Matching the Statistical Model to the Research Question for Dental Caries Indices with Many Zero Counts

John S. Preisser^a D. Leann Long^c John W. Stamm^b

Department of ^aBiostatistics and ^bDental Ecology, University of North Carolina at Chapel Hill, Chapel Hill, NC, and

^cDepartment of Biostatistics, University of Alabama at Birmingham, Birmingham, AL, USA

Keywords

Dental surveys · Excess zeros · Oral epidemiology · Poisson regression · Zero inflation

Abstract

Marginalized zero-inflated count regression models have recently been introduced for the statistical analysis of dental caries indices and other zero-inflated count data as alternatives to traditional zero-inflated and hurdle models. Unlike the standard approaches, the marginalized models directly estimate overall exposure or treatment effects by relating covariates to the marginal mean count. This article discusses model interpretation and model class choice according to the research question being addressed in caries research. Two data sets, one consisting of fictional dmft counts in 2 groups and the other on DMFS among schoolchildren from a randomized clinical trial comparing 3 toothpaste formulations to prevent incident dental caries, are analyzed with negative binomial hurdle, zero-inflated negative binomial, and marginalized zero-inflated negative binomial models. In the first example, estimates of treatment effects vary according to the type of incidence rate ratio (IRR) estimated by the model. Estimates of IRRs in the analysis of the randomized clinical trial were similar despite their distinctive interpreta-

tions. The choice of statistical model class should match the study's purpose, while accounting for the broad decline in children's caries experience, such that dmft and DMFS indices more frequently generate zero counts. Marginalized (marginal mean) models for zero-inflated count data should be considered for direct assessment of exposure effects on the marginal mean dental caries count in the presence of high frequencies of zero counts.

© 2017 S. Karger AG, Basel

Childhood dental caries, which can be associated with adverse health, social, and economic consequences, is a major research focus of oral health scientists. Count regression models for dental caries indices (e.g., DMFS, dmft) are useful for understanding risk factors and describing variations in caries prevalence and severity in cross-sectional dental surveys or in caries increment in prospective epidemiological investigations and in randomized clinical trials. In all these cases, count regression models are used to model a single count outcome for individuals whose responses are assumed to be independent of one another. Extensions of count regression models to clustered or repeated measures outcomes will be briefly covered in the Discussion. The purpose of this ar-

ticle is to describe and illustrate several common choices of count regression models for independently distributed counts.

The oral health research analyst or investigator has several regression models for counts from which to choose. Negative binomial regression [Hilbe, 2011] is sometimes used for modeling dental caries indices because caries counts tend to exhibit overdispersion, i.e., excess variation relative to the Poisson distribution [Grainger and Reid, 1954]. However, as caries rates have declined in populations over time [Campus et al., 2009], distributions of caries counts are increasingly characterized by a large number of zero counts, for which Poisson and negative binomial distributions frequently provide inadequate fits. Zero-inflated Poisson regression (ZIP) models [Mullahy, 1986; Lambert, 1992] were initially applied to dental caries to account for excess zeros [Böhning et al., 1999]. To address excess zeros and overdispersion simultaneously, zero-inflated negative binomial (ZINB) regression [Lewsey and Thompson, 2004] and negative binomial hurdle (NBH) models [Levin et al., 2009] have more recently been applied in the analysis of dental caries count data.

A review of the application of ZIP and ZINB models to dental caries counts published in 2012 found that interpretations provided in published caries research are often imprecise or inadvertently misleading, particularly with respect to failing to discriminate between inference for the class of susceptible persons defined by such models and inference for the sampled population in terms of overall exposure effects [Preisser et al., 2012]. Recommendations were provided to enhance the use as well as the interpretation and reporting of results of count regression models when applied to dental caries. The present article updates the earlier recommendations following the recent introduction of marginalized ZIP (MZIP) [Long et al., 2014] and marginalized ZINB models (MZINB) [Preisser et al., 2016] that directly specify the effect of exposures and covariates on the marginal (overall) mean instead of the latent class mean of susceptible persons.

Whereas ZIP and ZINB models are often useful in dental research for understanding the mechanisms of disease [Gilthorpe et al., 2009], insufficient emphasis has been given to the effects of caries risk factors on the overall population from which the study sample was drawn [Albert et al., 2014]. Recently, MZIP and MZINB regression models were introduced for the direct estimation of overall exposure effects on a count outcome. With several model classes for the potential analysis of zero-in-

flated count outcomes, this article aims to provide guidance to researchers in the choice among ZINB, NBH, and MZINB models. For this task, an understanding of the different model classes for count outcomes with many zero counts, and the research questions addressed by them, is essential.

In this article, ZINB, NBH, and MZINB models are applied to two data sets to illustrate how the different models are used to address distinct research questions. The first is a fictional data set of 1,000 observations generated from a ZINB distribution for each of 2 groups. The second data set is from a 3-year double-blind caries incidence trial in Lanarkshire, Scotland, conducted from 1988 to 1992 that randomized 4,294 children aged 11–12 years to 3 active toothpaste formulations to compare their respective anticaries efficacy [Stephen et al., 1994]. The different interpretations provided by the models are emphasized to support the assertion that drawing valid conclusions in caries research relies on matching the statistical analysis to a well-articulated research question [Long et al., 2012].

Methods

The ZINB, NBH, and MZINB Distributions

In the absence of covariates, there exists equivalent ZINB, NBH, and MZINB distributions for count data in the sense that they give identical values for the probabilities $P(Y = y)$ for counts $y = 0, 1, 2$, etc. The distributions are distinguished only by their parameterization or, in other words, the feature or features of the overall distribution of the counts that are the focus for a chosen model. Each of the ZINB, NBH, and MZINB distributions has 2 parts so as to account for a greater fraction of zeros than can be accommodated by the negative binomial distribution $NB(\mu, \phi)$ where μ is the mean count and ϕ (where $\phi > 0$) is the dispersion parameter, with larger values of ϕ corresponding to greater variation in the counts (see online suppl. equation A.1; for all online suppl. material, see www.karger.com/doi/10.1159/000452675). Definitions of the various distributions for zero-inflated counts are given in the online supplementary file.

As an illustrative example, Figure 1a–c shows 3 different representations of a distribution of caries counts with a large number of zeros, as is often encountered in caries research. For example, the counts could be dmft from a population of children on a normal diet compared to dmft from children on a high sugar diet shown in Figure 1d–f. Figure 1a depicts a $ZINB(\psi, \mu, \phi)$ distribution (see online suppl. equation A.2), which presupposes that the population of children under study is a mixture of 2 latent classes, i.e., hypothetical, unobserved groupings of children. The white area is the fraction $\psi = 0.40$ of the total area of the bars that represents the subpopulation of “nonsusceptible” children with only zero counts who are considered to be not-at-risk for caries due perhaps to some unknown environmental or genetic factors. The black bars show the relative frequencies of caries counts based on

Fig. 1. The top half of the figure presents equivalent representations of counts generated from a negative binomial distribution with added zeros describing dmft among a fictional population of children on a normal diet. **a** ZINB distribution ($\psi = 0.40$, $\mu = 1.00$, $\phi = 0.50$). **b** NBH distribution ($\pi = 0.33$, $\mu = 1.0$, $\phi = 0.50$). **c** MZINB distribution ($\psi = 0.40$, $\nu = 0.60$, $\phi = 0.50$). The bottom half presents equivalent representations of counts for dmft among children on a high sugar diet. **d** ZINB distribution ($\psi = 0.10$, $\mu = 3.00$, $\phi = 0.50$). **e** NBH distribution ($\pi = 0.76$, $\mu = 3.00$, $\phi = 0.50$). **f** MZINB distribution ($\psi = 0.10$, $\nu = 2.70$, $\phi = 0.50$).

a negative binomial distribution with mean $\mu = 1.00$ and overdispersion parameter $\phi = 0.50$ for the “susceptible” subpopulation of children at risk for dental caries; these probabilities are scaled so that the total area of the black bars is 0.60, the probability of being in the at-risk class. The ZINB distribution assumes that children without caries consist of those children who are susceptible but happened not to have any caries observed (i.e., 0.27 random zeros) and children who are believed to be practically not-at-risk for caries (0.40). The challenges that dental researchers face in understanding zero-inflated models are related to the fact that the composition of the 2 respective latent classes is a theoretical and mathematical construct such that the specific group membership of any given subject in a study with a zero count is unknown. Moreover, the ZINB parameters ψ and μ provide only indirect information on the population caries prevalence π and mean caries extent ν that are often of interest.

An alternative representation to the ZINB distribution is provided by the NBH distribution (π , μ , ϕ) (see online suppl. equation A.3) shown in Figure 1b for children on a normal diet. The height of the white bar is 0.67 (relative frequency), which is the probability of not having caries. The black area of Figure 1b is the prevalence $\pi = 0.33$ and corresponds to the group of children with any caries, whose count distribution is characterized with a zero-truncated negative binomial distribution scaled so that the sum of probabilities for nonzero counts when scaled equals π . The prevalence π cannot be greater than the probability of not being an excess zero, $1 - \psi$

(proportion of total area of bars that is black in Fig. 1a). The NBH parameter π relates to the ZINB(ψ , μ , ϕ) parameters:

$$\pi = (1 - \psi)[1 - g(0|\mu, \phi)],$$

where g is the probability of a zero count under the negative binomial probability function $NB(\mu, \phi)$. As in the ZINB distribution, the NBH parameter μ is the mean of the untruncated negative binomial distribution, not of the black zero-truncated area in Figure 1b. So while Figure 1b provides a graphical representation of the NBH data-generating mechanism in online supplementary equation A.3, its connection to μ is not readily apparent. Rather, the mean caries count among children with any caries $\tau = E(Y|Y > 0)$ is given by

$$\tau = \mu/[1 - g(0|\mu, \phi)].$$

Next, a third equivalent representation of the frequency count distribution is shown in Figure 1c, which displays the overall, or marginalized, ZINB distribution for caries counts with a mean of all the counts denoted $\nu = E(Y)$. The mean of MZINB(ψ , ν , ϕ) distribution is defined by a transformation of the 2 parameters of the ZINB distribution, $\nu = (1 - \psi)\mu$, which for children on a normal diet is $0.60 \times 1.00 = 0.60$. Note that $\nu \leq \mu$ so that the mean caries count in the overall population (mean for the black area in Fig. 1c) cannot be greater than the mean count of the susceptible population (mean for the black area in Fig. 1a). Figure 1c emphasizes the overall population from which the study sample was drawn. For

example, in a cross-sectional study, v is caries severity or extent. The characterization shown in Figure 1c of a single overall distribution for the caries counts views the mixture distribution model in Figure 1a as a convenient explanation for a distribution of counts with excess zeros [Mwalili et al., 2008]. Finally, Figure 1d–f shows analogous representations of a distribution for the number of caries in a fictional population of children on a high sugar diet. Our interest is in comparing the distribution of dmft in children in the high sugar group to that in the normal diet group.

The ZINB, NBH, and MZINB Regression Models

Two-part model extensions of negative binomial regression models are used to characterize the effects of exposures, e.g., diet, and possibly covariates, on parameters of interest for count responses with many zeros. In choosing a 2-part count regression model for the assessment of exposure (treatment) effects and covariates, interpretations resulting from statistical analysis are greatly facilitated when the model chosen directly expresses the parameters of interest as generalized linear models, specifically as linear functions of explanatory variables, the “linear predictors”, through suitable link functions. Thus, a choice between ZINB, NBH, and MZINB models can be made through the identification of the distributional parameters of interest guided by the research question. Two examples are used to illustrate the application of the 3 models and to contrast their respective interpretations.

Example 1: Fictional dmft Data

The first example considers a fictional cross-sectional observational study evaluating the effect of a high sugar diet on dmft in children. Define a dichotomous exposure as $x = 0$ for a child that is not exposed (diet with normal sugar content) versus $x = 1$ for an exposed child (e.g., diet with a high sugar level). To illustrate the different regression models, dmft counts for 500 children on a normal sugar diet are randomly generated from a ZINB distribution with $\psi_0 = 0.40$, $\mu_0 = 1.00$, and $\phi_0 = 0.50$, and dmft counts for 500 children on a high sugar diet are generated from a ZINB distribution with $\psi_1 = 0.10$, $\mu_1 = 3.00$, and $\phi_1 = 0.50$; subscripts of “0” and “1” on the parameters indicate the nonexposed and exposed groups, respectively. In other words, children on a high sugar diet tend to have higher dmft than children on a normal diet. Three different sets of questions that may be asked from these data lead to a choice of (1) ZINB, (2)NBH, or (3)MZINB models.

(1) ZINB: Research Questions for Susceptibility to Caries

The first set of questions arise from the latent construct of susceptibility to dental caries within the context of a theoretical model of disease occurrence whereby observed zeros are believed to come from 2 sources, from children who are believed to be non-susceptible to the development of dental caries and from children believed to be capable but have not yet developed dental caries.

Question 1a. Do children with a high sugar diet have a greater susceptibility to dental caries than children with a normal sugar diet? Conversely, do children on a normal sugar diet have greater odds, i.e., $\psi/(1 - \psi)$, of being nonsusceptible to dental caries.

Question 1b. Among children that are believed to be at risk for dental caries, do children with a high sugar diet have higher mean dmft than children with a low sugar diet?

To address question 1a, ZINB models the log odds of an “excess zero” (ψ) or of being in the class of children that are nonsusceptible to caries

$$\log[\psi/(1 - \psi)] = \gamma_0 + \gamma_1 x \quad (1)$$

and to address question 1b, ZINB models the mean dmft of the at-risk class of children

$$\log(\mu) = \lambda_0 + \lambda_1 x. \quad (2)$$

The odds of being nonsusceptible to dental caries for a child exposed to a high sugar diet relative to the odds of being nonsusceptible for a child on a normal sugar diet is $\xi_{EZ} = \exp(\gamma_1)$. Among those children believed to be at risk for caries, the multiplicative increase in the mean dmft for an exposed child relative to a non-exposed child is $\theta_Z = \exp(\lambda_1)$.

(2) NBH: Research Questions for Susceptibility to Caries

Alternative questions arise from considering that once the hurdle of developing clinically manifested disease is crossed, a different mechanism may be involved in determining the extent of disease among those with any disease.

Question 2a. Do children on a high sugar diet have greater odds of having any dental caries (i.e., dmft > 0) than children on a normal sugar diet?

Question 2b. Is a high sugar diet associated with higher mean dmft compared to a normal sugar diet in the class of children believed to be at risk for caries?

Whereas the ZINB model specifies the relationship of the exposure to the susceptibility of disease, the NBH model specifies its relationship to the probability of having any caries (π)

$$\text{logit}(\pi) = \delta_0 + \delta_1 x \quad (3)$$

in combination with

$$\log(\mu) = \lambda_0 + \lambda_1 x, \quad (4)$$

the same model part as in equation 2. In other words, questions 1b and 2b are the same question. In the NBH model, $\xi_{OR} = \exp(\delta_1)$ is the odds of having any caries for an exposed child relative to the odds of having any caries for a nonexposed child. It examines whether the prevalence of caries differs between children on high sugar diets and those on normal diets. In the second model part given by equation 4, the effect of diet on the untruncated mean μ is given by the incidence rate ratio (IRR) $\exp(\lambda_1)$ as interpreted for the ZINB model. In other words, the second part of the hurdle model describes variations in μ (i.e., the mean count from the black area in Fig. 1a). Conversely, the mean of the truncated distribution shown in the black area of Figure 1b that is less than μ corresponds to the following question:

Question 2c. Among children that have any dental caries, is a high sugar diet associated with higher mean dmft compared to a normal sugar diet?

Research question 2c is not addressed directly by the parameter λ_1 , but it may be addressed indirectly by computing the IRR $\theta_H = \tau_1/\tau_0$ using postmodeling calculations where truncated mean $\tau_j = \mu_j/[1 - g(0|\mu_j, \phi_j)]$ for $j = 0$ and 1 , $\mu_0 = \exp(\lambda_0)$, and $\mu_1 = \exp(\lambda_0 + \lambda_1)$. Because diet is the only explanatory variable in equations 3 and 4, a single value for θ_H is obtained. In general, θ_H will depend on the value of covariates in models with multiple explanatory variables.

(3) *MZINB: Research Questions for the Effect of Exposure on the Overall Population Mean*

More often than not, investigators are interested in the effects of covariates on the marginal mean v of the overall population as depicted in Figure 1c. Unlike ZINB, MZINB models specify direct covariate effects on the overall mean v .

Question 3. Do children with a high sugar diet have higher mean dmft than children with a normal sugar diet?

Though not of primary interest, the first part of the model addresses question 1a. The MZINB model gives the same probability of an “excess zero” or of being in the class of children that are non-susceptible to caries as the ZINB

$$\log[\psi/(1 - \psi)] = Y_0 + Y_1x, \quad (5)$$

while $\xi_{EZ} = \exp(Y_1)$ gives the odds ratio corresponding to the effect of x on being an excess zero. However, all zeros, whether excess zeros or not, are included in the mean for the overall population depicted in Figure 1c. The MZINB model replaces equation 2 of the ZINB model with

$$\log(v) = \beta_0 + \beta_1x, \quad (6)$$

while simultaneously accounting for excess zeros in equation 5. Whereas equation 6 is of primary interest in the MZINB model, the purpose of equation 5 is to account for the many zeros in the data so that valid inference for the regression coefficients in equation 6 can be obtained. Notably, the β -coefficients have the same interpretations as those in negative binomial regression and Poisson regression. In particular, $\theta_M = \exp(\beta_1)$ is the IRR of exposure to a high sugar diet versus a normal diet or, in other words, the multiplicative increase in the marginal mean caries count for an exposed child relative to the marginal mean caries count for a non-exposed child in the overall population.

In summary, the 3 models together define 2 odds ratios, ξ_{EZ} and ξ_{OR} , as the relative odds of an excess zero (i.e., subscript “ez”) or any kind of zero (i.e., subscript “OR”), respectively. They also directly define 2 IRRs, θ_Z and θ_M , for the ratio of mean counts in the susceptible class of children (i.e., “Z” for ZINB) or for children overall (“M” for marginal or MZINB), respectively. Finally, a third IRR, θ_H , may be indirectly obtained from the hurdle (i.e., “H” subscript) model. Importantly, ξ_{OR} and θ_M are the prevalence odds ratio and overall IRR, respectively, the parameters most often of interest to oral health researchers.

Example 2: DMFS Increment in the Lanarkshire Clinical Trial

The second data set is from a 3-year double-blind caries incidence trial initiated in Lanarkshire, Scotland, in 1988 to compare 3 toothpaste formulations among schoolchildren [Stephen et al., 1994]. Dental examinations were conducted at baseline and after 1, 2, and 3 years. The current analysis compares the toothpastes with respect to mean increment DMFS (number of Decayed, Missing, and Filled surfaces) after 2 years ($n = 3,412$; 79%), while adjusting for baseline caries status and calculus. For illustration purposes, the 2-year instead of the 3-year increment data was chosen because they have more zeros. High baseline caries refers to having at least 1 Decayed, Missing, or Filled anterior tooth or premolar [Stephen et al., 1994]. Whereas the original authors specify 4 ordinal categories for baseline caries status according to the location of the affected teeth and caries severity, the middle 2 categories are combined into “medium” to give a 3-level variable (low, medium, high). Thus, 5 indicator variables are defined in both parts of the

ZINB, NBH, and MZINB 2-part models: medium and high baseline caries, respectively (with low baseline caries as the reference), and the presence of calculus, sodium fluoride (NaF, $n = 1,370$), and sodium fluoride plus sodium trimetaphosphate (NaF + TMP, $n = 680$), with sodium monofluorophosphate (SMFP, $n = 1,362$) as the reference treatment.

The ZINB model that generalizes equations 1 and 2 is fitted to estimate the odds of being nonsusceptible to new caries (after 2 years) for a child receiving each toothpaste with sodium fluoride (NaF and NaF + TMP, respectively) relative to the odds of being nonsusceptible for a child receiving SMFP. The second model part is used to estimate the multiplicative increase in the mean caries increment for a child in the at-risk class receiving NaF (or NaF + TMP) relative to the mean caries increment for a child in the at-risk class receiving SMFP.

Whereas the ZINB model specifies the relationship of covariates on ψ , the NBH model specifies their relationship to the probability of having any new caries (after 2 years) in equation 3. It estimates the odds of having any caries for a child receiving NaF (or NaF + TMP) relative to the odds of having any caries for a child receiving SMFP. The second model part of the NBH addresses the same question as the second model part of the ZINB as expressed in the paragraph immediately above.

Finally, the MZINB model is used to estimate the multiplicative increase in the marginal mean caries increment for a child receiving NaF (or NaF + TMP) relative to the marginal mean caries increment for a child receiving SMFP in the overall population while also modeling excess zeros. All models were fit with SAS software, version 9.4. SAS Proc GENMOD is used to fit negative binomial and ZINB models whereas SAS Proc NLMIXED is used to fit NBH and MZINB models. The supplementary material gives the SAS code for example 1; also, see Preisser et al. [2016a] for the SAS code for NBH models and Preisser et al. [2016b] for the SAS code for MZINB models.

Results

Example 1: Fictional dmft Data

Children on a high sugar diet had an observed mean dmft that was 4.07 times greater than the mean dmft of children on a normal diet (Table 1). Moreover, 72.8% of children on a high sugar diet had one or more dmft compared to only 35.6% on the normal diet, which gives an observed odds ratio of $(0.728)/(1 - 0.356)/(1 - 0.728) = 4.84$. Among children with 1 or more dmft, children on a diet high in sugar had 1.99 times higher dmft than children on a normal diet.

Among the four models fit to the data, NB regression had the poorest fit as indicated by the largest AIC (Table 2). On the other hand, the AICs for NBH, ZINB and MZINB were not only smaller, but identical to one another, which is always the case for a saturated model. In other words, there can be no richer model than each of these three 4-parameter models involving a single dichot-

Table 1. Descriptive statistics for dmft in a fictional population of children

	Normal diet	High sugar diet	Rate ratio	Prevalence odds ratio
Children, n	500	500		
$v \pm \text{SD}$	0.63 ± 1.06	2.58 ± 2.84	4.07	
Prevalence, π	0.36	0.73		4.84
$\tau \pm \text{SD}$	1.78 ± 1.07	3.54 ± 2.77	1.99	

Note that v is the mean dmft among the overall group of 500 children; $\tau = E(Y | Y > 0)$ is the mean dmft among the children in the group with positive dmft counts.

Table 2. Estimates of incidence rate ratios (IRRs) and odds ratios (ORs) for the effect of high sugar diet relative to normal diet on dmft in a fictional population of children

	AIC ²	IRR (95% CI)	OR (95% CI)
<i>NBH model</i> ¹	3,210.0		
Susceptible class mean, Θ_Z		3.09 (2.45, 3.90)	
logit (any caries), ξ_{OR}			4.84 (3.70, 6.34)
<i>ZINB model</i>	3,210.0		
Susceptible class mean, Θ_Z		3.09 (2.45, 3.90)	
logit (excess zero), ξ_{EZ}			0.25 (0.10, 0.58)
<i>MZINB model</i>	3,210.0		
Marginal mean, Θ_M		4.07 (3.41, 4.86)	
logit (excess zero), ξ_{EZ}			0.25 (0.10, 0.58)
<i>Negative Binomial model</i>	3,217.6		
Marginal mean, Θ_M		4.07 (3.42, 4.84)	

¹ The IRR for the children with positive counts Θ_H was determined indirectly using postmodeling calculations from the NBH model and is estimated as 1.990. ² AIC, Akaike Information Criterion.

omous variable in each model part. Indeed, the 3 models are equivalent because the 3 distributions for each group (row) of Figure 1 are equivalent. The regression parameter estimates obtained from the generated fictional data and their corresponding true values derived from ($\psi_0 = 0.40$, $\mu_0 = 1.00$, $\phi_0 = 0.50$) and ($\psi_1 = 0.10$, $\mu_1 = 3.00$, $\phi_1 = 0.50$) via mathematical formula are shown in online supplementary Table A.1. Differences in true values and estimates are due to sampling variability.

By comparing Table 2 to Table 1, we see that the MZ-INB model is the only 2-part model that produces direct estimates and, in this case, reproduces the observed rate (mean) ratio (θ_M) of 4.07. In addition to the MZINB model, negative binomial regression is the only other modeling approach that models the marginal mean. Notably, the negative binomial model has a confidence interval for θ_M with smaller width than MZINB; this is expected because negative binomial regression tends to underestimate the variance when there are excess zeros. Both the NBH and ZINB model provide direct estimates of the ratio of the susceptible class means (θ_Z) of 3.09; this quantity cannot be estimated from the observed data without parametric model assumptions.

Continuing with the comparisons of Table 2 to Table 1, we see that the prevalence odds ratio is $\xi_{OR} = 4.84$ and that the NBH model reproduces the observed odds ratio. Thus, children on a high sugar diet have 4.84 times the odds to have any new dental caries than children on a normal diet. Among children who have 1 or more dmft, the mean ratio dmft of high sugar to normal diet (θ_τ) is estimated with post-NBH modeling calculations to be 1.99 (online suppl. Table A.2), which equals the observed mean ratio (last row of Table 1). To summarize, the true values of θ_M , ξ_{OR} , and θ_τ as derived from inputs for the random generation of the data are shown in online supplementary Table A.2 alongside their corresponding estimate from the generated sample.

Whereas the NBH model directly estimates the prevalence odds ratio, the ZINB and MZINB models estimate the excess zero odds ratio ξ_{EZ} , an entity that is not observed in the data and cannot be estimated without parametric modeling assumptions involving the mixture of the 2 latent classes. Not surprisingly, because the models are equivalent for saturated models, ξ_{EZ} is estimated to be 0.25 in both models. In other words, the odds of being not susceptible to acquiring dental caries is one quarter among children on a high sugar diet relative to children on a normal diet. Equivalently, by inverting the odds ratio, the odds of being at risk for dental caries are 4 times higher among children on a high sugar diet than among children on a normal diet.

Example 2: DMFS Increment in the Lanarkshire Clinical Trial

The overall (marginal) mean DMFS for the entire study population is higher for SMFP ($v = 4.65$; $\text{SD} = 5.25$) compared to NaF ($v = 4.26$; $\text{SD} = 4.52$) and NaF + TMP ($v = 4.42$; $\text{SD} = 4.69$). While the increased efficacy of the toothpastes with sodium fluoride holds true for low and

Table 3. Descriptive statistics for 2-year DMFS increments in the Lanarkshire, Scotland, caries trial

	Low baseline caries			Medium baseline caries			High baseline caries		
	SMFP	NaF	NaF + TMP	SMFP	NaF	NaF + TMP	SMFP	NaF	NaF + TMP
Children, <i>n</i>	456	459	227	368	393	183	538	518	270
<i>v</i>	2.84	2.50	2.21	3.27	3.51	4.49	7.14	6.40	6.24
SD	3.68	3.03	2.67	3.53	3.68	5.36	6.29	5.29	4.77
IRR	1.00	0.88	0.78	1.00	1.07	1.37	1.00	0.90	0.87
π	0.72	0.69	0.67	0.80	0.77	0.80	0.90	0.92	0.94
$E(Y Y > 0)$	3.94	3.64	3.28	4.09	4.56	5.59	7.95	6.95	6.66

The marginal mean of the overall study population is given by *v*. The mean DMFS among those children with any caries (the truncated mean) is given by $E(Y | Y > 0)$. IRR is the incidence rate ratio based on marginal means with SMFP as the reference group. The incidence of any caries is given by π .

high baseline caries, the trend is reversed for medium baseline caries (Table 3). Similar trends exist with respect to the truncated means of DMFS $E(Y | Y > 0)$ for children with any caries. Table 3 does not show estimates of means (μ) for the susceptible class of children that are the focus of ZINB models because these are constructs which are not observed in the data; rather, these means can only be estimated based on a statistical model. Finally, the incidence of caries (π) at the 2-year follow-up suggests a consistent anticaries effect of NaF and NaF + TMP relative to SMFP only for children with low baseline caries.

In this data set, 19% of counts are zero, which suggests a 2-part model might be needed. Indeed, the AIC (Akaike information criterion) for the negative binomial model is larger than the AICs for the three 2-part models, indicating its poorer fit (Table 4). The AICs for the 2-part models are very similar, but not identical, indicating similar goodness-of-fit. The fits of NBH, ZINB, and MZINB models were very similar (with essentially overlapping curves), and better than the fit of the negative binomial model for the low baseline caries group (online suppl. Fig. A.1), which underestimated the observed proportion of zeros.

Table 4 reveals that estimates for the count data part of the model are similar for NBH and ZINB models, which both target the mean caries count in the latent class of children believed to be susceptible to caries. Second, estimates for the count data part of the model are similar for MZINB and negative binomial models, which both target the marginal mean caries count in the overall population of children from which the children participating in the clinical trial were drawn. Furthermore, Table 4 shows similar estimates on all covariates except calculus from the logistic

part of the model for ZINB and MZINB, both of which model the probability of an excess zero. The NBH model estimates for the logit of any caries are very different than the logit estimates for excess zeros in the ZINB and MZINB models, which is not surprising given their distinct interpretations.

Table 5 reports IRRs and odds ratios for NaF and NaF + TMP relative to SMFP. Unlike the fictional data example, estimates of the different IRRs were similar. From the MZINB (or negative binomial) model, the mean caries increment for NaF in the overall population of children was approximately 94% of the mean increment among those randomized to SMFP ($p = 0.08$). From the ZINB model, the mean caries increment for susceptible children receiving NaF was approximately 95% of the mean caries increment among susceptible children randomized to SMFP ($p = 0.21$). In all models, the effect for NaF was stronger than the effect for NaF + TMP, but none of the differences between treatment pairs reached statistical significance.

The NBH model is the only model among those considered that estimates prevalence odds ratios corresponding to the effect of toothpastes on any new caries development. In particular, a representative child from the overall population who receives NaF has an estimated 93% odds of developing caries after 2 years than a representative child from the overall population who receives SMFP. There was essentially no effect of NaF + TMP relative to SMFP for caries incidence. In summary, the 4 count data regression models applied to the Lanarkshire caries clinical trial data all found mild and nonsignificant effects of the fluoride toothpaste formulations relative to SMFP toothpaste.

Table 4. Results of alternative regression modeling approaches for 2-year DMFS increments in the Lanarkshire caries trial, 1988–1992

	AIC ¹	Intercept	Baseline caries ²		Calculus	Toothpaste ³		Overdispersion, φ
			medium	high		NaF	NaF + TMP	
<hr/>								
<i>NBH model</i>	17,168.7							
Susceptible class mean								
Estimate		1.209	0.261	0.775	−0.195	−0.064	−0.052	0.614
SE		0.045	0.050	0.045	0.042	0.040	0.049	0.034
Z		27.1	5.20	17.4	−4.69	−1.60	−1.06	
logit (any caries)								
Estimate		0.896	0.472	1.521	−0.084	−0.071	0.007	
SE		0.092	0.103	0.119	0.096	0.100	0.124	
Z		9.76	4.60	12.8	−0.87	−0.71	0.06	
<hr/>								
<i>ZINB model</i>	17,167.7							
Susceptible class mean								
Estimate		1.201	0.256	0.771	−0.201	−0.050	−0.028	0.620
SE		0.045	0.050	0.045	0.041	0.040	0.049	0.036
Z		26.8	5.10	17.3	−4.88	−1.27	−0.58	
logit (excess zero)								
estimate		−1.799	−0.630	−2.546	−0.304	0.231	0.192	
SE		0.239	0.261	0.785	0.285	0.268	0.330	
Z		−7.53	−2.42	−3.24	−1.06	0.86	0.58	
<hr/>								
<i>MZINB model</i>	17,168.4							
Marginal mean								
Estimate		1.046	0.325	0.918	−0.182	−0.067	−0.042	0.618
SE		0.041	0.048	0.042	0.039	0.038	0.047	0.035
Z		25.3	6.81	21.8	−4.64	−1.76	−0.90	
logit (excess zero)								
Estimate		−1.871	−0.563	−2.506	−0.124	0.252	0.202	
SE		0.263	0.255	0.796	0.267	0.275	0.351	
Z		−7.11	−2.21	−3.15	−0.46	0.92	0.57	
<hr/>								
<i>Negative binominal model</i>	17,214.1							
Marginal mean								
Estimate		1.042	0.328	0.920	−0.174	−0.066	−0.041	0.789
SE		0.041	0.047	0.042	0.040	0.039	0.048	0.028
Z		25.6	7.04	21.8	−4.35	−1.68	−0.85	
<hr/>								

¹ AIC, Akaike Information Criterion. ² Reference is low. ³ Reference is SMFP.

¹ AIC, Akaike Information Criterion. ² Reference is low. ³ Reference is SMFP.

Discussion

Three types of 2-part models for counts with many zeros were described and applied to two data sets with emphasis given to choosing the class of model to match the research question. The example with fictional data illustrated the possibility of large numerical differences in the different types of rate ratios and odds ratios that may be estimated for exposure effects according to the model chosen. In the Lanarkshire caries trial, treatment effect

estimates were similar across the models, despite distinct interpretations among some of the effects.

The similarity of ZINB and MZINB model estimates of IRRs summarizing the toothpaste comparisons in the Lanarkshire caries trial are the result of small predicted probabilities of excess zeros (ranging from 1% for children with calculus and high baseline caries who received SMFP to 17% for children without calculus and low baseline caries who received NaF) and the fact that estimates for toothpaste effects in the excess zero model were small.

Table 5. Estimates of incidence rate ratios (IRRs) and odds ratios (ORs) for 2-year DMFS increments in the Lanarkshire caries trial, 1988–1992

	NaF vs. SMFP	NaF + TMP vs. SMFP
<i>NBH model</i>		
Susceptible class mean, Θ_Z	0.94 (0.86, 1.01)	0.95 (0.86, 1.04)
logit (any caries), ξ_{OR}	0.93 (0.75, 1.11)	1.01 (0.76, 1.25)
<i>ZINB model</i>		
Susceptible class mean, Θ_Z	0.95 (0.88, 1.03)	0.97 (0.88, 1.07)
logit (excess zero), ξ_{EZ}	1.26 (0.74, 2.13)	1.21 (0.64, 2.31)
<i>MZINB model</i>		
Marginal mean, Θ_M	0.94 (0.87, 1.00)	0.96 (0.87, 1.05)
logit (excess zero), ξ_{EZ}	1.29 (0.59, 1.98)	1.22 (0.38, 2.06)
<i>Negative binomial model</i>		
Marginal mean, Θ_M	0.94 (0.87, 1.01)	0.96 (0.87, 1.05)

Values in parentheses are 95% CI.

If estimates for toothpaste effects in the excess zero model had been zero, then the IRRs from ZINB and MZINB for the toothpaste effect would have been identical; see equation 4 of Preisser et al. [2012]. Conversely, when there are moderate to large proportions of excess zeros, and when the exposure effects of interest are strong in both model parts, the marginalized zero-inflated count regression models can give notably different estimates compared to traditional zero-inflated models as demonstrated in the fictional data example.

Even though MZINB gave improved fits relative to negative binomial regression, the 2 models produced similar estimates of overall IRRs (θ_M) in both data examples. In a simulation study based on the Lanarkshire clinical trial, MZINB gave only slightly improved power over negative binomial regression when empirical standard errors were used in the latter to account for variance misspecification through failure to model the excess zeros [Preisser et al., 2014]. However, another simulation study found that MZINB gave less biased overall exposure effects, improved type I error, and coverage of 95% confidence intervals closer to the nominal level than negative binomial regression in models with continuous covariates having skewed (log-normal) distributions [Preisser et al., 2016b].

Hurdle models are used for count data when the proportion of individuals with any caries is of interest, which could be prevalence in a cross-sectional study or incidence in a prospective study or trial. They may also be

used to compare the truncated-at-zero means between groups through postmodeling calculations, which was illustrated in the fictional data example. It would also be possible to estimate a parameter akin to θ_τ using appropriate adjustments for covariates by either fixing covariates at their mean values or by employing an average-predicted-value method [Albert et al., 2014].

The NBH model reported in this article has distinct parameters in its 2 model parts, and as a result, it gives results for prevalence that are identical to those from ordinary logistic regression. A shared-parameter NBH model, also referred to as zero-altered negative binomial model for logistic regression (ZANB-logist), may be used when interest is in the dichotomized outcome of any caries [Preisser et al., 2016a]. In simulations, ZANB-logist models were shown to give large gains in statistical power relative to ordinary logistic regression for dichotomized counts.

All of the models discussed in this article, which were models for independent data estimated using maximum likelihood methods, have extensions to longitudinal or clustered data. Random effects have been included in ZIP [Hall, 2000], ZINB [Yau et al., 2003], Poisson hurdle [Min and Agresti, 2005] and MZIP [Long et al., 2015] models. Generalized estimating equations methodology has been employed for ZIP models [Hall and Zhang, 2004] and ZINB models [Kong et al., 2015] to obtain population-averaged interpretations. A possible extension of generalized estimating equations methodology would be to the analysis of correlated outcomes with MZIP and MZINB models.

In some situations a reasonable alternative approach to modeling counts with or without excess zeros is to use a 1-part cumulative logits model applied to an ordinal outcome created by clumping adjacent count categories together to form a small number of categories (e.g., 3, 4, or 5). Such models for cross-sectional and longitudinal data have been applied to patient outcomes in a randomized clinical trial of orthognathic surgery patients [Preisser et al., 2011].

Finally, some general recommendations on the use of MZINB, ZINB, NBH, and negative binomial regression are offered to encourage discussion on the uses and relative merits of the classes of count data models reviewed in this article. Foremost, the choice between a latent class model (ZINB), a model for prevalence (NBH) and a model for the marginal mean (negative binomial, MZINB) should be based on the study's purpose (Fig. 2). When there is specific interest in the processes of disease, the latent class effects of ZINB models may be of interest.

Fig. 2. Flowchart for count regression model choice. ZIP and ZINB models are chosen when interest in latent classes for the study of disease processes. Otherwise, MZIP and MZINB, like Poisson regression and negative binomial regression, have use for random samples in population-based oral health surveys and epidemiological studies or convenience samples for clinical trials where overall effects of treatment and exposures on marginal means are of interest. Hurdle models are of interest in modeling prevalence in populations or, secondarily (dashed line), at-risk class means.

When interest is in modeling the probability of any caries, e.g., caries prevalence in a cross-sectional study or caries incidence in a prospective study, logistic regression or NBH models may be of interest. In most other settings, modeling the marginal mean will probably be of interest. In dental surveys, for example, zero-inflated marginal-ized models often correspond to study goals that are basically descriptive in nature. Additionally, MZIP and MZINB models as well as other members of a general class of marginal mean models for zero-inflated counts [Todem et al., 2016] should find wide use in epidemiological studies and in some clinical trials such as the caries trial presented here. However, MZIP and MZINB models are prone to convergence problems to a degree shared by ZIP and ZINB models, which may preclude their use in settings where analyses must be prespecified such as for confirmatory randomized clinical trials in pharmaceutical regulatory environments [Long et al., 2014]. In such settings, Poisson or negative binomial regression with empirical variance estimation (for robustness against violation of variance assumptions) may be a better choice than 2-part models for their simplicity and relative stability (i.e., higher convergence rates) of their computational algorithms. Moreover, the practical performance in terms of statistical bias and efficiency of Poisson and negative binomial (1-part) models as shown in simulation studies may be almost as good as correctly specified 2-part models [Preisser et al., 2016b].

Conclusions

Investigators are often interested in estimating the effect of exposures and risk factors on the mean dental caries indices or increment in the overall sampled population as opposed to the index/increment for some unobserved subpopulation of individuals believed to be at risk for dental caries. In data with many zero counts (e.g., no dental caries), 2-part count models are widely recognized as an effective tool for achieving adequate fits of regression models to count data. This article emphasized the distinct interpretations of NBH, ZINB, and MZINB models and argued that the statistical model class chosen should match the study's purpose. Moreover, illustrative examples showed that regression parameter estimates for exposure effects on the different types of mean counts estimated by the respective models may have substantial differences.

Acknowledgments

The authors thank Unilever Oral Care for the caries data. John S. Preisser acknowledges support from NC TraCS NIH UL1 TR001111-01. D. Leann Long acknowledges support from WV IDEA-CTR U54GM104942. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. This paper benefited from discussion following its presentation to participants at the Methodological Issues in Oral Health Research conference in Adelaide, Australia, April 1–3, 2014, organized by Dr. Gloria Mejia.

Disclosure Statement

There are no conflicts of interest with respect to the authorship and/or publication of this article.

Author Contributions

J.S.P. conceived the paper and analyzed the data. The paper was written by J.S.P., D.L.L., and J.W.S.

References

- Albert J, Wang W, Nelson S: Estimating overall exposure effects for zero-inflated regression models with application to dental caries. *Stat Methods Med Res* 2014;23:257–278.
- Böhning 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. *J R Stat Soc Ser A* 1999;162:195–209.
- Campus G, Solinas G, Strohmer L, Cagetti MG, Senna A, Minelli L, Majori S, Montagna MT, Reali D, Castiglia P: National pathfinder survey on children's oral health in Italy: pattern and severity of caries disease in 4-year-olds. *Caries Res* 2009;43:155–162.
- Gilthorpe MS, Frydenberg M, Cheng Y, Baelum V: Modelling count data with excessive zeros: the need for class prediction in zero-inflated models and the issue of data generation in choosing between zero-inflated and generic mixture models for dental caries data. *Stat Med* 2009;28:3539–3553.
- Grainger RM, Reid DBW: Distribution of dental caries in children. *J Dent Res* 1954;33:613–623.
- Hall DB: Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics* 2000;56:1030–1039.
- Hall DB, Zhang Z: Marginal models for zero inflated clustered data. *Stat Modeling* 2004;4:161–180.
- Hilbe JM: *Negative Binomial Regression*, ed 2. Cambridge, Cambridge University Press, 2011.
- Kong M, Xu S, Levy SM, Datta S: GEE type inference for clustered zero-inflated negative binomial regression with application to dental caries. *Comput Stat Data Anal* 2015;85:54–66.
- Lambert D: Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 1992;34:1–14.
- Levin KA, Davies CA, Topping GVA, Assaf AV, Pitts NB: Inequalities in dental caries of 5-year-old children in Scotland, 1993–2003. *Eur J Public Health* 2009;19:337–342.
- 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 Dent Oral Epidemiol* 2004;32:183–189.
- Long DL, Preisser JS, Herring AH, Golin CE: A marginalized zero-inflated Poisson regression model with overall exposure effects. *Stat Med* 2014;33:5151–5165.
- Long DL, Preisser JS, Herring AH, Golin C: A marginalized zero-inflated Poisson regression model with random effects. *J R Stat Soc Ser C Appl Stat* 2015;64:815–830.
- Long DL, Preisser JS, Stamm JS: Statistical analysis of dental caries: different methods for different outcomes. *Caries Res* 2012;46:424–426.
- Min Y, Agresti A: Random effect models for repeated measures of zero-inflated count data. *Stat Modeling* 2005;5:1–19.
- Mullahy J: Specification and testing of some modified count data models. *J Econom* 1986;33:341–365.
- Mwalili SM, Lesaffre E, Declerck D: The zero-inflated negative binomial regression model with correction for misclassification: an example in caries research. *Stat Methods Med Res* 2008;17:123–139.
- Preisser JS, Das K, Benecha H, Stamm JW: Logistic regression for dichotomized counts. *Stat Methods Med Res* 2016a;25:3038–3056.
- Preisser JS, Das K, Long DL, Divaris K: A marginalized zero-inflated negative binomial regression model with overall exposure effects. *Stat Med* 2016b;35:1722–1735.
- 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. Working Paper 43. 2014. <http://biostats.bepress.com/uncbiostat/papers/art43>.
- Preisser JS, Phillips C, Perin J, Schwartz TA: Regression models for patient-reported measures having ordered categories recorded on multiple occasions. *Community Dent Oral Epidemiol* 2011;39:154–163.
- 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 Res* 2012;46:413–423.
- Stephen KW, Chestnutt IG, Jacobson APM, McCall DR, Chesters RK, Huntingdon, E, Schaffer F: The effect of NaF and SMFP toothpastes on three-year caries increments in adolescents. *Int Dent J* 1994;44:287–295.
- Todem D, Kim K, Hsu W: Marginal mean models for zero-inflated count data. *Biometrics* 2016;72:986–994.
- Yau K, Wang K, Lee A: Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. *Biom J* 2003;45:437–452.