletion of large UAVI counts. Because the traditional ZIP with delta method is limited by the substitution of specific levels of the extraneous covariates, the overall effect of SafeTalk is difficult to summarize briefly. However, the marginalized ZIP model gives one IDR of SafeTalk intervention, adjusted for all the other covariates. In terms of model fit, the full likelihood values for the marginalized ZIP and traditional ZIP models are −291.28 and −288.51, indicating that the two models have similar fit to the SafeTalk data.

6. Conclusion

A marginalized ZIP model was proposed for population-averaged inference of count data with many zeros. The new statistical approach directly models the marginal means of mixtures of two discrete distributions, one consisting of Poisson counts and the other of structural zeros. This model formulation offers meaningful statements about an exposure effect on an entire population in contrast to the traditional ZIP model whose regression parameters have interpretations for unobservable latent classes. Whereas an ‘average’ effect of an exposure in a population can be determined with additional computations following the fit of a traditional ZIP model, the modeling approach proposed in this article provides direct estimates of a homogeneous exposure effect that does not require post-modeling computations. Indeed, the proposed model’s marginal effects of interest are given by log incidence density ratios that have the same interpretations as in Poisson regression. Also, an offset term can easily be included in the marginalized ZIP model to allow more flexibility through modeling incidence densities. The logistic model part for excess zeros in the new formulation is not of primary interest, but rather its role is to provide adjustment for overdispersion due to excess zeros. Based on the research question and perceived data structure, analysts may choose to specify different covariates in Z_i and X_i, but the marginal mean interpretations of α do not change with different specifications of Z_i. In a simulation study of Poisson generated counts with extra zeros, the marginalized ZIP model had percent relative bias of three percent or less with as few as 100 observations. In all scenarios considered, the marginalized ZIP had smaller percent relative bias than Poisson regression and provided appropriate Type I error; each Poisson regression, including after overdispersion adjustment either through Pearson scaling or robust standard error estimation, yielded inflated Type I error under each scenario.

Despite the increasing popularity of the ZIP model in health-related fields, the idea of latent class effects can be troublesome for many investigators to communicate, often yielding misleading or incorrect statements. For example, Preisser *et al*. [5] found that many dental caries researchers interpreted the Poisson parameters of the ZIP model with respect to the overall caries incidence, rather than the correct model-based interpretation relating to caries incidence within the *at-risk* population. This pattern of misinterpretation suggests that investigators when genuinely interested in marginal inference for count data in the presence of many zeros may sometimes be led to use ZIP models simply because of goodness-of-fit considerations. Generally, the research goal should lead to the identification of a class of models that can address the question of interest; only when considering competing models within the identified class should goodness-off-it considerations prevail. This approach to model selection based on collaboration between investigators and biostatistical scientists

discourages purely empirical model fitting exercises. The marginalized ZIP model is viewed as belonging to a different model class than the traditional ZIP model and so choosing between them should be based on the research question.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported in part by National Institute of Health (NIH) grants T32ES007018 (NIEHS), T32HD007237 (NICHD), R01MH069989 (NIMH), R01ES020619 (NIEHS), U54GM104942 (NIGMS), and University of North Carolina Center for AIDS Research AI50410. The content is solely the responsibility of the authors and does not necessarily represent the official views of NIH. This work was conducted as part of the first author's doctoral dissertation in the Department of Biostatistics at the University of North Carolina at Chapel Hill.

Appendix

A. SAS NLMIXED Code

The following SAS NLMIXED code fits the marginalized ZIP model to the SafeTalk data from the motivating example in section 5. In order utilize this code, a statistical analyst would need to specify the forms of logit_psi and log_nu, $\text{logit}(\psi_i)=\boldsymbol{Z}_i'\boldsymbol{\gamma}$ and $\log(\nu_i)=\boldsymbol{X}_i'\boldsymbol{\alpha}$ respectively. Additionally, the null initial parameter values statement would require alteration depending on the number of parameters required for estimation.

```
proc nlmixed data=work.for_analysis seed=31415 maxiter=500 qpoints=50 cov
hess;
/* null initial parameter values */
parms g0 0 g1 0 g2 0 g3 0 g4 0
a0 0 a1 0 a2 0 a3 0 a4 0;
/* linear predictor for the zero-inflation probability */
/* logit(psi)=Z\gamma */
logit_psi = g0 + g1*arm + g2*site2 + g3*site3 + g4*baseline_uavi;
/* Useful functions of psi */
psi1 = exp(logit_psi)/(1+exp(logit_psi)); /*psi = exp(Z\gamma)/(1+exp(Z
\gamma)) */
psi2 = 1/(1+exp(logit_psi)); /*1−psi = (1+exp(Z\gamma))ˆ−1 */
/* Overall mean \nu */
/* log(nu) = X\alpha */
log_nu = a0 + a1*arm + a2*site2 + a3*site3 + a4*baseline_uavi;
delta = log(psi2**(−1)) + log_nu;
/* Build the mZIPlog likelihood */
if outcome=0 then
ll = log(psi1 + psi2*(exp(−exp(delta))));
else ll = log(psi2) − exp(delta) + outcome*(delta) − lgamma(outcome + 1);
```

```
model outcome ~ general(ll);
run;
```

Note that this SAS NLMIXED code does not provide the robust standard error estimates presented in Table 5, for which separate SAS IML programming would be needed (see online supplementary material for example code).

B. Likelihood Derivations

First, we focus on the derivation of the MLE of $(\boldsymbol{\gamma}, \boldsymbol{\alpha})$ by constructing the likelihood. From Equation (1), we can derive the MLE's of $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ using Newton-Raphson algorithm, as well as derive the analytic variance of these MLE's.

$$L(\boldsymbol{\gamma},\boldsymbol{\delta}|\boldsymbol{y})=\prod_{y_i=0}[(e^{\boldsymbol{Z}_i'\boldsymbol{\gamma}}+e^{-\exp(\delta_i)})(1+e^{\boldsymbol{Z}_i'\boldsymbol{\gamma}})^{-1}]\prod_{y_i>0}[(1+e^{\boldsymbol{Z}_i'\boldsymbol{\gamma}})^{-1}e^{-\exp(\delta_i)}e^{\delta_i y_i}/(y_i!)]=\prod_{y_i}(1+e^{\boldsymbol{Z}_i'\boldsymbol{\gamma}})^{-1}\prod_{y_i=0}(e^{\boldsymbol{Z}_i'\boldsymbol{\gamma}}+e^{-\exp(\delta_i)})\prod_{y_i>0}[$$

$$L(\boldsymbol{\gamma},\boldsymbol{\alpha}|\boldsymbol{y})=\prod_{y_i}(1+e^{\boldsymbol{Z}_i'\boldsymbol{\gamma}})^{-1}\prod_{y_i=0}(e^{\boldsymbol{Z}_i'\boldsymbol{\gamma}}+e^{-(1+\exp(\boldsymbol{Z}_i'\boldsymbol{\gamma}))\exp(\boldsymbol{X}_i'\boldsymbol{\alpha})})$$

$$\prod_{y_i>0}[e^{-(1+\exp(\boldsymbol{Z}_i'\boldsymbol{\gamma}))\exp(\boldsymbol{X}_i'\boldsymbol{\alpha})}(1+e^{\boldsymbol{Z}_i'\boldsymbol{\gamma}})^{y_i}e^{\boldsymbol{X}_i'\boldsymbol{\alpha}y_i}/(y_i!)]$$

$$l(\boldsymbol{\gamma},\boldsymbol{\alpha}|\boldsymbol{y})=-\sum_i\log(1+e^{\boldsymbol{Z}_i'\boldsymbol{\gamma}})+\sum_{y_i=0}\log(e^{\boldsymbol{Z}_i'\boldsymbol{\gamma}}+e^{-\exp(\boldsymbol{X}_i'\boldsymbol{\alpha})(1+\exp(\boldsymbol{Z}_i'\boldsymbol{\gamma}))})+\sum_{y_i>0}(-(1+e^{\boldsymbol{Z}_i'\boldsymbol{\gamma}})e^{\boldsymbol{X}_i'\boldsymbol{\alpha}}+y_i\log(1+e^{\boldsymbol{Z}_i'\boldsymbol{\gamma}})+\boldsymbol{X}_i'\boldsymbol{\alpha}y_i-\mathrm{lc}$$

Using this log-likelihood, the score equations are

$$\frac{\partial l(\gamma,\alpha)}{\partial\boldsymbol{\gamma}}=\sum_{y_i=0}\frac{e^{\boldsymbol{Z}_i'\boldsymbol{\gamma}}\boldsymbol{Z}_i'+e^{-\exp(\boldsymbol{X}_i'\boldsymbol{\alpha})(1+e^{\boldsymbol{Z}_i'\boldsymbol{\gamma}})}(-e^{\boldsymbol{X}_i'\boldsymbol{\alpha}+\boldsymbol{Z}_i'\boldsymbol{\gamma}})\boldsymbol{Z}_i'}{e^{\boldsymbol{Z}_i'\boldsymbol{\gamma}}+e^{-\exp(\boldsymbol{X}_i'\boldsymbol{\alpha})(1+\exp(\boldsymbol{Z}_i'\boldsymbol{\gamma}))}}+\sum_{y_i>0}\frac{y_ie^{\boldsymbol{Z}_i'\boldsymbol{\gamma}}}{1+e^{\boldsymbol{Z}_i'\boldsymbol{\gamma}}}\boldsymbol{Z}_i'-e^{\boldsymbol{X}_i'\boldsymbol{\alpha}+\boldsymbol{Z}_i'\boldsymbol{\gamma}}\boldsymbol{Z}_i'-\sum_i\frac{e^{\boldsymbol{Z}_i'\boldsymbol{\gamma}}\boldsymbol{Z}_i'}{1+e^{\boldsymbol{Z}_i'\boldsymbol{\gamma}}}=\sum_i\left[\frac{I(y_i=0)e^{\boldsymbol{Z}_i'\boldsymbol{\gamma}}(}{e^{\boldsymbol{Z}_i'\boldsymbol{\gamma}}e^{}}\right.$$

$$\frac{\partial l(\gamma,\alpha)}{\partial\boldsymbol{\alpha}}=\frac{\partial\log L}{\partial\boldsymbol{\alpha}}=-\sum_{y_i=0}\frac{(1+e^{\boldsymbol{Z}_i'\boldsymbol{\gamma}})e^{\boldsymbol{X}_i'\boldsymbol{\alpha}}e^{-\exp(\boldsymbol{X}_i'\boldsymbol{\alpha})(1+\exp(\boldsymbol{Z}_i'\boldsymbol{\gamma}))}\boldsymbol{X}_i'}{e^{\boldsymbol{Z}_i'\boldsymbol{\gamma}}+e^{-\exp(\boldsymbol{X}_i'\boldsymbol{\alpha})(1+\exp(\boldsymbol{Z}_i'\boldsymbol{\gamma}))}}+\sum_{y_i>0}(y_i-e^{\boldsymbol{X}_i'\boldsymbol{\alpha}}(1+e^{\boldsymbol{Z}_i'\boldsymbol{\gamma}}))\boldsymbol{X}_i'=\sum_i\left[(y_i-e^{\boldsymbol{X}_i'\boldsymbol{\alpha}}(1+e^{\boldsymbol{Z}_i'}\right.$$

Substituting the link functions $\mathrm{logit}(\psi_i)=\boldsymbol{Z}_i'\boldsymbol{\gamma}$ and $\log(\nu_i)=\boldsymbol{X}_i'\boldsymbol{\alpha}$, these expressions of the score equations are equivalent to those presented in Section 2.2. The matrix of second derivatives of the log-likelihood has the form

$$\begin{bmatrix}\frac{\partial^2 l}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}'} & \frac{\partial^2 l}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\alpha}'}\\ \frac{\partial^2 l}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\gamma}'} & \frac{\partial^2 l}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}'}\end{bmatrix}$$

where

$$\frac{\partial^2 l}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}'}=\frac{\partial}{\partial\boldsymbol{\gamma}}\left[\frac{\partial l}{\partial\boldsymbol{\gamma}}\right]'=\frac{\partial}{\partial\boldsymbol{\gamma}}\left\{\sum_i \boldsymbol{Z}_i\left[\frac{I(y_i=0)e^{\boldsymbol{Z}_i'\boldsymbol{\gamma}}(e^{\exp(\boldsymbol{X}_i'\boldsymbol{\alpha})(1+\exp(\boldsymbol{Z}_i'\boldsymbol{\gamma}))}-e^{\boldsymbol{X}_i'\boldsymbol{\alpha}})}{e^{\boldsymbol{Z}_i'\boldsymbol{\gamma}}e^{\exp(\boldsymbol{X}_i'\boldsymbol{\alpha})(1+\exp(\boldsymbol{Z}_i'\boldsymbol{\gamma}))}+1}+\frac{e^{\boldsymbol{Z}_i'\boldsymbol{\gamma}}(y_i-1)}{1+e^{\boldsymbol{Z}_i'\boldsymbol{\gamma}}}-I(y_i>0)e^{\boldsymbol{X}_i'\boldsymbol{\alpha}+\boldsymbol{Z}_i'\boldsymbol{\gamma}}\right]\right\}=-\sum_i$$

$$\frac{\partial^2 l}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}'}=\frac{\partial}{\partial\boldsymbol{\alpha}}\left[\frac{\partial l}{\partial\boldsymbol{\alpha}}\right]'=\frac{\partial}{\partial\boldsymbol{\alpha}}\left\{\sum_i \boldsymbol{X_i}\left[(y_i-e^{\boldsymbol{X}_i'\boldsymbol{\alpha}}(1+e^{\boldsymbol{Z}_i'\boldsymbol{\gamma}}))I(y_i>0)-\frac{(1+e^{\boldsymbol{Z}_i'\boldsymbol{\gamma}})e^{\boldsymbol{X}_i'\boldsymbol{\alpha}}I(y_i=0)}{e^{\boldsymbol{Z}_i'\boldsymbol{\gamma}}e^{\exp(\boldsymbol{X}_i'\boldsymbol{\alpha})(1+\exp(\boldsymbol{Z}_i'\boldsymbol{\gamma}))}+1}\right]\right\}=\sum_i \boldsymbol{X_i}\left[(-e^{\boldsymbol{X}_i'\boldsymbol{\alpha}}\right.$$

$$\frac{\partial^2 l}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\alpha}'}=\frac{\partial}{\partial\boldsymbol{\gamma}}\left[\frac{\partial l}{\partial\boldsymbol{\alpha}}\right]'=\sum_i \boldsymbol{X_i}\frac{\partial}{\partial\boldsymbol{\gamma}}\left[(y_i-e^{\boldsymbol{X}_i'\boldsymbol{\alpha}}(1+e^{\boldsymbol{Z}_i'\boldsymbol{\gamma}}))I(y_i>0)-\frac{(1+e^{\boldsymbol{Z}_i'\boldsymbol{\gamma}})e^{\boldsymbol{X}_i'\boldsymbol{\alpha}}I(y_i=0)}{e^{\boldsymbol{Z}_i'\boldsymbol{\gamma}}e^{\exp(\boldsymbol{X}_i'\boldsymbol{\alpha})(1+\exp(\boldsymbol{Z}_i'\boldsymbol{\gamma}))}+1}\right]=-\sum_i \boldsymbol{X_i}\left[e^{\boldsymbol{X}_i'\boldsymbol{\alpha}+\boldsymbol{Z}_i'\boldsymbol{\gamma}}I\right.$$

In order to obtain the Fisher information matrix, we calculate the negative expectations of the above second derivatives. First, we note that

$$\begin{aligned}P(Y_i=0) &= \psi_i+(1-\psi_i)e^{-\nu_i(1-\psi_i)^{-1}}=\frac{1-\psi_i}{e^{\nu_i(1-\psi_i)^{-1}}}\left(\frac{\psi_i}{1-\psi_i}e^{\nu_i(1-\psi_i)^{-1}}+1\right)\\ P(Y_i>0) &= (1-\psi_i)(1-e^{-\nu_i(1-\psi_i)^{-1}})\\ E(Y_i) &= \nu_i\end{aligned}$$

Then

References

1. Böhning D, Dietz E, Schlattmann P, Mendonça L, Kirchner U. The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. Journal of the Royal Statistical Society. Series A (Statistics in Society). 1999; 162:195–209.
2. Mullahy J. Specification and testing of some modified count data models. Journal of Econometrics. 1986; 33:341–365.
3. Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics. 1992; 34:1–14.
4. Albert JM, Wang W, Nelson S. Estimating overall exposure effects for zero-inflated regression models with application to dental caries. Statistical Methods in Medical Research. 2014; 23(3):257–278. [PubMed: 21908419]
5. Preisser JS, Stamm JW, Long DL, Kincade ME. Review and recommendations for zero-inflated count regression modeling of dental caries indices in epidemiological studies. Caries Research. 2012; 46(4):413–423. [PubMed: 22710271]
6. Heagerty P. Marginally specified logistic-normal models for longitudinal binary data. Biometrics SEP. 1999; 55(3):688–698.
7. Lee K, Joo Y, Song J, Harper D. Analysis of zero-inflated clustered count data: A marginalized model approach. Computational Statistics & Data Analysis. 2011; 55(1):824–837.
8. Iddi S, Molenberghs G. A combined overdispersed and marginalized multilevel model. Computational Statistics & Data Analysis. 2012; 56:1944–1951.
9. Golin C, Davis R, Przybyla S, Fowler B, Parker S, Earp J, Quinlivan E, Kalichman S, Patel S, Grodensky C. Safetalk, a multicomponent, motivational interviewing-based, safer sex counseling program for people living with HIV/AIDS: A qualitative assessment of patients' views. AIDS Patient Care and STDs. 2010; 24(4):237–245. [PubMed: 20377435]
10. Heilbron D. Zero-altered and other regression models for count data with added zeros. Biometrical Journal. 1994; 36:531–547.

11. Ghosh P, Tu W. Assessing sexual attitudes and behaviors of young women: a joint model with nonlinear time effects, time varying covariates, and dropouts. Journal of the American Statistical Association. 2009; 104(486):474–485.

12. White H. Maximum likelihood estimation of misspecified models. Econometrica. 1982; 50(1):1–25.

13. Golin CE, Earp JA, Grodensky CA, Patel SN, Suchindran C, Parikh M, Kalichman S, Patterson K, Swygard H, Quinlivan EB, et al. Longitudinal effects of Safetalk, a motivational interviewing-based program to improve safer sex practices among people living with HIV/AIDS. AIDS and Behavior. 2012; 16(5):1182–1191. [PubMed: 21964975]

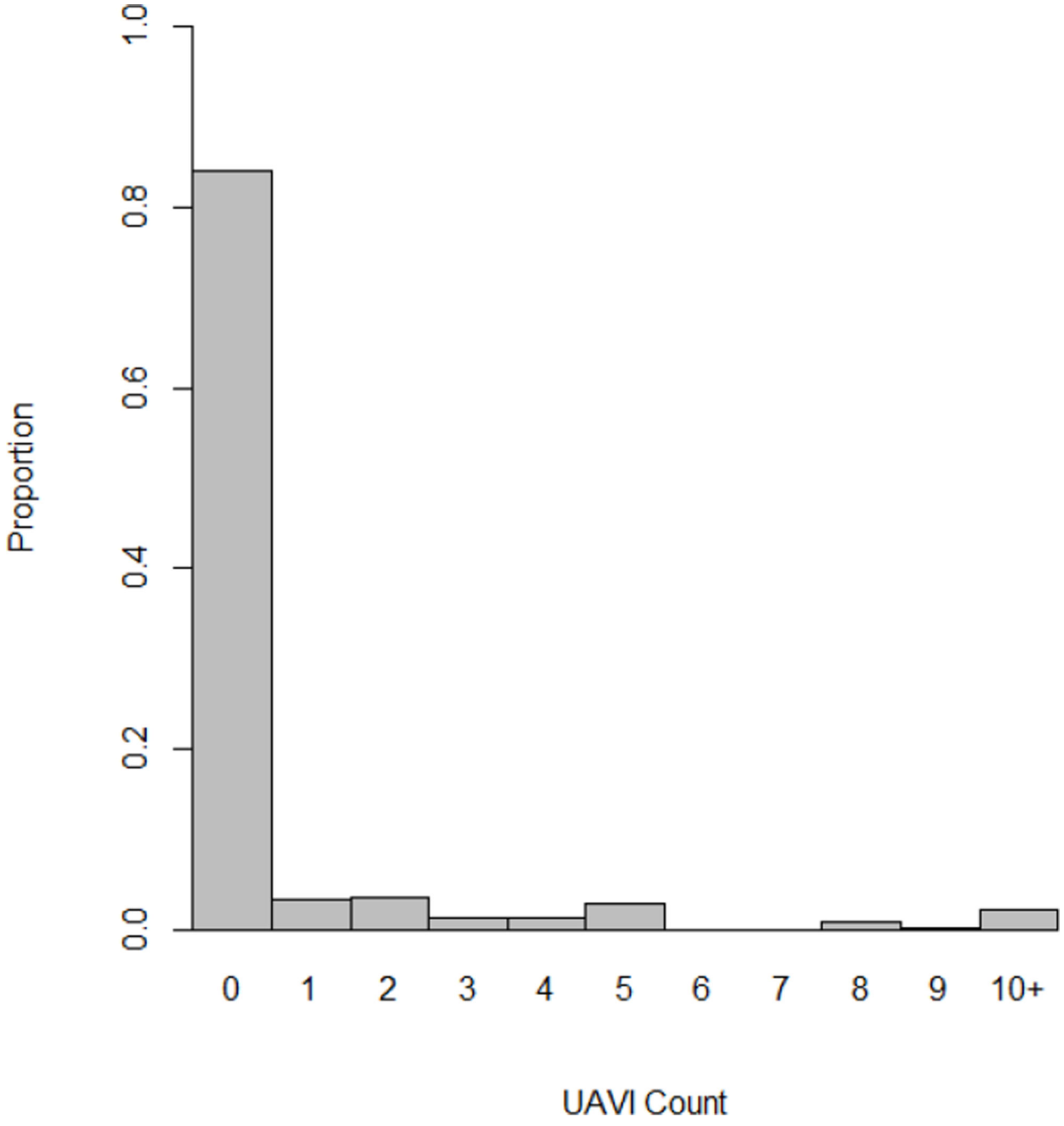

Figure 1.
Histogram of UAVI Counts

Stat Med. Author manuscript; available in PMC 2015 December 20.

Table 1

Marginalized ZIP model performance with 10,000 simulations and varying sample size

Sample Size	Parameter	Percent Relative Median Bias	Simulation Std Dev	Median Model-Based Std Error	Median Robust Std Error	Model-Based Coverage Probability	Robust Coverage Probability
100	γ_0	−0.57	0.381	0.353	0.348	0.941	0.937
	γ_1	2.64	0.484	0.447	0.434	0.940	0.929
	γ_2	0.87	0.127	0.099	0.085	0.891	0.830
	α_0	−1.35	0.288	0.270	0.266	0.948	0.941
	α_1	0.99	0.299	0.282	0.278	0.950	0.947
	α_2	1.25	0.087	0.076	0.069	0.908	0.874
200	γ_0	−1.57	0.259	0.247	0.244	0.942	0.937
	γ_1	1.04	0.322	0.306	0.295	0.943	0.930
	γ_2	1.67	0.081	0.066	0.059	0.897	0.855
	α_0	0.57	0.195	0.190	0.188	0.950	0.949
	α_1	−0.13	0.202	0.198	0.196	0.949	0.947
	α_2	0.03	0.057	0.053	0.050	0.927	0.903
1000	γ_0	−0.37	0.110	0.110	0.109	0.950	0.949
	γ_1	0.35	0.137	0.134	0.130	0.943	0.936
	γ_2	0.55	0.032	0.029	0.027	0.926	0.902
	α_0	0.67	0.085	0.085	0.085	0.949	0.949
	α_1	0.21	0.089	0.088	0.088	0.949	0.948
	α_2	−0.16	0.024	0.023	0.023	0.942	0.934

True parameter values: $\{\gamma_0 = 0.60, \alpha_0 = 0.25, \gamma_1 = -2, \alpha_1 = \log(1.5), \gamma_2 = \alpha_2 = 0.25\}$

Percent relative median bias: $[(\hat{\theta} - \theta)/\theta] \times 100\%$, $\boldsymbol{\theta} = \{\gamma_0, \gamma_1, \gamma_2, \alpha_0, \alpha_1, \alpha_2\}$

Table 2

Comparison of percent relative median biases for estimation of overall exposure effects α_1 with marginalized ZIP & Poisson models

True IDR (e^{α_1})	Sample Size	Marginalized ZIP		Poisson	
		Log-IDR	IDR	Log-IDR	IDR
1	100	−0.06	−0.06	6.75	6.98
	200	0.36	0.36	8.06	8.39
	1000	0.22	0.22	3.14	3.19
1.25	100	1.47	0.33	35.36	8.21
	200	1.01	0.23	31.99	7.40
	1000	0.80	0.18	16.99	3.86
1.5	100	0.99	0.40	17.58	7.39
	200	−0.13	−0.05	17.54	7.37
	1000	0.21	0.09	7.99	3.29
2	100	0.07	0.05	11.44	8.25
	200	0.69	0.48	11.11	8.01
	1000	0.12	0.09	5.42	3.83

Percent relative median bias ($\alpha_1 > 0$): $[(\hat{\theta} - \theta)/\theta] \times 100\%$, $\boldsymbol{\theta} = \{\alpha_1, e^{\alpha_1}\}$

Percent median bias ($\alpha_1 = 0$): $[(\hat{\theta} - \theta)] \times 100\%$, $\boldsymbol{\theta} = \{\alpha_1, e^{\alpha_1}\}$

Stat Med. Author manuscript; available in PMC 2015 December 20.

Table 3

Comparison of coverage probabilities for estimation of overall exposure effects α_1 with marginalized ZIP & Poisson models

		Marginalized ZIP		Poisson		Overdispersed Poisson
True IDR (e^{α_1})	Sample Size	Model	Robust	Model	Robust	Model
1	100	0.947	0.942	0.479	0.897	0.835
	200	0.949	0.946	0.414	0.900	0.797
	1000	0.945	0.944	0.303	0.886	0.696
1.25	100	0.949	0.943	0.459	0.901	0.838
	200	0.947	0.944	0.392	0.896	0.775
	1000	0.950	0.950	0.288	0.886	0.684
1.5	100	0.950	0.947	0.448	0.898	0.827
	200	0.949	0.947	0.390	0.898	0.776
	1000	0.949	0.948	0.273	0.886	0.666
2	100	0.952	0.948	0.436	0.901	0.825
	200	0.948	0.946	0.371	0.901	0.764
	1000	0.950	0.948	0.263	0.892	0.663

Table 4

Comparison of power for estimation of overall exposure effects α_1 with marginalized ZIP & Poisson models

		Marginalized ZIP		Poisson		Overdispersed Poisson
True IDR (e^{α_1})	Sample Size	Model	Robust	Model	Robust	Model
1	100	0.053	0.058	0.521	0.103	0.165
	200	0.051	0.054	0.586	0.100	0.203
	1000	0.055	0.056	0.697	0.114	0.304
1.25	100	0.103	0.109	0.641	0.179	0.256
	200	0.180	0.185	0.741	0.246	0.379
	1000	0.725	0.728	0.861	0.435	0.610
1.5	100	0.277	0.286	0.774	0.316	0.404
	200	0.535	0.544	0.871	0.448	0.582
	1000	0.998	0.998	0.956	0.654	0.830
2	100	0.745	0.755	0.927	0.609	0.692
	200	0.972	0.973	0.956	0.741	0.837
	1000	1.000	1.000	0.998	0.840	0.989

Table 5

Marginalized ZIP Model Results: SafeTalk Example

		Parameter	Model-Based Std Error	Robust Std Error
Zero-Inflation Model				
Intercept	γ_0	1.8485	0.2373	0.2444
Treatment	γ_1	−0.0242	0.2905	0.3488
Site 2	γ_2	0.1055	0.3141	0.3396
Site 3	γ_3	−0.1856	0.5824	0.6183
Baseline UAVI 4	γ_4	−0.1679	0.0421	0.0476
Marginalized Mean Model				
Intercept	α_0	−0.7338	0.2189	0.2335
Treatment	$\boldsymbol{\alpha_1}$	**−0.0666**	0.2630	0.3837
Site 2	α_2	0.3146	0.2863	0.3648
Site 3	α_3	1.4169	0.4974	0.5487
Baseline UAVI	α_4	0.1169	0.0266	0.0378