A Zero-inflated overdispersed hierarchical Poisson model

Peer-reviewed author version

KASSAHUN, Wondwosen; NEYENS, Thomas; FAES, Christel; MOLENBERGHS, Geert & VERBEKE, Geert (2014) A Zero-inflated overdispersed hierarchical Poisson model. In: STATISTICAL MODELLING, 14(5), p. 439-456.

DOI: 10.1177/1471082X14524676
Handle: http://hdl.handle.net/1942/16614

A Zero-Inflated Overdispersed Hierarchical Poisson Model

Wondwosen Kassahun[1] **Thomas Neyens**[2] **Christel Faes**[2]

Geert Molenberghs[2,3] **Geert Verbeke**[3,2]

[1] *Department of Epidemiology and Biostatistics, Jimma University, Ethiopia*

[2] *I-BioStat, CenStat, Universiteit Hasselt, B-3590 Diepenbeek, Belgium*

[3] *I-BioStat, L-BioStat, Katholieke Universiteit Leuven, B-3000 Leuven, Belgium*

Abstract

Count data are most commonly modeled using the Poisson model, or by one of its many extensions. Such extensions are needed for a variety of reasons: (1) a hierarchical structure in the data, e.g., due to clustering, the collection of repeated measurements of the outcome, etc.; (2) the occurrence of overdispersion (or underdispersion), meaning that the variability encountered in the data is not equal to the mean, as prescribed by the Poisson distribution; and (3) the occurrence of extra zeros beyond what a Poisson model allows. The first issue is often accommodated through the inclusion of random subject-specific effects. Though not always, one conventionally assumes such random effects to be normally distributed. Overdispersion is often dealt with through a model developed for this purpose, such as, for example, the negative-binomial model for count data. This can be conceived through a random Poisson parameter. Excess zeros are regularly accounted for using so-called zero-inflated models, which combine either a Poisson or negative-binomial model with an atom at zero. The novelty of this paper is that it combines all these features. The work builds upon the modeling framework defined by Molenberghs *et al.* (2010) in which clustering and overdispersion are accommodated for through two separate sets of random effects in a generalized linear model.

Some Keywords: Poisson model, Clustering, Overdispersion, Zero-inflation

1 Introduction

Count data are encountered in a wide range of applications, including medical and biomedical research. As we will now explain, they may exhibit a variety of features that somehow need to be taken into account in the modeling process: zero-inflation, overdispersion, and correlation. In this contribution, we will accommodate them all at once, within a likelihood framework, allowing for easy implementation in standard statistical software.

Univariate count data is typically modeled within the class of generalized linear models (GLM, Nelder

and Wedderburn 1972, McCullagh and Nelder 1989, Agresti 2002) and the exponential family is used to formulate distributional assumptions (McCullagh and Nelder 1989). In this regard, Poisson regression models provide a standard basis for the analysis of count data. Nevertheless, it has been clear for several decades that a key feature of the GLM framework and many of the exponential family members, the so-called *mean-variance relationship*, may be overly restrictive. By this relationship, we indicate that the variance is a deterministic function of the mean; for the Poisson model $v(\mu) = \mu$. In several cases, one can expect that the Poisson assumption, however, is not satisfied: (1) when the data are hierarchically structured, as happens in longitudinal studies or when individuals are grouped into clusters; (2) when overdispersion is present in the data, for example, due to unobserved confounding; (3) when an excess of zeros is present in the data. The novelty of our work is that it combines these three features in into a single, flexible framework. As such, it goes beyond what is available in the literature. A general model framework for counts will be presented in which hierarchy, overdispersion, and zero-inflation can be modeled, extending the work by Molenberghs *et al.* (2010). While a relatively straightforward extension given earlier work, it has practical relevance as our data analysis will illustrate. It will be investigated whether the different components are conveniently and jointly estimable. Indeed, zero-inflation is a special case of overdispersion, but then it is of a very particular kind. We will also examine how failure to account for at least one of the features affects the results.

There is a lot of literature on Poisson-model extensions. Breslow (1984) targets overdispersion in the Poisson model. One of the important models in this respect is the *negative-binomial model*, where the natural parameter is assumed to follow a gamma distribution, which has the effect of relaxing the mean-variance relationship. Lawless (1987) also contributed to this class of extensions.

When focusing on hierarchical data, the so-called generalized linear mixed model (GLMM, Engel and Keen 1994, Breslow and Clayton 1993, Wolfinger and O'Connell 1993) has gained popularity as a tool to accommodate overdispersion and/or hierarchy-induced association for outcomes that are not necessarily of a Gaussian type. Booth *et al.* (2003) extended the negative binomial log-linear model to the case of dependent counts, where dependence among the counts is handled by including linear combinations of random effects.

Overdispersion and excess zeros for cross-sectional count data are studied by, for example, Lambert (1992) and Greene (1994). Multi-level zero-inflated Poisson regression is considered by Lee *et al.* (2006). Zero-inflated count models provide a way of modeling the excess zeros in addition to allowing

for overdispersion by using two simultaneously operating data generation processes; one generates only zeros and the other is either a Poisson or negative-binomial data generating process. The hurdle model, which is a two-part model, is also available to model excess zeros (Mullahy 1986). One part is a binary model for whether the response outcome is zero or positive. Conditional on a positive outcome, the second part uses a truncated Poisson or negative-binomial that modifies an ordinary distribution by conditioning on the positive outcomes. For zero-inflated correlated data, the hurdle model has been studied by Min and Agresti (2005).

The paper is organized as follows. In Section 2, two motivating case studies with count outcomes are described, with analyses reported in Section 7. In Section 3, a review is given of the so-called *combined model*, the general modeling framework proposed by Molenberghs *et al.* (2010) to model overdispersed hierarchical data. Section 4 proposes an extension of this modeling framework to also deal with zero-inflation. Avenues for parameter estimation and ensuing inferences are explored in Section 5, with particular emphasis on so-called partial marginalization. Section 6 deals with a simulation study to investigate the importance of accounting for clustering, overdispersion, and a preponderance of zero counts. The case studies are analyzed in Section 7.

2 Motivating Case Studies

2.1 The Jimma Infant Growth Study

The Jimma Infants Survival Differential Longitudinal Growth Study is a survey to study infant survival in Ethiopia. Risk factors, including socio-economic, maternal, and infant-rearing factors, were recorded to be able to study their relationship with the child's early survival. The study is described in detail by Asefa and Tessema (2002). Children born in Jimma, Keffa, and the Illubabor Zones, Southwestern Ethiopia were examined for their first-year growth characteristics. At baseline, there were a total of 7969 infants whereby 4317, 1494, and 2158 were from rural, urban, and semi-urban areas, respectively. The children were visited every two months starting from birth until the age of one year (Table 1).

One of the questions of interest in the survey is to assess the diarrheal disease burden. In this paper, it is investigated whether the number of days of diarrheal illness in the two-month period prior to each visit, changes over time (i.e., age), whether the evolution differs for gender (male or female), place of residence (urban or rural), medical care (medical help given or not) and breast feeding behavior

Table 1: *Jimma Infant Growth Study. The mean number of days of illness and standard deviation at each of the seven follow-up times.*

Time (months)	Mean	Std. Dev.
0	0.01	0.19
2	0.91	4.21
4	1.28	4.62
6	1.56	4.87
8	2.14	5.93
10	2.63	6.66
12	2.67	6.95

(breast or artificial feeding). Of the 49,000 observations in total, only about 8000 observations are non-zero, indicating that there is a non-negligible dominance of zero counts.

2.2 A Clinical Trial in Epileptic Patients

These data are obtained from a randomized, double-blind, parallel group multi-center study for the comparison of placebo with a new anti-epileptic drug (AED), in combination with one or two other AED's. The study is described in full detail in Faught *et al.* (1996) and Molenberghs and Verbeke (2005). In this study, 45 patients were randomized to the placebo group and 44 to the new treatment group. The number of epileptic seizures was recorded on a weekly basis during a 16-week period. Thereafter, patients were entered into a long-term open-extension study, which contains follow-up measurements of patients up to 27 weeks. The key research question is whether or not the additional new treatment reduces the number of epileptic seizures.

Zero counts represent 33% of the measurements; the sample average and standard deviation are 3.18 and 6.14, respectively. Thus, there is a large proportion of zeros, as well as evidence of overdispersion and correlation stemming from the longitudinal aspect of the data.

3 Review of the Combined Model

Molenberghs *et al.* (2010) proposed a unified modeling framework for the analysis of overdispersed and hierarchical non-Gaussian data by bringing together normal random effects and conjugate random effects within the generalized linear model framework. On the one hand, the generalized linear mixed model (Engel and Keen 1994, Breslow and Clayton 1993, Wolfinger and O'Connell 1993, GLMM) with normally distributed subject-specific random effects is likely the most frequently used model in the context of non-Gaussian repeated measurements. On the other hand, an elegant way to accommodate overdispersion is through a two-stage approach, in which a conjugate measurement-specific random effect on the scale of the natural parameter is used, leading to models such as the negative-binomial model for count data. These two features are brought together.

We apply the following notational convention. The model that brings both features together, i.e., the combined mode, is denoted as (PNG), where the first symbol 'P' refers to basic Poisson model, the second symbol 'N' is for normal random effects and the final one for gamma random effects. Three special cases follow by leaving out one or more of the random-effects structures: (P--) for the Poisson model, (PN-) for the Poisson-normal GLMM, and (P-G) for the negative-binomial model.

Let Y_{ij} be the jth outcome measured for subject i, with $i = 1, \ldots, N$ and $j = 1, \ldots, n_i$. In general, the combined family is given by

$$f_i(y_{ij}|\boldsymbol{\xi}, \theta_{ij}, \phi) = \exp\left\{\phi^{-1}[y_{ij}\eta_{ij} - \psi(\eta_{ij})] + c(y_{ij}, \phi)\right\},$$

with η and ϕ the natural and dispersion parameter, respectively, and $\psi(\cdot)$ and $c(\cdot)$ known functions specifying a particular member of the exponential family. The conditional mean is modeled as $E(Y_{ij}|\boldsymbol{b_i}, \boldsymbol{\xi}, \theta_{ij}) = \theta_{ij}\kappa_{ij}$ with $\kappa_{ij} = g(\boldsymbol{x}'_{ij}\boldsymbol{\xi} + \boldsymbol{z}'_{ij}\boldsymbol{b_i})$ for a known inverse-link function $g(\cdot)$, $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ p-dimensional and q-dimensional vectors of known covariate values, respectively, and $\boldsymbol{\xi}$ a p-dimensional vector of unknown fixed regression coefficients. The measurement-specific parameters θ_{ij} follow a conjugate distribution $\theta_{ij} \sim \mathcal{G}_{ij}(\vartheta_{ij}, \sigma^2_{ij})$ and the subject-specific parameters κ_{ij} follow a normal distribution $\boldsymbol{b_i} \sim N(\boldsymbol{0}, D)$, with D a variance-covariance matrix reflecting the structure assumed for the random effects.

It is computationally convenient, but not strictly necessary, to assume that the two sets of random effects, $\boldsymbol{\theta}_i$ and $\boldsymbol{b_i}$, are independent of each other. The components θ_{ij} of $\boldsymbol{\theta}_i$ can be independent on the one end, identical on the other, or different but correlated as a compromise in between these extremes. In line with Molenberghs *et al.* (2010) we take the first of these three routes.

For count data, we assume that

$$Y_{ij} \sim \text{Poi}(\lambda_{ij} = \theta_{ij}\kappa_{ij}), \quad (1)$$

$$\kappa_{ij} = \exp\left(\boldsymbol{x}'_{ij}\boldsymbol{\xi} + \boldsymbol{z}'_{ij}\boldsymbol{b_i}\right), \quad (2)$$

$\boldsymbol{b_i} \sim N(\boldsymbol{0}, D)$, and $\theta_{ij} \sim \text{Gamma}(\alpha, \beta)$. Further, α and β are shape and scale parameters, respectively. For identifiability reasons it is assumed that $\beta = 1/\alpha$. The measurement-specific gamma random effect θ_{ij} is used to accommodate overdispersion, while the subject-specific normal random effect $\boldsymbol{b_i}$ is used to model the correlation coming from the hierarchy in the data. This model is denoted by (PNG), in line with our aforementioned notational conventions.

4 Zero-inflated Models

In zero-inflated count models, it is assumed that there are two processes that can generate zeros: zeros may come from both a point mass (process 1) as well as from the count component (process 2). It is assumed that for observation i at time j, process 1 is chosen with probability π_{ij} and process 2 with probability $1 - \pi_{ij}$ (Hinde and Demétrio 1998ab). Process 1 generates only zeros, whereas process 2, $f_i(y_{ij}|\boldsymbol{b}_{1i}, \boldsymbol{\xi}, \theta_{ij})$, generates counts from a Poisson, a negative-binomial model, a Poisson-normal GLMM, or a Poisson-gamma-normal combined model. In its most general form, the zero-inflated Poisson-gamma-normal model is given as the following mixture:

$$Y_{ij} \sim \begin{cases} 0 & \text{with probability } \pi_{ij}, \\ f_i(y_{ij}|\boldsymbol{b}_{1i}, \boldsymbol{\xi}, \theta_{ij}) & \text{with probability } 1 - \pi_{ij}, \end{cases} \quad (3)$$

leading to the probabilities $p(Y_{ij} = y_{ij}|\boldsymbol{b}_{1i}, \boldsymbol{\xi}, \theta_{ij}, \pi_{ij})$ given by

$$p(Y_{ij} = y_{ij}|\boldsymbol{b}_{1i}, \boldsymbol{\xi}, \theta_{ij}, \pi_{ij}) = \begin{cases} \pi_{ij} + (1 - \pi_{ij})f_i(0|\boldsymbol{b}_{1i}, \boldsymbol{\xi}, \theta_{ij}) & \text{if } y_{ij} = 0, \\ (1 - \pi_{ij})f_i(y_{ij}|\boldsymbol{b}_{1i}, \boldsymbol{\xi}, \theta_{ij}) & \text{if } y_{ij} > 0. \end{cases} \quad (4)$$

The zero-inflation component $\pi_{ij} = \pi(\boldsymbol{x}'_{2ij}\boldsymbol{\gamma} + \boldsymbol{z}'_{2ij}\boldsymbol{b}_{2i})$ is modeled using a Bernouilli model: in the simplest case with only an intercept, but potentially containing known regressors $\boldsymbol{x}_{2ij}$ and $\boldsymbol{z}_{2ij}$, a vector of zero-inflation coefficients $\boldsymbol{\gamma}$ to be estimated, as well as random effects $\boldsymbol{b}_{2i}$. Common link functions, such as the logit or probit, can be used. Note that $\boldsymbol{x}_{ij}$, $\boldsymbol{z}_{ij}$, and $\boldsymbol{b}_i$ in Section 3 are now replaced by $\boldsymbol{x}_{1ij}$, $\boldsymbol{z}_{1ij}$, and $\boldsymbol{b}_{2i}$, respectively, for the non-zero count part. The regressors in the count and zero-inflation component can either be overlapping, a subset of the regressors can be used for

the zero-inflation, or entirely different regressors for the two parts can be used. In many cases, but of course not always, a simple random-intercept model is adequate, where $\boldsymbol{b}_{1i} = b_{1i}$, $\boldsymbol{b}_{2i} = b_{2i}$, and $\boldsymbol{z}_{1ij} = \boldsymbol{z}_{2ij} = 1$. Assuming that the random effects are normally distributed and possibly correlated with correlation parameter ρ, the variance-covariance matrix is

$$\boldsymbol{D} = \begin{pmatrix} d_1 & \rho\sqrt{d_1}\sqrt{d_2}) \\ \rho\sqrt{d_1}\sqrt{d_2} & d_2 \end{pmatrix}.$$

The model is denoted as ZI(PNG), as an obvious extension with earlier notational conventions. Three obvious special cases are ZI(PN-), ZI(P-G), and ZI(P--). Also, all four models without zero inflation are special cases as well. The conditional mean and variance of the ZI(PNG) are:

$$\begin{aligned} \mathsf{E}(Y_{ij}|\boldsymbol{b}_{1i}, \boldsymbol{\xi}, \theta_{ij}) &= \theta_{ij}\kappa_{ij}(1-\pi_{ij}), & (5) \\ \mathsf{Var}(Y_{ij}|\boldsymbol{b}_{1i}, \boldsymbol{\xi}, \theta_{ij}) &= \theta_{ij}\kappa_{ij}(1-\pi_{ij})[1+\theta_{ij}\kappa_{ij}(\pi_{ij}+1/\alpha)]. & (6) \end{aligned}$$

It can be seen that the conditional variance is inflated as a result of either overdispersion in the data (parameter α), or as a result of zero-inflation (parameter π_{ij}), or both.

5 Estimation

Likelihood estimation of the (PNG) is done by integrating over the random effects, assembling the marginal likelihood, and maximizing it in the usual way. Molenberghs, Verbeke, and Demétrio (2007) and Molenberghs *et al.* (2010) marginalized analytically over the gamma random effect, with then further numerical integration over the normal random effects. This enables the use of a flexible normal random-effects tool such as the SAS procedure NLMIXED. Example code can be found in the Appendix The partially marginalized (PNG) takes the form:

$$\begin{aligned} f(y_{ij}|\boldsymbol{b}_{1i}, \boldsymbol{\xi}) &= \int f(y_{ij}|\boldsymbol{b}_{1i}, \boldsymbol{\xi}, \theta_{ij}) f(\theta_{ij}|\alpha_j, \beta_j) d\theta_{ij} & (7) \\ &= \begin{pmatrix} \alpha_j + y_{ij} - 1 \\ \alpha_j - 1 \end{pmatrix} \cdot \left(\frac{\beta_j}{1+\kappa_{ij}\beta_j}\right)^{y_{ij}} \cdot \left(\frac{1}{1+\kappa_{ij}\beta_j}\right)^{\alpha_j} \kappa_{ij}^{y_{ij}}. & (8) \end{aligned}$$

This idea extends in a straightforward fashion to the ZI(PNG):

$$\begin{aligned} f(y_{ij}|\boldsymbol{b}_{1i}, \boldsymbol{\xi}, \boldsymbol{b}_{2i}, \boldsymbol{\gamma}) &= I(y_{ij}=0)\pi_{ij} \\ &\quad +(1-\pi_{ij}) \begin{pmatrix} \alpha_j + y_{ij} - 1 \\ \alpha_j - 1 \end{pmatrix} \cdot \left(\frac{\beta_j}{1+\kappa_{ij}\beta_j}\right)^{y_{ij}} \cdot \left(\frac{1}{1+\kappa_{ij}\beta_j}\right)^{\alpha_j} \kappa_{ij}^{y_{ij}}, \end{aligned}$$

with $\pi_{ij} = \pi(\boldsymbol{x}'_{2ij}\boldsymbol{\gamma} + \boldsymbol{z}'_{2ij}\boldsymbol{b}_{2i})$.

6 Simulation Study

In this section, we report on a simulation study set up to examine the bias in estimating the regression parameters when dealing with overdispersed, longitudinal count data with excess zeros. For such data, the bias is likely to result from not appropriately accounting for the excess zero counts, misspecification of the overdispersion, which is a very common situation for count data in a way that the prescribed mean-variance link is violated and misspecification of the correlation results from the repeated-measurements nature of the data.

6.1 Simulation Setting

Data are generated along a design inspired by the Jimma Infant Study. Age in months, status of getting medical help, and breast feeding behavior were among the covariates of interest in the study, and are used in the simulation study as well.

We randomly generated 200 data sets from the zero-inflated combined model for 2000 subjects with 10 measurements per subject. The response vector $\boldsymbol{y}_i$ for the i^{th} subject was generated as a correlated and overdispersed count from a negative-binomial process subject to zero-inflation. That is, for each subject, $Y_{ij} \sim \text{NB}(\psi_{ij}, \theta)$, where $\theta = 1$ with $\psi_{ij} = (1 + \kappa_{ij}/\theta)^{-1}$ and where $\kappa_{ij} = \exp\{\xi_0 + b_i + \xi_1 t_{ij} + \xi_2 H_{ij}\}$ for $i = 1, \ldots, 2000$ and $j = 1, \ldots, 10$. Further, t_{ij} represents the time point at which the j^{th} measurement is recorded for the i^{th} subject and H_{ij} denotes whether or not the i^{th} subject is given any medication help at the j^{th} measurement occasion, generated from a Bernoulli process with $p = 0.9$. Correlation is induced via a subject-specific random intercept b_i generated from a normal distribution with mean 0 and variance 0.8. Then, zero inflation is added by defining the final response vector $\boldsymbol{Y}_i^*$ to have components ${Y_{ij}}^* = (1 - u_{ij})Y_{ij}$, where the u_{ij} are Bernoulli random variables with parameters π_{ij} and $\text{logit}(\pi_{ij}) = \gamma_0 + \gamma_1 t_{ij}$.

Three different scenarios were considered for data generation: S_1: without excess zeros; S_2: with an excess of zeros of around 20%; S_3: with an excess of zeros of roughly 40%. The corresponding total zero percentages are 48%, 68%, and 88%, respectively. This was achieved, for each scenario, by appropriately choosing the zero-inflation coefficients. The true parameter values used to generate the data were $\xi = (1.12, 0.13, -1.89)^T$. Similarly, for the zero-inflation part, $\gamma = (-1, -1)^T$,

$\gamma = (1, -0.25)^T$ and $\gamma = (1.8, -0.1)^T$ were used for S_1, S_2, and S_3, respectively.

6.2 Simulation Results

The simulated data are analyzed by the ZI(PNG), ZI(P-G), ZI(PN-), and ZI(P--), as well as by their non-zero-inflated counterparts. Mean, relative bias (rbias) and predicted probabilities of zero counts are summarized for the three scenarios in Tables 2–4, respectively.

Parameter estimates of the ZI(PNG) were in agreement with their true model in all scenarios. This shows that the different components: zero-inflation, overdispersion, and correlation, can be well separated in practice, in settings like the ones considered here. The zero-inflated model converged for almost all simulated sets of data, apart from one perhaps idiosyncratic failure to do so.

Under S_1, as shown in Table 2, the ZI(PNG) and the (PNG) performed well and fairly similar in terms of relative bias, except for the intercept ξ_0 for which a larger bias is observed in the (PNG). The percentage of zero counts (48%) is nearly equally predicted in both cases. But, severe impact starts to emerge in the non zero-inflation models when excess zero counts are present, but not accounted for, as evidenced in Tables 3 and 4. The predicted number of zero counts is largely underestimated in the non-zero-inflated models. When many zeros are allowed for, as in S_3, the effect is more pronounced in the intercept term and the negative-binomial parameter α as compared to S_2. Moreover, the bias in the standard deviation of the random-effects, for instance, in the 'true' model tends to increase in S_3, which gets substantially higher for models with neglected zero-inflation component, such as the (PNG) and (PN-).

The impact of omitting the overdispersion is remarkable. This can be clearly observed, for example, from the considerable increment in the relative bias of the ZI(PN-). When overdispersion is omitted, the zero-inflation component will try to recover part of the overdispersion.

When the correlation stemming from the repeated measurements is misspecified, substantial impact appears in inferences of the ZI(P-G), which gets even worse in the (P-G), as evidenced quite clearly from the larger relative bias of the intercept term. When correlation is omitted from the model, the overdispersion term will try to recover for this misspecification.

Unlike in S_1, the ZI(PNG) significantly beats the (PNG), confirming the importance of accounting for the excess zeros in addition to the repeated measures nature and the overdispersion.

We conclude that failure to account for excess zeros, overdispersion, and/or correlation has a substan-

tial impact on bias and predicted probabilities. This was clearly shown on such key model parameters as the intercept term, the overdispersion parameter, and the variance of the random effects. All scenarios suggest that the zero-inflated combined model is the preferred one in terms of relative bias and predicted probabilities of zeros.

7 Analysis of Case Studies

7.1 The Jimma Infant Growth Study

We will fit the ZI(PNG) to the data, introduced in Section 2.1, and compare it to its special cases: (P--), (P-G), (PN-) (PNG), ZI(P--), ZI(PN-), and ZI(P-G). We model κ_{ij} as

$$\begin{aligned} \ln(\kappa_{ij}) &= \xi_0 + b_{1i} + \xi_1 R_i + \xi_2 U_i + \xi_3 T_{ij} + \xi_4 G_i + \xi_5 B_{ij} + \xi_6 H_{ij} + \xi_7 R_i T_{ij} \\ &\quad + \xi_8 U_i T_{ij} + \xi_9 G_i T_{ij} + \xi_{10} B_{ij} T_{ij} + \xi_{11} H_{ij} T_{ij} \end{aligned}$$

and the zero-inflation probability (π_{ij}) as

$$\text{logit}(\pi_{ij}) = \gamma_0 + b_{2i} + \gamma_1 R_i + \gamma_2 U_i + \gamma_3 T_{ij} + \gamma_4 G_i + \gamma_5 B_{ij} + \gamma_6 H_{ij},$$

with R_i an indicator for rural residence and U_i for urban residence. The semi-urban residence category is taken as the reference. Further, G_i is a gender indicator and T_{ij} is the time point at which the j^{th} measurement is taken for the i^{th} subject; B_{ij} and H_{ij} denote, respectively, whether or not the i^{th} infant is breastfed and given any medication between the $(j-1)^{st}$ and j^{th} measurement occasions.

Clearly, as can be observed from Tables 5 and 6, the zero-inflated models performed much better than their respective non-zero-inflated counterparts, resulting in a substantial improvement in fit, thence implying that the extra zeros need to be accommodated, which is expected given the excessive zero counts in these data as shown in Section 2.1.

The ZI(PN-) model is an important improvement, in terms of likelihood, relative to the ZI(P--), while much more improvement is gained in the case of the ZI(P-G) relative to the ZI(P–). Moreover, considering the ZI(PNG), there is a strong improvement in fit when the gamma and normal random effects, in addition to zero-inflation, are simultaneously included. A similar observation can be made for the non-zero-inflated models.

There is a very strong improvement in fit of the ZI(P-G), when compared to the ZI(PN-). It points to the fact that overdispersion is more important an effect than the repeated-measures nature, hence

Table 2: *Simulation study under scenario S_1. Mean, standard error, and relative bias of the parameter estimates in ZI(PNG), ZI(P-G), ZI(PN-), ZI(P--), and its non-zero-inflated counterparts.*

			ZI(PNG)		(PNG)		ZI(P-G)		(P-G)	
Effect	Parameter	True	mean (s.e.)	rbias	mean (s.e.)	rbias	mean (s.e.)	rbias	mean (s.e.)	rbias
Intercept	ξ_0	1.12	1.068(0.003)	0.046	0.991(0.004)	0.115	1.277(0.003)	0.139	2.404(0.005)	1.147
Time	ξ_1	0.13	0.125(0.001)	0.040	0.136(0.001)	0.046	0.125(0.001)	0.040	0.133(0.001)	0.026
Help	ξ_2	-1.89	-1.794(0.002)	0.051	-1.796(0.002)	0.049	-1.705(0.002)	0.098	-1.708(0.002)	0.096
Negative-binomial parameter	α	1.00	0.953(0.002)	0.047	0.995(0.002)	0.005	1.774(0.003)	0.774	0.552(0.001)	0.448
Std. dev random effect	$\sqrt{d}$	0.80	0.780(0.001)	0.025	0.779(0.001)	0.026	–	–	–	–
Inflation intercept	γ_0	-1.00	-0.856(0.099)	0.104	–	–	-0.265(0.123)	0.725	–	–
Inflation time	γ_1	-1.00	-1.049(0.098)	0.049	–	–	-1.698(0.122)	0.687	–	–
Predicted prob. zeros		0.48	0.493		0.481		0.359		0.291	
Frequency of convergence			199		200		200		200	
			ZI(PN-)		(PN-)		ZI(P--)		(P--)	
Effect	Parameter	True	mean (s.e.)	rbias	mean (s.e.)	rbias	mean (s.e.)	rbias	mean (s.e.)	rbias
Intercept	ξ_0	1.12	1.216(0.003)	0.086	0.892(0.003)	0.204	1.661(0.003)	0.483	1.250(0.003)	0.116
Time	ξ_1	0.13	0.101(0.001)	0.225	0.127(0.001)	0.026	0.089(0.001)	0.318	0.124(0.001)	0.043
Help	ξ_2	-1.89	-1.467(0.002)	0.224	-1.693(0.002)	0.104	-1.275(0.002)	0.326	-1.682(0.002)	0.109
Std. dev random effect	$\sqrt{d}$	0.80	0.796(0.001)	0.005	0.861(0.001)	0.076	–	–	–	–
Inflation intercept	γ_0	-1.00	-0.386(0.005)	0.614	–	–	0.247(0.003)	1.247	–	–
Inflation time	γ_1	$-.00$	-0.094(0.001)	0.906	–	–	-0.094(0.001)	0.906	–	–
Predicted prob. zeros		0.48	0.473		0.365		0.483		0.255	
Frequency of convergence			200		200		200		200	

Table 3: *Simulation study under scenario S_2. Mean, standard error, and relative bias of the parameter estimates in ZI(PNG), ZI(P-G), ZI(PN-), ZI(P--), and its non-zero-inflated counterparts.*

			ZI(PNG)		(PNG)		ZI(P-G)		(P-G)	
Effect	Parameter	True	mean (s.e.)	rbias	mean (s.e.)	rbias	mean (s.e.)	rbias	mean (s.e.)	rbias
Intercept	ξ_0	1.12	1.079(0.004)	0.037	1.833(0.005)	0.637	1.089(0.005)	0.027	2.796(0.006)	1.497
Time	ξ_1	0.13	0.123(0.001)	0.052	0.239(0.001)	0.839	0.125(0.001)	0.040	0.225(0.001)	0.730
Help	ξ_2	-1.89	$-1.766(0.003)$	0.066	$-1.776(0.003)$	0.060	$-1.671(0.003)$	0.116	$-1.703(0.003)$	0.099
Negative-binomial parameter	α	1.00	0.908(0.004)	0.093	0.372(0.001)	0.628	2.379(0.008)	1.379	0.266(0.001)	0.734
Std. dev random effect	$\sqrt{d}$	0.80	0.772(0.002)	0.035	0.754(0.002)	0.058	–	–	–	–
Inflation intercept	γ_0	1.00	1.056(0.005)	0.056	–	–	0.993(0.006)	0.003	–	–
Inflation time	γ_1	-0.25	$-0.246(0.001)$	0.014	–	–	$-0.354(0.001)$	0.416	–	–
Predicted prob. zeros		0.68	0.696		0.398		0.549		0.367	
Frequency of convergence			200		200		200		200	
			ZI(PN-)		(P-N)		ZI(P--)		(P--)	
Effect	Parameter	True	mean (s.e.)	rbias	mean (s.e.)	rbias	mean (s.e.)	rbias	mean (s.e.)	rbias
Intercept	ξ_0	1.12	1.183(0.004)	0.056	$-0.235(0.004)$	1.210	1.666(0.004)	0.488	0.215(0.004)	0.808
Time	ξ_1	0.13	0.099(0.001)	0.235	0.212(0.001)	0.633	0.087(0.001)	0.329	0.210(0.001)	0.613
Help	ξ_2	-1.89	$-1.444(0.003)$	0.236	$-1.679(0.003)$	0.112	$-1.261(0.002)$	0.420	$-1.664(0.003)$	0.120
Std. dev random effect	$\sqrt{d}$	0.80	0.834(0.001)	0.042	0.976(0.001)	0.220	–	–	–	–
Inflation intercept	γ_0	1.00	1.473(0.004)	0.473	–	–	1.816(0.003)	0.816	–	–
Inflation time	γ_1	-0.25	$-0.209(0.001)$	0.163	–	–	$-0.202(0.001)$	0.193	–	–
Predicted prob. zeros		0.68	0.677		0.520		0.682		0.422	
Frequency of convergence			200		200		200		200	

Table 4: *Simulation study under scenario S_3. Mean, standard error, and relative bias of the parameter estimates in ZI(PNG), ZI(P-G), ZI(PN-), ZI(P--), and its non-zero-inflated counterparts.*

			ZI(PNG)		(PNG)		ZI(P-G)		(P-G)	
Effect	Parameter	True	mean (s.e.)	rbias	mean (s.e.)	rbias	mean (s.e.)	rbias	mean (s.e.)	rbias
Intercept	ξ_0	1.12	1.076(0.007)	0.039	4.005(0.009)	2.828	0.980(0.009)	0.125	4.494(0.008)	3.012
Time	ξ_1	0.13	0.125(0.001)	0.042	0.216(0.001)	0.658	0.121(0.001)	0.070	0.202(0.001)	0.554
Help	ξ_2	-1.89	$-1.757(0.005)$	0.070	$-1.765(0.005)$	0.067	$-1.676(0.005)$	0.113	$-1.701(0.005)$	0.100
Negative-binomial parameter	α	1.00	0.887(0.007)	0.112	0.088(0.001)	0.912	3.041(0.034)	2.041	0.076(0.001)	0.924
Std. dev random effect	$\sqrt{d}$	0.80	0.765(0.003)	0.043	0.609(0.004)	0.239	–	–	–	–
Inflation intercept	γ_0	1.80	1.862(0.006)	0.034	–	–	1.487(0.009)	0.174	–	–
Inflation time	γ_1	-0.10	$-0.102(0.001)$	0.017	–	–	$-0.118(0.001)$	0.177	–	–
Predicted prob. zeros		0.88	0.884		0.590		0.807		0.604	
Frequency of convergence			200		200		200		200	
			ZI(PN-)		(P-N)		ZI(P--)		(P--)	
Effect	Parameter	True	mean (s.e.)	rbias	mean (s.e.)	rbias	mean (s.e.)	rbias	mean (s.e.)	rbias
Intercept	ξ_0	1.12	1.051(0.008)	0.061	$-1.515(0.007)$	2.353	1.660(0.006)	0.482	$-0.631(0.007)$	1.563
Time	ξ_1	0.13	0.104(0.001)	0.203	0.195(0.001)	0.502	0.088(0.001)	0.323	0.193(0.001)	0.487
Help	ξ_2	-1.89	$-1.473(0.005)$	0.221	$-1.669(0.005)$	0.117	$-1.257(0.004)$	0.335	$-1.661(0.005)$	0.121
Std. dev random effect	$\sqrt{d}$	0.80	0.941(0.002)	0.176	1.416(0.002)	0.769	–	–	–	–
Inflation intercept	γ_0	1.80	2.205(0.005)	0.225	–	–	2.629(0.004)	0.382	–	–
Inflation time	γ_1	-0.10	$-0.112(0.001)$	0.122	–	–	$-0.127(0.001)$	0.271	–	–
Predicted prob. zeros		0.88	0.876		0.756		0.877		0.675	
Frequency of convergence			200		200		200		200	

the ZI(P-G) is able to perform better from the start. It underscores, once more, that overdispersion with count data is a very common situation. Eventually, both are needed.

The zero-inflation regression coefficients are similar in all models, statistically significant, and can be interpreted as model coefficients for the proportion of extra zeros.

ZI(PNG) and ZI(P-G) exhibit similar fits, not only in terms of parameter estimates but also in inference, except that gender is significant in the former ($p = 0.0311$) while this is not the case for the latter ($p = 0.0922$). Both models suggest that medical help, breast feeding, main effect of rural place of residence are significant; the same is true for time interactions with breast feeding and urban place of residence.

7.2 Epilepsy Data

We analyze the epilepsy data, introduced in Section 2.2. Let Y_{ij} represent the number of epileptic seizures that patient i experiences during week j of the follow-up period. Also, let t_{ij} be the time-point at which Y_{ij} has been recorded. Consider the combined model (1)–(4), with parameterization similar to the one in Molenberghs *et al.* (2010), but now accounting for zero inflation, assuming that counts are generated from a (PN-) process with mean λ_{ij}:

$$\ln(\lambda_{ij}) = \begin{cases} (\xi_{00} + b_{1i}) + \xi_{01} t_{ij} & \text{if placebo,} \\ (\xi_{10} + b_{1i}) + \xi_{11} t_{ij} & \text{if treated,} \end{cases} \tag{9}$$

or from a (PNG) process with mean $\lambda_{ij} = \theta_{ij}\kappa_{ij}$:

$$\ln(\kappa_{ij}) = \begin{cases} (\xi_{00} + b_{1i}) + \xi_{01} t_{ij} & \text{if placebo,} \\ (\xi_{10} + b_{1i}) + \xi_{11} t_{ij} & \text{if treated,} \end{cases} \tag{10}$$

The zero-inflation probability (π_{ij}) is modeled as $\text{logit}(\pi_{ij}) = \gamma_0 + b_{2i} + \gamma_1 t_{ij}$. The data are analyzed with the ZI(PNG), ZI(P-G), ZI(PN-), ZI(P--). For the sake of comparison, also the non-zero-inflated counterparts are fitted. Parameter estimates and predicted probabilities of zeros are presented in Table 7. Clearly, in terms of likelihood comparison, the zero-inflated versions performed much better, resulting in a substantial improvement in fit.

The ZI(P-G) is an important improvement relative to the ZI(P--), while much more improvement is gained in the case of the ZI(PN-). Moreover, the ZI(PNG) leads to a substantially improved

fit. Further, we observe that, omitting either the overdispersion or the correlation underestimates the predicted probability of zeros, which becomes worse when both are omitted at the same time. The ZI(PNG), fitted without random effects in the zero-inflation part, results in -2log-likelihood of 5386.8, and predicted probability of zeros equal to 0.3271. This implies that inclusion of random effects in the zero-inflation part tends to have little impact on the predicted probability of zeros. However, based on likelihood comparison, model fit improves considerably. This same phenomenon is also evident in the ZI(PN-) fitted with random effects included only in the non-zero count part (-2log-likelihood is 5971.9, and predicted probability of zeros 0.3112).

None of the zero-inflated models suggests evidence of significance in slope difference and slope ratio, except for the ZI(P--), where significance is maintained for the slope difference ($p = 0.0004$). However, the latter, unrealistically, omits correlation and overdispersion. The zero-inflation regression coefficients can be interpreted as model coefficients for the proportion of extra zeros, and are statistically significant in all ZI models.

8 Concluding Remarks

There is quite a bit of research on longitudinal count data, with or without overdispersion, and with or without excess zeros. In particular, the combined model by Molenberghs, Verbeke, and Demétrio (2007) and Molenberghs *et al.* (2010) uses normal random effects to capture the hierarchy in the count data and some overdispersion, with gamma random effects to more flexibly capture overdispersion. Also, zero inflation has been studied in the literature. The novelty of our work is that all these features are combined into one model, with more conventional models following as special cases.

In terms of estimation, we have focused on maximum likelihood estimation, in such a way that standard statistical software, such as the SAS procedure NLMIXED, can be used. An example of such code is given in the Appendix.

Of course, with the considerations of not only one but multiple sets of random effects, comes the obligation to reflect on the precise nature of such latent structures. As underscored by Verbeke and Molenberghs (2010), full verification of the adequacy of a random-effects structure is not possible based on statistical considerations alone, because there is a many-to-one map from hierarchical models to the implied marginal model. Of course, this should not stop the user from considering

such models, but rather issues a word of caution.

In this sense, it would be of interest to study extensions of or alternative formulations for the normal random effects. For example, normal mixtures for the normal random effects could be used. Such mixtures can be generated by assuming normality conditional on the mean vector, which itself is assumed to be sampled from a discrete distribution with as many support points as the number of mixture components (see, for example, Verbeke and Molenberghs 2000, Ch. 12).

Acknowledgments

The authors are grateful to M. Assefa and F. Tessema for the permission to use the data. Financial support from the Institutional University Cooperation of the Council of Flemish Universities (VLIR-IUC) is gratefully acknowledged. The authors gratefully acknowledge support from IAP research Network P7/06 of the Belgian Government (Belgian Science Policy).

References

Agresti, A. (2002) *Categorical Data Analysis* (2nd ed.). New York: John Wiley & Sons.

Andy, H., Jane, A., Kelvin, K., and Geoffery, J. (2006) Multi-level zero-inflated modeling of correlated count data with excess zeros. *Statistical Methods in Medical Research*, **15**, 47–61

Asefa, M. and Tessema, F. (2002) Infant survivorship and occurrence of multiple-births: A longitudinal community-based study, south west Ethiopia. *Ethiopian Journal of Health Development*, **16**, 5–11.

Booth, J., Casella, G., Friedl, H. and Hobert, J. (2003) Negative binomial loglinear mixed models *Statistical Modelling*, **3**, 179-191.

Breslow, N. (1984) Extra-Poisson variation in log-linear models. *Applied Statistics*, **33**, 38–44.

Breslow, N.E. and Clayton, D.G. (1993) Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.

Breslow, N.E. and Lin, X. (1995) Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika*, **82**, 81–91.

Engel, B. and Keen, A. (1994) A simple approach for the analysis of generalized linear mixed models. *Statistica Neerlandica*, **48**, 1–22.

Faught, E., Wilder, B.J., Ramsay, R.E., Reife, R.A., Kramer, L.D., Pledger, G.W., and Karim, R.M. (1996) Topiramate placebo-controlled dose-ranging trial in refractory partial epilepsy using 200-, 400-, and 600-mg daily dosages, *Neurology*, **46**, 1684–1690.

Greene, W. (1994). Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. Working Paper EC- 94-10, Department of Economics, New York University.

Hinde, J. and Demétrio, C.G.B. (1998a) Overdispersion: Models and estimation. *Computational Statistics and Data Analysis*, **27**, 151–170.

Hinde, J. and Demétrio, C.G.B. (1998b) *Overdispersion: Models and Estimation.* São Paulo: XIII Sinape.

Lambert D. (1992) Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1–14.

Lawless, J. (1987) Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics*, **15**, 209–225.

Lee, AH., Wang, K., Scott, J., Yau, KKW. and McLachlan, GJ. (2006) Multi-level zero-inflated Poisson regression modelling of correlated count data with excess zeros. *Statistical Methods in Medical Research*, **15**, 47–61.

McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models*. London: Chapman & Hall/CRC.

Min, Y. and Agresti, A. (2005) Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling*, **5**, 1–19.

Molenberghs, G. and Verbeke, G. (2005) *Models for Discrete Longitudinal Data.* New York: Springer.

Molenberghs, G., Verbeke, G., and Demétrio, C. (2007) An extended random-effects approach to modeling repeated, overdispersed count data. *Lifetime Data Analysis*, **13**, 513–531.

Molenberghs, G., Verbeke, G., Demétrio, C., and Vieira, A. (2010) A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical Science*, **25**, 325–347.

Mullahy, J (1986) Specification and testing of some modified count data models. *Journal of Econometrics* **33**, 341–65.

Nelder, J.A. and Wedderburn, R.W.M. (1972) Generalized linear models. *Journal of the Royal Statistical Society, Series B*, **135**, 370–384.

Verbeke, G. and Molenberghs, G. (2000) *Linear Mixed Models for Longitudinal Data.* New York: Springer.

Verbeke, G. and Molenberghs, G. (2010) Arbitrariness of models for augmented and coarse data, with emphasis on incomplete-data and random-effects models. *Statistical Modelling*, **10**, 391–419.

Wolfinger, R. and O'Connell, M. (1993) Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, **48**, 233–243.

Table 5: *Jimma Infant Growth Study. Parameter estimates and standard errors for the regression coefficients in (P--), (P-G), (PN-), and (PNG).*

		(P--)	(PN-)	(P-G)	(PNG)
Effect	Parameter	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)
Intercept	ξ_0	3.4198(0.0648)	2.0652(0.0744)	3.6443(0.3573)	5.7541(0.4567)
Rural	ξ_1	0.2209(0.0229)	0.2209(0.0291)	0.1674(0.0906)	−0.0733(0.1231)
Urban	ξ_2	−0.1850(0.0331)	−0.5266(0.0399)	−0.1185(0.1157)	−0.3000(0.1600)
Time	ξ_3	−0.1477(0.0073)	−0.1307(0.0078)	−0.1870(0.0425)	−0.3287(0.0506)
Gender	ξ_4	0.1681(0.0182)	0.2478(0.0241)	0.2351(0.0767)	0.2444(0.1041)
Breast feeding	ξ_5	−1.5710(0.0614)	−1.4554(0.0664)	−1.8120(0.3066)	−3.1539(0.4151)
Help	ξ_6	−3.2198(0.0196)	−2.9870(0.0230)	−3.7025(0.1784)	−6.1493(0.1896)
Slope Rural	ξ_7	−0.0085(0.0027)	−0.0090(0.0029)	−0.0033(0.0139)	0.0182(0.0158)
Slope Urban	ξ_8	0.0461(0.0037)	0.0542(0.0039)	0.0397(0.0174)	0.0797(0.0202)
Slope Gender	ξ_9	−0.0011(0.0021)	−0.0061(0.0023)	−0.0033(0.0114)	0.0063(0.0129)
Slope Breast feeding	ξ_{10}	0.1583(0.0069)	0.1441(0.0072)	0.1988(0.0359)	0.3213(0.0453)
Slope Help	ξ_{11}	0.1641(0.0023)	0.1324(0.0081)	0.2326(0.0221)	0.3448(0.0219)
Std. dev random effect	$\sqrt{d}$	—	1.9612(0.0267)	—	1.6847(0.0433)
Negative-binomial parameter	α	—	—	0.0641(0.0009)	0.1045(0.0021)
−2log-likelihood		281,126	203,981	91,370	90,274

Table 6: *Jimma Infant Growth Study. Parameter estimates and standard errors for the regression coefficients in ZI(P--), ZI(P-G), ZI(PN-), and ZI(PNG).*

		ZI(P--)	ZI(PN-)	ZI(P-G)	ZI(PNG)
Effect	Parameter	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)
Intercept	ξ_0	2.2148(0.0636)	1.3877(0.1205)	2.2200(0.1571)	2.0388(0.1616)
Rural	ξ_1	0.2610(0.0252)	0.3880(0.0400)	0.2536(0.0577)	0.2804(0.0586)
Urban	ξ_2	−0.1049(0.0364)	−0.0945(0.0549)	−0.1096(0.0842)	−0.1301(0.0858)
Time	ξ_3	−0.0289(0.0072)	0.0331(0.0119)	−0.0302(0.0176)	−0.0213(0.0178)
Gender	ξ_4	0.0835(0.0199)	0.1338(0.0321)	0.0797(0.0473)	0.1027(0.0477)
Breast feeding	ξ_5	−0.3430(0.0593)	0.0644(0.1138)	−0.3370(0.1481)	−0.3384(0.1528)
Help	ξ_6	0.2378(0.0211)	0.3312(0.0298)	0.2028(0.0498)	0.2225(0.0507)
Slope Rural	ξ_7	−0.0047(0.0030)	−0.0202(0.0042)	−0.0043(0.0071)	− 0.0060(0.0070)
Slope Urban	ξ_8	0.0222(0.0041)	0.0178(0.0059)	0.0223(0.0096)	0.0227(0.0096)
Slope Gender	ξ_9	−0.0010(0.0023)	−0.0100(0.0032)	−0.0003(0.0056)	− 0.0035(0.0054)
Slope Breast feeding	ξ_{10}	0.0372(0.0066)	− 0.0011(0.0113)	0.0375(0.0164)	0.0345(0.0167)
Slope Help	ξ_{11}	0.0087(0.0059)	0.0019(0.0035)	0.0087(0.0059)	0.0084(0.0058)
Std. dev. non-zero part random effect	$\sqrt{d_1}$	—	0.5856(0.0075)	—	0.4311(0.0112)
Negative-binomial parameter	α	—	—	0.4797(0.0099)	0.2807(0.0086)
Inflation intercept	γ_0	−6.0412(0.6933)	−6.0163(0.5759)	−6.0608(0.6255)	−6.0241(0.5656)
Inflation Rural	γ_1	0.1231(0.0396)	0.1222(0.0467)	0.1331(0.0398)	0.1306(0.0469)
Inflation Urban	γ_2	−0.1380(0.0475)	−0.1578(0.0569)	−0.1368(0.0478)	−0.1578(0.0571)
Inflation Time	γ_3	−0.1835(0.0045)	−0.1941(0.0048)	−0.1834(0.0045)	−0.1942(0.0048)
Inflation Gender	γ_4	−0.1606(0.0328)	−0.1658(0.0388)	−0.1582(0.0329)	−0.1675(0.0389)
Inflation Breast feeding	γ_5	0.2056(0.0814)	0.2394(0.0940)	0.1960(0.0821)	0.2285(0.0945)
Inflation Help	γ_6	9.3894(0.6877)	9.6095(0.5680)	9.3833(0.6192)	9.6145(0.5576)
Std. dev. zero part random effect	$\sqrt{d_2}$	—	0.7575(0.0333)	—	0.7604(0.0335)
Correlation of random effects	ρ	—	−0.0907(0.0402)	—	−0.1127(0.0566)
−2log-likelihood		100,780	80,555	74,489	73,570

Table 7: *Epilepsy Study. Parameter estimates and standard error in ZI(P--), ZI(P-G), ZI(PN-), ZI(PNG), (P--), (P-G), (PN-), and (PNG).*

		ZI(PNG)	(PNG)	ZI(P-G)	(P-G)
Effect	Parameter	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)
Intercept placebo	ξ_{00}	0.9467(0.1665)	0.9113(0.1755)	1.2361(0.1100)	1.2594(0.1119)
Slope placebo	ξ_{01}	−0.0162(0.0075)	−0.0248(0.0077)	−0.0072(0.0113)	−0.0126(0.0111)
Intercept treatment	ξ_{10}	0.8361(0.1716)	0.6557(0.1782)	1.3974(0.1098)	1.4750(0.1093)
Slope treatment	ξ_{11}	−0.0061(0.0074)	−0.0118(0.0075)	−0.0219(0.0112)	−0.0352(0.0101)
Negative-binomial parameter	α_1	0.2449(0.0253)	2.4640(0.2113)	1.7874(0.1004)	0.5274(0.0255)
Std. dev. non-zero part random effect	$\sqrt{d_1}$	0.9974(0.0854)	1.0625(0.0871)	–	–
Inflation intercept	γ_0	−4.5813(0.6405)	–	−7.1064(1.3344)	–
Inflation slope	γ_1	0.0921(0.0339)	–	0.2921(0.0655)	–
Std. dev. zero part random effect	$\sqrt{d_2}$	2.5327(0.4396)	-	–	–
Correlation of random effects	ρ	−0.0961(0.1534)	–	–	–
Predicted prob. zeros		0.3522	0.3206	0.1849	0.1583
−2log-likelihood		5317.9	5417.0	6318.9	6326.1

		ZI(PN-)	(PN-)	ZI(P--)	(P--)
Effect	Parameter	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)
Intercept placebo	ξ_{00}	0.9027(0.1552)	0.8179(0.1677)	1.4205(0.0439)	1.2662(0.0424)
Slope placebo	ξ_{01}	−0.0042(0.0047)	−0.0143(0.0044)	0.0061(0.0045)	−0.0134(0.0043)
Intercept treatment	ξ_{10}	0.9078(0.1590)	0.6475(0.1701)	1.7608(0.0402)	1.4531(0.0383)
Slope treatment	ξ_{11}	−0.0074(0.0045)	−0.0120(0.0043)	−0.0153(0.0041)	−0.0328(0.0038)
Std. dev. non-zero part random effect	$\sqrt{d_1}$	0.9713(0.0824)	1.0755(0.0857)	–	–
Inflation intercept	γ_0	−3.7123(0.5003)	–	−1.2879(0.1203)	–
Inflation slope	γ_1	0.0952(0.0249)	–	0.0593(0.0109)	–
Std. dev. zero part random effect	$\sqrt{d_2}$	2.2215(0.3434)	–	–	–
Correlation of random effects	ρ	−0.1541(0.1574)	–	–	–
Predicted prob. zeros		0.3384	0.2627	0.3316	0.0459
−2log-likelihood		5845.1	6271.9	9760	11590

A Zero-Inflated Overdispersed Hierarchical Poisson Model

Wondwosen Kassahun[1] **Thomas Neyens**[2] **Christel Faes**[2]

Geert Molenberghs[2,3] **Geert Verbeke**[3,2]

[1] *Department of Epidemiology and Biostatistics, Jimma University, Ethiopia*

[2] *I-BioStat, CenStat, Universiteit Hasselt, B-3590 Diepenbeek, Belgium*

[3] *I-BioStat, L-BioStat, Katholieke Universiteit Leuven, B-3000 Leuven, Belgium*

Supplementary Material: SAS Procedure NLMIXED Code

A Sample PROC NLMIXED Code

```
/*
Analyses for the epilepsy Data
treatment 0= placebo
treatment 1= treatment
y=nseizw
time=studyweek
*/

/*(P--) Model */
proc nlmixed data=epilepsy qpoints=20;
title 'Univariate analyse';
parms int0=0.5 slope0=-0.1 int1=1 slope1=0.1;
if (trt = 0) then eta = int0 + slope0*time;
else if (trt = 1) then eta = int1  + slope1*time;
lambda = exp(eta);
loglik=-lambda+y*eta-log(fact(y));
model y~ general(loglik);
estimate 'difference in slope' slope1-slope0;
estimate 'ratio of slopes' slope1/slope0;
predict exp(-lambda) out=P;
run;

/*Average predicted probability of zeros*/
proc means data=P;
```

```
var pred;
run;


/*ZI(P--) Model */
proc nlmixed data=epilepsy qpoints=20;
title 'Univariate analyse';
parms int0=0.5 slope0=-0.1 int1=1 slope1=0.1 a0=0 a1=0;
eta_prob = a0+ a1*time ;
p_0 = exp(eta_prob) / (1 + exp(eta_prob));
if (trt = 0) then eta = int0 + slope0*time;
else if (trt = 1) then eta = int1  + slope1*time;
lambda = exp(eta);
if y = 0 then loglik = log(p_0 + (1 - p_0) * exp(-lambda));
else loglik = log(1 - p_0) + y * log(lambda)- lambda  - lgamma(y+1);
model y~ general(loglik);
estimate 'difference in slope' slope1-slope0;
estimate 'ratio of slopes' slope1/slope0;
predict p_0 +(1-p_0)*exp(-lambda) out=ZIP;
run;

/*Average predicted probability of zeros*/
proc means data=ZIP;
var pred;
run;

/*(PN-) Model */
proc nlmixed data=epilepsy qpoints=20;
title 'Poisson-normal met general likelihood';
parms int0=0.5 slope0=-0.1 int1=1 slope1=0.1 sigma=1;
if (trt = 0) then eta = int0 + b + slope0*time;
else if (trt = 1) then eta = int1 + b + slope1*time;
lambda = exp(eta);
loglik=-lambda+nseizw*eta-log(fact(y));
model nseizw ~ general(loglik);
random b ~ normal(0,sigma**2) subject = id;
estimate 'difference in slope' slope1-slope0;
estimate 'ratio of slopes' slope1/slope0;
predict exp(-lambda) out=PN;
run;

/*Average predicted probability of zeros*/
proc means data=PN;
```

```
var pred;
run;

/*ZI(PN-) Model*/
proc nlmixed data=epilepsy qpoints=20;
title 'Poisson-normal met general likelihood';
parms int0=0.8179 slope0=-0.014 int1=0.647 slope1=-0.012 d11=0.98
       rho=0 d22=1.10 a0=-3 a1=0.1;
eta_prob = a0+ a1*time+b2 ;
p_0 = exp(eta_prob) / (1 + exp(eta_prob));
if (trt = 0) then eta = int0 + b1 + slope0*time;
else if (trt = 1) then eta = int1 + b1 + slope1*time;
lambda = exp(eta);
if y = 0 then loglik = log(p_0 + (1 - p_0) * exp(-lambda));
else loglik = log(1 - p_0) + y * log(lambda) - lambda - log(fact(y));
random b1 b2 ~ normal([0,0], [d11**2,rho*d11*d22,d22**2]) subject = id;
model y ~ general(loglik);
estimate 'difference in slope' slope1-slope0;
estimate 'ratio of slopes' slope1/slope0;
predict p_0+(1-p_0)*exp(-lambda) out=ZIPN;
run;

/*Average predicted probability of zeros*/
proc means data=ZIPN;
var pred;
run;

/*(P-G) Model*/
proc nlmixed data=epilepsy qpoints=20;
title 'Poisson-gamma == negative-binomial - alpha*beta=1';
parms int0=0.5 slope0=-0.1 int1=1 slope1=0.1 alpha=2;
if (trt = 0) then eta = int0 + slope0*time;
else if (trt = 1) then eta = int1 + slope1*time;
lambda = exp(eta);
beta=1/alpha;
loglik=lgamma(alpha+y)-lgamma(alpha)+y*log(beta)-(y+alpha)*log(1+beta*lambda)
       +y*eta-lgamma(y+1);
model y ~ general(loglik);
estimate 'difference in slope' slope1-slope0;
estimate 'ratio of slopes' slope1/slope0;
estimate 'beta=1/alpha' 1/alpha;
predict (1/(1+lambda/beta))**beta out=PG;
run;
```

```
/*Average predicted probability of zeros*/
proc means data=PG;
var pred;
run;

/*ZI(P-G) Model*/
proc nlmixed data=epilepsy qpoints=20;
title 'Poisson-gamma == negative-binomial - alpha*beta=1';
parms int0=0.5 slope0=-0.1 int1=1 slope1=0.1 alpha=0.05 a0=-1 a1=0.1;
if (trt = 0) then eta = int0 + slope0*time;
else if (trt = 1) then eta = int1 + slope1*time;
lambda = exp(eta);
eta_prob=a0+a1*time;
p_0=exp(eta_prob)/(1+exp(eta_prob));
m = 1/alpha;
p = 1/(1+alpha*lambda);
if y=0 then
ll = log(p_0+ (1-p_0)*(p**m));
else ll = log(1-p_0) + log(gamma(m + y)) - log(gamma(y + 1))
     - log(gamma(m)) + m*log(p) + y*log(1-p);
model y ~ general(ll);
estimate 'difference in slope' slope1-slope0;
estimate 'ratio of slopes' slope1/slope0;
estimate 'beta=1/alpha' 1/alpha;
predict p_0 +(1-p_0)*(1/(1+lambda/alpha))**alpha out=ZIPG ;
run;

/*Average predicted probability of zeros*/
proc means data=ZIPG;
var pred;
run;

/*(PNG) Model*/
proc nlmixed data=epilepsy  qpoints=20;
parms int0=0.5 slope0=-0.1 int1=1 slope1=0.1 sigma=1 alpha=1 ;
if (trt = 0) then eta = int0 + b + slope0*time;
else if (trt = 1) then eta = int1 + b + slope1*time;
lambda = exp(eta);
beta=1/alpha;
loglik=lgamma(alpha+y)-lgamma(alpha)+y*log(beta)-(y+alpha)*log(1+beta*lambda)
       +y*eta-lgamma(y+1);
random b ~ normal(0,sigma**2) subject = id ;
```

```
model y~ general(loglik);
estimate 'difference in slope' slope1-slope0;
estimate 'ratio of slopes' slope1/slope0;
estimate 'beta=1/alpha' 1/alpha;
predict (1/(1+lambda/alpha))**alpha out=PNG;
run;

/*Average predicted probability of zeros*/
proc means data=PNG;
var pred;
run;

/*ZI(PNG) Model*/
proc nlmixed data=epilepsy qpoints=20;
title 'Poisson-combined -  alpha*beta=1';
parms int0= 0.8511 slope0=-0.01048 int1=0.8165 slope1=-0.008 alpha=0.2937
d11=1.0810 rho=0 d22=3.19 a0=-1.78 a1=0.052;
if (trt = 0) then eta = int0 + b1 + slope0*time;
else if (trt = 1) then eta = int1 + b1 + slope1*time;
lambda = exp(eta);
eta_prob = a0+a1*time+b2 ;
p_0=exp(eta_prob)/(1+exp(eta_prob));
m = 1/alpha;
p = 1/(1+alpha*lambda);
if y=0 then
ll = log(p_0 + (1-p_0)*(p**m));
else ll = log(1-p_0) + log(gamma(m + y)) - log(gamma(y + 1))
     - log(gamma(m)) + m*log(p) + y*log(1-p);
model y ~ general(ll);
random b1 b2 ~ normal([0,0], [d11**2,rho*d11*d22,d22**2]) subject = id;
estimate 'difference in slope' slope1-slope0;
estimate 'ratio of slopes' slope1/slope0;
estimate 'beta=1/alpha' 1/alpha;
predict p_0+(1-p_0)*(1/(1+lambda/m))**m out=ZIPNG;
run;

/*Average predicted probability of zeros*/
proc means data=ZIPNG;
var pred;
run;
```