Published in final edited form as:

Stat Med. 2016 May 10; 35(10): 1722–1735. doi:10.1002/sim.6804.

Marginalized zero-inflated negative binomial regression with application to dental caries

John S. Preisser[a,*], **Kalyan Das**[b], **D. Leann Long**[c], and **Kimon Divaris**[d]

[a]Department of Biostatistics, University of North Carolina

[b]Department of Statistics, University of Calcutta

[c]Department of Biostatistics, West Virginia University

[d]Departments of Epidemiology and Pediatric Dentistry, University of North Carolina

Abstract

The zero-inflated negative binomial regression model (ZINB) is often employed in diverse fields such as dentistry, health care utilization, highway safety, and medicine to examine relationships between exposures of interest and overdispersed count outcomes exhibiting many zeros. The regression coefficients of ZINB have latent class interpretations for a susceptible subpopulation at risk for the disease/condition under study with counts generated from a negative binomial distribution and for a non-susceptible subpopulation that provides only zero counts. The ZINB parameters, however, are not well-suited for estimating overall exposure effects, specifically, in quantifying the effect of an explanatory variable in the overall mixture population. In this paper, a marginalized zero-inflated negative binomial regression (MZINB) model for independent responses is proposed to model the population marginal mean count directly, providing straightforward inference for overall exposure effects based on maximum likelihood estimation. Through simulation studies, the finite sample performance of MZINB is compared to marginalized zero-inflated Poisson, Poisson, and negative binomial regression. The MZINB model is applied in the evaluation of a school-based fluoride mouthrinse program on dental caries in 677 children.

Keywords

caries prevention; count data; excess zeros; marginal models; overdispersion

1. Introduction

Zero-inflated count regression models are widely used to analyze count data that include many zeros in such diverse areas as health care utilization, highway safety, medicine, and population oral health. In dental research, the negative binomial distribution has historically been used for characterizing caries counts owing to the fact that they are routinely over-dispersed relative to the Poisson distribution [1]. As populations have become healthier over time, reported distributions of dental caries indices such as the number of decayed, missing

[*] Correspondence to: Department of Biostatistics, CB #7420 University of North Carolina, Chapel Hill NC 27599-7420. jpreisse@bios.unc.edu.

or filled teeth or surfaces have been increasingly characterized by a preponderance of zero counts in proportions greater than expected under either the Poisson or negative binomial distributions. To account for these excess zeros, zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB) regression models are frequently applied to caries indices [2, 3, 4, 5]. Other applications of ZINB models include modeling lesion counts in the analysis of magnetic resonance imaging for multiple sclerosis [6], mosquito counts in a study of Malaria [7], and vaccine adverse event count data [8].

The evaluation of the caries-preventive effects of a school-based weekly fluoride mouth-rinse (FMR) program on North Carolina (NC) schoolchildren provides a motivating example for the statistical analysis of zero-inflated count data [9]. The main exposure variable is the parental-reported number of years that the student participated in the FMR program. The study is based on a cross-sectional sample of students in grades 1 through 5 who had a clinical examination for caries detection and whose parents participated in the 2003-04 NC Oral Health Survey. The dental exam assessed the caries status of five surfaces (occlusal, mesial, distal, buccal, lingual) for each tooth. In the original published study [9] there were two caries-related outcomes of interest, the count of the number of carious surfaces among primary teeth and the count of the number of affected surfaces for the entire dentition (i.e., primary teeth plus any permanent teeth present, usually molars and incisors). We consider the first of these outcome variables $d_{23}fs$, the number of $\underline{d}$ecayed and $\underline{f}$illed $\underline{s}$urfaces among the primary teeth where subscripted numbers identify the severity of lesions classified as decayed; by convention, the lower case letters in ”$d_{23}fs$” indicate that the index pertains exclusively to the primary dentition [10]. As there are twenty teeth in the primary dentition, the possible range for $d_{23}fs$ is 0 to 100, whereas its range in the FMR data is 0 to 41. Moreover, the sample mean $d_{23}fs$ count is 4.03 and 49% of children have zero counts (Figure 1). Zero-inflated count regression models often provide good fits to counts such as these that exhibit excess zeros and overdispersion. Importantly, ZINB models are applicable when there is interest in a model for latent classes corresponding to a susceptible subpopulation at risk for the disease/condition under study with counts generated from a negative binomial distribution and a non-susceptible subpopulation that provides only zero counts.

Notwithstanding their model-fitting capabilities, the applicability of latent class interpretations of zero-inflated count models requires scrutiny in the search for solutions to substantive problems. The existence of a sub-population that is not at risk for caries (zeros emanating from the logistic part of the model) is questionable; at the very least, the characteristics that would render a child not at risk are unknown. In the rare instance where a child had no natural teeth or where all teeth had already experienced the disease, the child would fall outside the population of interest in a prospective intervention study of the benefits of fluoride mouthrinse and would not be enrolled because there could be no possible and thus quantifiable caries-preventive effect. In prospective as well as cross-sectional studies such as the North Carolina FMR study [9], the use of zero-inflated count regression models in dental caries research appears to be mainly due to their capacity to provide good fits to caries counts; in other words, they provide a convenient mechanism for describing caries counts distributions [4]. Often times their use does not match a study's purported aims as evidenced by the imprecise and misleading interpretations that often accompany their usage [5]. Even if the existence of ‘not-at-risk’ and ‘at-risk’ latent classes is presumed, the question should be asked as to whether there is interest in evaluating interventions (or identifying risk factors) in these two unobserved, hypothetical classes of children, or whether the focus of inference is the overall population from which the study sample is drawn. In the motivating example, the primary research question is whether there is a relationship between number of years of participation in the FMR program and d_{23}fs in the overall population of NC schoolchildren grades 1 to 5.

While the ZINB model regression coefficients have latent class interpretations for ‘at-risk’ and ‘not-at-risk’ subpopulations, researchers in the health sciences frequently seek to make inference on the marginal mean of the overall population. In the FMR example and for caries interventions generally, zero-inflated count models are not well-suited for generating inference concerning overall effects of risk factors and treatments on caries counts in the mixture population that is often of primary interest. In a study of dental caries in very low birth weight adolescents, Albert *et. al.* [11] proposed a causal inference estimator for overall effect of a binary variable in a zero-inflated count model as the average within-subject difference of the marginal means under exposed and unexposed conditions. While being able to flexibly define parameters of interest (e.g., as differences versus ratios of mean counts), their approach may give results that are not generalizable to populations with other configurations of explanatory variables and its extension to continuous exposures is not straightforward. With a similar goal for overall effect estimation, Long *et. al.* [12] proposed a marginalized zero-inflated Poisson (MZIP) regression model with maximum likelihood estimation in the framework of the ZIP model. Instead of modeling the Poisson mean in the ‘at-risk’ latent class, the marginal mean is modeled directly as a function of covariates in conjunction with a logistic model for the probability of an excess zero. As with (log-link) Poisson regression, exponentiating regression coefficients from the marginal mean component of the MZIP model gives incidence density ratios (IDRs) for the overall effects of exposures and covariates on the count outcome.

This article extends the MZIP model to the negative binomial case by drawing upon the marginalized model literature. Heagerty [13] proposed marginalized multilevel models for correlated binary data, which directly model the marginal means whose regression parameters are specified in a likelihood that links the marginal model with a flexibly specified conditional model with random effects. Lee *et al.* [14] proposed marginalized negative binomial hurdle models to analyze clustered data with excess zeros, marginalizing over the random effects. These methods for regression of correlated outcomes combine the desire for population average interpretations with the convenience of estimation with a likelihood function. In a comparatively simple implementation of the principle of marginalization, the marginalized models approach was adapted in the ZIP model by Long *et. al.* [12] 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, their approach marginalizes over the Poisson and degenerate components of the two-part ZIP model to obtain overall effects.

Extending the MZIP model of Long *et. al.* [12], this article introduces a marginalized zero-inflated negative binomial model (MZINB) to model the population mean count directly, allowing straightforward inference for overall exposure effects that accounts for both excess zeros and overdispersion. Section 2 reviews MZIP and the traditional ZIP and ZINB models, section 3 introduces the MZINB model, section 4 presents a simulation study, section 5 presents the statistical analysis of the school-based weekly fluoride mouthrinse program data with respect to caries in the primary dentition, and section 6 further discusses the MZINB model and FMR application.

2. Zero-inflated Count Data Regression Models

2.1. Zero-inflated Poisson and negative binomial models

In ZIP and ZINB models, caries counts arise from a mixture of two latent (i.e., unobserved) classes of subjects [15, 16]. The first model part (defined by a Bernoulli 0/1 process) selects subject *i* with probability ψ_i to be considered not at risk for caries where, conditional on being a member in this class, subject *i* has a zero count with probability one. The second model part provides a caries count with probability 1− ψ_i from a stochastic distribution with mean μ_i for the response counts in an “at-risk” class of subjects. Specifically, let Y_i be a random variable for the *i*-th individual's caries count. The zero-inflated count distribution of the *i*-th individual's caries counts, Y_i, is:

$$\begin{aligned} P(Y_i=0) &= \psi_i + (1-\psi_i)\, g(0|\theta_i) \\ P(Y_i=y_i) &= (1-\psi_i)\, g(y_i|\theta_i), \quad y_i>0, \end{aligned} \qquad (1)$$

where $g(y_i|\theta_i)$ is the either the Poisson probability function with $\theta_i = \mu_i$ so that (1) is the ZIP probability distribution, or the negative binomial with $\theta_i = (\mu_i, a)$ where a is the overdispersion parameter giving the ZINB distribution. Thus, the individuals with zeros include both fixed zeros from ‘not-at-risk’ subjects and random zeros from ‘at-risk’ subjects with group membership being unknown. While the latent classes are sometimes of intrinsic interest, their mixture has been viewed in the dental caries literature as a convenient construct that leads to a statistical distribution for caries counts that accounts for excess zeros [4, 5].

The joint distribution across all *n* individuals in the sample based on equation (1) is

$$f(\mathbf{y}|\boldsymbol{\psi},\boldsymbol{\mu}) = \prod_{y_i=0}\left[\left(\frac{\psi_i}{1-\psi_i}+g(0|\mu_i)\right)(1-\psi_i)\right]\prod_{y_i>0}\left[(1-\psi_i)\, g(y_i|\mu_i)\right] \qquad (2)$$

where y = (y_1, . . . , y_n), ψ = (ψ_1, . . . , ψ_n) and μ = (μ_1, . . . , μ_n). Lambert [15] and Mullahy [16] proposed the model

$$logit(\psi_i) = \boldsymbol{Z}_i'\boldsymbol{\gamma} \quad \text{and} \quad \mathbf{log}(\mu_{\mathbf{i}}) = \mathbf{X}_{\mathbf{i}}'\boldsymbol{\lambda} \qquad (3)$$

where Z_i and X_i are the covariate vectors for the *i*-th individual for excess zeros and Poisson (or negative binomial) processes, respectively. Insertion of equation (3) into equation (2) produces the likelihood function, say $L_{zip}(\gamma, \lambda|y)$. In the traditional ZIP or ZINB model in equation (3), γ and λ have latent class interpretations: γ_j, the element of γ corresponding to the *j*-th covariate, is the multiplicative increase in the log-odds of being an excess zero due to a unit increase in the covariate (z_{ij}) and λ_j is the multiplicative increase in the log-incidence density rate due to a unit increase in the covariate x_{ij} in the susceptible population. In practice, models often have specification $Z_i = X_i$. Occasionally, Z_i consists of a subset of the covariates in X_i.

Challenges arise in model choice because the latent class interpretations of ZIP and ZINB models are not well-suited for caries investigations like the FMR study where interest is in the marginal parameters of the mixture population, specifically, the probability the i-th child has any caries (caries prevalence), $\pi_i = P(Y_i > 0)$, and the marginal mean caries count $\nu_i = E(Y_i)$. In particular, the importance of γ and λ lies in their relationship to π_i and ν_i via $\pi_i = (1 - \psi_i)[1 - g(0|\theta_i)]$ and $\nu_i = (1 - \psi_i)\mu_i$ [11]. When interest is in the effects of explanatory variables on π_i, hurdle models [16, 17] provide an appropriate modeling choice. When interest is in covariate effects on the marginal mean, equation (3) with $Z_i = X_i$ implies $\nu_i = exp\left(X_i'\lambda\right) / \left[1 + exp\left(X_i'\gamma\right)\right]$. It follows that incidence density ratios defined as ratios of two such marginal means where one covariate is allowed to vary while the others are held fixed will generally depend upon γ [5, 11]. Thus, the elements of λ will not have interpretations for covariate effects as incidence density ratios in the mixture population. However, overall effects may be obtained indirectly with post-modelling calculations [2, 15], including counterfactual and log-log model approaches [11]. These approaches require variance estimation for functions of multiple parameters, for which statisticians have suggested the delta method and bootstrap resampling methods. Unfortunately, such indirect methods of overall effects estimation for a binary exposure variable are sufficiently complicated to be inaccessible to many dental research analysts, and they do not have obvious extensions to continuous explanatory variables. Hence, marginalized model approaches are proposed immediately below to provide easy, direct inference for overall effects for count data with many zeros through modeling ν_i instead of μ_i.

2.2. Marginalized zero-inflated Poisson regression model

When the overall mean caries increment $\nu_i = E(Y_i)$ is of primary interest, one may specify a marginalized zero-inflated count response model [12]

$$logit\left(\psi_i\right) = \boldsymbol{Z}_i'\boldsymbol{\gamma} \quad \text{and} \quad \mathbf{log}\left(\nu_\mathbf{i}\right) = \mathbf{X}_\mathbf{i}'\beta \qquad (4)$$

with a straightforward extension available for accommodating varying exposure times/units (e.g., N_i) through introduction of offsets ($\log(N_i)$) in the second model part. While $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_{q-1})'$ models the excess zeros as in the traditional ZIP model, the vector parameter of log(IDR)'s with the intercept β_0 included denoted by $\beta = (\beta_0, \beta_1, \ldots, \beta_{p-1})'$ represents the same overall effect of covariates on caries increment as in Poisson or negative binomial regression. In other words, $\exp(\beta_j)$ represents the multiplicative increase in log-mean count (more generally, log incidence density) for caries in the overall population corresponding to a one-unit increase in the covariate x_{ij}. Adopting ideas from marginalized longitudinal data model approaches [13], β in equation (4) is estimated, accounting for excess zeros, in a maximum likelihood framework via substitution of $\mu_i = \nu_i/(1 - \psi_i)$ into (2) giving

$$f\left(\mathbf{y}|\boldsymbol{\psi},\boldsymbol{\nu}\right)=\prod_{all\quad y_i}\left(1-\psi_i\right)\prod_{y_i=0}\left[\left(\frac{\psi_i}{1-\psi_i}+g\left(0|\psi_i,\nu_i\right)\right)\right]\prod_{y_i>0}\left[g\left(y_i|\psi_i,\nu_i\right)\right]. \quad (5)$$

Long *et. al.* [12] developed procedures for a marginalized zero-inflated Poisson (MZIP) regression model with maximum likelihood estimation of (γ, β) in model (4) where $g\left(y_i|\mu_i\right)=exp\left(-\mu_i\right)\mu_i^{y_i}/y_i!$ in (2) becomes $g(y_i|\psi_i, \nu_i)$ = exp[− $\nu_i/(1-\psi_i)$][$\nu_i/(1-\psi_i)$]$^{y_i}/y_i!$ in (5) with simplification for y_i = 0. Insertion of the model equations (4) into (5) yields the MZIP log-likelihood function

$$\begin{aligned}&L_{mzip}\left(\boldsymbol{\gamma},\beta|\mathbf{y}\right)\\&=\prod_{all\quad y_i}\left(1+e^{\boldsymbol{Z}'_i\boldsymbol{\gamma}}\right)^{-1}\prod_{y_i=0}\left(e^{\boldsymbol{Z}'_i\boldsymbol{\gamma}}+e^{-\left[1+exp\left(\boldsymbol{Z}'_i\boldsymbol{\gamma}\right)\right]exp\left(\mathbf{X}'_i\beta\right)}\right)\\&\times\prod_{y_i>0}\left[\left(1+e^{\boldsymbol{Z}'_i\boldsymbol{\gamma}}\right)^{y_i}e^{\mathbf{X}'_i\beta y_i}e^{-\left[1+exp\left(\boldsymbol{Z}'_i\boldsymbol{\gamma}\right)\right]exp\left(\mathbf{X}'_i\beta\right)}\right]/y_i!\end{aligned}$$

Generally, equations (3) and (4) are non-nested models and L_{zip}(γ, λ|y) is not the same as L_{mzip}(γ, β|y). Instances of equivalence arise when models (3) and (4) are null (no covariates) or saturated, i.e., all covariates are categorical with all possible interactions involving them included in both model parts. Choice of ZIP versus MZIP (or ZINB versus the proposed MZINB in section 3) should depend upon the desired parameter interpretations, as given by λ for the latent class of susceptible persons or by β for the overall population. Model (4) is easy to fit with SAS Proc NLMIXED [12]. The next section extends MZIP to MZINB allowing for extra-Poisson variation in addition to excess zeros.

3. Marginalized Zero-inflated Negative Binomial Regression Model

3.1. Estimation of the distinct-parameters MZINB model

A MZINB regression model for estimation of the (p+q)-vector of distinct parameters (γ, β) in model (4) is introduced within a likelihood framework where $\mu_i = \nu_i/(1-\psi_i)$ is substituted into the negative binomial probability function

$$g\left(y_i|\mu_i,\alpha\right)=\frac{\Gamma\left(y_i+\alpha\right)}{y_i!\Gamma\left(\alpha\right)}\left(\frac{\alpha}{\alpha+\mu_i}\right)^{\alpha}\left(\frac{\mu_i}{\alpha+\mu_i}\right)^{y_i},\quad \text{where}\quad y_i=0,1,\ldots$$

where $var\left(y_i\right)=\mu_i+\phi\mu_i^2$ and $\varphi = 1/\alpha > 0$ so that the dependence of the ZINB density function $g(y_i|\psi_i, \nu_i, \alpha)$ on the marginal mean ν_i is made explicit. It then replaces the ZIP density function $g(y_i|\psi_i, \nu_i)$ in the likelihood expression appearing in equation (5) with simplification for y_i = 0. Model (4) then gives the MZINB likelihood function

$$L_{mzinb}\left(\boldsymbol{\gamma},\beta,\alpha|\mathbf{y}\right)=\prod_{all\ \ y_i}\left(1+e^{\boldsymbol{Z}'_i\boldsymbol{\gamma}}\right)^{-1}\prod_{y_i=0}\left\{e^{\boldsymbol{Z}'_i\boldsymbol{\gamma}}+\left[1+\tfrac{1}{\alpha}\left(1+e^{\boldsymbol{Z}'_i\boldsymbol{\gamma}}\right)e^{\mathbf{X}'_i\beta}\right]^{-\alpha}\right\}$$
$$\prod_{y_i>0}\frac{\Gamma(y_i+\alpha)}{y_i!\Gamma(\alpha)}\left[1+\tfrac{1}{\alpha}\left(1+e^{\boldsymbol{Z}'_i\boldsymbol{\gamma}}\right)e^{\mathbf{X}'_i\beta}\right]^{-\alpha}\left[\frac{\left(1+e^{\boldsymbol{Z}'_i\gamma}\right)e^{\mathbf{X}'_i\beta}}{\alpha+\left(1+e^{\boldsymbol{Z}'_i\boldsymbol{\gamma}}\right)e^{\mathbf{X}'_i\beta}}\right]^{y_i}$$

The log-likelihood of the MZINB model is

$$\begin{aligned} l\left(\gamma,\beta,\alpha|\mathbf{y}\right) =& -\sum_i log\left(1+e^{\boldsymbol{Z}'_i\gamma}\right)+\sum_{y_i=0}log\left\{e^{\boldsymbol{Z}'_i\gamma}+\left[1+\tfrac{1}{\alpha}\left(1+e^{\boldsymbol{Z}'_i\gamma}\right)e^{\mathbf{X}'_i\beta}\right]^{-\alpha}\right\} \\ & -\sum_{y_i>0}log\,y_i!+\sum_{y_i>0}\sum_{j=0}^{y_i-1}log\left(j+\alpha\right)-\sum_{y_i>0}\alpha\,log\left[1+\tfrac{1}{\alpha}\left(1+e^{\boldsymbol{Z}'_i\gamma}\right)e^{\mathbf{X}'_i\beta}\right] \\ & +\sum_{y_i>0}y_i\left[log\left(1+e^{\boldsymbol{Z}'_i\gamma}\right)+\mathbf{X}'_i\beta\right]-\sum_{y_i>0}y_i\,log\left[\alpha+\left(1+e^{\boldsymbol{Z}'_i\gamma}\right)e^{\mathbf{X}'_i\beta}\right]\end{aligned}$$

using the relation $\frac{\Gamma(k+\alpha)}{\Gamma(\alpha)}=\prod_{j=0}^{k-1}(j+\alpha)$ for an integer k. The score equations are given in Appendix 1. Estimation is performed using nonlinear optimization by the quasi-Newton method, which may be implemented in SAS PROC NLMIXED [18].

3.2. A shared-parameter MZINB model

A second model considered is a shared-parameter marginalized zero-inflated negative binomial (SP-MZINB) regression model

$$logit\left(\psi_i\right)=\gamma_0^*+\gamma_1^*\left(\mathbf{X}'_i\beta\right)\quad\text{and}\quad log\left(\nu_i\right)=\mathbf{X}'_i\beta, \qquad (6)$$

which includes only the regression parameters γ_0^* and γ_1^* in addition to β, thus reducing the number of parameters that need to be estimated relative to model (4) with $Z_i = X_i$ nearly in half when p is much greater than 2. In this case, model (6) implies $\gamma_0=\gamma_0^*+\gamma_1^*\beta_0$ and $\gamma_j=\gamma_1^*\beta_j$, $j = 1, \ldots, p - 1$. Thus, the constraint

$$H_0{:}\frac{\gamma_1}{\beta_1}=\frac{\gamma_2}{\beta_2}=\cdots=\frac{\gamma_{p-1}}{\beta_{p-1}}$$

identifies model (6) as being nested within model (4), which allows hypothesis testing for model comparison. A Wald test statistic for testing whether the constraints hold is based on the $p - 2$ dimensional vector $g(\zeta) = (\gamma_1/\beta_1 - \gamma_2/\beta_2, \cdots, \gamma_1/\beta_1 - \gamma_{p-1}/\beta_{p-1})$, where $\zeta = (\gamma', \beta')'$. In particular, $W\sim\chi^2_{p-2}$ where $W=g\left(\hat{\zeta}\right)'\left[\widehat{var}\left(g\left(\hat{\zeta}\right)\right)\right]^{-1}g\left(\hat{\zeta}\right)$ with $var\left(g\left(\hat{\zeta}\right)\right)\approx\left(\partial g/\partial\phi\right)var\left(\hat{\zeta}\right)\left(\partial g/\partial\zeta\right)'$ computed using the delta method [17].

Shared-parameter models analogous to model (6) have been proposed for both traditional ZIP and ZINB models [15, 19] and hurdle models [17, 20]. While the inclusion of $\mathbf{X}'_i\beta$ in both MZINB model parts has its motivation from these other models, SP-MZINB is distinct in both form and purpose. In particular, there is primary interest in β while γ_0^* and γ_1^* are considered to be nuisance parameters much like φ in that they account for the extra-Poisson variation, one source emanating from excess zeros and the other from overdispersion. Nonetheless, interpretations for covariates in the excess zero (logistic) model part are available recognizing that $\gamma_1^*\beta_j$ assumes the role of γ_j in equation (4). It is not surprising that in practice it is almost always that case that $\hat{\gamma}_1^*<0$ because as the overall mean ν_i increases ψ_i tends to decrease. The same can be said for μ_i and ψ_i in equation (3).

4. Simulation Study

Simulations were performed to examine the finite sample properties of MZINB and other count data regression models in an observational study setting similar to the FMR study. Let Y_i be the count outcome of interest for the ith participant, x_{i1} a continuous covariate, and x_{i2} a binary covariate. Also, define x_{i3} as a rescaled count variable, which is the primary exposure variable of interest. In the fluoride mouthrinse example, Y_i is d$_{23}$fs, x_{i1} is family income in units of \$10,000, x_{i2} indicates whether a child had sealants placed, and x_{i3} is the number of years of participation divided by three years in a fluoride mouthrinse program. As minimum benefit would be expected from one year of participation, division by three years enables direct estimation of the effect of three years participation that is a typical duration for public health programs designed to reduce childhood caries. Data were generated from the MZINB model

$$\begin{aligned} logit\left(\psi_i\right) = &\quad \gamma_0+\gamma_1 x_{i1}+\gamma_2 x_{i2}+\gamma_3 x_{i3}, \\ log\left(\nu_i\right) = &\quad \beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3 x_{i3}, \end{aligned}$$

where x_{i1} is lognormally distributed as ln(x_{i1}) ~ N(1.0, (0.75)^2), x_{i2} is dichotomous and generated to be positively associated with x_{i1} as x_{i2} ~ Bernoulli($p_{x_{i1}}$) where $p_{x_{i1}}$ = exp(η_{i1})/[1 + exp(η_{i1})] with $\eta_{i1} = -0.7 + 0.2x_{i1}$; and x_{i3} is C_i/3 where C_i is generated as a count that is positively associated with x_{i2} having a negative binomial distribution with mean 1.5 + 0.5 $*$ x_{i2} and overdisperson parameter $\varphi_{x3} = 1.0$. Thus, dependencies are built into the generated variables to approximate correlations in the fluoride mouthrinse data (Table 1). Together with fixed vectors of β and γ, the independent variables were used to define ψ_i and μ_i, which were employed to randomly generate excess zeros and negative binomial counts, the latter through $\mu_i = \nu_i/(1 - \psi_i)$ and $\varphi = 1$. We assume $\beta_0 = 1.5$, $\beta_1 = -0.1$, $\beta_2 = 0.5$, $\beta_3 = -0.2$, $\gamma_0 = -1$, $\gamma_1 = 0.3$, $\gamma_2 = -1.0$ and $\gamma_3 = 0.1$, which are rounded estimates from the model fit to the FMR data; the only exception is that for the evaluation of Type I error for the hypothesis $H_0 : \beta_3 = 0$ versus $H_0 : \beta_3 \neq 0$, data were generated assuming $\beta_3 = 0$. This model generated counts of which 49% are zeros. Poisson, negative binomial, MZIP and MZINB models were fit to these simulated data using 10,000 simulations for sample sizes of 100, 200, 500 and 1000; each method correctly specified the model for ν_i and used both model-based and empirical standard errors to evaluate the suggestion by Long et al. (2014) that MZIP with empirical standard errors may perform well even when the true model is MZINB. Convergence rates were 100% for Poisson, negative binomial regression and MZIP, and 99.9% or above for MZINB scenarios with $n \geq 200$. For $n = 100$, MZINB convergence was 98.9% (99.0% when $\beta_3 = 0$).

Table 2 reports the percent relative median bias. In general, MZINB gives the least biased estimates and NB regression gives the most biased estimates. Bias decreases with increasing sample size for all models. The maximum likelihood estimator $\hat{\beta}_1$ corresponding to the effect on E(y_i) through the log incidence density ratio for an unit change in the continuous and skewed variate x_{i1} has the greatest amount of bias, whereas the intercept parameter estimate $\hat{\beta}_0$ and $\hat{\beta}_2$ for the binary variate x_{i2}, have the least amount of bias.

Table 3 shows the coverage probabilities of 95% Wald-type confidence intervals for the marginal mean model parameters based on the four methods with model-based and empirical standard errors. The MZINB model has coverage probabilities near the nominal 0.95 level when model-based standard errors are used, whereas it results in slight undercoverage for $n = 100$ when empirical standard errors are used. The Poisson and MZIP models with model-based standard errors result in gross undercoverage, whereas the NB model with model-based or empirical standard errors gives particularly poor coverage for the skewed covariate x_{i1}, and undercoverage, generally. Poisson regression with empirical standard errors gives good coverage for x_{i2} and x_{i3} with undercoverage for x_{i1}. MZIP with empirical standard errors gives coverage that is comparable to MZINB with empirical standard errors. Interestingly, MZIP with empirical standard errors performs better than Poisson regression with empirical standard errors for β_1 corresponding to the skewed continuous covariate, but performs worse for β_2 and β_3 when $n = 100$ or $n = 200$.

Table 4 reports type I errors of two-sided Wald tests at the nominal 0.05 significance level for the primary exposure variable of interest x_{i3}. In particular, test size was defined as the proportion of the 10,000 simulation replicates that incorrectly reject $H_0 : \beta_3 = 0$ in favor of $H_1 : \beta_3 \neq 0$. MZINB with model-based standard errors was the best method with respect to maintaining the nominal 0.05 significance level for all four sample sizes. For $n = 500$ and $n = 1000$, all four models with empirical standard errors performed acceptably well in addition to negative binomial regression with model-based standard errors. Poisson and MZIP models with model-based standard errors performed very poorly as illustrated by their grossly inflated Type I errors at all sample sizes.

Finally, Table 5 shows the finite sample efficiency of Poisson, NB and MZIP regression relative to the MZINB model. For $n = 500$ or $n = 1000$, all three models are inefficient with respect to MZINB. Interestingly, Poisson regression was more efficient than NB or MZIP regression, generally, and more efficient than MZINB for n=100.

Traditional ZIP and ZINB models are intentionally excluded from the simulation study because β does not exist in these models. In contrast to the "average" treatment effects that could be obtained from ZIP and ZINB from post-modeling calculations [11], components of the marginalized model parameter-vector β are exposure effects that are homogeneous across the levels of the other covariates in the model. The distinctive role of MZINB relative to ZINB is illustrated in the next section with emphasis on interpretations of exposure and treatment effects in the evaluation of a school-based fluoride mouthrinse program.

5. Application to a School-based Fluoride Mouthrinse Program

Several models for count outcomes are applied to the dental caries data from the FMR study. The data on *n* = 677 children were previously analyzed by Divaris, Rozier and King [9]; the sample size of 1363 children reported in their Table 2 was before those with missing covariate data were excluded. The main exposure variable is the parent report of years of participation in the FMR program that, prior to rescaling, ranges from 0 to 6 years with a mean of 1.04 years. As noted in the introduction, the outcome is d_{23}fs, the number of decayed and filled tooth surfaces among primary teeth (Figure 1). The independent variables vector is $\mathbf{X}_i'=(x_{i1},\ldots,x_{i11})'$ where x_{i1} is family income (in units of $10,000); x_{i2} is whether child had sealants placed (1=yes, 0=no); x_{i3} is the number of FMR years divided by 3; x_{i4}, x_{i5} and x_{i6} are the linear, quadratic and cubic effects, respectively, for child's age, mean-centered at 8.35 years; x_{i7} is dental home established (1=yes vs. 0=no); x_{i8} is 1 if child needed dental care but could not get it and 0 otherwise; x_{i9} is 0 if brushes less than once a day, 1 if once a day, and 2 if twice or more a day; and x_{i10} and x_{i11} are indicator variables for African American and other race, respectively, with whites as the reference.

In order to emphasize differences in model interpretation, the traditional ZINB model in equation (3) is initially considered, where $\boldsymbol{Z}_i'=\mathbf{X}_i'$ for the *i*-th child. Results obtained from Proc GENMOD in SAS v. 9.3 using the “dist=ZINB” option on the model statement are shown in Table 6. The estimated IDR comparing three years to zero years of participation in the FMR program for the ‘at-risk’ class of children is $exp\left(\hat{\lambda}_3\right)$ or exp(−0.077) = 0.926, with 95% CI (0.696, 1.232). Thus, with the other covariates held fixed, the mean caries index μ_i for a child in the ‘at-risk’ class participating for three years in the FMR program is approximately 92% the mean caries index for a child in the ‘at-risk’ class not participating in the FMR program. As the confidence interval contains 1.0, the result is not statistically significant at the 0.05 significance level.

Next, the MZINB model with $\boldsymbol{Z}_i'=\mathbf{X}_i'$ and the form shown in equation (4) is fitted in order to model the marginal mean caries index ν_i directly with β_j being the overall log incidence density ratio for the *j*-th factor, *j* = 1, . . . , 11. From the results in Table 6, the estimated IDR comparing three years to zero years of participation in the FMR program for the overall population of children is $exp\left(\hat{\beta}_3\right)$ or exp(−0.113) = 0.89 with 95% CI (0.63, 1.15). Thus, all other covariates fixed, the mean caries index ν_i for a child in the *overall* population with three years participation in the FMR program is approximately 89% of the mean caries index of a child with zero years participation. However, there are no statistically significant treatment differences since the IDR is not significantly different than 1.

The last model considered is the shared-parameter marginalized zero-inflated negative binomial (SP-MZINB) regression model in equation (6). Excepting the polynomial terms for age, the mean model parameter estimates have the opposite sign of their excess zero model counterparts for all eight covariates in both the ZINB and MZINB models. More importantly, results in Table 6 for SP-MZINB show that the IDR in the mixture population for the *overall* effect of three additional years of participation in the FMR program (exp(β_3))

is estimated as exp(−0.093) = 0.91 with 95% CI (0.70, 1.12). The SAS Proc NLMIXED code that produced these results is given in Appendix 2.

The smallest AIC among the models in Table 6 is for the SP-MZINB model indicating that it provides the best fit. Furthermore, $W = 7.2$ with 10 degrees of freedom is not statistically significant (p=0.71) supporting the goodness-of-fit of SP-MZINB. Although not shown, the corresponding (one-part) negative binomial regression model with −2logLik = 2922.9 and AIC=2972.9 had poorer fit than that of each of the models reported in Table 6. Note that standard errors for the $\hat{\beta}_j$, $j = 1, \ldots, 11$, are smallest for the SP-MZINB model illustrating that variance reduction may be achieved through use of a parsimonious model. However, the overall effect of FMR participation is also the smallest for this model, and correspondingly its p-value for FMR is not the smallest among the models considered. Finally, the MZINB model has a slightly smaller AIC than ZINB, while having similar model complexity as they have the same number of parameters.

6. Conclusion

The marginalized zero-inflated negative binomial regression model proposed in this article for count data exhibiting overdispersion and having many zeros directly models the marginal mean of a mixture of two discrete distributions, one consisting of negative binomial 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 ZINB 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 ZINB model, MZINB provides direct estimates of a homogeneous exposure effect that does not require post-modeling computations. Indeed, the marginal exposure effects of MZINB are given by log incidence density ratios that have the same interpretations as in negative binomial or Poisson regression. The logistic model part for excess zeros in MZINB is of secondary interest as its role is to provide adjustment for extra-negative binomial variation due to excess zeros.

In simulations reported in section 4, the correctly specified MZINB model had excellent finite sample performance as did MZIP with empirical standard errors. Surprisingly, Poisson regression with empirical standard errors performed almost as well. In particular, our simulation results for a binary covariate (x_{i2}) were in basic agreement with simulation results reported in a technical report ([18]) that demonstrated that Poisson regression with empirical standard errors may have good finite sample properties for categorical covariates even when the true model is MZINB. However, for a skewed covariate generated from a lognormal distribution (x_{i1}), Poisson regression with empirical standard errors resulted in biased regression coefficients and undercoverage of nominal 95% confidence intervals. Similarly, in a simulation study reported elsewhere that generated counts from an MZIP model, the MZIP model produced less biased regression parameter estimates and confidence interval coverage closer to the nominal level than Poisson regression with empirical standard errors for a log-normally distributed covariate [12].

The ZINB model was not considered in the simulation study of section 4 because (1) the estimand β_3 in the MZINB model, which is a homogeneous exposure effect of the FMR program across all possible combinations of levels of covariates, does not exist in the ZINB model; and (2) the exposure x_{i3} is a scaled count variable so that defining an ‘average’ effect is not straightforward. Unlike for a binary exposure, a measure of ’average’ effect obtained for a count or continuous exposure from post-ZINB modeling computations depends on the choice of the two levels to be compared, i.e., the ’average’ IDR comparing k versus $k + 1$ years of participation in the FMR program depends on k. In the supplemental file to Long et al. (2014), a simulation study involving a binary exposure and a skewed (lognormal) covariate compared MZIP to two ZIP-based estimators of ‘average’ effect. The study found that the ZIP estimators including the average predicted value method ([11]) resulted in increased absolute bias, undercoverage of 95% confidence intervals and inflated Type I error of the homogeneous exposure effect relative to the correctly specified MZIP model for $n =$ 1000. When the covariate was binary there was little difference in model performance between MZIP and ZIP methods.

Notwithstanding the distinct interpretations of the FMR program effect based on the IDR for the ’at-risk’ class and the IDR for the overall population, there was little difference in their respective estimates of 0.93 from the traditional ZINB model and 0.89 from the MZINB model in section 5. The reason for their similar values is because the effect of FMR program on excess zeros (γ_3) in either ZINB or MZINB models was small relative to the effects of other covariates on excess zeros (Table 6). This claim may be supported by defining the IDR for the at-risk class for the i-th individual as a function of the MZINB model parameters. Using the simulation set-up in section 4 for simplicity and noting that $\mu_i = \nu_i/(1 - \psi_i)$, one can write

$$\mu_i = [1+exp(\gamma_1+\gamma_1 x_{i1}+\gamma_2 x_{i2}+\gamma_3 x_{i3})]\ exp(\beta_0+\beta_1 x_{i1}+\beta_1 x_{i2}+\beta_1 x_{i3})\,.$$

Letting $s_i = 1$ indicate the at-risk class, the IDR corresponding to the FMR program effect for the at-risk class is

$$\frac{E(Y_i|s_i=1, x_{i1}, x_{i2}, x_{i3}=1)}{E(Y_i|s_i=1, x_{i1}, x_{i2}, x_{i3}=0)} = e^{\beta_3}\frac{[1+exp(\gamma_1+\gamma_1 x_{i1}+\gamma_2 x_{i2}+\gamma_3)]}{[1+exp(\gamma_1+\gamma_1 x_{i1}+\gamma_2 x_{i2})]} = e^{\beta_3}\frac{[1+exp(-1+0.3x_{i1}-x_{i2}+\gamma_3)]}{[1+exp(-1+0.3x_{i1}-x_{i2})]}$$

where the last expression is based on inserting excess zero model parameter values from section 4. The average IDR for the at-risk class can then be approximated by inserting the mean simulation values E(x_{i1}) = exp(1.0 + $(0.75)^2$/2) = 3.6 for x_{i1} and E($x_{i2}|x_{i1}$ = 3.6) = 0.50 for x_{i2}, reducing the last expression above to e^{β_3}[1 + exp(−0.42 + γ_3)]/1.657. It follows that when γ_3 = 0.1 as in the simulation study, the IDR for the at-risk is 0.854 compared to the overall IDR of e^{β_3} = 0.818. If γ_3 had equaled 0 instead of 0.1, the last fraction in the above equation would have equaled one and e^{β_3} would have represented both at-risk class and overall IDRs. In contrast, when γ_3 equals the relatively large values of 0.5 or 1.0, the IDR for the at-risk class becomes 1.031 or 1.379, respectively, which are much different from the overall IDR. Therefore, considering that the simulation design in section 4 was based on the application in section 5 that had a small value for γ_3, the estimated IDR for the at-risk class based on the traditional ZINB would have been similar to the IDR for the overall effect had the traditional ZINB model been included in the simulation study.

Future study could address the properties of estimators of overall exposure effects in the SP-MZINB model, which were not studied in the simulations nor in the asymptotic efficiency calculations in this article. Shared-parameter models for count outcomes, including those for traditional zero-inflated and hurdle models [15, 19, 20], have two advantages. First, by having fewer parameters to estimate, their use may circumvent computational problems of fitting distinct-parameter models. Second, shared-parameter count models often result in more efficient estimation than distinct-parameter models for the exposure effects, as demonstrated for a zero-altered negative binomial (hurdle) model [17]. As the SP-MZINB model is a reduced model relative to the distinct-parameter MZINB, it is expected to be more efficient when it holds. The smaller standard errors observed for the SP-MZINB compared to the MZINB for the FMR study reported in Table 6 are consistent with this conjecture.

In addition to the simulation results, goodness-of-fit considerations supported use of the SP-MZINB model in an application examining the effectiveness of a school-based fluoride mouthrinse program in children grades 1 through 5. These were retrospectively gathered observational data from a sample survey of schoolchildren carried out in North Carolina in 2003-2004 whereby children participated in an FMR program from between 0 and 6 years. Because the classroom or school to which a child belongs can change frequently, assignment of the child to a cluster for the purposes of conducting a multilevel statistical analysis was not straightforward and was not undertaken in the data analysis in this article. Where cluster membership can reasonably be assigned, marginalized zero-inflated count regression models can be extended to include random effects, as recently demonstrated for the MZIP model in a longitudinal data analysis of a motivational interviewing-based safer sex intervention [21]. An MZINB model with random effects could be useful in a cluster trial for caries prevention where failure to adjust for clustering may inflate the Type I error of tests for the intervention effect possibly resulting in spurious findings of significance [22].

Despite the increasing popularity of the ZINB 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, many dental caries researchers have interpreted the negative binomial regression parameters of the ZINB model with respect to the overall caries incidence, rather than the correct model-based interpretation relating to caries incidence within the presumed at-risk population [5]. This pattern of misinterpretation suggests that investigators when genuinely interested in marginal inference for count data may choose ZINB models simply because of goodness-of-fit considerations for data with many zeros. Unless post-modeling calculations are performed to obtain exposure effects on the marginal mean count, the research analyst may unwittingly alter the target of inference. Use of the MZINB model maintains the marginal mean as the target of inference through direct modeling while accounting for excess zeros.

Generally, the research goal should guide 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-of-fit considerations prevail. This approach to model selection based on collaboration between investigators and biostatistical scientists discourages purely empirical model fitting. Indeed, such exercises often reveal ZINB, MZINB, and negative binomial hurdle models of comparable complexity to have similar goodness-of-fit statistics as was the case in the FMR study [8]. The MZINB model belongs to a different model class than the traditional ZINB model and so choosing between them should be based on the research question. Future research could extend the MZINB model to include random effects, incorporate sampling weights, handle missing covariates or develop inferential methods for marginal mean regression models based on other finite mixture distributions.

Acknowledgement

The authors thank R. Gary Rozier, Department of Health Policy and Management, University of North Carolina, for providing the dental caries data. The original study of the fluoride mouthrinse program was supported by a grant from the Centers for Disease Control and Prevention U48/CCU415769. John Preisser acknowledges partial support from NC TraCS NIH UL1 TR001111-01. D. Leann Long acknowledges partial support from grant U54GM104942. Kimon Divaris acknowledges partial support from NIH grant U01DE025046. Part of this work was done while Kalyan Das was Pranab K. Sen Distinguished Visiting Professor in the Department of Biostatistics, University of North Carolina in 2013.

7. Appendix

Appendix 1. Score equations for the marginalized zero-inflated negative binomial model

The MZINB model score equations are obtained by differentiating the log likelihood in section 3 with respect to the model parameters (γ, β, α). Noting that $\mu_i = e^{\mathbf{X}'_i\beta}\left(1+e^{\boldsymbol{Z}'_i\gamma}\right)$ and defining $\theta_i = \alpha/(\alpha + \mu_i)$, the score equations are:

$$\begin{aligned}
\frac{\partial l}{\partial \beta} = & \sum_i \left\{ I\left(y_i > 0\right) \left[y_i - \mu_i \left(\frac{\alpha + y_i}{\alpha + \mu_i} \right) \right] - I\left(y_i = 0\right) \frac{\theta_i^{\alpha+1} \mu_i}{e^{\boldsymbol{Z}'_i\gamma} + \theta_i^{\alpha}} \right\} X'_i \\
\frac{\partial l}{\partial \alpha} = & \sum_i \left\{ I\left(y_i = 0\right) \left[\frac{\theta_i^{\alpha}\left(1 - \theta_i + ln\theta_i\right)}{e^{\boldsymbol{Z}'_i\gamma} + \theta_i^{\alpha}} \right] + I\left(y_i > 0\right) \left[ln\theta_i - \frac{y_i - \mu_i}{\alpha + \mu_i} + \sum_{j=0}^{y_i - 1} \frac{1}{j + \alpha} \right] \right\} \\
\frac{\partial l}{\partial \gamma} = & \sum_i \left[\left(y_i - 1\right) + I\left(y_i = 0\right) \left\{ \frac{1 + e^{\boldsymbol{Z}'_i\gamma} - \mu_i \theta_i^{\alpha+1}}{e^{\boldsymbol{Z}'_i\gamma} + \theta_i^{\alpha}} \right\} - I\left(y_i > 0\right) \mu_i \left(\frac{\alpha + y_i}{\alpha + \mu_i} \right) \right] \psi_i Z'_i
\end{aligned}$$

The model-based asymptotic covariance estimator of $\hat{\zeta}$ where $\zeta = (\gamma, \beta, \alpha)'$ is computed as the inverse of the Fisher information matrix $I(\zeta) = -\mathrm{E}(\partial^2 l/\partial\zeta\partial\zeta')$.

Appendix 2. SAS code for the shared-parameter marginalized zero-inflated negative binomial model

The following SAS code is used to fit the SP-MZINB model to the mouthrinse program data using model-based standard errors. The “parms” statement is used to specify starting values for the parameters b0, . . . , b11, which are the components of β; estimates for β and φ from the negative binomial regression model are used as initial estimates.

```
proc nlmixed data= fmr /* empirical */;
title “MZINB shared-parameters model”;
parms gamma0=0.5 gamma1=−0.5 phi=1
b0= 1.72 b1=−0.17 b2=0.65 b3=−0.15 b4=−0.01 b5=−0.04 b6=−0.02
b7= 0.46 b8= 0.46 b9=−0.19 b10=−0.37 b11=0.51;
nu = exp(b0+b1*income10+b2*sealants+b3*fmryears3+b4*age_c+b5*age2_c+b6*age3_c
+
b7*dhome+b8*noaccess+b9*brushing+b10*race_black+b11*race_other /* + c1 */);
>linpinfl = gamma0 + gamma1*(log(nu));
>psi = 1/(1+exp(-linpinfl));
>mu = nu/(1-psi);
>alpha = 1/phi;
>theta = 1/(1+(mu/alpha));
>if dfsd=0 then loglike =log(psi + (1-psi)*(theta**alpha));
>else loglike = log(1-psi) + lgamma(dfsd+alpha) - lgamma(alpha)
>+ dfsd*log(1-theta)+alpha*log(theta) - lgamma(dfsd+1);
>model dfsd ~ general(loglike);
>/* random c1 ~normal(0,0) SUBJECT=id; */
>estimate ’fmr years: d23fs density ratio for three additional years’
exp(b3);
run;
```

Note that “/*” and “*/” are used in combination for comments or, in this case, to de-activate code. Removing all instances of “/*” and “*/” will provide results based on empirical standard errors. For computation of empirical standard errors invoked with the “empirical” option, SAS Proc NLMIXED (version 9.3 or 9.4) requires specification of a **random** statement. However, empirical standard errors for the independent count data models considered in this article are obtained using the “trick” of specifying a random subject-level intercept with zero variance; it is necessary to define a subject level variable (“id”) in the dataset **fmr** even though there is only one observation per subject.

References

1. Grainger RM, Reid DBW. Distribution of dental caries in children. Journal of Dental Research. 1954; 33:613–623. [PubMed: 13201694]
2. Bohning D, Dietz E, Schlattmann P, Mendonca 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.
3. Lewsey J, Thomson W. The utility of the zero-inflated Poisson and zero-inflated negative binomial models: a case study of cross-sectional and longitudinal DMF data examining the effect of socio-economic status. Community Dentistry and Oral Epidemiology. 2004; 32:183–189. [PubMed: 15151688]

4. Mwalili SM, Lesaffre E, Declerck D. The zero-inflated negative binomial regression model with correction for misclassification: an example in caries research. Statistical Methods in Medical Research. 2008; 17:123–139. [PubMed: 17698937]

5. Preisser JS, Stamm JW, Long DL, Kincade M. Review and recommendations for zero-inflated count regression modeling of dental caries indices in epidemiological studies. Caries Research. 2012; 46:413–423. [PubMed: 22710271]

6. Francois M, Peter C, Gordon F. Dealing with excess of zeros in the statistical analysis of magnetic resonance imaging lesion count in multiple sclerosis. Pharmaceutical Statistics. 2012; 11:417–424. [PubMed: 22888087]

7. Lawal BH. Zero-inflated count regression models with applications to some examples. Quality & Quantity. 2012; 46:19–38.

8. Rose CE, Martin SW, Wannemuehler KA, Plikaytis BD. On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. Journal of Biopharmaceutical Statistics. 2006; 16:463–481. [PubMed: 16892908]

9. Divaris K, Rozier RG, King RS. Effectiveness of a school-based fluoride mouthrinse program. Journal of Dental Research. 2012; 91:282–287. [PubMed: 22202124]

10. Klein H, Palmer CE, Knutson JW. Studies on dental caries: I. Dental status and dental needs of elementary school children. Public Health Reports. 1938; 53:751–765.

11. Albert J, Wang W, Nelson S. Estimating overall exposure effects for zero-inflated regression models with application to dental caries. Statistical Methods in Medical Research. 2014; 23:257–278. [PubMed: 21908419]

12. Long DL, Preisser JS, Herring AH, Golin CE. A marginalized zero-inflated Poisson regression model with overall exposure effects. Statistics in Medicine. 2014; 33:5151–5165. [PubMed: 25220537]

13. Heagerty PJ. Marginally specified logistic-normal models for longitudinal binary data. Biometrics. 1999; 55:688–698. [PubMed: 11314994]

14. Lee K, Joo Y, Song JJ, Harper DW. Analysis of zero-inflated clustered count data: a marginalized model approach. Computational Statistics and Data Analysis. 2011; 55:824–837.

15. Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics. 1992; 34:114.

16. Mullahy J. Specification and testing of some modified count data models. Journal of Econometrics. 1986; 33:341–365.

17. Preisser JS, Das K, Benecha H, Stamm JW. Logistic regression for dichotomized counts. Statistical Methods in Medical Research. (published online May 26, 2014). DOI: 10.1177/0962280214536893.

18. Preisser JS, Das K, Long DL, Stamm JW. A marginalized zero-inflated negative binomial regression model with overall exposure effects. The University of North Carolina at Chapel Hill Department of Biostatistics Technical Report Series. 2014; 43 Working Paper http://biostats.bepress.com/uncbiostat/papers/art43.

19. Heilbron D. Zero-altered and other regression models for count data with added zeros. Biometrical Journal. 1994; 36:531–547.

20. Min Y, Agresti A. Random effects models for repeated measures of zero-inflated count data. Statistical Modelling. 2005; 5:1–19.

21. Long DL, Preisser JS, Herring AH, Golin CE. A Marginalized Zero-Inflated Poisson Regression Model with Random Effects. Journal of the Royal Statistical Society, Series C (Applied Statistics). (published online April 30, 2015). DOI: 10.1111/rssc.12104.

22. Burnside G, Pine CM, Williamson PR. Statistical aspects of design and analysis of clinical trials for the prevention of caries. Caries Research. 2006; 40:360–365. [PubMed: 16946602]

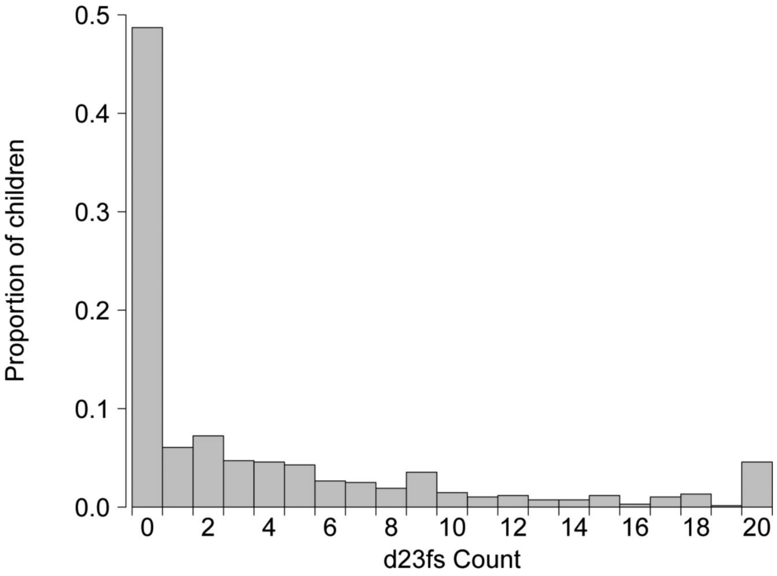

Figure 1.

Distribution of d_{23}fs for 677 children grades 1 to 5 participating a school-based fluoride mouthrinse program. The bar at the label of '20' on the x-axis represents 31 children (4.6%) with d_{23}fs ranging from 20 to 41.

Table 1

Descriptive statistics of simulated variables and variables observed in the school-based mouth rinse program.

Simulated data[†]			**Pearson correlations**[††]				**Observed data (n=677)**		
variable	**mean**	**sd**	$\boldsymbol{y}$	$\boldsymbol{x_1}$	$\boldsymbol{x_2}$	$\boldsymbol{x_3}$	**variable**	**mean**	**sd**
y	3.80	6.36	l	−0.10	0.18	−0.03	$\mathrm{d}_{23}\mathrm{fs}$	4.03	6.65
x_1	3.60	3.13	−0.11	1	0.17	−0.01	income	3.82	2.11
x_2	0.50	0.50	0.11	0.24	1	0.13	sealant	0.48	0.50
x_3	0.58	0.74	−0.07	0.03	0.11	1	(fmr years)/3	0.35	0.49

The upper triangular matrix contains correlations from the FMR study.

[†] Based on 10 million (10,000 replications of n=1000) generated observations.

[††] The lower triangular matrix entries contain correlations from simulated data.

Table 2

Percent relative median bias[†] of estimated regression coefficients in the marginal mean model.

Parameter	n	Poisson	NB	MZIP	MZINB
β_0	100	1.97	8.27	1.73	−0.29
	200	1.34	5.93	1.64	−0.10
	500	0.81	3.96	1.08	−0.13
	1000	0.69	3.06	1.09	0.16
β_1	100	18.84	56.17	26.35	2.93
	200	11.75	36.97	17.45	1.71
	500	7.47	23.49	12.08	0.86
	1000	4.87	16.86	9.16	0.11
β_2	100	1.50	9.86	7.76	−0.38
	200	1.54	6.97	5.60	0.74
	500	0.63	3.81	3.42	0.46
	1000	−0.02	2.26	2.58	−0.10
β_3	100	4.30	11.56	8.06	4.75
	200	3.20	7.98	4.25	2.97
	500	0.90	3.06	−0.58	1.34
	1000	0.49	1.87	−1.48	1.11

[†] $100\% \times med\left\{\left(\hat{\beta}_j - \beta_j\right)/\beta_j\right\}$ where $med\{\cdot\}$ is the median of the converged simulation replicates.

Table 3

Coverage of 95% Wald-type confidence intervals with model-based (MB) or empirical (Emp) standard errors for marginal mean regression model parameters.[†]

		Poisson		**NB**		**MZIP**		**MZINB**	
Parameter	**n**	**MB**	**Emp**	**MB**	**Emp**	**MB**	**Emp**	**MB**	**Emp**
β_1	100	0.424	0.894	0.742	0.704	0.696	0.918	0.936	0.911
	200	0.403	0.906	0.728	0.743	0.689	0.932	0.945	0.931
	500	0.379	0.915	0.691	0.780	0.678	0.935	0.949	0.943
	1000	0.361	0.920	0.661	0.793	0.665	0.939	0.948	0.943
β_2	100	0.465	0.936	0.929	0.909	0.714	0.926	0.939	0.933
	200	0.457	0.945	0.936	0.923	0.716	0.937	0.944	0.941
	500	0.456	0.948	0.943	0.932	0.719	0.944	0.947	0.947
	1000	0.461	0.947	0.946	0.935	0.720	0.946	0.949	0.947
β_3	100	0.502	0.920	0.916	0.860	0.707	0.908	0.939	0.907
	200	0.495	0.931	0.927	0.891	0.705	0.923	0.941	0.922
	500	0.494	0.942	0.938	0.921	0.713	0.939	0.948	0.936
	1000	0.490	0.947	0.943	0.933	0.715	0.945	0.952	0.944

[†] Based on 10,000 simulation replicates.

Table 4

Type I error[†] of model-based (MB) and empirical (EMP) Wald tests for β_3 in the marginal mean model.[††]

		Poisson		NB		MZIP		MZINB	
Parameter	**n**	**MB**	**Emp**	**MB**	**Emp**	**MB**	**Emp**	**MB**	**Emp**
β_3	100	0.565	0.096	0.094	0.152	0.358	0.119	0.062	0.095
	200	0.572	0.080	0.080	0.116	0.368	0.093	0.057	0.081
	500	0.571	0.066	0.068	0.082	0.361	0.072	0.053	0.067
	1000	0.571	0.062	0.064	0.070	0.369	0.064	0.051	0.059

[†] Type I error for testing $H_0 : \beta_3 = 0$ vs. $H_0 : \beta_3 \neq 0$.

[††] Based on 10,000 simulation replicates.

Table 5

Efficiency[†] of models for the marginal mean relative to marginalized negative binomial regression.

Parameter	n	Poisson	NB	MZIP
β_1	100	1.170	0.606	0.734
	200	1.020	0.555	0.695
	500	0.845	0.487	0.658
	1000	0.764	0.433	0.633
β_2	100	1.005	0.788	0.786
	200	0.989	0.795	0.792
	500	0.970	0.777	0.794
	1000	0.964	0.772	0.789
β_3	100	1.067	0.782	0.730
	200	0.999	0.784	0.741
	500	0.951	0.764	0.758
	1000	0.939	0.755	0.757

[†] Ratio of the mean squared error (mse) of the MZINB model divided by the mse of the model in the column heading; mse is computed as the sum of the monte carlo variance of $\hat{\beta}_j$ and its squared bias.

Table 6

ZINB and MZINB results for the analysis of the effect of a school-based fluoride mouthrinse (FMR) program and other covariates on dental caries ($d_{23}fs$) in first through fifth grade children (n=677).

	ZINB			MZINB			SP-MZINB	
Variable	**parm**	**est**	**mbse**	**parm**	**est**	**mbse**	**est**	**mbse**
		at-risk class mean			*marginal mean*		*marginal mean*	
Intercept	λ_0	1.565	$0.322^{\dagger}$	β_0	1.524	$0.336^{\dagger}$	1.845	$0.302^{\dagger}$
Fam. income	λ_1	−0.011	0.031	β_1	−0.118	$0.033^{\dagger}$	−0.161	$0.029^{\dagger}$
sealants	λ_2	0.341	$0.123^{\#}$	β_2	0.673	$0.133^{\dagger}$	0.647	$0.124^{\dagger}$
FMR (3 years)	λ_3	−0.077	0.146	β_3	−0.113	0.148	−0.093	0.118
age	λ_4	−0.110	0.082	β_4	−0.047	0.088	−0.001	0.069
age-sq	λ_5	−0.008	0.026	β_5	−0.047	0.028	−0.050	0.023^{*}
age-cu	λ_6	−0.001	0.015	β_6	−0.016	0.016	−0.015	0.012
dental home	λ_7	0.395	0.183^{*}	β_7	0.483	$0.188^{\#}$	0.279	0.152
No access	λ_8	0.275	0.157	β_8	0.413	$0.160^{\#}$	0.305	0.137^{*}
brushing freq.	λ_9	−0.047	0.110	β_9	−0.138	0.113	−0.127	0.096
African American	λ_{10}	−0.282	0.156	β_{10}	−0.419	$0.163^{\#}$	−0.320	0.137^{*}
Other race	λ_{11}	0.150	0.280	β_{11}	0.275	0.318	0.164	0.294
Dispersion	φ	0.980	0.143	φ	0.963	0.143	0.994	0.144
				logistic regression model for the probability of an excess zero				
Intercept	γ_0	−1.965	$0.686^{\#}$	γ_0	−2.116	$0.695^{\#}$		
Fam. income	γ_1	0.268	$0.061^{\dagger}$	γ_1	0.314	$0.062^{\dagger}$		
sealants	γ_2	−0.866	$0.227^{\dagger}$	γ_2	−0.979	$0.228^{\dagger}$		
FMR (3 γ_{ears})	γ_3	0.163	0.256	γ_3	0.152	0.235		
age	γ_4	−0.160	0.141	γ_4	−0.098	0.131		
age-sq	γ_5	0.095	0.045^{*}	γ_5	0.091	0.044^{*}		
age-cu	γ_6	0.034	0.023	γ_6	0.031	0.022		
dental home	γ_7	−0.292	0.325	γ_7	−0.255	0.302		
No access	γ_8	−0.364	0.301	γ_8	−0.297	0.276		
brushing freq.	γ_9	0.329	0.210	γ_9	0.309	0.206		
African American	γ_{10}	0.404	0.277	γ_{10}	0.524	0.277		
Other race	γ_{11}	−0.608	0.625	γ_{11}	−0.133	0.589		
Intercept, SP-MZINB	sgu						1.645	$0.345^{\dagger}$
Slope, SP-MZINB	sgu						−1.626	$0.274^{\dagger}$
−2logLik		2922.9			2918.4		2932.2	
AIC		2972.9			2968.4		2962.2	

est=estimate; mbse=model-based standard error

[*] $p < 0.05$

[#] $p < 0.01$

[†] $p < 0.001$.