Stat Biosci (2016) 8:181–203
DOI 10.1007/s12561-015-9136-6

Marginalized Zero-Altered Models for Longitudinal Count Data

Loni Philip Tabb[1] · **Eric J. Tchetgen Tchetgen**[2,4] · **Greg A. Wellenius**[3] · **Brent A. Coull**[4]

Received: 3 March 2015 / Accepted: 8 September 2015 / Published online: 22 September 2015

Abstract Count data often exhibit more zeros than predicted by common count distributions like the Poisson or negative binomial. In recent years, there has been considerable interest in methods for analyzing zero-inflated count data in longitudinal or other correlated data settings. A common approach has been to extend zero-inflated Poisson models to include random effects that account for correlation among observations. However, these models have been shown to have a few drawbacks, including interpretability of regression coefficients and numerical instability of fitting algorithms even when the data arise from the assumed model. To address these issues, we propose a model that parameterizes the marginal associations between the count outcome and the covariates as easily interpretable log relative rates, while including random effects to account for correlation among observations. One of the main advantages of this marginal model is that it allows a basis upon which we can directly compare the performance of standard methods that ignore zero inflation with that of a method that explicitly takes zero inflation into account. We present simulations of these various model formulations in terms of bias and variance estimation. Finally, we apply the

Electronic supplementary material The online version of this article (doi:10.1007/s12561-015-9136-6) contains supplementary material, which is available to authorized users.

✉ Loni Philip Tabb
lpp22@drexel.edu

1 Department of Epidemiology & Biostatistics, School of Public Health, Drexel University, Philadelphia, PA, USA

2 Department of Epidemiology, Harvard School of Public Health, Boston, MA, USA

3 Department of Community Health, Brown University, Boston, MA, USA

4 Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA

proposed approach to analyze toxicological data of the effect of emissions on cardiac arrhythmias.

Keywords Longitudinal data · Marginal regression · Negative binomial · Poisson · Zero inflation

1 Introduction

Count data often exhibit more zeros than that predicted by one of the common count distributions, such as the Poisson or negative binomial. Such data are common in many applications and can be difficult to analyze when many subjects have zero observations while others have much larger observations [16]. For example, in counting disease lesions on plants, a plant may have no lesions either because it is resistant to the disease, or simply because no disease spores have landed on it. This is often the distinction between structural zeros (i.e., zeros are inevitable) and sampling zeros (i.e., zeros occur by chance) [20]. There are a wide variety of application areas in which zero inflation occurs, such as road safety and accident frequencies [24], the manufacturing industry [12], and clinical trials [14].

A large literature now exists on longitudinal and other correlated data extensions of methods for zero-inflated Poisson (ZIP) counts [4,6,25,27]. This work, almost all of which builds on the standard ZIP models initially proposed [12], models the data as arising from a mixture of a point mass at zero and a Poisson distribution. In this framework, one typically relates the covariates to both the probability that an observation arises from the zero point mass (via logistic regression), as well as to the mean parameter associated with the Poisson component (via a log-linear model). In the event that the Poisson assumption of equality between the mean and variance fails, and zero inflation remains, zero-inflated negative binomial (ZINB) models have also been developed. In fact, score tests for ZIP regression versus ZINB regression have been proposed [5,21].

These ZIP and ZINB formulations have been the standard way to account for zero inflation in counts for quite sometime and have proven useful in many applications. On the other hand, these models have a few disadvantages as well. A drawback of zero inflation regression is that it can be susceptible to instability or even non-existence of the maximum likelihood estimates for the regression coefficients on the mixing probability. This problem is essentially one of complete separation [2,13] that occurs in logistic regression when the linear predictor specifies covariate patterns for which the responses are all either zero or non-zero. Second, the parameters are difficult to interpret, as one must distinguish whether an effect relates to a change in the probability of a zero count or to the mean count. Based on these considerations, two alternative approaches were proposed [16], a hurdle model and an ordinal regression model applied to a categorized version of the count response, that yield more parsimonious summaries of a given covariate effect. As noted in [16], there is a large literature on the hurdle formulation for zero-inflated distributions. The model itself is framed in two stages: (1) model the probability an outcome is equal to zero; (2) conditional that it is not zero, model the distribution of the count using a conditional distribution derived from a traditional distribution for count data.

Even if one uses the hurdle model formulation to obtain a more parsimonious description of covariate effects, interpretation issues still arise when one includes random effects to account for correlation among clustered responses. In standard Poisson or negative binomial models (i.e., no zero inflation) with a log link, inclusion of random effects does not affect the marginal interpretation of the fixed effects [28]. However, in zero-inflated extensions of the model, this is no longer true. The fixed effects now represent conditional (i.e., within-cluster) effects. As pointed out in the literature, this interpretation can be problematic for covariates that do not vary within a subject in that the data do not directly reflect this within-subject change [9]. Thus, it is of interest to formulate a version of the zero-inflated count models that estimates the marginal effect of each covariate directly. Although recent work explored a marginalized hurdle model and a marginalized zero-inflated model for correlated and overdispersed count data with excess zero observations [10], this research does not present the operating characteristics of their proposed model and how they compare to current modeling frameworks.

The primary purpose of this research is to introduce a marginalized zero-altered model [18] and compare its performance to existing regression models for count data. Section 2 reviews existing methodology for zero-inflated count data. Section 3 outlines our proposed model and maximum likelihood estimation of the model parameters. Section 4 presents the results of two simulation studies designed to compare the performance of the proposed model to that of existing modeling approaches for longitudinal count data, which includes Poisson and negative binomial mixed-effect models and generalized estimating equations. In Sect. 5, we present a case study that applies the proposed marginalized zero-altered model to toxicological data on the cardiovascular effects of exposure to coal-fired power plant emissions and compares the results to simpler approaches. We conclude with a discussion and our recommendations in Sect. 6.

2 Existing Methodology

Our proposed model combines the conditional zero-altered random effect model [16] with a marginal regression approach [8,9,23]. That is, the model accounts for possible zero inflation/deflation and correlation among counts while relating the marginal mean of the outcome directly to covariates of interest.

In this section, we briefly outline the zero-altered model proposed by [16]. Let Y_{ij} and X_{ij} denote the count response and $1 \times p$ vector of covariates, respectively, for subject $i\,(i = 1, \ldots, n)$ at time $j\,(j = 1, \ldots, n_i)$. The model is as follows:

$$\log[\mu_{ij}^b] = X_{ij}\beta + b_i$$
$$\log[-\log[P(Y_{ij} = 0|X_{ij}, b_i)]] = \gamma_1 + \gamma_2\left(X_{ij}\beta\right) + b_i$$
$$P(Y_{ij} = y_{ij}|X_{ij}, b_i) = I(y_{ij} > 0)[1 - P(Y_{ij} = 0|X_{ij}, b_i)]\frac{g(y_{ij}; \mu_{ij}^b)}{1 - g(0; \mu_{ij}^b)},$$

where μ_{ij}^b is the mean of a random variable with un-truncated distribution defined by density g, conditional on X_{ij} and b_i; $b_i \overset{iid}{\sim} N(0, \sigma^2)$ is a subject-specific random intercept; $\gamma_2 \geq 0$, γ_1, β are unrestricted; and $g(\cdot;\cdot)$ is assumed to be a count density, such as the Poisson or negative binomial. The first equation relates the expectation of the response variable to both measured covariates and random effects. The next two equations model the probability of having a zero count and having a count greater than zero, respectively, conditional on the covariates and the random effects. The appeal of this model is that if $g(\cdot;\cdot)$ is assumed to be Poisson, the standard Poisson model is a special case when $\gamma_1 = 0$ and $\gamma_2 = 1$, and it directly parameterizes zero inflation/deflation via γ_1 and γ_2. However, as noted previously [16], one drawback of this model is that there is no single parameter describing the association between the covariates of interest, X_{ij}, and the unconditional mean count, $E(Y_{ij}|X_{ij})$. In the next section, we propose a marginalized zero-altered model that specifies a simple marginal relationship between the response variable and covariates of interest while adjusting for possible zero inflation/deflation.

3 Marginalized Zero-Altered Count Model

Section 3.1 outlines the marginalized zero-altered formulation. Section 3.2 describes maximum likelihood estimation and software implementation of the proposed model.

3.1 Statistical Model

Again let Y_{ij} and X_{ij} be the response variable and vector of covariates, respectively, for subject $i (i = 1, \ldots, n)$ at time $j (j = 1, \ldots, n_i)$. We model the marginal mean, $\mu_{ij}^Y = E(Y_{ij}|X_{ij})$, conditional on covariates using a log link function,

$$\log[\mu_{ij}^Y] = X_{ij}\beta, \tag{1}$$

where β is a $p \times 1$ vector of estimated coefficients and X_{ij} is a $1 \times p$ vector of covariates. To capture the dependence among the longitudinal measurements, we model the conditional mean response, μ_{ij}^b, of a random variable distributed according to count density $g(\cdot;\cdot)$, using the conditional model,

$$\log[\mu_{ij}^b] = \Delta_{ij} + b_i, \tag{2}$$

where the $\{b_i\}$ are subject-specific random effects. The random effects are assumed to have an independent structure, such that $b_i \overset{iid}{\sim} N(0, \sigma^2)$. Δ_{ij} is a function of the model parameters, $\Delta_{ij} = \Delta_{ij}(\beta, \gamma_1, \gamma_2, \sigma)$, where σ is the standard deviation of the subject-specific random effects. To account for zero inflation/deflation, we explicitly model the probability of having a zero count and a count greater than zero:

$$P(Y_{ij} = 0|X_{ij}, b_i) = h\left[\gamma_1 + \gamma_2\left(\Delta_{ij}\right) + b_i\right] \tag{3}$$

$$P(Y_{ij} = y_{ij}|X_{ij}, b_i) = [1 - P(Y_{ij} = 0|X_{ij}, b_i)]\frac{g(y_{ij}; \mu_{ij}^b)}{1 - g(0; \mu_{ij}^b)}, y = 1, 2, \cdots \tag{4}$$

where $\gamma_2 \geq 0$ and γ_1, β are unrestricted. Collectively, γ_1 and γ_2 measure the zero inflation/deflation, such that $\gamma_1 < 0$ and $\gamma_2 < 1$ introduce zero inflation in an additive and multiplicative fashion, respectively (and zero deflation in a counter fashion). The ratio in (4) is the conditional probability of a response greater than zero, where $g(\cdot; \mu_{ij}^b)$ again represents the density for, say, a Poisson random variable with mean μ_{ij}^b. This framework can also accommodate overdispersion in the non-zero counts via the negative binomial assumption, in which case we use $g(y_{ij}; \mu_{ij}^b) = \frac{\Gamma(y_{ij}+1/\alpha)}{\Gamma(y_{ij}+1)\Gamma(1/\alpha)}\left(\frac{1/\alpha}{1/\alpha+\mu_{ij}^b}\right)^{1/\alpha}\left(\frac{\mu_{ij}^b}{1/\alpha+\mu_{ij}^b}\right)^{y_{ij}}$, and $\alpha > 0$ is the overdispersion parameter.

The quantities $\Delta_{ij}(\theta)$ are indirectly defined and can be obtained as the solution to the convolution equation, which links the marginal and conditional means,

$$\begin{aligned}\mu_{ij}^Y &= \exp(X_{ij}\beta) \\ &= \int [1 - P(Y_{ij} = 0|X_{ij}, b_i)]\frac{\mu_{ij}^b}{[1 - g(0; \mu_{ij}^b)]}\phi(b_i|\sigma)\mathrm{d}b_i,\end{aligned} \tag{5}$$

where the integral defined is one-dimensional (see Appendix 1 for details). Given a fixed value of θ, we solve for $\Delta_{ij}(\theta)$ after approximating (5) using numerical quadrature, such as Gauss-Hermite quadrature [15,23]. Equation (5) expresses the marginal mean of the outcome as the conditional mean (given the random effects b_i) marginalized over the assumed distribution of the random effects. Because these solutions are analytically intractable, we use a simple Newton–Raphson algorithm to solve for them numerically.

We propose the choice of the h function vary depending on the chosen count distribution g. If we let $h(x) = g(0; x)$, then the model given by (1–4) reduces to the standard count distribution g in the special case of $\gamma_1 = 0$ and $\gamma_2 = 1$, which has the advantage that the presence of zero inflation can be formally tested against a standard count distribution via likelihood ratio testing. [16] used this strategy for the Poisson distribution with

$$P(Y_{ij} = 0|X_{ij}, b_i) = \exp\left\{-\exp\left[\gamma_1 + \gamma_2\left(\Delta_{ij}\right) + b_i\right]\right\},$$

the inverse of which yields the log-negative-log link function in their original model formulation. In the negative binomial model, the link that results from this strategy depends on the overdispersion parameter α. The zero-altered component of the model becomes

$$P(Y_{ij} = 0|X_{ij}, b_i) = \left(1/[1 + \alpha\exp(\gamma_1 + \gamma_2(\Delta_{ij}) + b_i)]\right)^{1/\alpha}.$$

We refer to the general formulation given by (1) - (4) as the marginalized zero-altered count (MZAC) model, with model parameters $\theta = (\beta, \gamma_1, \gamma_2, \sigma)$. Because the MZAC model can assume either a Poisson or a negative binomial distribution, we will also refer to either a MZAP or a MZANB model, respectively, throughout the remainder of the paper.

The above formulation follows that proposed by [16] in that the parameter γ_2 scales the component, in this case Δ_{ij}, that involves the covariates X_{ij} only. This structure corresponds to a scaling factor of 1 for the random effects, which assumes that the scales of the random effects in the zero-altered and non-zero components of the model are equal. We also consider alternative scalings that relax this assumption. Specifically, we also consider models in which the full predictor $\Delta_{ij} + b_i$ is scaled by γ_2,

$$h^{-1}[P(Y_{ij} = 0|X_{ij}, b_i)] = \gamma_1 + \gamma_2(\Delta_{ij} + b_i),$$

which assumes that the scaling of covariate component Δ_{ij} and the random effects b_i are scaled equally for the zero-altered component of the model. We also consider the more flexible formulation

$$h^{-1}[P(Y_{ij} = 0|X_{ij}, b_i)] = \gamma_1 + \gamma_2(\Delta_{ij}) + \gamma_3(b_i),$$

which makes no assumptions about how the scalings of either the covariate component Δ_{ij} and b_i are related. We use standard likelihood-based fit statistics, such as Akaike information criterion (AIC) [1], to choose between these three scaling models.

A key feature of this model is the relationship given in (1), which specifies a simple log-linear relationship between the response and corresponding covariates of interest. Moreover, the mean model is separate from the association model. As noted in previous research [9], this separation typically yields inferences for the regression coefficient parameters that are less sensitive to misspecification of the association model for the repeated measures. Another advantage of this model lies in (3), which explicitly parameterizes zero inflation/deflation through γ_1 and γ_2. To provide an interpretation of the γ_1 and γ_2 parameters governing the amount of zero inflation in the model, Fig. 1 presents a simulated dataset from each of six values of (γ_1, γ_2), for an intercept only model with $\beta_0 = 1.50$ and $\sigma^2 = 0.10$. Data were generated from a MZAP model with the following values of $\gamma = (\gamma_1, \gamma_2)$: (0, 1), (−0.25, 1), and (−0.50, 1) in column 1, and (0, 0.25), (−0.25, 0.25), and (−0.5, 0.25) in column 2. Therefore, the first histogram in Fig. 1 represents data generated from the Poisson distribution with $\mu = \exp(1.50) = 4.48$, and the other five plots show how this distribution changes as one increases the amount of zero inflation either additively (i.e., changing γ_1), multiplicatively (i.e., changing γ_2), or both. Moving down the first column, which represents increasing amounts of zero inflation on the additive scale, the frequency of zeros increases to approximately 15 and 30, respectively, while the range of the distribution decreases slightly and the spike at zero becomes evident in the $\gamma = (\gamma_1, \gamma_2) = (\text{-}0.5, 1)$ case. The first plot in the second column represents zero inflation on the multiplicative scale only, and the frequency of zeros is about 30. When $\gamma = (\gamma_1, \gamma_2) = (\text{-}0.25, 0.25)$, zero inflation is present on both the additive and multiplicative scales, and, not only does the frequency of zeros increase, but the

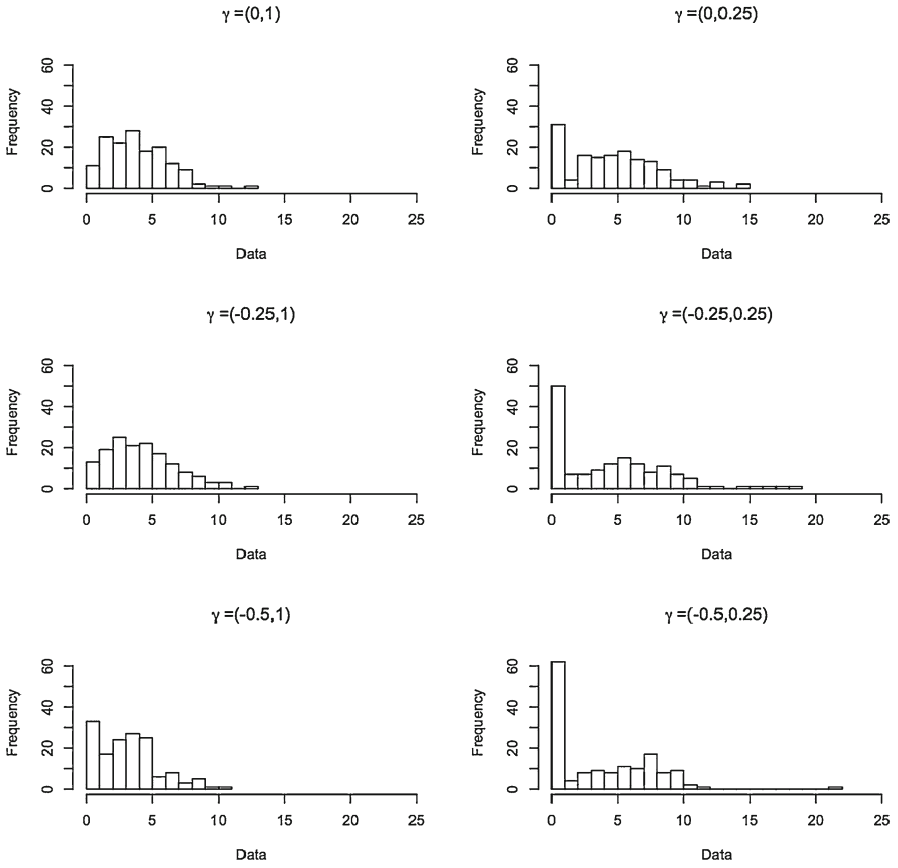

Fig. 1 Histograms of zero-inflated simulated data, where $\gamma = (\gamma_1, \gamma_2)$ is (0, 1), (−0.25, 1), and (−0.5, 1) in column 1 and (0, 0.25), (−0.25, 0.25), and (−0.5, 0.25) in column 2

distribution becomes more skewed to the right, with the maximum count close to 20. The last histogram, when $\gamma = (\gamma_1, \gamma_2) = (\text{-}0.5, 0.25)$, shows an even larger increase in the frequency of zeros with a much thicker right tail. Therefore, when zero inflation is present on the additive scale only, the frequency of zeros increases, whereas zero inflation on both the additive and multiplicative scales increases both the amount of zero inflation and the skewness of the distribution, allowing for more extreme observations. Sections 4 and 5 present some evidence that this property of the MZAC distribution may result in regression estimates that are less sensitive to outliers.

3.2 Maximum Likelihood Estimation

We estimate the parameters of the MZAC model via maximum likelihood estimation. The likelihood contribution from subject i is constructed under the standard assump-

tions that the response counts are independent given the random effects. The marginal likelihood contribution for subject i is

$$
\begin{aligned}
L_i &= \int \prod_{j=1}^{n_i} P(Y_{ij}|X_{ij}, \beta, \gamma_1, \gamma_2, b_i)\phi(b_i|\sigma)\mathrm{d}b_i \\
&= \int \prod_{j=1}^{n_i} [P(Y_{ij} = 0|X_{ij}, b_i)]^{I(Y_{ij}=0)} \\
&\quad \times \left[[1 - P(Y_{ij} = 0|X_{ij}, b_i)] \frac{g(y_{ij}; \mu_{ij}^b)}{1 - g(0; \mu_{ij}^b)} \right]^{I(Y_{ij}>0)} \phi(b_i|\sigma)\mathrm{d}b_i
\end{aligned}
$$

where $\phi(b_i|\sigma)$ is the probability density function for a normal random variable with mean zero and standard deviation σ. The conditional (on the random effects) likelihood $P(Y_{ij}|X_{ij}, \beta, \gamma_1, \gamma_2, b_i)$ is the product of two terms. The first term depends on the probability that Y_{ij} is equal to zero and the second term corresponds to the probability that Y_{ij} is greater than zero. Because the integral in this marginal likelihood can not be evaluated analytically, numerical methods are required for approximation. Any number of software packages that allow fitting of nonlinear mixed models can maximize this likelihood. We use SAS PROC NLMIXED (sample code provided in Appendix 2), which uses fixed Gaussian quadrature to approximate $\log(L_i)$ and a dual quasi-Newton algorithm to maximize this approximation summed over all subjects, to obtain estimates for all model parameters. The standard errors associated with these parameters are based on the Hessian matrix computed using the quadrature approximation of the marginal likelihood.

Our primary interest is in estimating the exposure effect, along with other potential confounders. Because the MZAC model allows separate marginal and conditional mean structures, the interpretation of the β coefficients is straightforward. In general, an element of β represents the log relative change in response mean related to a one unit difference in the corresponding predictor. For example, if we consider a single binary covariate, such as exposure level, then $\beta = \log[E(Y_{ij}|\mathrm{Exp}_i = 1)] - \log[E(Y_{ij}|\mathrm{Exp}_i = 0)]$ measures the average change in, say, frequency of our outcome between subjects exposed as opposed to subjects unexposed.

4 Simulation Studies

There is no previous work on the impact of zero inflation on estimators that naively ignore this feature of the count data. The primary reason for this gap in the literature is the fact that the parameters in the most common models for zero-inflated count data do not have a marginal interpretation, and so it is difficult to compare results from one of these models to those from a marginal one. One advantage of our marginalized parameterization of the zero-altered count model is that the resulting regression coefficient estimators can be directly compared to those from other marginal models. We conducted two simulation studies to investigate how the estimators from the

MZAP model perform as compared to those from more traditional models, namely the negative binomial mixed-effect (NBME) model, Poisson mixed-effect (PME) model and generalized estimating equations (GEE). The NBME, PME, and GEE models are commonly used to analyze longitudinal count data, but do not explicitly adjust and/or estimate zero inflation. The NBME and PME models are a special case of a generalized linear mixed model, and takes the form

$$\log[\mu_{ij}^{b}] = X_{ij}\beta + b_i.$$

In this model, the regression coefficients have a marginal interpretation, and so coefficient estimates are directly comparable to their MZAC counterparts.

Because the likelihood function in NBME and PME models include integrals that are analytically intractable, maximum likelihood estimation for this model also depends on quadrature, and we implemented this approach in SAS PROC NLMIXED. To provide consistent results in the presence of misspecification [3], we also utilized the sandwich estimation option for the standard errors for these two approaches. The GEE approach [28], on the other hand, solves the set of estimating equations,

$$s(\beta) = \sum_{i=1}^{n} D_i^{'} V_i^{-1}(Y_i - \mu_i^m) = 0,$$

where $Y_i = (Y_{i1}, \ldots, Y_{in_i})^{'}$, $\mu_i^m = (\mu_{i1}^m, \ldots, \mu_{in_i}^m)^{'}$, $D_i = \partial\mu_i/\partial\beta$, and $V_i = A_i^{\frac{1}{2}} R A_i^{\frac{1}{2}}$, with A_i being the diagonal matrix of marginal variances for the ith cluster and $R = R(\alpha)$ being the working correlation matrix. We fit the log-linear GEE

$$\log[\mu_{ij}^m] = X_{ij}\beta,$$

to each simulated dataset.

The first simulation study assessed the performance of each method, in terms of bias, efficiency, and variance estimation, while increasing the amount of zero inflation in the data. Motivated by an empirical observation that GEE approaches may be overly sensitive to a single outlying count (see Sect. 5), we conducted a second simulation study that assessed the sensitivity of the four methods to response outliers. In carrying out these simulations, we note that although the GEE is the correct model, both the PME and NBME are misspecified but with consistent score equations.

4.1 Simulation 1

We simulated datasets containing 100 clusters, each with a cluster size of 5, from the proposed MZAP model. Each dataset contained 50 "exposed" subjects and 50 "unexposed" subjects. The marginal model was $\mu_{ij}^{Y} = \exp(\beta_0 + \beta_1 \text{Time}_j + \beta_2 \text{Exp}_i + \beta_3 \text{Exp}_i * \text{Time}_j)$. Taken to closely match the results from a similar data analysis presented in Sect. 5, the true β values were $(\beta_0, \beta_1, \beta_2, \beta_3) = (1.12, -0.87, 0.29, 1.10)$

and the covariates of interest are exposure (Exp_i : 1 = exposed, 0 = unexposed) and time ($\text{Time}_{ij} = (j-1)*0.25$). The subject-specific random effects, b_i, were generated from a normal distribution with mean 0 and variance 0.10. Equivalently, in a setting where there are only two exposure groups and 5 observation points, there will be 10 different values of Δ_{ij} for each time and exposure combination; therefore, the values of Δ_{ij} that correspond to the assumed β and σ^2 values, as well as assuming no zero inflation, are Δ_{ij} = (1.36, 1.42, 1.47, 1.53, 1.59, 1.07, 0.85, 0.63, 0.42, 0.20), where the first five values are for the exposed subject and the last five are for the unexposed subject. To introduce different amounts of zero inflation, we varied γ_1 and γ_2, resulting in three different cases: (A) for $\gamma_2 = 1$, $\gamma_1 = (0, -0.25, -0.50, -0.75)$, (B) for $\gamma_2 = 0.50$, $\gamma_1 = (0, -0.25, -0.50, -0.75)$, and (C) for $\gamma_2 = 0.25$, $\gamma_1 = (0, -0.25)$. Recall that γ_1 and γ_2 measure zero inflation/deflation in an additive and multiplicative fashion, respectively; therefore, the smaller γ_1 (i.e., $\gamma_1 < 0$) and γ_2 (i.e., $\gamma_2 < 1$), the more zero inflation is present. The amounts of zero inflation increase within each case as γ_1 gets smaller, and, overall, the amount of zero inflation increases from case A to case B to case C, because γ_2 is also getting smaller. Only two values for γ_1 are reported for case C due to the fact that, when $\gamma_2 = 0.25$, larger negative values of γ_1 yielded a distribution with such an overwhelming proportion of zeros that none of the four estimation methods considered were able to reliably estimate the model parameters.

Table 1 presents the results, which are based on 500 simulated datasets, for only three specific zero inflation settings (no zero inflation, $\gamma = (\gamma_2, \gamma_1 = (1, 0))$, moderate zero inflation, $\gamma = (\gamma_2, \gamma_1 = (1, -0.75))$, and severe zero inflation, $\gamma = (\gamma_2, \gamma_1 = (-0.25, 0.25))$). A detailed table, with 10 different zero inflation settings, is presented in Sect. 8 (Online Resource 1). To compare the four different modeling approaches for the three different cases (A–C), we focus on the simulated mean and variability (standard deviation) of the estimates, as well as the mean standard errors. We also present the 95% coverage rates for the β coefficients. Across all four methods, the β coefficients display little to no bias, even as the amount of zero inflation increases; similarly, the coverage is approximately the same across the four methods. As the amount of zero inflation increases, though, the mean of the PME estimator of σ^2 does not match the true value, but this is to be expected as it is estimating cluster heterogeneity on the log-linear scale and not within the zero-altered Poisson framework. As the zero inflation moderately increases, the variance component for the NBME performs somewhat similar to the variance component for the PME in Case A, in that the simulated variance is larger than the true value. However, as the amount of zero inflation significantly increases, the mean of the NBME estimator of σ^2 decreases, approaching 0.02 in the most extreme case. For all three simulation settings, the MZAP estimators for γ_1, γ_2, and the between-subject variability, σ^2, display little bias, suggesting that the proposed software implementation yields adequate inferences when the model is correctly specified.

4.2 Simulation 2

To compare the sensitivity of the four estimators to outlying observations, we simulated 30 clusters, with 5 observations each. The marginal mean model for each subject was

Table 1 Simulation 1

γ_2	γ_1	Model	$\beta_{time} = -0.87$	$\beta_{exp} = 0.29$	$\beta_{time \times exp} = 1.10$	$\sigma^2 = 0.10$
1	0	GEE	−0.87 (0.13, 0.13; 0.94)	0.33 (0.08, 0.11; 0.98)	1.10 (0.15, 0.15; 0.95)	
		PME	−0.87 (0.13, 0.13; 0.94)	0.32 (0.08, 0.11; 0.98)	1.10 (0.15, 0.15; 0.95)	0.10 (0.02, 0.02)
		NBME	−0.87 (0.13, 0.13; 0.94)	0.32 (0.08, 0.11; 0.98)	1.10 (0.15, 0.15; 0.96)	0.10 (0.02, 0.02)
		MZAP	−0.87 (0.13, 0.13; 0.95)	0.32 (0.08, 0.11; 0.98)	1.09 (0.15, 0.15; 0.95)	0.10 (0.02, 0.02)
1	−0.75	GEE	−0.86 (0.16, 0.16; 0.95)	0.34 (0.11, 0.13; 0.96)	1.09 (0.19, 0.19; 0.94)	
		PME	−0.86 (0.16, 0.16; 0.95)	0.35 (0.11, 0.13; 0.97)	1.10 (0.19, 0.19; 0.94)	0.18 (0.03, 0.04)
		NBME	−0.87 (0.16, 0.16; 0.95)	0.34 (0.11, 0.13; 0.97)	1.10 (0.20, 0.19; 0.95)	0.15 (0.03, 0.04)
		MZAP	−0.86 (0.16, 0.16; 0.95)	0.34 (0.11, 0.13; 0.97)	1.09 (0.19, 0.18; 0.95)	0.10 (0.02, 0.02)
0.25	−0.25	GEE	−0.87 (0.18, 0.18; 0.93)	0.33 (0.13, 0.15; 0.97)	1.10 (0.24, 0.23; 0.93)	
		PME	−0.87 (0.18, 0.18; 0.93)	0.32 (0.14, 0.15; 0.97)	1.10 (0.24, 0.23; 0.94)	0.21 (0.04, 0.04)
		NBME	−0.87 (0.18, 0.18; 0.94)	0.33 (0.14, 0.15; 0.97)	1.10 (0.24, 0.23; 0.94)	0.02 (0.03, 0.05)
		MZAP	−0.87 (0.14, 0.15; 0.96)	0.32 (0.10, 0.12; 0.97)	1.10 (0.18, 0.18; 0.93)	0.10 (0.02, 0.02)

Average estimates (standard deviation, standard error; $\beta_{95\%\text{ CI coverage}}$) for generalized estimating equations (GEE), Poisson mixed-effect (PME), negative binomial mixed-effect (NBME), and marginalized zero-altered Poisson (MZAP) models under no ($\gamma = (\gamma_2, \gamma_1) = (1, 0)$), moderate ($\gamma = (1, -0.75)$), and severe ($\gamma = (0.25, -0.25)$) zero inflation. 500 replications

Table 2 Simulation 2

y^*	Model	$\beta_{exp} = 0.29$	$\sigma^2 = 0.10$
y+0	GEE	0.13 (0.15, 0.19; 0.92)	
	PME	0.14 (0.15, 0.19; 0.93)	0.18 (0.08, 0.07)
	NBME	0.13 (0.15, 0.18; 0.93)	0.03 (0.05, 0.09)
	MZAP	0.15 (0.13, 0.16; 0.98)	0.07 (0.04, 0.05)
y+20	GEE	0.22 (0.14, 0.21; 0.99)	
	PME	0.21 (0.15, 0.21; 0.99)	0.23 (0.08, 0.09)
	NBME	0.24 (0.15, 0.21; 0.99)	0.07 (0.07, 0.08)
	MZAP	0.24 (0.12, 0.18; 0.98)	0.13 (0.05, 0.06)
y+40	GEE	0.31 (0.14, 0.25; 1.00)	
	PME	0.25 (0.15, 0.22; 1.00)	0.29 (0.08, 0.12)
	NBME	0.29 (0.15, 0.23; 1.00)	0.14 (0.07, 0.09)
	MZAP	0.29 (0.12, 0.21; 0.98)	0.20 (0.06, 0.08)
y+60	GEE	0.40 (0.13, 0.29; 1.00)	
	PME	0.27 (0.15, 0.24; 1.00)	0.34 (0.08, 0.15)
	NBME	0.32 (0.15, 0.25; 1.00)	0.20 (0.07, 0.11)
	MZAP	0.30 (0.13, 0.23; 0.98)	0.26 (0.06, 0.09)
y+80	GEE	0.48 (0.13, 0.34; 1.00)	
	PME	0.29 (0.15, 0.25; 1.00)	0.39 (0.09, 0.18)
	NBME	0.34 (0.15, 0.26; 1.00)	0.24 (0.07, 0.13)
	MZAP	0.29 (0.13, 0.24; 0.99)	0.31 (0.07, 0.11)
y+100	GEE	0.55 (0.13, 0.38; 1.00)	
	PME	0.30 (0.15, 0.26; 1.00)	0.42 (0.09, 0.20)
	NBME	0.35 (0.15, 0.27; 1.00)	0.28 (0.07, 0.15)
	MZAP	0.26 (0.15, 0.25; 0.99)	0.35 (0.07, 0.11)

Contamination of datasets. Average estimates (standard deviation, standard error; $\beta_{95\%\text{ CI coverage}}$) for generalized estimating equations (GEE), Poisson mixed-effect (PME), negative binomial mixed-effect (NBME), and marginalized zero-altered Poisson (MZAP) models. 500 replications

$\mu_i^m = \exp(\beta_0 + \beta_1 \text{Time}_i + \beta_2 \text{Exp}_i)$, where the true β values are $(\beta_0, \beta_1, \beta_2) = (1.12, -0.87, 0.29)$. The random effects were normally distributed with zero mean and variance equal to 0.10. The true γ values were $\gamma_1 = -0.25$ and $\gamma_2 = 0.25$, corresponding to a significant amount of zero inflation. To induce contamination, we increased the last response count of one exposed subject, such that, $y_{ij}^* = y_{ij} + c, c\epsilon(0, 20, 40, 60, 80, 100)$, where $i = 1$ and $j = 5$. Table 2 presents the results of this contamination simulation. Due to the small number of clusters, there is a small amount of finite sample bias. Focusing on the coefficient estimates for the exposure covariate, as the final observation in this exposed subject becomes more extreme, the bias in the GEE estimator is approximately linear in the amount of contamination present in the response, agreeing with established results on the form of the influence function for GEE estimators [19]. MZAP, NBME, and PME estimators incurs less bias as compared to its GEE counterpart, with the amount of this bias being relatively small. These simulation results suggest that, in the presence of zero inflation, the MZAP estimators can be more robust to extreme observations as compared to their GEE counterparts.

5 Analysis of Arrhythmia Data

We use the proposed MZANB and MZAP models to analyze data collected as part of the toxicological evaluation of realistic emissions of source aerosols (TERESA) study [26]. In this study, rats with an induced acute myocardial infarction were exposed to either stack emissions ($n = 15$) or filtered air ($n = 14$) for up to five hours. The exposed rats were assessed while in photochemical chambers filled with stack emissions composed of SO_2, NO, and primary particulate matter. The control group of rats was exposed to room air filtered through a high-efficiency particulate air (HEPA) filter. The outcome of interest was arrhythmia frequency per hour, with electrocardiogram output inspected manually by an investigator blinded to the exposure status of each animal. Because each rat did not have a measured outcome for all five hours, we assumed that the data are missing completely at random, which is reasonable due to the fact that shorter exposures were likely due to technical issues unrelated to the outcome. Figure 2 shows the histogram of the frequency of arrhythmias per hour for the rats in the study. This histogram shows that there are potentially more observed zeros than expected from a standard distribution for count data, as indicated by the significant spike at zero. It also shows that the distribution of the frequency of arrhythmias per hour is extremely skewed to the right.

We compare the results from the MZANB and MZAP models to several traditional count models and a hurdle model with random effects. Preliminary analyses showed no evidence of an interaction between exposure and time, so to assess the association between the frequency of arrhythmias with exposure level and time, we considered the following marginal model:

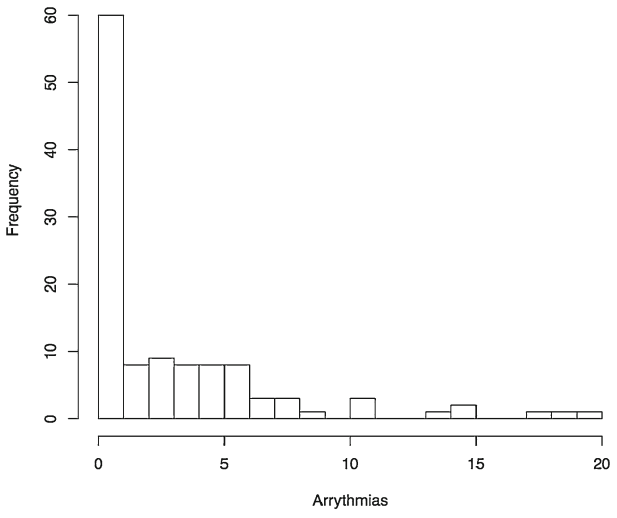

Fig. 2 Histogram of frequency of arrhythmias for each rat in the TERESA study

$$\log[\mu_{ij}^{Y}] = \beta_0 + \beta_1 \text{Exp}_i + \beta_2 I(\text{Time}_{ij} = 1) + \beta_3 I(\text{Time}_{ij} = 2) + \beta_4 I(\text{Time}_{ij} = 3) + \beta_5 I(\text{Time}_{ij} = 4),$$

for subject $i (i = 1, \ldots, 29)$ at hour $j (j = 1, 2, 3, 4, 5)$. To complete the MZANB and MZAP model specification, we assume

$$\log[\mu_{ij}^{b}] = \Delta_{ij} + b_i,$$

where $b_i \sim N(0, \sigma^2)$, and considered all three possible scaling formulations outlined in Sect. 3.

Table 3 presents the maximum likelihood estimates, the corresponding standard errors, and p values from the MZANB, MZAP, NBME, PME, and GEE models, as well as the AIC and likelihood ratio test statistics to compare model fit. A more detailed table along with results from the hurdle model with random effects as presented in [16] is given in Sect. 9 (Online Resource 2) , even though the parameters in this mixture model formulation do not directly compare to those from the other marginal models considered. Here, we only present results from the most flexible scaling model,

$$h^{-1}[P(Y_{ij} = 0 | X_{ij}, b_i)] = \gamma_1 + \gamma_2(\Delta_{ij}) + \gamma_3(b_i),$$

because both the MZANB and the MZAP models had the lowest AIC values in comparison to the other two scaling models for these data. Both the MZANB and MZAP results suggest that there is a significant (at the $\alpha = 0.05$ level) effect of exposure ($\beta_E \approx 0.7$), with the MZANB producing the best fit as indicated by the lowest AIC. For both of the MZANB and MZAP models, the estimated relative risk, $\exp(0.7) = 2.01$, of arrhythmias is doubled for those rats exposed to stack emissions as opposed to those rats exposed to HEPA filtered air. Although the MZANB maximum likelihood estimate for the heterogeneity parameter is $\sigma^2 = 0.187$, indicating a small amount of subject-to-subject variation in the frequency of arrhythmias, the γ_3 parameter that measures the overall effect of the subject-specific random effects is large. It appears that, in this case, because the NBME and PME models do not account for zero inflation, they underestimate the effect of exposure and overestimate the corresponding standard error. The GEE approach yields an estimated effect of exposure similar to those from the MZANB and MZAP models, but as we see in the next table this estimate can be sensitive to the inclusion or exclusion of a few extreme observations.

Due to the small sample size of the toxicity study, the zero-altered parameters, γ_1, γ_2, and γ_3, exhibit large standard errors. Although these quantities indicate evidence of high uncertainty, these marginalized zero-altered models consistently perform better than their standard counterpart models, NBME and PME, as denoted by the AIC values as well as a formal likelihood ratio test. More specifically, we have compared the MZANB against the NBME as well as the MZAP against the PME. The likelihood ratio test statistic for both tests are 15.3 and 39, respectively, highlighting significant zero inflation as well. Both tests also indicate better model fit for our proposed marginalized models, as opposed to the models that ignore zero inflation.

Table 3 Model estimates, standard errors (SE), and p values (p) for the frequency of arrhythmias, with -2 log likelihood ($-2\ln L$) and Akaike information criterion (AIC) values for the marginalized zero-altered negative binomial (MZANB), marginalized zero-altered Poisson (MZAP), negative binomial mixed-effect (NBME), Poisson mixed-effect (PME), and generalized estimating equations (GEE) models

Coefficient	MZANB		MZAP		NBME		PME		GEE	
	Estimate (SE)	p	Estimate (SE)	p	Estimate (SE)	p	Estimate (SE)	p	Estimate (SE)	p
β_E	0.705 (0.266)	0.013	0.667 (0.300)	0.034	0.327 (0.490)	0.511	0.304 (0.503)	0.550	0.657 (0.326)	0.044
σ^2	0.187 (0.173)	0.289	0.409 (0.208)	0.059	1.290 (0.577)	0.033	1.561 (0.583)	0.012		
a	0.433 (0.187)	0.028			0.636 (0.192)	0.003				
γ_1	0.102 (0.731)	0.890	−0.373 (0.516)	0.476						
γ_2	0.215 (0.553)	0.700	0.343 (0.427)	0.428						
γ_3	3.501 (1.982)	0.088	1.512 (0.543)	0.010						
$-2\ln L$	486.9		512.8		502.2		551.8			
AIC	508.9		532.8		518.2		565.8			

Table 4 Model estimates (standard errors) for exposure and % change for the marginalized zero-altered negative binomial (MZANB), marginalized zero-altered Poisson (MZAP), negative binomial mixed effect (NBME), Poisson mixed-effect (PME), and generalized estimating equations (GEE) models with and without an outlier

	β_E w/outlier	β_E w/o outlier	% Change
MZANB	0.588 (0.283)	0.705 (0.266)	16.596
MZAP	0.658 (0.354)	0.667 (0.300)	1.349
NBME	0.233 (0.523)	0.327 (0.490)	28.746
PME	0.139 (0.546)	0.304 (0.503)	54.276
GEE	0.214 (0.446)	0.657 (0.326)	67.428

The original data also included an extreme observation, which was removed from the dataset prior to analysis. The largest number of arrhythmias did not exceed 20 for any rat, with the exception of this one outlier, which was 59. In addition to the primary analyses above, we analyzed the dataset with this extreme observation using all marginal models presented as well as the hurdle model with random effects. Table 4 displays the results for the exposure coefficient only, and a detailed table including the hurdle results is presented in Sect. 10 (Online Resource 3). The results show that the MZANB and MZAP models are more robust in the presence of this extreme observation compared to the NBME, PME, GEE, and hurdle approaches (considering the percent change across both portions of this mixture model). The MZANB and MZAP have the smallest percent change, 16.596% and 1.349%, respectively, in analyzing the toxicity dataset with and without the outlier, and the overall conclusions of these analyses do not change. In contrast, the GEE has the largest percent change, 67.428%, coupled with a large change in the strength of this association as well.

6 Discussion

We proposed marginalized zero-altered count models for longitudinal data with excessive zeros. This class of models allowed us to address the issue of zero inflation in a simple, clear, and parsimonious way. Current methods, such as zero-inflated regression models, adjust for excessive zeros but often at the expense of interpretability.

Our statistical framework has a few advantages. First, because our model includes a marginal mean structure, model parameters are easily interpretable. Unlike zero-inflated regression models, the interpretation of our regression coefficients is of the marginal form. Second, we are able to adjust for the zero inflation/deflation in the data via the parameters γ_1 and γ_2. Estimates of these parameters reflect the degree of the zero inflation/deflation. Third, the MZANB and MZAP models simplify to a negative binomial and Poisson model, respectively, when $\gamma_1 = 0$ and $\gamma_2 = 1$, which allows us to test whether γ_1 is different from zero and γ_2 (and γ_3 in models containing this additional scale parameter) is different from one. We also carried out two simulation studies to test the performance of our proposed MZAP model to that of standard approaches

for longitudinal data, such as the negative binomial and Poisson mixed-effect models and generalized estimating equations. Results suggested that, when compared to GEE, the MZAP model was more robust in the presence of outlying response observations and, compared to the NBME and PME models, more accurately reflected estimation uncertainty in the presence of zero inflation.

We also presented the use of the marginalized zero-altered models in a toxicity study where the study population exhibited a significant number of zero outcomes in the arrhythmia response. Because our interest moves beyond simply modeling whether or not the arrhythmias exists as well as a more parsimonious environment for interpretation, our marginalized zero-altered models proved more appropriate than either traditional random effect models as well as the hurdle model with random effects. Our proposed models displayed better model fit in the presence of significant zero inflation when compared to the use of these other statistical models for longitudinal data.

Finally, [10] considers marginalized versions of the hurdle model that specify two regression equations, one for the probability of a zero count and one for the truncated non-zero distribution, with distinct model parameters. The zero-altered model that we consider assumes that a single set of regression coefficients, $\boldsymbol{\beta}$, appear in both regression equations but are additively and multiplicatively scaled by γ_1 and γ_2, respectively. Just as importantly, our zero-altered model uses a complimentary log–log function in the model related to the probability of a zero count. The combination of these two features (γ_1 and γ_2 scaling plus log–log link) yield the important feature of the model that an ordinary Poisson mixed model is nested within this model, yielding tests of zero-alteration. Moreover, we derive and provide the analogous novel link that leads to this nesting property when the truncated count distribution is the negative binomial model, thereby extending the methods to cases in which the non-zero counts exhibit overdispersion. To our knowledge this link function has not been previously considered in the literature.

As with any statistical framework, our proposed model has a few limitations. Because each of our parameters of interest rely on the evaluation of the implicit Δ_{ij} function, computing time increases significantly as the number of clusters and observations per cluster increases. Also, if the correlated data structure has multiple nested levels, our marginal likelihood could be multi-dimensional, and model fitting will be computational expensive. However, the simulations fitting the proposed marginalized zero-altered count models suggested the models were not prohibitively slow for a reasonable number of clusters (i.e., $N = 100$). We estimate that for this sample size the models took approximately 80 minutes per 100 datasets, suggesting the model is feasible for larger problems.

Current research focuses on extending this framework to analyze spatially correlated zero-inflated count data. Although one could fit this model via maximum likelihood estimation, we develop Bayesian methods of model fitting. One primary reason for turning to a Bayesian framework is because spatial data often have a multilevel structure. For instance, breast cancer incidence rates may be nested within census tracts, which are then nested within neighborhoods, which are then nested within cities. Because of this structure, there is a possible correlation at multiple levels, which then implies the use of multiple sets of random effects in a conditional mean model. In a similar fashion, the zero-altered models considered in [16] have the advantage that

they are more parsimonious than alternative versions that use distinct parameters for the two parts of the model, but this can also be a source of lack of fit. In some situations, it would be of interest to include distinct sets of random effects in which case alternative estimation methods, such as Bayesian approaches, may be preferred.

Also, semi-continuous longitudinal data, whereby data are either continuous or zero with observations that are correlated, presents similar statistical challenges when in the presence of large zeros and strong skewness. Current methods include modeling this type of data via two-part models [11,17]. Although different, there are close parallels; therefore, extending our proposed marginalized approach to this semi-continuous longitudinal data is of interest.

Finally, it is well known that due to the linearity of the influence functions [7,22] associated with a GEE that is non-robust to outliers, that mixed models have the potential to be more robust because the influence functions have nonlinear form. Although we do not have any specific analytic results in this regard for the proposed class of models considered in this paper, we now acknowledge this as one possible reason for the difference in performance between the two methods and note that this is a possible future research direction.

Acknowledgments The authors would like to acknowledge the following NIH Grants: ES07142, CA134294, ES012044, and ES00002.

Compliance with Ethical Standards

Conflicts of interest No conflict of interest.

Appendix 1: Calculation of Δ_{ij}

In solving for Δ_{ij}, we need to solve the convolution equation that links the marginal (i.e, $\mu_{ij}^Y = \mathrm{E}(Y_{ij}|X_{ij}) = \exp(\beta X_{ij})$) and conditional means, where, assuming the MZAP setting,

$$\begin{aligned}\mu_{ij}^Y &= \exp(X_{ij}\beta)\\ &= \int [1 - P(Y_{ij} = 0|X_{ij}, b_i)]\frac{\mu_{ij}^b}{[1-\exp(-\mu_{ij}^b)]}\phi(b_i|\sigma)\mathrm{d}b_i\\ &= \int \frac{[1-\exp[-\exp(\gamma_1 + \gamma_2(\Delta_{ij} + b_i))]]}{[1-\exp[-\exp(\Delta_{ij} + b_i)]]}\exp(\Delta_{ij} + b_i)\phi(b_i|\sigma)\mathrm{d}b_i.\end{aligned}$$

Estimates of Δ_{ij} can be obtained using a Newton–Raphson algorithm, such that

$$\Delta_{ij}^{(t+1)} = \Delta_{ij}^{(t)} - \left(\frac{\partial f(\Delta_{ij})}{\partial \Delta_{ij}}\right)^{-1} \times f(\Delta_{ij}),$$

where $\Delta_{ij} = \Delta_{ij}(\beta, \gamma_1, \gamma_2, \sigma)$ and $f(\Delta_{ij})$ refers to the convolution equation above. The derivative needed for the Newton–Raphson algorithm is as follows

$$\begin{aligned}\frac{\partial}{\partial \Delta_{ij}}\mu_{ij}^{Y} &= \frac{\partial}{\partial \Delta_{ij}}\int \frac{[1-\exp[-\exp(\gamma_1+\gamma_2(\Delta_{ij}+b_i))]]}{[1-\exp[-\exp(\Delta_{ij}+b_i)]]}\exp(\Delta_{ij}+b_i)\phi(b_i|\sigma)\mathrm{d}b_i \\ &= \int \left\{\frac{\partial}{\partial \Delta_{ij}}\frac{[1-\exp[-\exp(\gamma_1+\gamma_2(\Delta_{ij}+b_i))]]}{[1-\exp[-\exp(\Delta_{ij}+b_i)]]}\exp(\Delta_{ij}+b_i)\right\}\phi(b_i|\sigma)\mathrm{d}b_i.\end{aligned}$$

After using the chain rule, $\diamond = \left\{\frac{\partial}{\partial \Delta_{ij}}\frac{[1-\exp[-\exp(\gamma_1+\gamma_2(\Delta_{ij}+b_i))]]}{[1-\exp[-\exp(\Delta_{ij}+b_i)]]}\exp(\Delta_{ij}+b_i)\right\}$ results in

$$\begin{aligned}\diamond &= \{1-e^{-e^{\gamma_1+\gamma_2(\Delta_{ij}+b_i)}}\} * \left\{\frac{[1-e^{-e^{\Delta_{ij}+b_i}}]e^{\Delta_{ij}+b_i} - e^{\Delta_{ij}+b_i}[-e^{-e^{\Delta_{ij}+b_i}}][-e^{\Delta_{ij}+b_i}]}{[1-e^{-e^{\Delta_{ij}+b_i}}]^2}\right\} \\ &\quad + \left\{\frac{e^{\Delta_{ij}+b_i}}{1-e^{-e^{\Delta_{ij}+b_i}}}\right\} * \{e^{-e^{\gamma_1+\gamma_2(\Delta_{ij}+b_i)}}[e^{\gamma_1+\gamma_2(\Delta_{ij}+b_i)}]\gamma_2\} \\ &= \frac{e^{\Delta_{ij}+b_i}}{1-e^{-e^{\Delta_{ij}+b_i}}} * [\{1-e^{-e^{\gamma_1+\gamma_2(\Delta_{ij}+b_i)}}\} * \left\{1-\frac{[e^{-e^{\Delta_{ij}+b_i}}][e^{\Delta_{ij}+b_i}]}{1-e^{-e^{\Delta_{ij}+b_i}}}\right\} \\ &\quad + \{e^{-e^{\gamma_1+\gamma_2(\Delta_{ij}+b_i)}}[e^{\gamma_1+\gamma_2(\Delta_{ij}+b_i)}]\gamma_2\}].\end{aligned}$$

Gauss–Hermite quadrature can be used to evaluate this one-dimensional integral.

Appendix 2: SAS Execution via PROC NLMIXED

Marginalized zero-altered Poisson (MZAP) and marginalized zero-altered negative binomial (MZANB) models

```
/*MZAP: gamma2 only scaling delta*/
/*dataset must be in long format (i.e., each row contains
  one observation) as opposed to wide*/
proc nlmixed data = toxicity itdetails;
        parms psi0=-2.71148 psi1=0.884423 psi2=0 psi3=0
        psi4=0 psi5=0 sigma=1.578935 gamma1=-1.34175
             gamma2=0.574837;
        mu=exp(psi0 + psi1*exp + psi2*hour1 + psi3*hour2 +
        psi4*hour3 + psi5*hour4);
        bounds gamma2>0;
        array absc  aa1-aa10; /*fixed quadrature nodes
       obtained prior to analysis*/
        array weight ww1-ww10; /*fixed quadrature weights\\
       obtained prior to analysis*/

        delta=log(mu);
        do s=1 to 10;
            denum0=0;
```

```
        denum1=0;
        denum2=0;
        num=-mu;
        do t=1 to 10; /*looping over quadrature points*/
            num=num+weight[t]*(2*3.14)**(-.5)*exp
                    (absc[t]**2)
                *exp(-absc[t]**2/(2))
                *exp(delta+absc[t]*sigma**.5)*(1-exp
                (-(exp(gamma1+gamma2*(delta)+absc[t]
                *sigma**.5))))
                /(1-exp(-(exp(delta+absc[t]*sigma**
                .5))));

            denum0=denum0+weight[t]*(2*3.14)**(-.5)*exp
            (absc[t]**2)
                *exp(-absc[t]**2/(2))*gamma2*exp
               (gamma1\\+gamma2*(delta)+absc[t]*sigma**
               .5)
                *exp(-(exp(gamma1+gamma2*(delta)+absc[t]
                *sigma**.5)))*exp(delta+absc[t]*sigma**
                 .5)/(1-exp(-(exp(delta+absc[t]*sigma**
               .5))));

            denum1=denum1-weight[t]*(2*3.14)**(-.5)*exp
             (absc[t]**2)
                *exp(-absc[t]**2/(2))*(exp(delta+absc
                [t]))**2*exp(-(exp(delta+absc[t]*sigma**
                .5)))
                *(1-exp(-(exp(gamma1+gamma2*(delta)+absc
                [t]*sigma**.5))))
                /(1-exp(-(exp(delta+absc[t]*sigma**
                .5))))**2;

            denum2=denum2+weight[t]*(2*3.14)**(-.5)*exp
            (absc[t]**2)
                *exp(-absc[t]**2/(2))
                *(1-exp(-(exp(gamma1+gamma2*(delta)+absc
                [t]*sigma**.5))))*exp(delta+absc[t]
                *sigma**.5)/(1-exp(-(exp(delta+absc[t]
                *sigma**.5))));
        end;
        delta=delta-num/(denum0+denum1+denum2);
    end;
    prob_0=exp(-(exp(gamma1+gamma2*(delta)+u)));
    mu_c=exp(delta+u);
    ll=r*log(prob_0)+(1-r)*(log(1-prob_0)+y*log(mu_c)
    -mu_c-lgamma(y+1)-log(1-exp(-mu_c)));
    model y~general(ll);
```

```
             random u~normal(0,sigma)  subject=rat;
            ods output ParameterEstimates=Parms_cond3;
predict u OUT=predu1;
title 'MZAP: gamma2 only scaling delta';
run;

/*MZANB: gamma2 only scaling delta*/
/*dataset must be in long format (i.e., each row contains
one observation) as opposed to wide*/
proc nlmixed data = toxicity itdetails;
        parms gamma1=0 gamma2=1 psi0=-0.3988 psi1=0.6565
        psi2=-.0477 psi3=0 psi4=0 psi5=0
        sigma=1.11 a=1;
        mu = exp(psi0 + psi1*exp + psi2*hour1 + psi3*hour2
        + psi4*hour3 + psi5*hour4);
        bounds a>0;
        bounds gamma2>0;
        bounds sigma>0;
        array absc  aa1-aa10; /*fixed quadrature nodes
        obtained prior to analysis*/
        array weight ww1-ww10; /*fixed quadrature weights
        obtained prior to analysis*/
        delta=log(mu);
        do s=1 to 10;
            denum=0;
            num=-mu;
            do t=1 to 10; /*looping over quadrature points*/
                num=num+weight[t]*(2*3.14)**(-.5)*exp
                (absc[t]**2)
                    *exp(-absc[t]**2/(2))
                    *exp(delta+absc[t]*sigma**.5)*(1-(1/(1
                       + a*exp(gamma1+gamma2*(delta)
                    +absc[t]*sigma**.5)))**(1/a))
                    / (1 - (1 / (1 + a * exp(delta+absc[t]
                    *sigma**.5)))**(1/a));

                denum=denum+weight[t]*(2*3.14)**(-.5)*exp
                (absc[t]**2)
                    *exp(-absc[t]**2/(2))
                    *((1-(1/(1 + a*exp(gamma1+gamma2*(delta)
                    +absc[t]*sigma**.5)))**(1/a))
                    *(((1 - (1/(1 + a * exp(delta+absc[t]
                    *sigma**.5)))**(1/a))
                    * exp(delta+absc[t]*sigma**.5)
                    - exp(delta+absc[t]*sigma**.5) * (1/a)
                    * (1/(1 + a * exp(delta+absc[t]*sigma**
                       .5)))**((1/a) + 1)
                    * a * exp(delta+absc[t]*sigma**.5))
```

```
                    /(1 - (1/(1 + a * exp(delta+absc[t]
                    *sigma**.5)))**(1/a))**2)
                    +((exp(delta+absc[t]*sigma**.5)
                    / (1 - (1/(1 + a * exp(delta+absc[t]
                    *sigma**.5)))**(1/a)))
                    *((1/a) * ((1/(1 + a * exp(gamma1
                    +gamma2*(delta)+absc[t]*sigma**.5)))
                    **((1/a) + 1)) * a
                    * exp(gamma1+gamma2*(delta)+absc[t]
                    *sigma**.5) * gamma2)));
end;
            delta=delta-num/(denum);
end;
        prob_0 = (1/(1+a*exp(gamma1+gamma2*(delta)+u)))
        **(1/a);
        mu_c = exp(delta+u);
        p = a*mu_c/(1+a*mu_c);
        ll=r*log(prob_0)+(1-r)*(log(1-prob_0) + lgamma(y +
            (1/a)) - lgamma(y+1) - lgamma(1/a)
           + (1/a)*log(1-p) + y*log(p) - log(1 - (1-p)**
               (1/a)) );
        model y~general(ll);
             random u~normal(0, sigma)  subject=rat;
             ods output ParameterEstimates=Parms_cond1;
predict u OUT=predu3;
title 'MZANB: gamma2 only scaling delta';
run;
```

References

1. Akaike H (1987) Factor analysis and AIC. Psychometrika 52(3):317–332
2. Albert A, Anderson JA (1984) On the existence of maximum likelihood estimates in logistic regression models. Biometrika 71(1):1–10
3. Everitt BS (1998) The cambridge dictionary of statistics. Cambridge University Press, Cambridge
4. Hall DB (2000) Zero-inflated poisson and binomial regression with random effects: a case study. Biometrics 56(4):1030–1039
5. Hall DB, Berenhaut KS (2002) Score tests for heterogeneity and overdispersion in zero-inflated poisson and binomial regression models. Can J Stat 30(3):415–430
6. Hall DB, Zhang Z (2004) Marginal models for zero inflated clustered data. Stat Model 4(3):161–180
7. Hampel FR (1974) The influence curve and its role in robust estimation. J Am Stat Assoc 69(346):383–393
8. Heagerty PJ (1999) Marginally specified logistic-normal models for longitudinal binary data. Biometrics 55(3):688–698
9. Heagerty PJ, Zeger SL (2000) Marginalized multilevel models and likelihood inference (with comments and a rejoinder by the authors). Stat Sci 15(1):1–26
10. Kassahun W, Neyens T, Molenberghs G, Faes C, Verbeke G (2014) Marginalized multilevel hurdle and zero-inflated models for overdispersed and correlated count data with excess zeros. Stat Med 33(25):4402–4419
11. Lachenbruch PA (2002) Analysis of data with excess zeros. Stat Methods Med Res 11(4):297–302

12. Lambert D (1992) Zero-inflated poisson regression, with an application to defects in manufacturing. Technometrics 34(1):1–14
13. Lesaffre E, Albert A (1989) Partial separation in logistic discrimination. J R Stat Soc Ser B 51:109–116
14. Lu S-E, Lin Y, Shih W-CJ (2004) Analyzing excessive no changes in clinical trials with clustered data. Biometrics 60(1):257–267
15. Miglioretti DL, Heagerty PJ (2004) Marginal modeling of multilevel binary data with time-varying covariates. Biostatistics 5(3):381–398
16. Min Y, Agresti A (2005) Random effect models for repeated measures of zero-inflated count data. Stat Model 5(1):1–19
17. Olsen MK, Schafer JL (2001) A two-part random-effects model for semicontinuous longitudinal data. J Am Stat Assoc 96(454):730–745
18. Philip LP (2010) Multilevel models for zero-inflated count data in environmental health and health disparities research. Ph.D. thesis, Harvard University
19. Qu A, Song PX-K (2004) Assessing robustness of generalised estimating equations and quadratic inference functions. Biometrika 91(2):447–459
20. Ridout M, Demétrio CGB, Hinde J (1998) Models for count data with many zeros. Proceedings of the sixth international biometric conference, vol. 19, pp 179–192
21. Ridout M, Hinde J, DemeAtrio CGB (2001) A score test for testing a zero-inflated poisson regression model against zero-inflated negative binomial alternatives. Biometrics 57(1):219–223
22. Rousseeuw FR, Hampel EM, Ronchetti PJ, Stahel WA (1986) Robust statistics: the approach based on influence functions. Wiley, New York
23. Schildcrout JS, Heagerty PJ (2007) Marginalized models for moderate to long series of longitudinal binary response data. Biometrics 63(2):322–331
24. Shankar V, Milton J, Mannering F (1997) Modeling accident frequencies as zero-altered probability processes: an empirical inquiry. Accid Anal Prev 29(6):829–837
25. Tooze JA, Grunwald GK, Jones RH (2002) Analysis of repeated measures data with clumping at zero. Stat Methods Med Res 11(4):341–355
26. Wellenius Gregory A, Diaz Edgar A, Gupta Tarun, Ruiz Pablo A, Long Mark, Kang Choong Min, Coull Brent A, Godleski John J (2011) Electrocardiographic and respiratory responses to coal-fired power plant emissions in a rat model of acute myocardial infarction: results from the toxicological evaluation of realistic emissions of source aerosols study. Inhal Toxicol 23(S2):84–94
27. Yau KKW, Lee AH (2001) Zero-inflated poisson regression with random effects to evaluate an occupational injury prevention programme. Stat Med 20(19):2907–2920
28. Zeger SL, Liang K-Y, Albert PS (1988) Models for longitudinal data: a generalized estimating equation approach. Biometrics 44:1049–1060