



# Zero-inflated models for adjusting varying exposures: a cautionary note on the pitfalls of using offset

Cindy Feng <sup>a,b</sup>

<sup>a</sup>School of Epidemiology and Public Health, Faculty of Medicine, University of Ottawa, Ottawa, Canada;

<sup>b</sup>School of Public Health, University of Saskatchewan, Saskatoon, Canada

## ABSTRACT

Zero-inflated count data are frequently encountered in public health and epidemiology research. Two-parts model is often used to model the excessive zeros, which are a mixture of two components: a point mass at zero and a count distribution, such as a Poisson distribution. When the rate of events per unit exposure is of interest, offset is commonly used to account for the varying extent of exposure, which is essentially a predictor whose regression coefficient is fixed at one. Such an assumption of exposure effect is, however, quite restrictive for many practical problems. Further, for zero-inflated models, offset is often only included in the count component of the model. However, the probability of excessive zero component could also be affected by the amount of ‘exposure’. We, therefore, proposed incorporating the varying exposure as a covariate rather than an offset term in both the probability of excessive zeros and conditional counts components of the zero-inflated model. A real example is used to illustrate the usage of the proposed methods, and simulation studies are conducted to assess the performance of the proposed methods for a broad variety of situations.

## ARTICLE HISTORY

Received 18 January 2020

Accepted 12 July 2020

## KEYWORDS

Count data; zero-inflated models; exposure; offset



## 2010 MATHEMATICS SUBJECT CLASSIFICATION


62-07

## 1. Introduction

In public health and epidemiology research, count data with a large proportion of zeros are often encountered. For example, in health services utilization study, the number of service utilization often includes a large number of zeros representing the patients with no utilization during the study period. A common feature of this type of data is that the count measure tends to have excessive zeros beyond a common count distribution that can accommodate, such as Poisson or negative binomial (NB).

To overcome the issue with excessive zeros, the so-called zero-inflated (ZI) models [19] can be specified, which are a mixture of two components: a point mass at zero and a count distribution, such as a Poisson or negative binomial distribution. An alternative modeling strategy is hurdle model [17,23], which assumes all zero data are from one ‘structural’

**CONTACT** Cindy Feng  [cindy.feng@uottawa.ca](mailto:cindy.feng@uottawa.ca)  School of Epidemiology and Public Health, Faculty of Medicine, University of Ottawa, 600 Peter Morand Crescent, Ottawa, Ontario, K1G5Z3, Canada; School of Public Health, University of Saskatchewan, Saskatoon, Saskatchewan, S7N2Z4, Canada.

 Supplemental data for this article can be accessed here. <https://doi.org/10.1080/02664763.2020.1796943>

source with one part of the model being a binary model for modeling whether the response variable is zero or positive, and another part using a zero truncated model, such as a zero truncated Poisson or a zero truncated NB distribution for the positive data. Both types of models were rapidly embraced by a large number of areas, from population and epidemiological studies to ecological studies [7,10,11,20,22,24,25,27,34,35].

Despite the widespread application of both types of models in the literature, one question arises of how to address the effect of varying *exposure* (underlying population or duration at risk) in such models. For example, for geographically distributed disease count data, the number of cases in region  $i$  might be larger than that in the region  $j$  because region  $i$  had a substantially larger population at risk than the region  $j$ . As a result, a higher number of disease counts in one region compared to another does not necessarily imply subjects in this region have a higher susceptibility to this disease. Similarly, in longitudinal studies, the number of repeated events may depend on the follow-up time for the patients. Event rates can be calculated as events per unit time, which allows the observation window to vary for each unit. In these examples, exposure is respectively unit area, person-years and unit time.

In literature, a commonly used method for incorporating a population size at risk and/or the amount of exposure time is through the introduction of an offset term, i.e. the log of exposure, as an explanatory variable whose coefficient is fixed at one [1]. For example, for a Poisson log-linear model with expected mean  $\mu_i$  and covariate  $X_i$ , the model can be written as  $\log(\mu_i) = X_i^T \alpha + \log(E_i)$ , where  $\log(E_i)$  is referred to as an offset. This implies

$$\mu_i = e^{X_i^T \alpha + \log(E_i)}, \quad \text{i.e.} \quad \mu_i = E_i e^{X_i^T \alpha}. \quad (1)$$

This means that the mean count  $\mu_i$  has a proportionality constant for  $E_i$  that depends on the values of the explanatory variables. However, such proportionality assumption may not be plausible. For example, for modeling the incidence of infectious disease, the heterogeneity of the underlying population may have a varying effect on the likelihood and intensity of disease transmission. As a result, the number of events in the response variable may increase non-proportionally with the population at risk. When such sophisticated exposure effects arise from applications, the assumption embedded in the offset term becomes inadequate. Hence, using offset to adjust the varying extent of exposure as a widely accepted common practice should be carefully examined and used with caution.

For zero-inflated models, offset is often incorporated only in the count component of the model ([13,20,21,32,38], for example). Hall [16] considered relaxing the assumption in offset by setting the coefficient on the logarithm of exposure as an unknown parameter in the count component to be estimated in the model-fitting procedure. However, the probability of observing excessive zeros can also be impacted by varying exposure in many situations; that is, the probability of excessive zeros is expected to decrease with increasing exposure. Baetschmann *et al.* [4] proposed a modified zero-inflated count model where the probability of extra zero is derived from an underlying duration model with Weibull hazard rate. However, to the best of our knowledge, no attempt has been made to extend the zero-inflated model to adjust for the extent of exposure as a covariate in both excess zeros and the count components, particularly in the context of modeling disease incidence collected over a geographical region. Further, simulation studies to explore the impact of misspecification of modeling the effect of varying exposures are limited and therefore warranted.

The overall goal of this study is to discuss the impact of misspecification of the modeling method for varying exposures for zero-inflated models. We explored two potential types of misspecification:

- The effect of varying exposure may differ from one. In the situation when the mean count is not proportional to the population at risk, imposing the offset term by constraining the effect of the population at risk as one can be very restrictive. Such a constraint may lead to a biased estimation of the parameter estimates and prediction through a biased estimation of the effect of exposure.
- Both excess zero and count components may depend on varying exposures. In literature, the zero-inflated model typically only includes the exposure as an offset in the count component of the zero-inflated model. We show that ignoring varying exposures for the binary part can lead to biased parameter estimates and can be sensitive to the degree of the effect of exposures.

The results for the hurdle models are consistent with zero-inflated models, so for the ease of presentation, we chose to focus on the zero-inflated models. The remainder of the paper is organized as follows. In Section 2, a review of zero-inflated models without and with varying extent of exposure is presented. Section 3 describes the methods for model selection and diagnosis check for zero-inflated models. To demonstrate the pitfalls of using an offset term for zero-inflated models, in Section 4, a real example of a health care utilization study is given. Simulation studies comparing the finite sample performance of the approaches for accounting for varying extent of exposure for zero-inflated models are presented in Section 5. Concluding remarks are given in Section 6.

## 2. Statistical models

### 2.1. Zero-inflated model

In a zero-inflated (ZI) model [19], zero observations have two different origins: ‘structural’ and ‘sampling’. The sampling zeros are due to the usual Poisson or negative binomial (NB) distribution, which assumes that those zero observations happened by chance.

Let  $Y_i$  denote the response for the  $i$ th subject,  $i = 1, \dots, n$ . The zero-inflated Poisson (ZIP) model is given by:

$$Y_i \sim \begin{cases} 0 & \text{with probability } \pi_i \\ \text{Poisson}(\mu_i) & \text{with probability } 1 - \pi_i \end{cases}, \tag{2}$$

where  $\pi_i$  denotes the probability of the observation arising from the degenerated distribution at zero and  $\mu_i$  represents the mean of the Poisson distribution. This formulation allows for more zeros than permitted under the Poisson assumption when  $\pi_i > 0$ . The probability distribution function of the ZIP model can be written as

$$P(Y_i = y_i | \pi_i, \mu_i) = \begin{cases} \pi_i + (1 - \pi_i) e^{-\mu_i} & \text{if } y_i = 0, \\ (1 - \pi_i) \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} & \text{if } y_i > 0 \end{cases}. \tag{3}$$

ZIP model can include covariates for modeling both  $\mu_i$  and  $\pi_i$ . Generally,  $\pi_i$  is modeled with a logistic regression and  $\mu_i$  is modeled as a log-linear regression. The ZI model can be written as,

$$\begin{aligned}\text{logit}(\pi_i) &= \mathbf{Z}_i^T \boldsymbol{\beta} \\ \log(\mu_i) &= \mathbf{X}_i^T \boldsymbol{\alpha}\end{aligned}\quad (4)$$

where  $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_{p_1})^T$  is the  $(p_1 \times 1)$  column vector of parameters associated with the excess zeros and  $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \dots, \alpha_{p_2})^T$  is a  $(p_2 \times 1)$  column vector of parameters associated with the Poisson process and  $\mathbf{Z}_{i(1 \times p_1)}^T$  and  $\mathbf{X}_{i(1 \times p_2)}^T$  are the vectors of covariates for the  $i$ th study subject for the excess zeros and Poisson processes, respectively. Note that the explanatory variables describing the  $\mu_i$  do not need to be the same as those describing  $\pi_i$ . To account for the unobserved heterogeneity, one can assume the Poisson process as  $\log(\mu_i) = \mathbf{X}_i^T \boldsymbol{\alpha} + b_i$ , where  $b_i$  is a random effect term, which can follow an independent Normal(0,  $\sigma_b^2$ ). In this article, we consider the common logit and log link functions for the binary and count outcomes. Other choices of link functions are possible, such as probit or complementary log-log link functions for the binary component.

The ZIP model can be regarded as a mixture of Poisson and a degenerate component with all of its mass at zero. As such, this model posits an unobserved and latent binary variable  $\delta_i$  with  $\delta_i = 1, y_i = 0$  and  $\delta_i = 0, y_i$  is a Poisson ( $\mu_i$ ) variate. The marginal mean of the ZIP model can be then derived as

$$E(Y_i) = E[E(Y_i | \delta_i)] = p(\delta_i = 1)E(Y_i | \delta_i = 1) + p(\delta_i = 0)E(Y_i | \delta_i = 0) = (1 - \pi_i)\mu_i. \quad (5)$$

The second equality holds because  $\delta_i = 1$  implies  $y_i = 0$ . The variance can be derived as

$$\text{Var}(Y_i) = E[\text{Var}(Y_i | \delta_i)] + \text{Var}[E(Y_i | \delta_i)] \quad (6)$$

$$= [\pi_i \cdot 0 + (1 - \pi_i)\mu_i] + \{\pi_i[0 - (1 - \pi_i)\mu_i]^2 + (1 - \pi_i)[\mu_i - (1 - \pi_i)\mu_i]^2\} \quad (7)$$

$$= (1 - \pi_i)\mu_i[1 + \pi_i\mu_i]. \quad (8)$$

As a result, zero-inflated model can accommodate overdispersion relative to a Poisson model, since  $\text{Var}(Y_i) > E(Y_i)$ . This also indicates that model misspecification of either the binary or Poisson component of a ZIP model can lead to biased predicted mean and variance estimations.

## 2.2. Zero-inflated models with varying exposures

In modeling zero-inflated count data, the population at risk and/or the amount of time of exposure are often heterogeneous among the study subjects. In this article, we refer to the population at risk and/or amount of time of exposure as ‘exposure’. The variable exposure for the positive count process is handled typically through an offset term, as is typically done in a log-linear Poisson regression. For example, let  $E_i$  denote the population at risk for the disease of interest. An offset term,  $\log(E_i)$  is often incorporated into the count

component of a ZI model to account for variable exposure, so the model can be written as,

$$\begin{aligned} \text{logit}(\pi_i) &= \mathbf{Z}_i^T \boldsymbol{\beta} \\ \log(\mu_i) &= \mathbf{X}_i^T \boldsymbol{\alpha} + \log(E_i) \end{aligned} \tag{9}$$

This model implicitly assumes that all subjects who belong to the excessive zero component with the same covariates profiles are at the same risk of experiencing the outcome regardless of the size of the population at risk. This may not be plausible, as the probability of observing excessive zero is likely to decrease as the exposure size increases. Ignoring the differential exposure may result in biased estimates for both the binary and Poisson components of the ZIP model. A direct adaptation of the model ZIP-O<sup>c</sup> is to introduce an offset term in the binary component of the model as well; that is,

$$\begin{aligned} \text{logit}(\pi_i) &= \mathbf{Z}_i^T \boldsymbol{\beta} + \log(E_i) \\ \log(\mu_i) &= \mathbf{X}_i^T \boldsymbol{\alpha} + \log(E_i) \end{aligned} \tag{10}$$

Nevertheless, this model can be implausible, since zero inflation and the conditional model work in opposite directions. (i.e. a higher expected value for the zero inflation ( $\pi_i$ ) leads to a lower response, but a higher value for the conditional model ( $\mu_i$ ) leads to a higher response). Further, restricting the effect of exposure as one can be inconsistent with the true extent of association between exposure and the outcome.

Therefore, we propose modifying the ZIP model to addresses the variable exposure not only for the count component but also for the binary process by incorporating exposure as a covariate in both binary and count components of the model. The modified ZIP model can be expressed as,

$$\begin{aligned} \text{logit}(\pi_i) &= \mathbf{Z}_i^T \boldsymbol{\beta} + g_1(E_i) \\ \log(\mu_i) &= \mathbf{X}_i^T \boldsymbol{\alpha} + g_2(E_i) \end{aligned} \tag{11}$$

where  $g_1(E_i)$  and  $g_2(E_i)$  represent the functional effects of the extent of exposure ( $E_i$ ) for the binary and count components of a ZIP model, respectively, which can take on any form, such as polynomials or spline functions, or can be modeled as:

$$g_1(E_i) = \xi_1 \log(E_i), \quad g_2(E_i) = \xi_2 \log(E_i), \tag{12}$$

where  $\xi_1$  and  $\xi_2$  are the regression coefficients for the logarithm transformed  $E_i$ . This model relaxes the assumption made on the offset term by allowing  $\xi_1$  and  $\xi_2$  to deviate from one and also have opposite signs.

### 2.3. Statistical inference

The log-likelihood function for the proposed ZIP model as presented in Equation (11) is given by:

$$\begin{aligned} \log L(\boldsymbol{\beta}, \boldsymbol{\alpha}) &= \sum_{i=1}^n \left( I(y_i = 0) \log[\pi_i + (1 - \pi_i) e^{-\mu_i}] \right. \\ &\quad \left. + I(y_i > 0) [\log(1 - \pi_i) - \mu_i + y_i \log(\mu_i) - \log(y_i!)] \right), \end{aligned}$$

where  $\pi_i = \exp[\mathbf{Z}_i^T \boldsymbol{\beta} + g_1(E_i)] / \{1 + \exp[\mathbf{Z}_i^T \boldsymbol{\beta} + g_1(E_i)]\}$  and  $\mu_i = \exp[\mathbf{X}_i^T \boldsymbol{\alpha} + g_2(E_i)]$ . For the model accounting for the unobserved regional variation, the marginal log likelihood function of the ZIP model with random effects can be written as,

$$\log L(\boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma) = \sum_{i=1}^n \log \left[ \int_{-\infty}^{+\infty} P(Y_i = y_i | b_i, \boldsymbol{\beta}, \boldsymbol{\alpha}, \sigma) \Phi(b_i) db_i \right], \quad (13)$$

where the unobserved heterogeneity is quantified by the random effect  $b_i$ , which is assumed to be Gaussian on the scale of the linear predictor with mean zero and standard error  $\sigma_b$ .  $\Phi(\cdot)$  is the standard normal density function. Lambert [19] expressed the log-likelihood in terms of latent variables  $\delta_i$  and used the EM algorithm for maximum likelihood fitting by treating  $\delta_i$  as missing values. For a model with random effect terms, numerical integration techniques, such as Gauss-Hermite quadrature or Markov chain Monte Carlo (MCMC) can be used. Nevertheless, those methods can be computationally intensive. Alternatively, the models can be fitted using `glmmTMB` R package [8], which performs maximum likelihood estimation via TMB (Template Model Builder) [18]. To maximize computational efficiency, TMB uses the Laplace approximation to integrate over random effects and automatic differentiation to estimate the first and second derivatives of the log likelihood function [18]. This package is more flexible than other packages available for estimating zero-inflated models via maximum likelihood estimation and is faster than packages that use MCMC sampling for estimation [8].

### 3. Model selection and diagnostic checks

To inform model selection, we use the Akaike Information Criterion (AIC) [2] and Bayesian information criteria (BIC) [28]. AIC and BIC are defined as  $AIC = D + 2p$ ,  $BIC = D + m \log(n)$ , where  $m$  is the number of parameters in the model,  $n$  is number of observations,  $D$  is the deviance defined as twice of negative log likelihood in the ZIP model.  $D = \sum_{i=1}^n d_i$ , where

$$d_i = -2 \begin{cases} \log [\pi_i + (1 - \pi_i) e^{-\mu_i}] & \text{if } y_i = 0, \\ \log \left[ (1 - \pi_i) \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} \right] & \text{if } y_i > 0 \end{cases}. \quad (14)$$

The smaller the values of AIC and BIC, the better a model fits the data. Examining residuals is a standard tool for assessing the adequacy of regression models. For discrete response, Pearson or deviance residuals are far from normality; graphical and quantitative inspection of these residuals provides little information for model diagnosis [14]. Hence, the adequacy of the ZIP models is examined on the basis of randomized quantile residuals (RQR), as developed by Dunn and Smyth [14]. RQR can be defined as follows. Suppose  $F(y_i; \pi_i, \mu_i)$  denote the CDF for the response variable  $y_i$  following ZIP distribution given the set of covariates  $\mathbf{Z}_i$  and  $\mathbf{X}_i$ , for the binary and count component, respectively, where  $\pi_i = \exp(\mathbf{Z}_i^T \boldsymbol{\beta} + g_1(E_i)) / \{1 + \exp(\mathbf{Z}_i^T \boldsymbol{\beta} + g_1(E_i))\}$  and  $\mu_i = \exp(\mathbf{X}_i^T \boldsymbol{\alpha} + g_2(E_i))$ . Let  $d(y_i; \pi_i, \mu_i)$  be the corresponding probability mass function of  $F(y_i; \pi_i, \mu_i)$ . Since  $F$  is discrete, it is then randomized into a uniform random number, which is defined as a function with a random number  $u_i$  from the uniform distribution on  $(0, 1]$  as an additional

argument,

$$F(y_i; \hat{\pi}_i, \hat{\mu}_i, u_i) = F(y_i-; \hat{\pi}_i, \hat{\mu}_i) + u_i d(y_i; \hat{\pi}_i, \hat{\mu}_i), \tag{15}$$

where  $F(y_i-; \hat{\pi}_i, \hat{\mu}_i)$  is the lower limit of  $F$  at  $y_i$ , i.e.  $\sup_{y < y_i} F(y; \hat{\pi}_i, \hat{\mu}_i)$ , the lower limit in the ‘gap’ of  $F(\cdot, \hat{\pi}_i, \hat{\mu}_i)$  at  $y_i$ . RQR for  $y_i$  is the standard normal quantile corresponding to the random lower tail probability with  $\pi_i$  and  $\mu_i$  estimated from the sample,

$$r_i = \Phi^{-1}[F(y_i; \hat{\pi}_i, \hat{\mu}_i, u_i)], \quad i = 1, \dots, n, \tag{16}$$

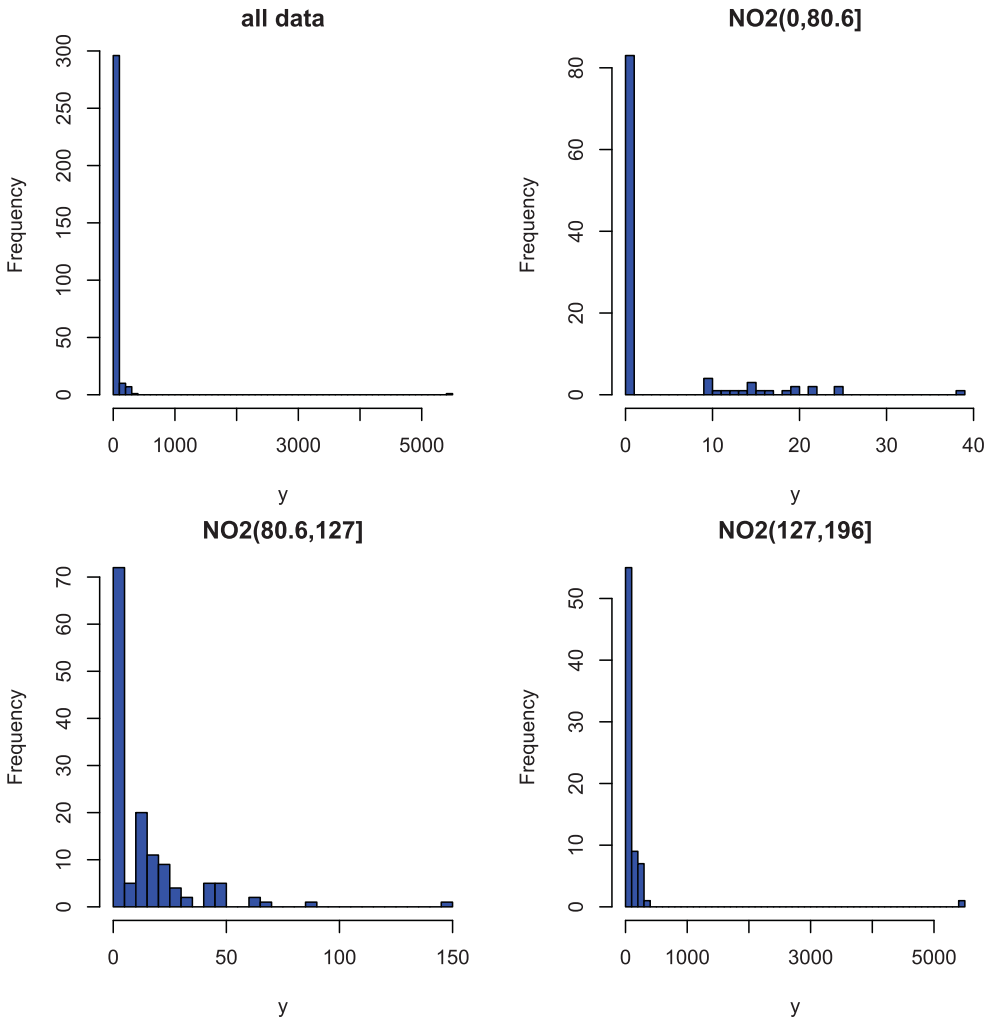
where  $\Phi^{-1}(\cdot)$  is the quantile function of the standard normal distribution, and  $u_i$  is a random number uniformly distributed on  $(0, 1]$ . These residuals are expected to approximately follow a standard normal distribution, if the model is correctly specified. Hence, the validity of the model can be assessed by graphing the RQRs versus the predicted response variable. If the model fits the data well, RQRs should be randomly scattered between  $-3$  and  $3$  without a discernible pattern. The normality of the RQRs can be examined using Q-Q plots, which should lie along the straight diagonal line, if the model is correctly specified.

#### 4. Motivating example: respiratory hospital admissions data

The adverse effect of ambient air pollution has drawn considerable attention over the past decade and has been shown to be associated with respiratory morbidity and mortality. In this motivating example, our goal is to study the relationship between nitrogen dioxide ( $\text{NO}_2$ ) and the number of hospital admissions for respiratory causes in Turin province (Italy) in 2004, while accounting for the differential size of the population at risk from the study area. The dataset records the number of observed hospitalizations for respiratory causes and population size at the municipality level as well as the average  $\text{NO}_2$  for the same period and the same areas. The data were obtained from the data repository of the *Spatial and Spatial-temporal Bayesian Models with R-INLA* [6].

Of the 315 municipalities, 173 (54.9%) had zero hospitalizations. We categorize  $\text{NO}_2$  according to its tertiles, i.e. 80.57 and 126.54 into three levels  $\text{NO}_2 \leq 80.57$  (reference category),  $80.57 < \text{NO}_2 \leq 126.54$  and  $\text{NO}_2 > 126.54$ . As shown in Figure 1, the distribution of the hospitalization counts is highly positively skewed and the distribution depends on the values of  $\text{NO}_2$  with higher response values occurring at a higher level of  $\text{NO}_2$ . Let  $Y_i$  and  $n_i$  denote the number of observed hospitalizations and size of the population at risk at the  $i$ th municipality, respectively. In the context of disease mapping, the expected number is often included as an offset term, which is often expressed as the number of cases defined by an epidemiologic ‘null model’ of incidence, i.e. the product of  $n_i$ , the number of individuals at risk in region  $i$ , and  $r$ , a constant ‘baseline’ risk per individual defined as  $r = \sum_{i=1}^n Y_i / \sum_{i=1}^n n_i$ , the global observed disease rate. The following four competing models are considered, which are expressed as

$$\text{ZIP-W}^b \begin{cases} \logit(\pi_i) = \beta_0 + \beta_1 \text{NO}_{2i}^{(1)} + \beta_2 \text{NO}_{2i}^{(2)} + \xi_1 \log(E_i) \\ \log(\mu_i) = \alpha_0 + \alpha_1 \text{NO}_{2i}^{(1)} + \alpha_2 \text{NO}_{2i}^{(2)} + b_i + \xi_2 \log(E_i), \end{cases} \tag{17}$$



**Figure 1.** Distributions of number of hospital admissions for respiratory causes over 315 municipalities in Turin province (Italy) in 2004. The top left panel is for the whole data and rest of the panels are for the data stratified by three categories of  $\text{NO}_2$ .

$$\text{ZIP-}W^c \begin{cases} \text{logit}(\pi_i) = \beta_0 + \beta_1 \text{NO}_{2i}^{(1)} + \beta_2 \text{NO}_{2i}^{(2)} \\ \log(\mu_i) = \alpha_0 + \alpha_1 \text{NO}_{2i}^{(1)} + \alpha_2 \text{NO}_{2i}^{(2)} + b_i + \xi_2 \log(E_i), \end{cases} \quad (18)$$

$$\text{ZIP-}O^b \begin{cases} \text{logit}(\pi_i) = \beta_0 + \beta_1 \text{NO}_{2i}^{(1)} + \beta_2 \text{NO}_{2i}^{(2)} + \log(E_i) \\ \log(\mu_i) = \alpha_0 + \alpha_1 \text{NO}_{2i}^{(1)} + \alpha_2 \text{NO}_{2i}^{(2)} + b_i + \log(E_i), \end{cases} \quad (19)$$

$$\text{ZIP-}O^c \begin{cases} \text{logit}(\pi_i) = \beta_0 + \beta_1 \text{NO}_{2i}^{(1)} + \beta_2 \text{NO}_{2i}^{(2)} \\ \log(\mu_i) = \alpha_0 + \alpha_1 \text{NO}_{2i}^{(1)} + \alpha_2 \text{NO}_{2i}^{(2)} + b_i + \log(E_i), \end{cases} \quad (20)$$

where  $\pi_i$  is the probability of no hospital admissions at the  $i$ th municipality and  $\mu_i$  is the expected mean number of hospitalizations of the Poisson distribution;  $\text{NO}_2^{(1)}$  and



$\text{NO}_2^{(2)}$  denote the dummy variables for  $80.57 < \text{NO}_2 \leq 126.54$  and  $\text{NO}_2 > 126.54$ , respectively. To account for unobservable heterogeneity, the area-specific random effect,  $b_i \sim \text{Normal}(0, \sigma_b^2)$ , is included in the model. In the model ZIP- $W^b$ , the coefficient on  $\log(E_i)$  is considered as an unknown parameter in both the binary and Poisson processes to be estimated in the model-fitting procedure, where the superscript  $b$  refers to ‘both’ components; ZIP- $W^c$  includes  $\log(E_i)$  as an explanatory variable only in the count component, where the superscript  $c$  refers to ‘count’ component. In contrast, ZIP- $O^b$  considers the population at risk as an offset term in both the binary and count components of the ZIP model and ZIP- $O^c$  only includes the offset term in the count component. The parameter estimations were carried out in *R* (R Core Team, 2019) via `glmmTMB` package [8]. For the binary component of the candidate models,  $\text{NO}_2^{(1)}$  and  $\text{NO}_2^{(2)}$  are not significantly associated with the probability of excessive zeros and therefore were removed from the binary component of the models.

From Table 1, ZIP- $W^b$  gave the smallest values of AIC and BIC, suggesting that they provided the best fit to the data compared to the other competing models. The intercept for the binary component of the ZIP- $W^b$  model is significantly different from zero with  $\hat{\beta}_0 = 4.24$  ( $p$ -value  $< 0.001$ ) and is positive. Under model ZIP- $O^b$ , the estimated intercept is negative and significantly different from zero  $\hat{\beta}_0 = -3.31$  ( $p$ -value  $< 0.001$ ). By comparison, the estimated intercept of the binary component is not significantly different from zero under models ZIP- $W^c$  and ZIP- $O^c$ . The opposite signs of the estimated intercept under ZIP- $O^b$  compared to ZIP- $W^b$  and non-significance under models ZIP- $W^c$  and ZIP- $O^c$  are due to the model misspecification of the exposure effect in the binary component, so  $\beta_0$  is trying to recover from this misspecification.

The effect of  $\log(E_i)$  for the binary component of the model ZIP- $W^b$  is estimated as  $\xi_1 = -2.27$  ( $p$ -value  $< 0.001$ ), which models the fact that the probability of observing an excess zero count decreases as the extent of exposure increases. In other words, odds of observing an excess zero count are inversely proportional to  $E_i$ . In contrast, the effect of  $\log(E_i)$  for the count component of the ZIP- $W^b$  model is estimated as  $\xi_2 = 0.73$  ( $p$ -value  $< 0.001$ ), which reflects that the conditional mean count increases as the extent of exposure increases.

**Table 1.** Parameter estimates (Est), standard error (SE),  $p$ -value, AIC and BIC values for the ZIP- $W^b$ , ZIP- $W^c$ , ZIP- $O^b$  and ZIP- $O^c$  models for modeling the number of hospital admissions for respiratory causes over 315 municipalities in Turin province (Italy) in 2004.

	ZIP- $W^b$			ZIP- $W^c$			ZIP- $O^b$			ZIP- $O^c$		
	Est	SE	$p$ -value	Est	SE	$p$ -value	Est	SE	$p$ -value	Est	SE	$p$ -value
<i>Binary component</i>												
$\beta_0$ (intercept)	4.24	0.52	0.00	0.03	0.12	0.81	-3.31	0.20	0.00	-0.11	0.13	0.39
$\xi_1$ ( $\log(E)$ )	-2.27	0.27	0.00									
<i>Count component</i>												
$\alpha_0$ (intercept)	1.04	0.13	0.00	0.79	0.14	0.00	-0.55	0.15	0.00	0.21	0.13	0.10
$\alpha_1$ ( $\text{NO}_2^{(1)}$ )	0.24	0.12	0.04	0.34	0.12	0.00	0.85	0.17	0.00	0.36	0.14	0.01
$\alpha_2$ ( $\text{NO}_2^{(2)}$ )	0.43	0.12	0.00	0.48	0.13	0.00	0.91	0.17	0.00	0.26	0.14	0.07
$\xi_2$ ( $\log(E)$ )	0.73	0.04	0.00	0.78	0.04	0.00						
$\sigma_b^2$	0.15	0.38	0.00	0.15	0.39	0.00	0.48	0.69	0.00	0.23	0.48	0.00
<i>Model fit</i>												
AIC		1358.04			1560.22			1856.33			1585.32	
BIC		1384.30			1582.74			1875.10			1604.08	

Thus, including the population at risk as an offset term by imposing its effect as one results in an incorrect assumption, particularly for the binary component. This is evident from the model comparison between ZIP- $W^b$  (AIC = 1358.04, BIC = 1384.30) and ZIP- $O^b$  (AIC = 1856.33, BIC = 1875.10) or ZIP- $W^c$  (AIC = 1560.22, BIC = 1582.74) and ZIP- $O^c$  (AIC = 1585.32, BIC = 1604.08). For covariates included in the Poisson component of the ZIP models,  $NO_2^{(1)}$  and  $NO_2^{(2)}$  are significantly and positively related to the expected counts of hospital admissions under the best fitting model ZIP- $W^b$  with the estimated effects equal to  $\hat{\alpha}_1 = 0.24$  for  $NO_2^{(1)}$  and  $\hat{\alpha}_1 = 0.43$  for  $NO_2^{(2)}$ . The effects of  $NO_2^{(1)}$  and  $NO_2^{(2)}$  are estimated to be slightly larger under model ZIP- $W^c$  and much larger under ZIP- $O^b$  as compared to ZIP- $W^b$ . A notable difference between ZIP- $W^b$  and ZIP- $O^b$  is that the two methods provide estimates of opposite signs for the intercept. Under model ZIP- $O^c$ , only  $NO_2^{(1)}$  is positively and significantly associated with the conditional mean number of hospitalizations with  $\hat{\alpha}_1 = 0.36$ , but  $NO_2^{(2)}$  is not significant at the 5% level of significance. The variances of the random effect  $\sigma_b^2$  are all estimated significantly different from zero, suggesting the importance of accounting for unobserved regional effect. We also observe that the variance component under the model ZIP- $O^b$  ( $\hat{\sigma}_b^2 = 0.48$ ) is much larger than the model ZIP- $O^c$  ( $\hat{\sigma}_b^2 = 0.23$ ). Models ZIP- $W^b$  and ZIP- $W^c$  give an almost identical estimate for the variance component ( $\hat{\sigma}_b^2 = 0.15$ ). These results suggest that including exposure as a covariate in either or both of the binary and count components of the ZIP model can explain the residual variation of the response variable.

We have so far only considered  $NO_2$  and extent of exposure have an ‘additive’ effect; however,  $NO_2$  and exposure may have an interaction effect. We, therefore, extended ZIP- $W^b$  and ZIP- $W^c$  models by including the interactions between  $\log(E)$  and  $NO_2^{(1)}$  and  $NO_2^{(2)}$  in both the binary and count components, named as ZIP- $W^{b2}$  and only in the count component, named as ZIP- $W^{c2}$ . Our model fit shows that  $NO_2^{(1)}$  and  $NO_2^{(2)}$  have no main and interaction effect with  $\log(E)$  for the binary component of the ZIP model and therefore were excluded from the model ZIP- $W^{b2}$ . Nevertheless,  $NO_2^{(1)}$  and  $NO_2^{(2)}$  have significant interaction effect with exposure for the Poisson component of the ZIP- $W^{b2}$  model, but not for ZIP- $W^{c2}$ , as shown in Table 2. ZIP- $W^{b2}$  also gives the best model fit among all the competing models (AIC = 1325.53, BIC = 1359.30). These results indicate that (i) failing to include  $\log(E)$  in the binary component of the model may lead to incorrect inference for the parameters in the count component (ii) covariates and extent of exposure can have a significant interaction effect on the outcome.

We also considered fitting Poisson and NB models including the random effect term at the municipality level with and without interactions between  $NO_2$  and  $\log(E)$  to examine if a simpler model can adequately describing the data. Our results indicate that Poisson and NB models fit to the data worse than the candidate ZIP models, i.e. for Poisson without interaction (AIC = 1771.9, BIC = 1786.9) and NB without interaction (AIC = 1720.7, BIC = 1739.5) and Poisson with interaction (AIC = 1710.0, BIC = 1736.3) NB with interaction (AIC = 1711.1, BIC = 1741.1). These results suggest the importance of accounting for excess zeros in this application.

In many applications, interest focuses on estimating and predicting marginal means given the explanatory variables. Misspecification of the extent of the population at risk may also have an impact on the marginal mean response, since the marginal means depend on the estimated parameters in both binary and Poisson processes. In this application, we are

**Table 2.** Parameter estimates (Est), standard error (SE),  $p$ -value, AIC and BIC values for the ZIP- $W^{b2}$  and ZIP- $W^{c2}$  models including the interaction between exposure and  $\text{NO}_2$  for modeling the number of hospital admissions for respiratory causes over 315 municipalities in Turin province (Italy) in 2004.

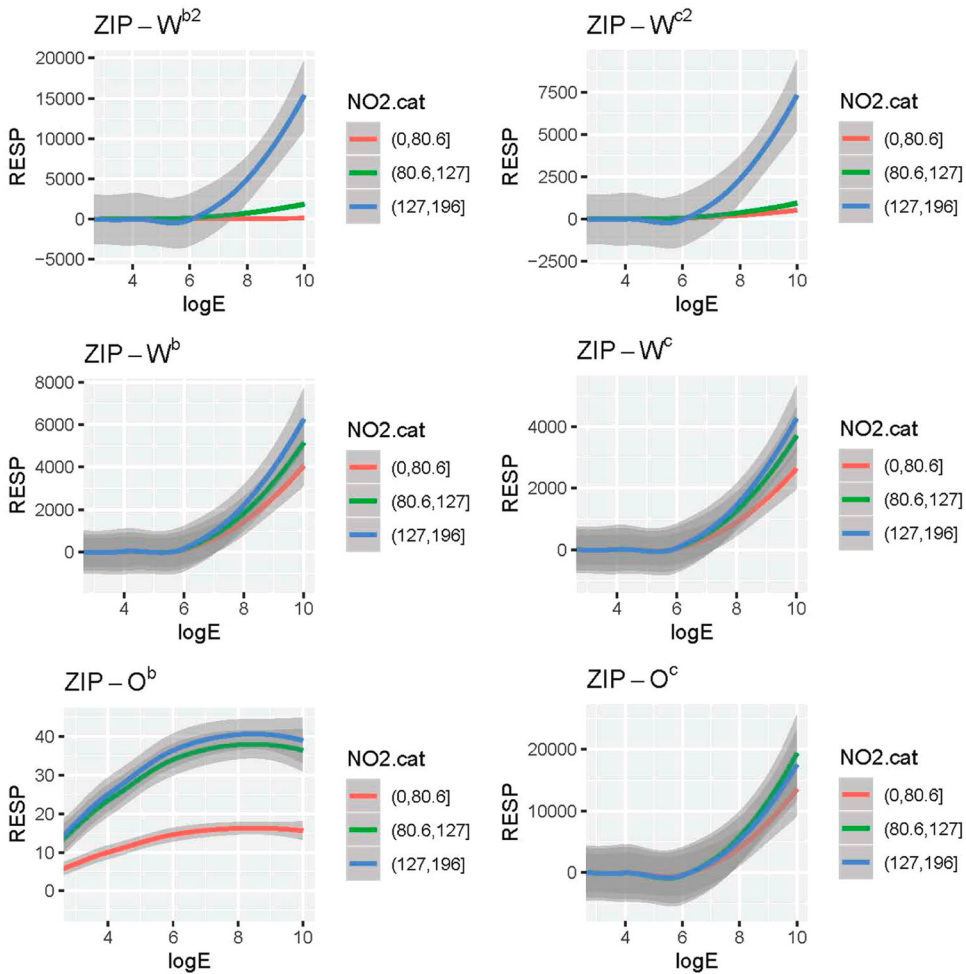
	ZIP- $W^{b2}$			ZIP- $W^{c2}$		
	Est	SE	$p$ -value	Est	SE	$p$ -value
	<i>Binary component</i>					
$\beta_0$ (Intercept)	4.25	0.52	0.00	0.13	0.15	0.39
$\xi_1(\log(E))$	-2.28	0.27	0.00			
	<i>Count component</i>					
$\alpha_0$ (Intercept)	2.21	0.26	0.00	1.37	0.72	0.06
$\alpha_1(\text{NO}_2^{(1)})$	-0.57	0.31	0.06	0.24	0.73	0.74
$\alpha_2(\text{NO}_2^{(2)})$	-1.26	0.31	0.00	-0.44	0.73	0.55
$\alpha_3(\text{NO}_2^{(1)} : \log(E))$	0.34	0.12	0.00	0.04	0.28	0.90
$\alpha_4(\text{NO}_2^{(2)} : \log(E))$	0.63	0.11	0.00	0.31	0.27	0.25
$\xi_2(\log(E))$	0.25	0.10	0.02	0.57	0.27	0.04
$\sigma_b^2$	0.11	0.33	0.00	0.12	0.35	0.00
	<i>Model fit</i>					
AIC		1325.53			1550.40	
BIC		1359.30			1580.42	

interested in estimating and predicting the population-level mean counts of the number of hospitalizations over the studied area in association with the  $\text{NO}_2$  and size of the population at risk. For ease of computation without marginalizing and conditioning on the random effects, the population-level predicted values of the response variable are calculated as the predicted unconditional counts at the mode (i.e. area-level effect  $b_i = 0$ ), as suggested by Brooks *et al.* [8]. For example, the predicted response value under model ZIP- $W^{b2}$  can be calculated as,

$$\begin{aligned}
 E(Y_i) &= E[E(Y_i | \delta_i)] = (1 - \pi_i)\mu_i \\
 &= \left\{ 1 - \text{logit}^{-1} \left[ \hat{\beta}_0 + \hat{\xi}_1 \log(E_i) \right] \right\} \\
 &\quad \times \exp \left[ \hat{\alpha}_0 + \hat{\alpha}_1 \text{NO}_{2i}^{(1)} + \hat{\alpha}_2 \text{NO}_{2i}^{(2)} + \hat{\alpha}_3 \text{NO}_{2i}^{(1)} \log(E_i) \right. \\
 &\quad \left. + \hat{\alpha}_4 \text{NO}_{2i}^{(2)} \log(E_i) + \hat{\xi}_2 \log(E_i) \right].
 \end{aligned}$$

For the standard errors of the predicted values, posterior predictive simulations were used by drawing multivariate normal samples from the parameters for the fixed effects, given that the resulting estimators follow asymptotically normal distributions [8].

Figure 2 displays the predicted response value and 95% point-wise confidence intervals of the respiratory hospitalization counts at the mode (i.e. area-specific random effect  $b_i = 0$ ) against  $\log(E)$  by the three categories of  $\text{NO}_2$  for all the considered ZIP models. The results based on models ZIP- $W^{b2}$  and ZIP- $W^{c2}$  clearly demonstrate the significance of the interaction effect between  $\text{NO}_2$  and  $\log(E)$ , with the number of hospitalizations increases more sharply as  $\log(E)$  increases for  $\text{NO}_2$  at the highest category (127, 196] as compared to the other  $\text{NO}_2$  levels. For the models without interaction terms, i.e. ZIP- $W^b$ , ZIP- $W^c$ , and ZIP- $O^c$ , the predicted marginal means did not differ substantially at various levels of

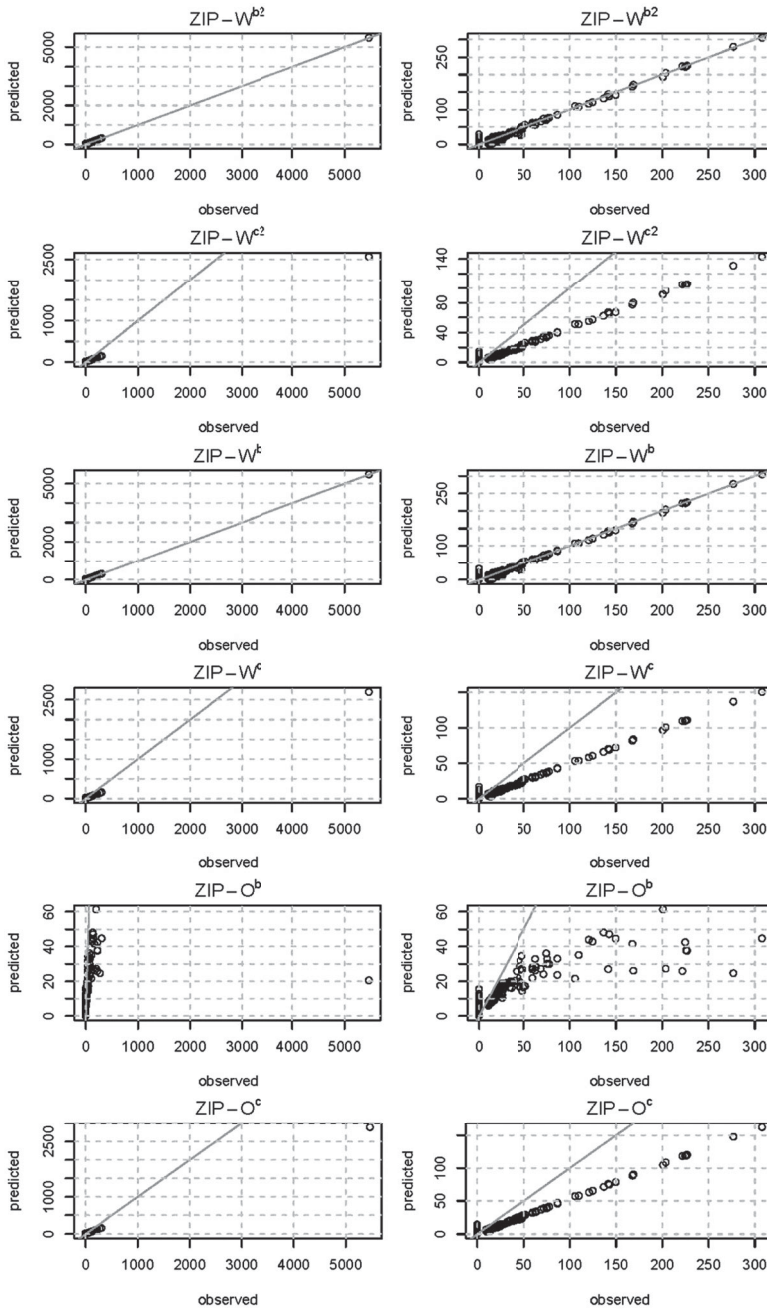


**Figure 2.** The predicted number of respiratory hospitalization count at the mode, i.e. area-specific random effect  $b_i = 0$  (labeled as 'RESP') and 95% point-wise confidence intervals against the log of the population at risk, i.e.  $\log(E)$ , by the categories of NO<sub>2</sub>.

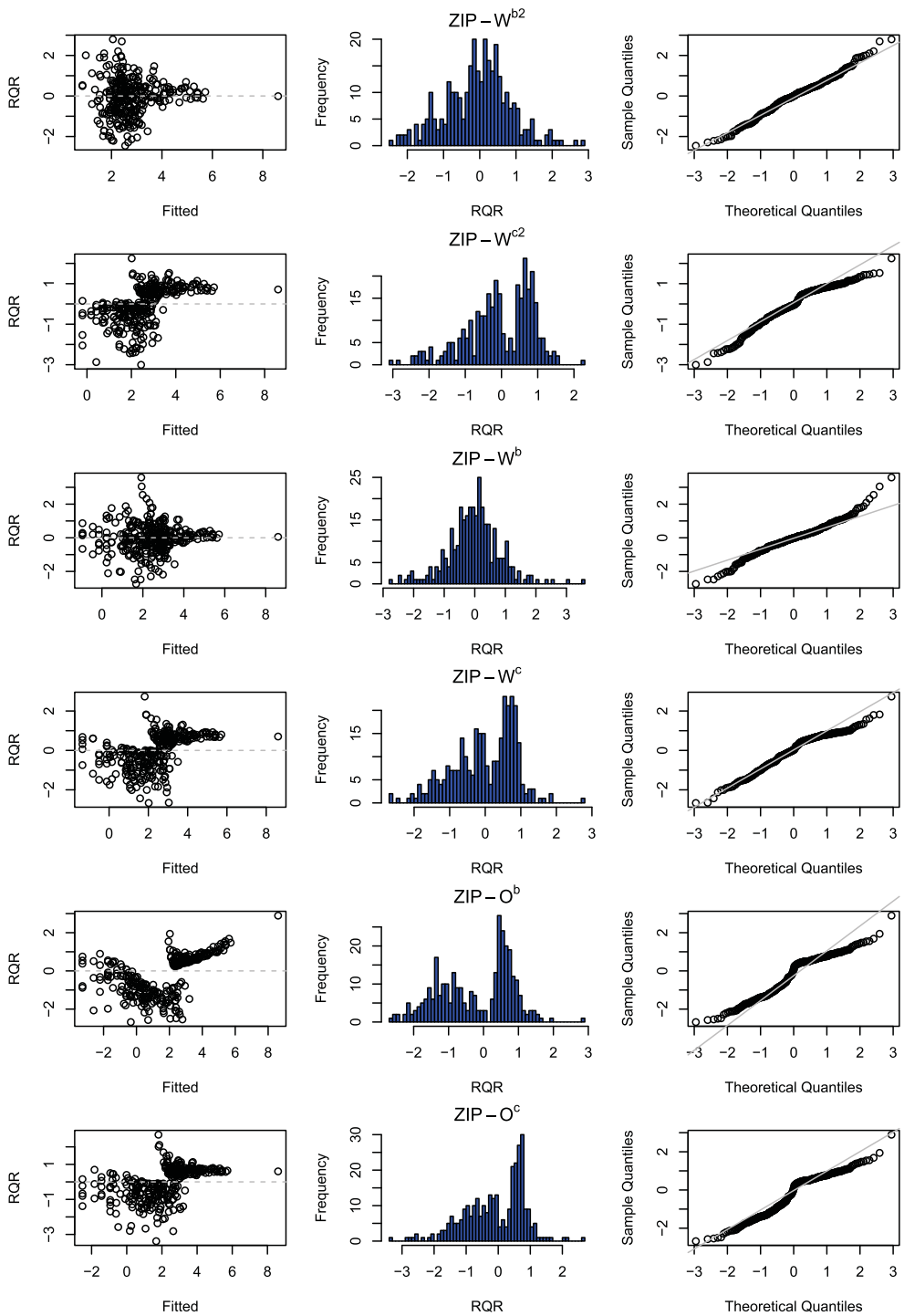
NO<sub>2</sub> and appeared to be averaged out across the levels. ZIP- $O^b$  yielded unreasonably low predicted values compared to other models.

We also examined the model goodness of fit by comparing the predicted and the observed response variable, as shown in Figure 3. The panels in the left column are for the entire data and the panels in the right column are for displaying the data after excluding the largest observed response value for the ease of visualization. It is evident that the predicted values under models ZIP- $W^{b^2}$  and ZIP- $W^b$  are very close to the observed values. The other competing models result in underestimated response values. The results suggest that the models ZIP- $W^{b^2}$  and ZIP- $W^b$  predict the number of hospitalizations better than the other candidate models.

Despite the aforementioned numerical comparisons of model fit, a careful residual analysis indicates that model ZIP- $W^{b^2}$  has the best model fit with RQRs nearly normally



**Figure 3.** Fitted vs. observed response variable values based on the models ZIP-W<sup>b<sup>2</sup></sup> (panels in row 1), ZIP-W<sup>c<sup>2</sup></sup> (panels in row 2), ZIP-W<sup>k</sup> (panels in row 3), ZIP-W<sup>c</sup> (panels in row 4), ZIP-O<sup>b</sup> (panels in row 5) and ZIP-O<sup>c</sup> (panels in row 6). The panels in the left column are for the full data. For ease of visualization, the panels in the right column display the results after suppressing the largest value of the observed response variable.



**Figure 4.** Scatter plots of RQRs vs. fitted values (first column), histograms of RQRs (second column) and QQ plots of RQRs (third column) based on the models ZIP- $W^{b2}$  (panels in row 1), ZIP- $W^{c2}$  (panels in row 2), ZIP- $W^b$  (panels in row 3), ZIP- $W^c$  (panels in row 4), ZIP- $O^b$  (panels in row 5) and ZIP- $O^c$  (panels in row 6).

distributed and no discernible pattern, as shown in the top panels of Figure 4. The distributions of the RQRs under the models ZIP- $W^{c2}$ , ZIP- $W^c$ , ZIP- $O^b$  and ZIP- $O^c$  exhibit bimodal pattern separated at zeros, which suggest the importance of including exposure as an exploratory variable in the binary component of the ZIP model in this application. Altogether, failure to properly model the effect of exposure in both the binary and count components of the ZIP model can have a serious impact on the marginal parameter estimates. Inevitably, incorrect inferences may follow.

## 5. Simulation studies

Simulations were carried out to investigate the performances of the proposed method relative to the traditional methods specifying the underlying population at risk as an offset term in the model under a wide range of scenarios.

### 5.1. Data-generating mechanism

In the simulations, zero-inflated counts were generated from ZIP- $W^b$  model of sizes  $n = 250, 500$  and  $1000$ , defined as

$$\text{ZIP-}W^b \begin{cases} \text{logit}(\pi_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \xi_1 \log(E_i) \\ \log(\mu_i) = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \xi_2 \log(E_i), \end{cases} \quad (21)$$

where  $(\beta_0, \beta_1, \beta_2) = (5, -1, -1)$ ,  $(\alpha_0, \alpha_1, \alpha_2) = (1, 1, 1)$ ,  $x_{1i} \sim \text{Bernoulli}(0.3)$  and  $x_{2i} \sim \text{Normal}(\tau, \sigma^2)$  with  $\tau = 0$  and  $\sigma^2 = 0.25$ . We considered  $\xi_1 = -2$  and  $\xi_2 = 0.8$ , respectively, to reflect the fact that the probability of excessive zeros decreases with increasing extent of exposure and the conditional mean count increases over the increased extent of exposure. The values are chosen to mimic our motivating example presented in Section 4. The intercept for the binary component  $\beta_0$  is set as 5 to yield about 53% zeros in the simulated datasets. The expected population at risk  $E_i$  is simulated from a zero truncated negative binomial distribution with probability mass function  $[\Gamma(x+r)p^r(1-p)^x]/[\Gamma(n)x!(1-p)^r]$  where  $r$  is set as 0.1 and  $p$  is set as 0.0005, which gives the mean about 40, median 10 and variance 6000. For simplicity of presentation, we assume  $x_{i1}$ ,  $x_{i2}$  and  $\log(E_i)$  do not interact.

For each simulation scenario, we generated 200 random samples from the true model and fitted ZIP- $W^c$ , ZIP- $O^b$  and ZIP- $O^{b2}$  to determine the impact of mismodelling the effect of exposure on the parameter estimates in both parts of the ZIP model and the overall model fits. To simulate zero inflated data, we firstly simulate the latent variable  $\delta_i$  from Bernoulli ( $\pi_i$ ) with  $\pi_i = \exp[\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \xi_1 \log(E_i)] / \{1 + \exp[\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \xi_1 \log(E_i)]\}$ . Then, if  $\delta_i = 1$ ,  $y_i = 0$ ; otherwise, simulate  $y_i$  from Poisson with mean  $\mu_i = \exp[\alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \xi_2 \log(E_i)]$ .

To assess how biased the parameter estimates can be related to the increased or decreased effect of exposures, additional simulation studies were conducted by setting  $\xi_1 = -2$ ,  $\xi_2 = 2$  and  $\xi_1 = -0.5$ ,  $\xi_2 = 0.5$ , respectively. We also considered simulating data from ZIP- $O^c$ , which includes the exposure as an offset term only in the count component of the model, to examine the impact of over parametrization of our proposed model ZIP- $W^b$  on statistical inference.

### 5.2. Performance measures

The goals of the simulation study are to determine how parameter estimates, standard errors, coverage probability and overall model fits are affected by misspecification of the effect of the extent of exposures. To this end, we present the bias (the mean of the estimated parameter minus the true value), mean square error (MSE, the average of the sum of the squared differences between the estimated parameter and the true value), the coverage probabilities (CP) of 95% confidence intervals of the estimated parameters and average values of AIC and BIC over repeated samples.

### 5.3. Simulation results

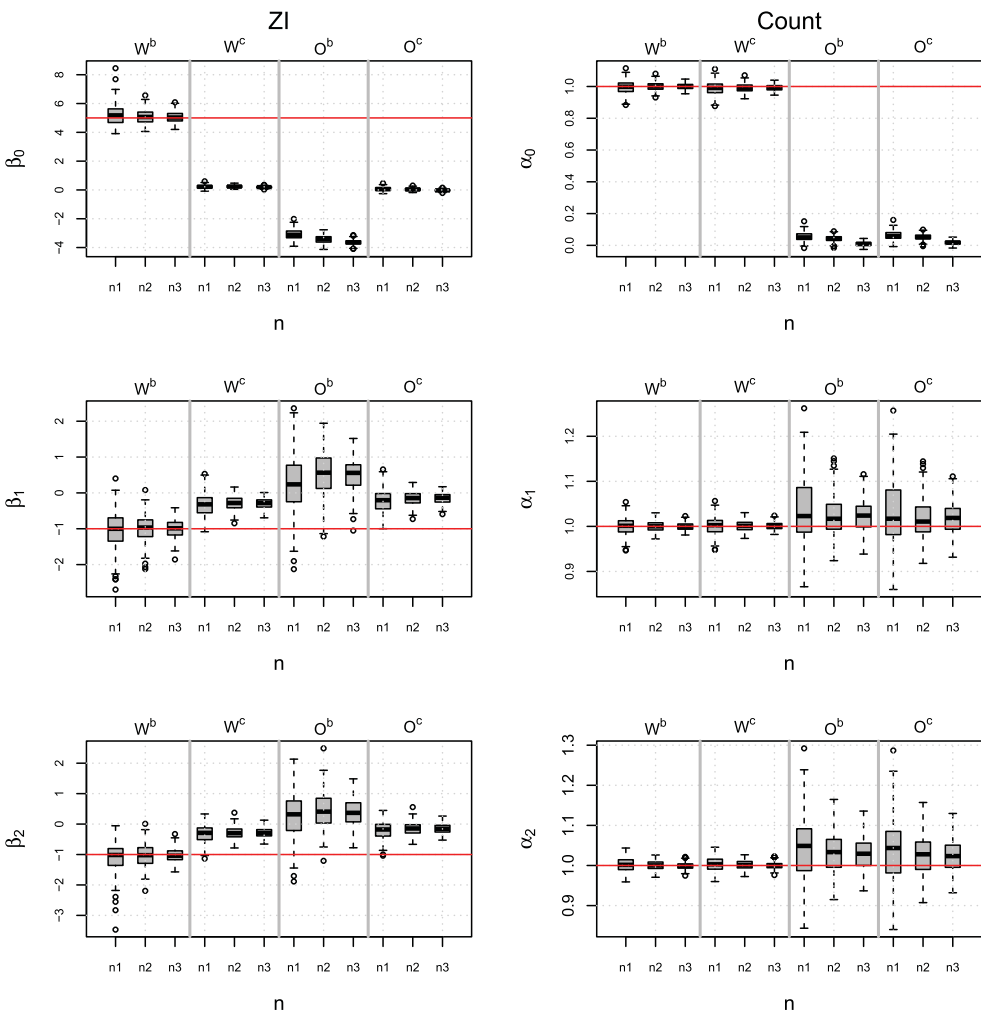
Table 3 reports the bias, MSE, and CP of the 95% confidence intervals for the parameter estimates from ZIP- $W^b$ , ZIP- $W^c$ , ZIP- $O^b$  and ZIP- $O^c$  fitted to the 200 simulated datasets generated from model ZIP- $W^b$  of sample size  $n = 250, 500$  and  $1000$ , respectively. For the binary component, only ZIP- $W^b$  results in unbiased estimates with CPs very close to the

**Table 3.** Bias, MSE and coverage probability (CP) of the 95% confidence intervals for the parameter estimates from models ZIP- $W^b$ , ZIP- $W^c$ , ZIP- $O^b$  and ZIP- $O^c$  fitted to the 200 simulated datasets generated from the model ZIP- $W^b$  of sample size  $n = 250, 500$  and  $1000$ , respectively.

		Binary component				Count component				
		ZIP- $W^b$	ZIP- $W^c$	ZIP- $O^b$	ZIP- $O^c$	ZIP- $W^b$	ZIP- $W^c$	ZIP- $O^b$	ZIP- $O^c$	
$n = 250$										
Bias	$\beta_0$	0.229	-4.786	-8.100	-4.944	$\alpha_0$	-0.005	-0.012	-0.946	-0.938
	$\beta_1$	-0.033	0.667	1.248	0.778	$\alpha_1$	0.000	0.001	0.030	0.025
	$\beta_2$	-0.097	0.674	1.273	0.787	$\alpha_2$	0.002	0.003	0.039	0.033
MSE	$\beta_0$	0.583	22.918	65.716	24.464	$\alpha_0$	0.002	0.002	0.896	0.881
	$\beta_1$	0.252	0.539	2.246	0.705	$\alpha_1$	0.000	0.000	0.006	0.005
	$\beta_2$	0.240	0.524	2.166	0.695	$\alpha_2$	0.000	0.000	0.008	0.008
CP	$\beta_0$	0.965	0.000	0.000	0.000	$\alpha_0$	0.950	0.940	0.000	0.000
	$\beta_1$	0.935	0.345	0.285	0.265	$\alpha_1$	0.940	0.945	0.360	0.365
	$\beta_2$	0.955	0.275	0.240	0.230	$\alpha_2$	0.975	0.970	0.260	0.290
$n = 500$										
Bias	$\beta_0$	0.088	-4.764	-8.437	-4.966	$\alpha_0$	-0.001	-0.010	-0.958	-0.949
	$\beta_1$	0.003	0.707	1.528	0.845	$\alpha_1$	0.000	0.001	0.021	0.015
	$\beta_2$	-0.024	0.714	1.436	0.845	$\alpha_2$	0.000	0.001	0.033	0.027
MSE	$\beta_0$	0.255	22.703	71.240	24.668	$\alpha_0$	0.001	0.001	0.919	0.900
	$\beta_1$	0.128	0.542	2.721	0.759	$\alpha_1$	0.000	0.000	0.002	0.002
	$\beta_2$	0.121	0.546	2.396	0.754	$\alpha_2$	0.000	0.000	0.004	0.003
CP	$\beta_0$	0.960	0.000	0.000	0.000	$\alpha_0$	0.955	0.935	0.000	0.000
	$\beta_1$	0.955	0.090	0.100	0.020	$\alpha_1$	0.955	0.950	0.425	0.475
	$\beta_2$	0.940	0.040	0.070	0.020	$\alpha_2$	0.970	0.975	0.285	0.300
$n = 1000$										
Bias	$\beta_0$	0.051	-4.811	-8.644	-5.028	$\alpha_0$	-0.001	-0.009	-0.991	-0.982
	$\beta_1$	-0.007	0.703	1.494	0.849	$\alpha_1$	0.000	0.001	0.023	0.018
	$\beta_2$	-0.027	0.713	1.390	0.848	$\alpha_2$	-0.001	0.000	0.027	0.022
MSE	$\beta_0$	0.137	23.146	74.749	25.284	$\alpha_0$	0.000	0.000	0.982	0.965
	$\beta_1$	0.057	0.513	2.424	0.741	$\alpha_1$	0.000	0.000	0.002	0.002
	$\beta_2$	0.052	0.529	2.129	0.742	$\alpha_2$	0.000	0.000	0.002	0.002
CP	$\beta_0$	0.950	0.000	0.000	0.000	$\alpha_0$	0.950	0.935	0.000	0.000
	$\beta_1$	0.950	0.000	0.020	0.000	$\alpha_1$	0.955	0.935	0.290	0.310
	$\beta_2$	0.950	0.000	0.005	0.000	$\alpha_2$	0.935	0.920	0.255	0.270



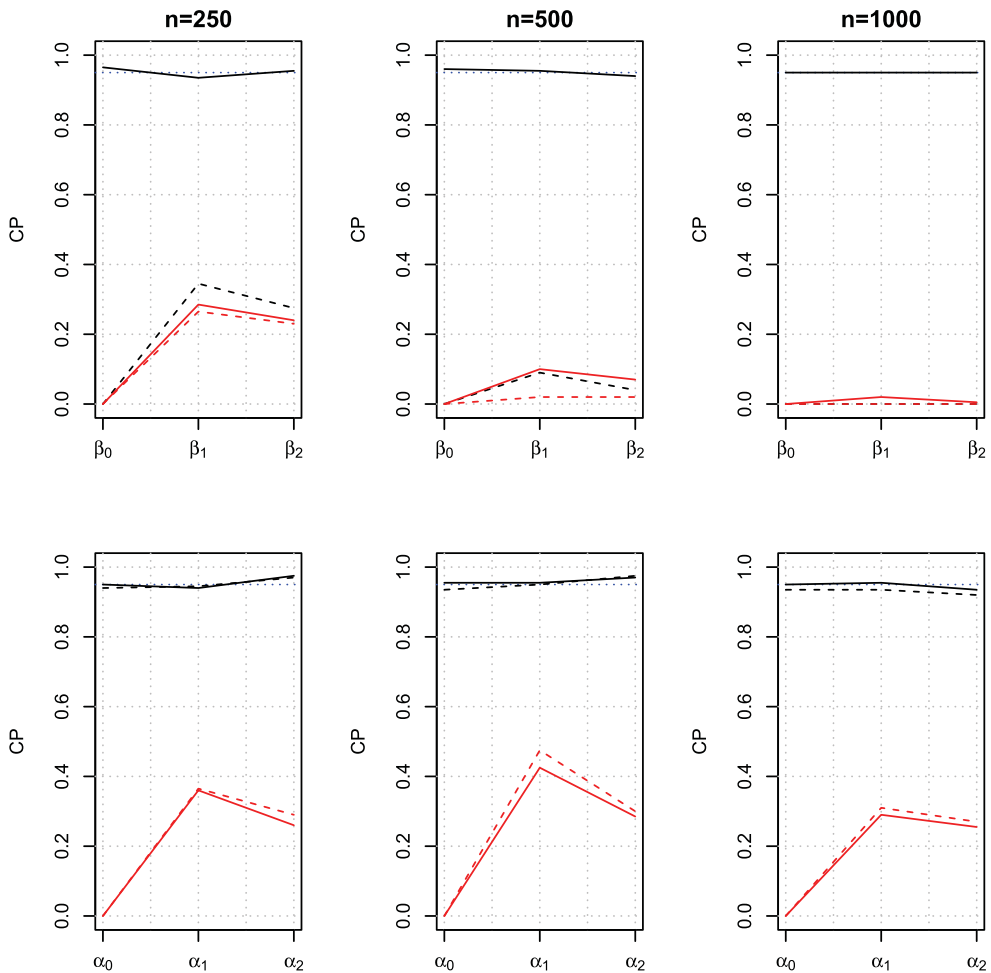
nominal level of 0.95. In contrast, fitting the misspecified model results in severely biased parameter estimates with CPs far from the nominal level. CP also decreases towards zero as the sample size increases from 250 to 1000. In particular, the intercept ( $\beta_0$ ) is highly affected by model misspecification yielding very large bias and there is also a substantial bias in the estimated regression coefficients  $\beta_1$  and  $\beta_2$ . The biased parameter estimates of the ZIP- $W^c$  model indicates that omitting exposure as a covariate in the binary component of the model results in biased estimates and invalid inference. This result is consistent with the finding from the literature, which showed that if a covariate is removed from a Poisson model, both the estimated regression coefficient and the standard error are the same as the results based on the full model [5]. ZIP- $O^b$  gives the worst model fit, yielding the largest bias, MSE, and lowest CP. This result is not surprising, since ZIP- $O^b$  constrains the regression coefficients



**Figure 5.** Estimated regression coefficients for the binomial process for modeling the probability of excess zeros (left panels) and the Poisson process (right panels) over 200 simulated datasets from the model  $ZI-W^b$  of sample size  $n_1 = 250, n_2 = 500$  and  $n_3 = 1000$ .

for the exposure in both model components equal to one; nevertheless, both the binary and count processes are influenced by the exposure, but in opposite directions. Hence, the degree of model misspecification is higher compared to other candidate models.

For the count component of the ZIP models, ZIP- $W^b$  and ZIP- $W^c$  result in nearly unbiased estimates with CPs very close to the nominal level of 0.95. The results based on the ZIP- $W^c$  model indicate that omitting exposure as a covariate in the binary component has minimal impact on the parameter estimates for the count component. This result is consistent with the finding from the literature that if a covariate is removed from a Poisson model, both the estimated regression coefficient and the standard error are the same as those of the full model [26]. Nevertheless, under ZIP- $O^b$  and ZIP- $O^c$  models, the intercept is severely underestimated, the estimated regression coefficients are biased upward and the



**Figure 6.** Coverage probabilities of the 95% confidence intervals of the estimated regression coefficients for the binomial process for modeling the probability of excess zeros (top panels) and the Poisson process (bottom panels) over 200 simulated datasets from model ZI- $W^b$  of sample size  $n = 250, 500$  and  $1000$ . The black solid and dashed lines are for the models ZI- $W^b$  and ZI- $W^c$ , respectively; the red solid and dashed lines are for the models ZI- $O^b$  and ZI- $O^c$ , respectively.

**Table 4.** Comparison of model fit of ZIP- $W^b$ , ZIP- $W^c$ , ZIP- $O^b$  and ZIP- $O^c$  in terms of average AIC and BIC across 200 simulated datasets generated from the model ZIP- $W^b$ .

$n$	ZIP- $W^b$	ZIP- $W^c$	ZIP- $O^b$	ZIP- $O^c$
			AIC	
250	976.520	1162.257	2064.654	1732.514
500	1900.825	2288.361	4130.966	3448.865
1000	3887.104	4671.542	8960.393	7597.910
			BIC	
250	1004.692	1186.908	2085.782	1753.643
500	1934.542	2317.863	4156.254	3474.153
1000	3926.366	4705.896	8989.840	7627.356

CPs decline as sample size increases. For the ease of comparison of the results among the considered models, we also graphed the estimated regression coefficients over the 200 simulated samples in Figure 5 and the CPs of the 95% confidence intervals for the regression coefficients in Figure 6. Table 4 presents the average AIC and average BIC over 200 simulated datasets for model comparison. The results indicate that ZIP- $W^b$  consistently gives the smallest AIC and BIC compared to other competing models in all scenarios, followed by ZIP- $W^c$ , ZIP- $O^c$  and ZIP- $O^b$ . Model ZIP- $O^b$  gives the worst model fit, since it forces a model fit that is inconsistent with the binary and Poisson processes of the ZIP model.

Additional simulation studies were conducted by increasing or decreasing the values of  $\xi_1$  and  $\xi_2$ . For the scenario when  $\xi_1 = -2, \xi_2 = 2$ , the results are presented in Figures S1 and S2 in the web supplementary materials, which indicate that the results for the binary component remain roughly consistent with the results from the previous simulation setting when  $\xi_1 = -2$  and  $\xi_2 = 0.8$ . However, for the count component, the performances of ZIP- $O^b$  and ZIP- $O^c$  become worse with increased bias and much lower CPs as compared to Figures 5 and 6, since models ZIP- $O^b$  and ZIP- $O^c$  constrain  $\xi_2 = 1$  and  $\xi_2 = 2$  deviates more from one than  $\xi_2 = 0.8$ . As a comparison, when  $\xi_1 = -0.5$  and  $\xi_2 = 0.5$ , as shown in Figures S3 and S4, the parameter estimates of the regression coefficients in the binary component are less biased and CPs are closer to the nominal level compared to the previous setting, since  $\xi_1 = -0.5$  is less deviated from 1 as compared to  $\xi_1 = -2$ . The results of the model fit for these additional simulation studies are presented in Tables S1 and S2, with the results being consistent with the previous simulation setting showing model ZIP- $W^b$  outperforms the other competing models. In summary, the results based on the model ZIP- $W^c$  suggest that ignoring modeling the effect of varying exposure in the binary component of the ZIP model can bias the estimation of the covariate effect in the binary component. Such bias becomes more severe as the effect of varying exposure increases. The results based on the ZIP- $O^b$  and ZIP- $O^c$  indicate that incorporating the varying exposure as an offset term can lead to biased and inefficient parameter estimates in both the binary and count components of the ZIP model. The simulation results confirmed that the degree of bias and variation of the estimated regression coefficients depend on the effect of the exposure variable.

In another set of the simulation study, we simulate data from ZIP- $O^c$  model. Our results (Figures S5 and S6) indicate that both ZIP- $W^b$  and ZIP- $W^c$  models provide parameter estimates with negligible bias, low MSE and nominal coverage probabilities reasonably close to 95%. Only ZIP- $O^b$  yielded biased estimates for the binary component but not for the

count component, since it imposes unreasonable assumption on the effect of exposure in the binary component. Overall, it appears over parametrization by estimating  $\xi_1$  and  $\xi_2$  rather than restricting them equal to one as specified in the offset terms for both binary and count component has negligible effects on inference.

We also run another set of simulations to ascertain whether the percentage of zeros is an important feature in determining the impact of misspecification of the effect of varying exposure. We generated data from the ZIP- $W^b$  model with about 30%, 60%, and 90% of zeros. Our simulation results are comparable to the results presented earlier.

## 6. Conclusion and future work

In this study, we reviewed zero-inflated regression models with a focus on investigating the extent to which misspecification of modeling underlying population at risk on the estimation of the regression coefficients and overall model fit for the zero-inflated model. We showed that including an offset term could be very restrictive in the sense that it forces the effect of exposure as one, which can be inconsistent with the data, as shown in our motivating example. Therefore, we formulated and developed a framework to understand the nature of zero-inflated models by allowing the extent of exposure to be included in both parts of the binary and count components of the ZI model as a regular covariate.

The evidence provided in this paper serves as a warning not to make strong assumptions about the effect of exposure, like those embodied in using offset in a Poisson distribution. It is wise at least to make a sensitivity check by estimating the effect of the varying exposure. Also, the probability of excessive zero may also depend on the population at risk. The relationship between exposure and the probability of excessive zero component, therefore, needs to be carefully assessed and properly incorporated in the model.

In our motivating example, the ZIP model with varying exposure being included in both the binary and count components as a covariate fits this particular data set well. However, in some situations, after accounting for zero-inflation and adjusting for the effects of the covariates and varying exposure, the data may still suggest additional overdispersions. The proposed modeling approach could then be applied to other zero-inflated models to account for additional overdispersion, such as zero-inflated negative binomial model and zero-inflated generalized Poisson model [12,33]. Score test could be conducted to help determine whether a more complex model is appropriate, without fitting a more complex model [33].

In addition, including varying exposure as a covariate in both the binary and count components of the ZI model leads to an increase in the number of parameters to be estimated. In the situation where many covariates are involved, the variable selection needs to be conducted to address the potential over parametrization problem, especially when the sample size is small. Traditional variable selection procedures, such as the automated variable selection methods, may result in models that are unstable and not reproducible [3]. Penalized regression methods are popular for selecting variables, which keep all the variables in the model but constrain the regression coefficients by shrinking them toward zero. A variety of penalty functions can be considered, such as Least Absolute Shrinkage and Selection Operator (LASSO) [29], Smoothly Clipped Absolute Deviation penalty (SCAD) [15], and minimax concave penalty (MCP) [37]. Penalized regression methods have been extended for selecting parsimonious zero-inflated models [9,30,31,36]. Future studies will

be conducted to evaluate the performance of the proposed modeling strategy with these variable selection methods under different ratios of the number of candidate covariates to the sample size.

Note that zero-inflated and hurdle models have been extended to model longitudinal or clustered count measures with excess zeros by linking the binary and count components using a shared subject-specific random effect term or bivariate normal distribution. The linkage of the model components allows the dependence between the binary and count components of the model [24,25]. As a result, the binary and count processes will not act independently; that is, model misspecification for one component may have an impact on the other component passed through the shared random effect terms. Future work will be conducted to investigate the impact of misspecification of the exposure effect on such correlated random effects models.

In our simulations and empirical studies, we considered the linear effect of the log of the exposure variable. However, more flexible modeling of exposure-outcome associations can be applied to avoid constraining a priori functional form of this relationship to a particular parametric family of functions, such as conventionally used linear functions.

## Acknowledgments

This research was supported by the discovery grant from the Natural Sciences and Engineering Research Council of Canada. The author is also grateful to the Editor, Associate Editor, and two anonymous referees for their very valuable and constructive comments, which greatly helped to improve the quality of this paper.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

This research was supported by the discovery grant from the Canadian Network for Research and Innovation in Machining Technology, Natural Sciences and Engineering Research Council of Canada [RGPIN 07212-2019].

## ORCID

Cindy Feng  <http://orcid.org/0000-0003-4030-7413>

## References

- [1] A. Agresti, *Foundations of Linear and Generalized Linear Models*, Wiley, New York, 2015.
- [2] H. Akaike, *Information theory as an extension of the maximum likelihood principle*, in *Second International Symposium on Information Theory*, B.V. Petrov and B.F. Csaki, eds., Akademiai Kiado; Budapest, 1973.
- [3] P. Austin and J. Tu, *Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality*, *J. Clin. Epidemiol.* 57 (2004), pp. 1138–1146.
- [4] G. Baetschmann and R. Winkelmann, *Modeling zero-inflated count data when exposure varies: With an application to tumor counts*, *Biom. J.* 55 (2013), pp. 679–686.
- [5] M.D. Begg and S. Lagakos, *Loss in efficiency caused by omitting covariates and misspecifying exposure in logistic regression models*, *J. Am. Stat. Assoc.* 88 (1993), pp. 166–170.

- [6] M. Blangiardo and M. Cameletti, *Spatial and Spatio-temporal Bayesian Models with R-INLA*, Wiley, New York, 2015.
- [7] D. Bohning, E. Dietz, P. Schlattmann, L. Mendonca, and U. Kirchner, *The zero-inflated poisson model and the decayed, missing and filled teeth index in dental epidemiology*, J. R. Stat. Soc. Ser. A 162 (1999), pp. 195–209.
- [8] M.E. Brooks, K. Kristensen, K.J. van Benthem, A. Magnusson, C.W. Berg, A. Nielsen, H.J. Skaug, M. Maechler, and B.M. Bolker, *glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling*, R J. 9 (2017), pp. 378–400. Available at <https://journal.r-project.org/archive/2017/RJ-2017-066/index.html>.
- [9] A. Buu, N.J. Johnson, R. Li, and X. Tan, *New variable selection methods for zero-inflated count data with applications to the substance abuse field*, Stat. Med. 30 (2011), pp. 2326–2340.
- [10] A. Buu, R. Li, X. Tan, and R. Zucker, *Statistical models for longitudinal zero-inflated count data with applications to the substance abuse field*, Stat. Med. 31 (2012), pp. 4074–4086.
- [11] Y. Cheung, *Zero-inflated models for regression analysis of count data: A study of growth and development*, Stat. Med. 21 (2002), pp. 1461–1469.
- [12] C. Czado, V. Erhardt, A. Min, and S. Wagner, *Zero-inflated generalized Poisson models with regression effects on the mean, dispersion and zero-inflation level applied to patent outsourcing rates*, Stat. Modelling 7 (2007), pp. 125–153.
- [13] L. Dai, M.D. Sweat, and M. Gebregziabher, *Modeling excess zeros and heterogeneity in count data from a complex survey design with application to the demographic health survey in sub-Saharan Africa*, Stat. Methods Med. Res. 27 (2018), pp. 208–220.
- [14] P.K. Dunn and G.K. Smyth, *Randomized quantile residuals*, J. Comput. Graph. Stat. 5 (1996), pp. 236–244.
- [15] J. Fan and R. Li, *Variable selection via nonconcave penalized likelihood and its oracle properties*, J. Amer. Statist. Assoc. 96 (2001), pp. 1348–1360.
- [16] D.B. Hall, *Zero-inflated poisson and binomial regression with random effects: A case study*, Biometrics 56 (2000), pp. 1030–1039.
- [17] D.C. Heilbron, *Zero-altered and other regression models for count data with added zeros*, Biom. J. 36 (1994), pp. 531–547.
- [18] K. Kristensen, A. Nielsen, C.W. Berg, H. Skaug, and B.M. Bell, *TMB: Automatic differentiation and Laplace approximation*, J. Stat. Softw. 70 (2016), pp. 1–21.
- [19] D. Lambert, *Zero-inflated Poisson regression with an application to defects in manufacturing*, Technometrics 34 (1992), pp. 1–14.
- [20] A.H. Lee, K. Wang, and K.K. Yau, *Analysis of zero-inflated Poisson data incorporating extent of exposure*, Biom. J. 43 (2001), pp. 963–975.
- [21] O. Loquiha, N. Hens, L. Chavane, M. Temmerman, N. Osman, C. Faes, and M. Aerts, *Mapping maternal mortality rate via spatial zero-inflated models for count data: A case study of facility-based maternal deaths from Mozambique*, PLoS ONE 13 (2018), e0202186.
- [22] Y. Min and A. Agresti, *Random effect models for repeated measures of zero-inflated count data*, Stat. Modelling 5 (2005), pp. 1–19.
- [23] J. Mullahy, *Specification and testing of some modified count data models*, J. Econom. 33 (1986), pp. 341–365.
- [24] B. Neelon, P. Ghosh, and P. Loebis, *A spatial Poisson hurdle model for exploring geographic variation in emergency department visits*, J. R. Stat. Soc. Ser. A 176 (2013), pp. 389–413.
- [25] B. Neelon, A. O'Malley, and S. Normand, *A Bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use*, Stat. Modelling 10 (2010), pp. 421–439.
- [26] M.R. Petersen and J.A. Deddens, *Effects of omitting a covariate in poisson models when the data are balanced*, Canad. J. Statist. 28 (2000), pp. 439–445.
- [27] C. Rose, S. Martin, K. Wannemuehler, and B. Plikaytis, *On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data*, J. Biopharm. Stat. 16 (2006), pp. 463–481.
- [28] G. Schwarz, *Estimating the dimension of a model*, Ann. Statist. 6 (1978), pp. 461–464.
- [29] R. Tibshirani, *Regression shrinkage and selection via the LASSO*, J. R. Stat. Soc. 58 (1996), pp. 267–288.

- [30] Z. Wang, S. Ma, and C. Wang, *Variable selection for zero-inflated and overdispersed data with application to health care demand in germany*, *Biom. J.* 57 (2015), pp. 867–884.
- [31] Z. Wang, S. Ma, C. Wang, M. Zappitelli, P. Devarajan, and C. Parikh, *EM for regularized zero-inflated regression models with applications to postoperative morbidity after cardiac surgery in children*, *Stat. Med.* 33 (2014), pp. 5192–5208.
- [32] Y. Xia, J. Sun, and D.G. Chen, *Modeling zero-inflated microbiome data*, in *Statistical Analysis of Microbiome Data with R*, Springer Singapore, Singapore, 2018, pp. 453–496.
- [33] Z. Yang, J.W. Hardin, and C.L. Addy, *Testing overdispersion in the zero-inflated Poisson model*, *J. Stat. Plan. Inference* 139 (2009), pp. 3340–3353.
- [34] K. Yau and A. Lee, *Zero-inflated poisson regression with random effects to evaluate an occupational injury prevention programme*, *Stat. Med.* 20 (2001), pp. 2907–2920.
- [35] A. Zeileis, C. Kleiber, and S. Jackman, *Regression models for count data in R*, *J. Stat. Softw.* 27 (2008), pp. 1–25.
- [36] P. Zeng, Y. Wei, Y. Zhao, J. Liu, L. Liu, R. Zhang, J. Gou, S. Huang, and F. Chen, *Variable selection approach for zero-inflated count data via adaptive lasso*, *J. Appl. Stat.* 41 (2014), pp. 879–894.
- [37] C.H. Zhang, *Nearly unbiased variable selection under minimax concave penalty*, *Ann. Statist.* 38 (2010), pp. 894–942.
- [38] Z. Zhen, L. Shao, and L. Zhang, *Spatial hurdle models for predicting the number of children with lead poisoning*, *Int. J. Environ. Res. Public Health* 15 (2018), p. 1792.