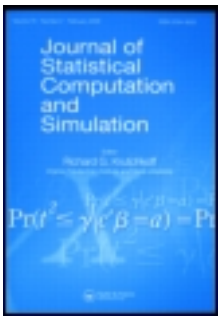

Journal of Statistical Computation and Simulation

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/gscs20

A joint model for hierarchical continuous and zero-inflated overdispersed count data

Wondwosen Kassahun[a], Thomas Neyens[b], Geert Molenberghs[bc], Christel Faes[b] & Geert Verbeke[bc]

[a] Department of Epidemiology and Biostatistics, Jimma University, Addis Ababa, Ethiopia;

[b] I-BioStat, CenStat, Universiteit Hasselt, B-3590 Diepenbeek, Belgium;

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

Published online: 28 Aug 2013.

To cite this article: Wondwosen Kassahun, Thomas Neyens, Geert Molenberghs, Christel Faes & Geert Verbeke , Journal of Statistical Computation and Simulation (2013): A joint model for hierarchical continuous and zero-inflated overdispersed count data, Journal of Statistical Computation and Simulation, DOI: 10.1080/00949655.2013.829058

To link to this article: http://dx.doi.org/10.1080/00949655.2013.829058

Journal of Statistical Computation and Simulation, 2013
http://dx.doi.org/10.1080/00949655.2013.829058

Taylor & Francis
Taylor & Francis Group

A joint model for hierarchical continuous and zero-inflated overdispersed count data

Wondwosen Kassahun[a], Thomas Neyens[b], Geert Molenberghs[b,c*], Christel Faes[b] and Geert Verbeke[b,c]

[a]*Department of Epidemiology and Biostatistics, Jimma University, Addis Ababa, Ethiopia;* [b]*I-BioStat, CenStat, Universiteit Hasselt, B-3590 Diepenbeek, Belgium;* [c]*I-BioStat, L-BioStat, Katholieke Universiteit Leuven, B-3000 Leuven, Belgium*

(*Received 16 March 2013; final version received 23 July 2013*)

Many applications in public health, medical and biomedical or other studies demand modelling of two or more longitudinal outcomes jointly to get better insight into their joint evolution. In this regard, a joint model for a longitudinal continuous and a count sequence, the latter possibly overdispersed and zero-inflated (ZI), will be specified that assembles aspects coming from each one of them into one single model. Further, a subject-specific random effect is included to account for the correlation in the continuous outcome. For the count outcome, clustering and overdispersion are accommodated through two distinct sets of random effects in a generalized linear model as proposed by Molenberghs et al. [A family of generalized linear models for repeated measures with normal and conjugate random effects. Stat Sci. 2010;25:325–347]; one is normally distributed, the other conjugate to the outcome distribution. The association among the two sequences is captured by correlating the normal random effects describing the continuous and count outcome sequences, respectively. An excessive number of zero counts is often accounted for by using a so-called ZI or hurdle model. ZI models combine either a Poisson or negative-binomial model with an atom at zero as a mixture, while the hurdle model separately handles the zero observations and the positive counts. This paper proposes a general joint modelling framework in which all these features can appear together. We illustrate the proposed method with a case study and examine it further with simulations.

Keywords: clustering; conjugate random effect; hurdle model; joint model; normal random effect; overdispersion; zero-inflation

1. Introduction

Many longitudinal studies involve collecting data on more than one outcome from a given subject repeatedly over time. These outcomes include, but are not limited to, continuous, count, and binary data. For example, in HIV studies, sero-positive patients are monitored until they develop AIDS or die, and their immune system is regularly measured using markers such as the CD4 lymphocyte count, the estimated viral load, or whether viral load is below detectable limits. In the case of the Jimma Infant Growth Study described in the next section is a continuous outcome, such as body weight, is measured repeatedly from each infant. At the same time, the health condition of a child was also assessed to see if the child has experienced a specific disease, like diarrhoea, whereby the number of days of illness, as a count outcome, were recorded so as to measure the magnitude of the disease burden.

*Corresponding author. Email: geert.molenberghs@uhasselt.be

Extensive literature is available on the analysis of single longitudinal outcomes. For a Gaussian longitudinal response, the linear mixed model is very popular.[1,2] Subject-level random effects that are of a Gaussian type are introduced in such a model to capture the within-subject correlation.

In the literature, several models are available to jointly model various longitudinal and possibly also time-to-event outcomes.[3] Horrocks and van den Heuvel [4] consider the problem of predicting the achievement of successful pregnancy, in a population of women undergoing treatment for infertility, based on longitudinal measurements of adhesiveness. For this purpose, they used a joint model consisting of a linear mixed-effects submodel for the longitudinal adhesion outcome and a generalized linear submodel for the primary binary end point. Molenberghs and Verbeke [5] discuss a number of techniques that jointly model continuous and discrete outcomes. The correlation between the sequences is then captured by allowing the normal random effects to be correlated.

Molenberghs et al. [6] proposed a unified modelling framework for the analysis of overdispersed and hierarchical non-Gaussian data by bringing together normal random effects to capture correlation over time and conjugate random effects and to capture overdispersion all within the generalized linear model framework. The resulting model is referred to as the *combined model*. For a count outcome, overdispersion is accommodated by using observation-specific Gamma random-effects, leading to the negative-binomial model. Booth et al. [7] extended the negative binomial log-linear model to the case of dependent counts, where dependence among the counts is handled by including also normal random effects. This particular instance of the combined model can be termed the Poisson-gamma-normal model.

Zero-inflated (ZI) count models provide a way of modelling 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 a negative-binomial data-generating process. Overdispersion and excess zeros for count data in a univariate setup are studied by, for example, Lambert [8] and Greene.[9] Multi-level ZI Poisson regression is considered by Lee et al.[10] The hurdle model, which is a two-part model, is also available to model excess zeros.[11] 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 a positive outcome. The two models do not constrain each other in any way. For ZI correlated data, the hurdle model has been studied by Min and Agresti.[12]

Our joint modelling approach for longitudinal continuous and count sequences, the latter possibly overdispersed and ZI, requires to assemble aspects coming from each one of them in one single model. These include the correlation from the continuous, as well as the correlation, overdispersion, and zero-inflation features from the count sequence. In spite of this apparent complexity, the model can be implemented in standard software, such as the Statistical Analysis System (SAS) procedure NLMIXED.

The rest of the paper is organized as follows. In Section 2, a motivating case is described with analysis reported in Section 5. In Section 3, a review is given on the single longitudinal continuous model and the so-called *combined model* and its extensions to account for an excess of zeros, followed by formulation of a joint modelling framework to also deal with correlation, overdispersion, and zero-inflation. Avenues for parameter estimation and ensuing inferences are explored in Section 4. Section 6 deals with a simulation study to investigate the importance of accounting for an excess of zeros and overdispersion. Some concluding statements are given in Section 7.

2. The Jimma Infant Growth Study

The study was conducted in Ethiopia in the administrative zones of Jimma, Keffa, and the Illubabor, and described in detail by Asefa and Tessema.[13] Children were examined for their first-year

Table 1. Jimma Infant Growth Study.

Age (months)	Mean weight (s.d.)	Mean days of illness (s.d.)
0	3.11(0.52)	0.01(0.19)
2	4.88(0.78)	0.91(4.24)
4	5.97(0.99)	1.28(4.62)
6	6.67(1.12)	1.56(4.87)
8	7.13(1.21)	2.14(5.93)
10	7.50(1.26)	2.63(6.66)
12	7.84(1.28)	2.67(6.95)

Mean and standard deviation of weight and days of illness at each of the seven follow-up times.

growth characteristics. At baseline, there were a total of 7969 infants who were visited every two months starting from birth until the age of one year. Risk factors, including socio-economic, maternal, and infant-rearing factors, were recorded in order to be able to study their relationship with the child's early survival.

For our purposes, two outcome variables will be considered, namely (1) body weight (kg), measured longitudinally from each infant and (2) number of days of diarrhoeal illness recorded at each visit to assess the diarrhoeal disease burden (Table 1). For the latter, of the nearly 49,000 total observations, only about 8000 observations are non-zero (i.e. roughly 85%), indicating that there is a dominance of zero counts. It is then useful to assess the connection between both by studying their association which can be addressed in a joint modelling context.

3. Methodology

In this section, existing models will be reviewed, extended, and combined. To describe these without ambiguity and to ensure that the links between them are clearly seen, we will make a few terminology conventions. Indeed, we have different data types (continuous and count), different types of random effects (normal and conjugate) that may or may not be combined into a single model, the possibility to accommodate zero inflation, and the possibility to analyse two or more outcomes jointly. All of these options create a large number of model variations. We can effectively describe them by introducing a symbol of the form (XYZ) for each outcome model, with X being the outcome type, Y the presence of absence of normal random effects, and Z the presence or absence of a conjugate random effect. For example, (NN-) stands for the linear mixed model, as the first 'N' is for the normal outcome and the second one for the normal random effect. In this case, there is no need for a conjugate random effect. For counts, we may see such models as (PNG) for the Poisson model with normal and gamma random effects. The generalized linear mixed model corresponding to it would be (PN-), whereas the negative binomial model is denoted as (P-G). When two outcomes are jointly analysed, we will write (NN-)(PNG). Should there be correction for zero inflation, then we might write ZI(NN-)(PNG) or H(NN-)(PNG) when hurdle model ideas are used. We observe that the notation obviates the need to use the terms 'joint' or 'combined', given that these aspects are now unambiguously clear from notation.

3.1. *A model for longitudinal continuous data*

For a longitudinal Gaussian outcome, the linear mixed model (NN-) provides a general and flexible modelling framework based on a random-effects approach.[2] Suppose that Y_{ij} are the jth continuous outcome measured for subject $i = 1, \ldots, N, j = 1, \ldots, n_i$. A linear mixed-effects model

can be represented as $Y_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i + \varepsilon_{ij}$. The $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ are p-dimensional and q-dimensional vectors of known covariate values, $\boldsymbol{\beta}$ is a p-dimensional vector of unknown fixed regression coefficients, $\boldsymbol{b}_i$ is a q-dimensional vector of the random effects, and $\boldsymbol{\varepsilon}_i$ an n_i-dimensional vector of residual variation. The subject-specific random effect $\boldsymbol{b}_i$ and the residual error $\boldsymbol{\varepsilon}_i$ are independent and assumed to follow normal distributions $\boldsymbol{b}_i \sim N(\boldsymbol{0}, D)$ and $\boldsymbol{\varepsilon}_i \sim N(\boldsymbol{0}, \Sigma_i)$, respectively.

3.2. *A combined model for count data*

Molenberghs et al. [6] proposed a unified modelling framework for the analysis of overdispersed and hierarchical non-Gaussian data by bringing together normal random effects and overdispersion random effects within the generalized linear model framework.

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

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

with η and ϕ being the natural and dispersion parameter, respectively, and $\psi(\cdot)$ and $c(\cdot)$ known functions specifying a particular distribution of the exponential family distribution.

For count data, we assume that $Y_{ij} \sim \text{Poi}(\lambda_{ij} = \theta_{ij}\kappa_{ij})$, with the conditional mean λ_{ij} modelled as $\theta_{ij}\kappa_{ij}$ with

$$\kappa_{ij} = \exp(\boldsymbol{x}'_{ij}\boldsymbol{\xi} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i), \tag{1}$$

$\boldsymbol{b}_i \sim N(\boldsymbol{0}, D)$, $\theta_{ij} \sim \text{Gamma}(\alpha, \beta)$, and $\boldsymbol{\xi}$ a p-dimensional vector of unknown fixed regression coefficients. Further, α and β are shape and scale parameters, respectively. These can be generalized as well, for example, when the gamma process parameters would be measurement occasion specific. The measurement-specific gamma random-effect θ_{ij} is used to accommodate residual overdispersion, while the subject-specific normal random effect $\boldsymbol{b}_i$ is used to model the correlation coming from the hierarchy in the data as well as some overdispersion. Using the notational conventions from the beginning of this section, this model is denoted as (PNG).

3.3. *ZI combined model*

Count data exhibiting an excess number of zeros beyond what can be allowed for by a standard probability distribution are very common. Such data are often modelled by using the so-called ZI models, which assume two independent data generation processes as sources of zeros. Assume that for observation i, one can account for the excess of zeros by introducing the probability π_i for the point mass (process 1) and the probability $1 - \pi_i$ for the count component (process 2; [14,15]). Process 1 generates only zeros, whereas process 2, $f_i(y_{ij} \mid \boldsymbol{b}_i, \boldsymbol{\xi}, \theta_{ij})$, generates counts from, say, a (PNG) model. In its most general form, the ZI(PNG) is then given as the following mixture:

$$Y_{ij} \sim \begin{cases} 0 & \text{with probability } \pi_{ij}, \\ f_i(y_{ij} \mid \boldsymbol{b}_i, \boldsymbol{\xi}, \theta_{ij}) & \text{with probability } 1 - \pi_{ij}, \end{cases} \tag{2}$$

leading to the probabilities $p(Y_{ij} = y_{ij} \mid \boldsymbol{b}_i, \boldsymbol{\xi}, \theta_{ij}, \pi_{ij})$ given by

$$p(Y_{ij} = y_{ij} \mid \boldsymbol{b}_i, \boldsymbol{\xi}, \theta_{ij}, \pi_{ij}) = \begin{cases} \pi_{ij} + (1 - \pi_{ij})f_i(0 \mid \boldsymbol{b}_i, \boldsymbol{\xi}, \theta_{ij}) & \text{if } y_{ij} = 0, \\ (1 - \pi_{ij})f_i(y_{ij} \mid \boldsymbol{b}_i, \boldsymbol{\xi}, \theta_{ij}) & \text{if } y_{ij} > 0. \end{cases} \tag{3}$$

The Bernoulli model is used to model the zero-inflation component $\pi_{ij} = \pi(\boldsymbol{x}'_{2ij}\boldsymbol{\gamma})$ by using commonly used links such as the logit or probit. The model may contain known regressors $\boldsymbol{x}_{2ij}$,

a vector of zero-inflation coefficients $\boldsymbol{\gamma}$ to be estimated. Note that the $\boldsymbol{x}_{ij}$ in Section 3.2 are now replaced by $\boldsymbol{x}_{1ij}$ 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 different regressors for the two parts can be used.

The conditional mean and variance of the ZI(PNG) are as follows:

$$\mathrm{E}(Y_{ij} \mid \boldsymbol{b}_i, \boldsymbol{\xi}, \theta_{ij}) = \theta_{ij}\kappa_{ij}(1-\pi_{ij}), \tag{4}$$

$$\mathrm{Var}(Y_{ij} \mid \boldsymbol{b}_i, \boldsymbol{\xi}, \theta_{ij}) = \theta_{ij}\kappa_{ij}(1-\pi_{ij})\left[1+\theta_{ij}\kappa_{ij}\left(\pi_{ij}+\frac{1}{\alpha}\right)\right]. \tag{5}$$

It can be seen that the conditional variance is inflated as a result of the overdispersion in the data (parameter α) and as a result of zero-inflation (parameter π).

3.4. *Hurdle combined model*

The hurdle model, as an alternative approach to account for an excess of zeros, is a two-part model with a truncated count distribution for the positive counts and a hurdle component for the zero counts.[11] Suppose that occurrence of zeros is governed by a probability π_i, and non-zero counts and follows a truncated-at-zero probability mass function, such as a truncated (PNG). The H(PNG) leads to a probability mass function of the form as in the following:

$$p(Y_{ij}=y_{ij} \mid \boldsymbol{b}_i, \boldsymbol{\xi}, \theta_{ij}, \pi_{ij}) = \begin{cases} \pi_{ij} & \text{if } y_{ij}=0, \\ (1-\pi_{ij})\dfrac{f_i(y_{ij} \mid \boldsymbol{b}_i, \boldsymbol{\xi}, \theta_{ij})}{1-f_i(0 \mid \boldsymbol{b}_i, \boldsymbol{\xi}, \theta_{ij})} & \text{if } y_{ij}>0. \end{cases} \tag{6}$$

where the π_{ij} are parameterized similarly to what was presented in Section 3.3.

3.5. *A joint combined model for continuous and ZI count data*

Let Y_{ij} denote a longitudinal continuous outcome, and Z_{ik} an overdispersed count outcome, potentially with excess zeros and overdispersion with densities $f_{1i}(y_{ij})$ and $f_{2i}(z_{ik})$, respectively ($i=1,\ldots,N$, $j=1,\ldots,n_{1i}$, and $k=1,\ldots,n_{2i}$). Note that the number of measurements per sequence can but does not have to be the same.

Formulation of a joint model will be based on a random-effects approach. Precisely, each sequence will incorporate normal random effects, which are potentially correlated between them. This means that the correlation between both sequences for the same subject is induced by the shared or correlated normal random effects. At the same time, the normal random effects induce correlation within the sequences as well. We still allow for additional overdispersion by potentially integrating gamma random effects into the count sequence. This leads to the base model (NN-)(PNG).

The normal random effects, $\boldsymbol{b}_{1i}$ and $\boldsymbol{b}_{2i}$, are incorporated into the continuous and count sequences, respectively. They follow the joint distribution

$$\boldsymbol{b}_i = (\boldsymbol{b}_{1i}, \boldsymbol{b}_{2i})' \sim N\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix} D_{11} & D_{12} \\ D_{12}' & D_{22}\end{bmatrix}\right).$$

In this general formulation, the two sets of random effects are different but correlated. The sequences would be independent only if $D_{12}=0$. It is possible to let some components be shared between $\boldsymbol{b}_{1i}$ and $\boldsymbol{b}_{2i}$. If all are common, this is a pure shared random-effects model. The modeller can assess and then choose whether a shared random effects or a correlated random effects approach is more plausible, or even a hybrid of both.

Further allowing for zero inflation, the resulting model becomes ZI(NN-) (PNG) or H(NN-) (PNG), with joint distribution:

$$f_i(y_{ij}, z_{ik} \mid \boldsymbol{b_{1i}}, \boldsymbol{b_{2i}}, \boldsymbol{\beta}, \boldsymbol{\xi}, \theta_{ik}, \pi_{ik}) = f_{1i}(y_{ij} \mid \boldsymbol{b_{1i}}, \boldsymbol{\beta}) \times f_{2i}(z_{ik} \mid \boldsymbol{b}_{2i}, \boldsymbol{\xi}, \theta_{ik}, \pi_{ik}). \tag{7}$$

For the continuous outcome, $f_{1i}(y_{ij} \mid \boldsymbol{b_{1i}}, \boldsymbol{\beta})$ is the linear mixed model (NN-), as discussed in Section 3.1, and for the count sequence, $f_{2i}(z_{ik} \mid \boldsymbol{b}_{2i}, \boldsymbol{\xi}, \theta_{ik}, \pi_{ik})$ is the ZI(PNG), as presented in Section 3.3. Similarly, for the hurdle version, we bring together the (NN-) and the H(PNG) into the H(NN-)(PNG), where $f_{2i}(z_{ik} \mid \boldsymbol{b}_{2i}, \boldsymbol{\xi}, \theta_{ik}, \pi_{ik})$ is now defined as in Section 3.4.

4. Estimation

Let us consider the count component. We will make use of the partial marginalization for parameter estimation, as presented in [6]. By this we refer to integrating the likelihood over the gamma random effects only, leaving the normal random effects untouched. The corresponding conditional probability for the (PNG) is as follows:

$$\begin{aligned} f(y_{ij} \mid \boldsymbol{b}_i, \boldsymbol{\xi}, \theta_{ij}) &= \int f(y_{ij} \mid \boldsymbol{b}_i, \boldsymbol{\xi}, \theta_{ij}) f(\theta_{ij} \mid \alpha_j, \beta_j)\, \mathrm{d}\theta_{ij} \\ &= \binom{\alpha_j + y_{ij} - 1}{\alpha_j - 1} \cdot \left(\frac{\beta_j}{1 + \kappa_{ij}\beta_j}\right)^{y_{ij}} \cdot \left(\frac{1}{1 + \kappa_{ij}\beta_j}\right)^{\alpha_j} \kappa_{ij}^{y_{ij}}, \end{aligned}$$

where κ_{ij} is as in Equation (1). For the ZI(PNG) case:

$$\begin{aligned} f(y_{ij} \mid \boldsymbol{b}_i, \boldsymbol{\xi}, \theta_{ij}, \pi_{ij}) &= I(y_{ij} = 0)\pi_{ij} \\ &\quad + (1 - \pi_{ij}) \binom{\alpha_j + y_{ij} - 1}{\alpha_j - 1} \cdot \left(\frac{\beta_j}{1 + \kappa_{ij}\beta_j}\right)^{y_{ij}} \cdot \left(\frac{1}{1 + \kappa_{ij}\beta_j}\right)^{\alpha_j} \kappa_{ij}^{y_{ij}}, \end{aligned}$$

with $\pi_{ij} = \pi(\boldsymbol{x}'_{2ij}\boldsymbol{\gamma})$. Note that, with this approach, we assume that the gamma random effects are independent within a subject. This is fine, given that the correlation is induced by the normal random effects.

Applying the above principle to the ZI(NN-)(PNG) leads to

$$\begin{aligned} f_i(y_{ij}, z_{ik} \mid \boldsymbol{b}_{1i}, \boldsymbol{b}_{2i}, \boldsymbol{\beta}, \boldsymbol{\xi}, \theta_{ik}, \pi_{ik}) &= \frac{1}{(2\pi)^{n_i/2}|\Sigma_i|^{1/2}} \mathrm{e}^{-(1/2)(\boldsymbol{y}_i - X_i\boldsymbol{\beta} - Z_i\boldsymbol{b}_{1i})'\Sigma_i^{-1}(\boldsymbol{y}_i - X_i\boldsymbol{\beta} - Z_i\boldsymbol{b}_{1i})} \\ &\quad \times \prod_k f(z_{ik} \mid \boldsymbol{b}_{2i}, \boldsymbol{\xi}, \theta_{ik}, \pi_{ik}). \end{aligned}$$

Similarly, the H(PNG), say $f^*(y_{ij})$, as discussed in Section 3.4, can be spelled out as

$$\begin{aligned} &f^*(y_{ij} \mid \boldsymbol{b}_i, \boldsymbol{\xi}, \theta_{ij}, \pi_{ij}) \\ &\quad = I(y_{ij} = 0)\pi_{ij} + (1 - \pi_{ij}) \binom{\alpha_j + y_{ij} - 1}{\alpha_j - 1} \cdot \left(\frac{\beta_j}{1 + \kappa_{ij}\beta_j}\right)^{y_{ij}} \cdot \left(\frac{1}{1 + \kappa_{ij}\beta_j}\right)^{\alpha_j} \kappa_{ij}^{y_{ij}} \\ &\qquad \times \frac{1}{1 - (1/(1 + \kappa_{ij}\beta_j))^{\alpha_j}}. \end{aligned}$$

Consequently, the partially marginalized form for the H(NN-)(PNG) is

$$f_i(y_{ij}, z_{ik} \mid \boldsymbol{b}_{1i}, \boldsymbol{b}_{2i}, \boldsymbol{\beta}, \boldsymbol{\xi}, \theta_{ik}, \pi_{ik}) = \frac{1}{(2\pi)^{n_i/2}|\Sigma_i|^{1/2}} \mathrm{e}^{-(1/2)(\boldsymbol{y}_i - X_i\boldsymbol{\beta} - Z_i\boldsymbol{b}_{1i})' \Sigma_i^{-1} (\boldsymbol{y}_i - X_i\boldsymbol{\beta} - Z_i\boldsymbol{b}_{1i})} \times \prod_k f^*(z_{ik} \mid \boldsymbol{b}_{2i}, \boldsymbol{\xi}, \theta_{ik}, \pi_{ik}).$$

For all of these, it is straightforward to obtain the fully marginalized probability and to optimize the resulting likelihood, by numerically integrating over the normal random effects. This can be done using a tool, such as the SAS procedure NLMIXED, that allows for normal random effects in arbitrary, user-specified models. While the SAS procedure NLMIXED is equipped with default starting values, it is advisable to provide user-defined starting values instead. These can be obtained, for example, from models without random effects, with some trial and error. It is equally wise to ensure that both the outcomes values as well as the covariates have magnitudes that are neither very large nor extremely small, because this may jeopardize stability of the iterative process. Also, it is useful, for example, to first fit the individual models and use the output as starting values for the joint model. Against this background, our data analysis proceeded without difficulty. Example NLMIXED code is provided in Appendix 1, for both the data analysis and the simulation study.

5. Analysis of the Jimma Infant Growth Study

We analyse the Jimma Infant data as introduced in Section 2, where body weight as well as number of days of diarrhoeal illnesses were measured repeatedly for each infant. The two outcomes will be modelled jointly to capture association between them. Denote by Y_{ij} and Z_{ik} weight and number of days of illness measurements for the ith infant at the jth and kth visit. We formulate a ZI(NN-)(PNG) or a H(NN-)(PNG) model for these data. The means are μ_{ij} and κ_{ik}, respectively. We model these as $\mu_{ij} = \beta_0 + b_{1i} + \beta_1 A_{ij} + \beta_2 {A_{ij}}^2$, and κ_{ik} as $\ln(\kappa_{ik}) = \xi_0 + b_{2i} + \xi_1 A_{ik}$. The quadratic effect of age, β_2, is included to improve the model's fit. A graphical inspection of age versus average weight revealed a nonlinear trend and hinted on a quadratic one. To account for excess zeros, the zero-inflation probability π_{ik} is written as $\mathrm{logit}(\pi_{ik}) = \gamma_0 + \gamma_1 A_{ik}$. Here, A_{ij} is the age of the ith infant at the jth visit. Further, b_{1i} and b_{2i} represent subject-specific intercepts, assumed normally distributed and possibly correlated with mean and variance–covariance matrix given by

$$(b_{1i}, b_{2i})' \sim N\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix} d_1 & \rho\sqrt{d}_1\sqrt{d}_2) \\ \rho\sqrt{d}_1\sqrt{d}_2 & d_2 \end{pmatrix}\right].$$

We examine the zero-inflation as well as the overdispersion aspect. The first issue can be addressed using a special type of the hurdle model, also known as the zero-altered model.[12] These authors consider the testing problem for zero-inflation. This model requires the same covariates as well as the same distributional forms in the two parts. Explicitly, we assume,

$$\ln[-\ln(1 - \pi_{ik})] = a_1 + a_2(\xi_0 + \xi_1 A_{ik}) + b_{2i},$$
$$\ln(\kappa_{ik}) = \xi_0 + \xi_1 A_{ik} + b_{2i}.$$

By setting, $a_2 = 1$, and testing whether $a_1 = 0$, one can test for zero-inflation. If $a_1 < 0$, then the data are ZI; If $a_1 > 0$, the data are zero-deflated. Fitting the zero-altered combined model showed strong evidence of zero-inflation ($a_1 = -2.0689$, likelihood ratio test statistic $= 3343$, on one degree of freedom).

Table 2. Jimma Infant Growth Study.

Effect	Parameter	(NN-) & (PNG) Estimate (s.e.)	(NN-)(PNG) Estimate (s.e.)
Continuous process (weight)			
Intercept	β_0	3.2767(0.0112)	3.2768(0.0112)
Age	β_1	0.7680(0.0026)	0.7680(0.0026)
Age × Age	β_2	−0.0335(0.0002)	−0.0335(0.0002)
Std. dev. error	σ	0.6298(0.0022)	0.6298(0.0022)
Std. dev. random effect	$\sqrt{d_1}$	0.8298(0.0072)	0.8299(0.0072)
Count process (days of illness)			
Intercept	ξ_0	−1.1567(0.0557)	−1.2094(0.0551)
Age	ξ_1	0.2246(0.0072)	0.2279(0.0072)
Std. dev. random effect	$\sqrt{d_2}$	0.3654(0.0507)	0.4611(0.0435)
Negative-binomial parameter	α	17.9605(0.2914)	17.6345(0.2858)
Common parameter			
Corr. random effect	ρ	–	−0.6282(0.0565)
−2log-likelihood		172,241	172,054

H(NN-), (PNG), and (NN-)(PNG) models.

We also fitted the H(NN-)(PN-) and compared it with the H(NN-)(PNG). The difference in deviance is $94,181 - 86,379$, which evidently is extremely significant ($p < .0001$). This strongly underscores the presence of the overdispersion parameter α.

Parameter estimates for the (NN-), (PNG), and (NN-)(PNG) models are presented in Table 2. Technically, the separate models were fitted for the two outcomes together anyway, but assuming that $\rho = 0$, which is entirely equivalent to fitting the models separately. Clearly, body weight and number of days of diarrhoeal illness show a strong inverse relationship as evidenced by the correlation of the random effects in the (NN-)(PNG). In addition, likelihood comparison shows a convincing improvement in model fit, when random effects are allowed to correlate. Comparing the separate and joint models, while parameter estimates for the continuous outcome remain the same, small changes are observed in the count part. All parameters are statistically significant in all models considered.

Table 3 gives the hurdle counterparts to the models in Table 2, i.e. (NN-), H(PNG), and H(NN-)(PNG). Evidently, the linear mixed model (NN-) is left unaltered because the hurdle aspect applies to the count process only. Similarly, Table 4 shows the ZI counterparts, (NN-), ZI(PNG), and ZI(NN-)(PNG). We deduce that the fit is improved quite a bit further, implying that the excess zeros need to be accommodated in the model. The aforementioned inverse relationship remains. While accounting for the excess zeros does not bring changes in the parameters for the continuous process, this is not the case for the counts, where the estimates change with their corresponding standard errors getting relatively smaller. When the excess zeros are accounted for, the negative binomial parameter α gets much smaller in Tables 3 and 4, as compared to the corresponding value in Table 2. This underscores, once again, the connection between the zero-inflation and overdispersion phenomena. Indeed, when ZI is unaccounted for the overdispersion aspect of the models captures a certain amount of this effect.

Turning to Tables 2–4 with an eye on the overdispersion parameter α, we see that it drops drastically when comparing Table 2 with Tables 3 and 4. Because in our parameterization, larger values for α imply more pronounced overdispersion effects, it is clear that accounting for excess zeros explains a good amount of apparent overdispersion. The amount explained is relatively invariant to whether the two processes are estimated jointly or rather separately. The reduction also holds regardless of whether either a ZI or a hurdle correction is made. Finally, even after correction

Table 3. Jimma Infant Growth Study.

Effect	Parameter	(NN-) & (PNG) Estimate (s.e.)	H(NN-)(PNG) Estimate (s.e.)
Continuous process (weight)			
Intercept	β_0	3.2767(0.0112)	3.2767(0.0112)
Age	β_1	0.7680(0.0026)	0.7680(0.0026)
Age × Age	β_2	−0.0335(0.0002)	−0.0335(0.0002)
Std. dev. error	σ	0.6298(0.0022)	0.6298(0.0022)
Std. dev. random effect	$\sqrt{d_1}$	0.8298(0.0072)	0.8299(0.0072)
Count process (days of illness)			
Intercept	ξ_0	2.0437(0.0251)	2.0225(0.0251)
Age	ξ_1	0.0185(0.0028)	0.0199(0.0028)
Std. dev. random effect	$\sqrt{d_2}$	0.4374(0.0118)	0.4392(0.0118)
Negative-binomial parameter	α	0.3271(0.0104)	0.3259(0.0103)
Inflation intercept	γ_0	2.8383(0.0290)	2.8383(0.0290)
Inflation Age	γ_1	−0.1687(0.0035)	−0.1687(0.0035)
Common parameter			
Corr. random effect	ρ	–	−0.2255(0.0246)
−2log-likelihood		165,475	165,395

(NN-), H(PNG), and H(NN-)(PNG) models.

Table 4. Jimma Infant Growth Study.

Effect	Parameter	(NN-) & ZI(PNG) Estimate (s.e.)	ZI(NN-)(PNG) Estimate (s.e.)
Continuous process (weight)			
Intercept	β_0	3.2767(0.0112)	3.2767(0.0112)
Age	β_1	0.7680(0.0026)	0.7680(0.0026)
Age × Age	β_2	−0.0335(0.0002)	−0.0335(0.0002)
Std. dev. error	σ	0.6298(0.0022)	0.6298(0.0022)
Std. dev. random effect	$\sqrt{d_1}$	0.8298(0.0072)	0.8299(0.0072)
Count process (days of illness)			
Intercept	ξ_0	2.0270(0.0259)	2.0020(0.0260)
Age	ξ_1	0.0190(0.0029)	0.0205(0.0029)
Std. dev. random effect	$\sqrt{d_2}$	0.4464(0.0123)	0.4500(0.0123)
Negative-binomial parameter	α	0.3259(0.0104)	0.3261(0.0104)
Inflation intercept	γ_0	2.8071(0.0292)	2.8051(0.0292)
Inflation age	γ_1	−0.1686(0.0035)	−0.1685(0.0035)
Common parameter			
Corr. random effect	ρ	–	−0.2409(0.0245)
−2log-likelihood		165,461	165,369

(NN-), ZI(PNG), and ZI(NN-)(PNG) models.

for excess zeros, there convincingly remains an amount of overdispersion. This underscores that both overdispersion as well as excess zeros need to be accounted for.

From Tables 3 and 4, we further observe that the H(NN-)(PNG) and the ZI(NN-)(PNG) are very similar, not only in terms of parameter estimates but also as far as resulting inferences go. In count data modelling, the choice among ZI and hurdle models should be based not only on model fit but also on assumptions about the underlying data generation process.[16] If zeros are expected to come from both the point mass and the count component, then ZI models may be preferable. In addition, in the presence of strong evidence of zero-inflation, zero-inflation models may provide better fit. On the other hand, Min and Agresti [12] discuss why the hurdle model

is, in general, preferred to the ZI model: it allows to test for zero-inflation, it works also in the zero-deflation setting and its two parts are separate.

6. Simulation study

A simulation study is conducted to assess the impact of not appropriately accounting for the excess zero counts as well as mis-specification of the overdispersion in joint modelling of hierarchical continuous and count outcome. We choose to conduct this study following three different settings.

6.1. *Simulation settings*

Data are generated in the spirit of the design and outcomes of the data in Section 2, which consist of body weight and counts of the number of days of diarrhoeal disease illnesses among infants measured repeatedly over time. Age in months is considered as the time variable.

A random sample of 250 data sets are generated under three scenarios. S_1: from a ZI(NN-)(PNG); S_2 from a ZI(NN-)(PN-); S_3 from a (NN-)(PNG). Model fitting is based on these three models, supplemented with others: ZI(NN-)(PNG), H(NN-)(PNG), or (NN-)(PNG); also, the versions without overdispersion are considered: ZI(NN-)(PN-), H(NN-)(PN-), or (NN-)(PN-).

We consider 200 subjects with 10 measurements per subject. The continuous response Y_{ij} is modelled as $Y_{ij} = \beta_0 + \beta_1 A_{ij} + \beta_2 {A_{ij}}^2 + b_{1i} + \varepsilon_{ij}$. The subject-specific random intercept b_{1i} and the residual error $\boldsymbol{\varepsilon}_i$ are assumed independent, and generated from normal distribution with mean 0 and standard deviations 2 and 0.6, respectively. The count outcome Z_{ik} is modelled using predictor function $\kappa_{ik} = \exp\{\xi_0 + b_{2i} + \xi_1 A_{ik}\}$. Whenever we want to place overdispersion into the model simulated from the outcome is generated directly from a negative-binomial process with $Z_{ik} \sim \text{NB}(\psi_{ik}, \theta)$, where $\theta = 1$ and $\psi_{ik} = (1 + \kappa_{ik}/\theta)^{-1}$. As before, A_{ij} represents the age at which the jth measurement is taken for the ith subject. Practically, age is generated from the empirical distribution observed in the Jimma Infant Growth Study. The random intercept b_{2i} follows a mean-zero normal with variance 1.5. When zero-inflation is present, this is added by defining the final response vector $\boldsymbol{Z}_i^*$ with components ${Z_{ik}}^* = (1 - u_{ik})Z_{ik}$, where the u_{ik} are Bernoulli random variables with parameters π_{ik} and $\text{logit}(\pi_{ik}) = \gamma_0 + \gamma_1 A_{ik}$. To correlate the two processes, the random intercepts b_{1i} and b_{2i} are allowed to correlate with one another, with $\rho = -0.5$. When generating data, the true parameter values were $\boldsymbol{\beta} = (3.3, 0.77, -0.03)^{\text{T}}$, $\boldsymbol{\xi} = (2, 0.02)^{\text{T}}$, and $\boldsymbol{\gamma} = (2, -0.2)^{\text{T}}$.

6.2. *Simulation results*

The results under S_1 are summarized in Tables 5 and 6. Clearly, the ZI(NN-)(PNG) and the H(NN-)(PNG) result in estimates very close to the true values. However, as can be seen from the (NN-)(PNG), omitting zero-inflation highly affects the estimates in the count component, with a non-negligible amount of bias loaded on the correlation parameter. Further, when both zero-inflation and overdispersion are mis-specified, by fiting the (NN-)(PN-), as given in Table 6, a similar phenomenon is evident, where now the random-effects variance tries to recover from mis-specifying the overdispersion.

Under scenario S_2, the results of which are presented in Table 7, the impact of omitting the extra zeros is still evident, though the overdispersion parameter α in the (NN-)(PNG) seems to help recover from mis-specification. Further, we also note that correlation is overestimated as a result of the mis-specification. These results, once more, underscore the necessity of models appropriately accounting for the excessive zeros.

Table 5. Simulation study under scenario S_1.

Effect	Parameter	True	ZI(NN-)(PNG) Mean (RB)	H(NN-)(PNG) Mean (RB)	(NN-)(PNG) Mean (RB)
Continuous process					
Intercept	β_0	3.3	3.2967(−0.0010)	3.2965(−0.0011)	3.2965(−0.0011)
Age	β_1	0.77	0.7716(0.0021)	0.7716(0.0021)	0.7716(0.0021)
Age × Age	β_2	−0.03	−0.0301(0.0033)	−0.0301(0.0033)	−0.0301(0.0033)
Std. dev. error	σ	0.6	0.5995(−0.0008)	0.5995(−0.0008)	0.5995(−0.0008)
Std. dev. random effect	$\sqrt{d_1}$	2	1.9875(−0.0063)	1.9873(−0.0063)	1.9875(−0.0063)
Count process					
Intercept	ξ_0	2	1.9771(−0.0115)	2.1452(0.0726)	−0.1694(−1.0847)
Age	ξ_1	0.02	0.0229(0.1450)	0.0220(0.1000)	0.2177(9.8850)
Std. dev. random effect	$\sqrt{d_2}$	1.5	1.4779(−0.0147)	1.3409(−0.1061)	1.3107(−0.1262)
Negative-binomial parameter	α	1	0.9922(−0.0078)	1.0509(0.0509)	11.8950(10.8950)
Inflation intercept	γ_0	2	1.9956(−0.0022)	2.2092(0.1046)	–
Inflation age	γ_1	−0.2	−0.1982(−0.0090)	−0.1876(−0.0620)	–
Common parameter					
Corr. random effect	ρ	−0.5	−0.5028(0.0056)	−0.5005(0.0010)	−0.5629(0.1258)
Frequency of convergence			250	250	250

Mean and relative bias (RB) of the parameter estimates in the ZI(NN-)(PNG), H(NN-)(PNG), and (NN-)(PNG).

Table 6. Simulation study under scenario S_1.

Effect	Parameter	True	ZI(NN-)(PN-) Mean (RB)	H(NN-)(PN-) Mean (RB)	(NN-)(PN-) Mean (RB)
Continuous process					
Intercept	β_0	3.3	3.2843(−0.0048)	3.2920(−0.0024)	3.2959(−0.0012)
Age	β_1	0.77	0.7716(0.0021)	0.7717(0.0022)	0.7716(0.0021)
Age × Age	β_2	−0.03	−0.0302(0.0067)	−0.0301(0.0033)	−0.0301(0.0033)
Std. dev. error	σ	0.6	0.5994(−0.0010)	0.5995(−0.0008)	0.5995(−0.0008)
Std. dev. random effect	$\sqrt{d_1}$	2	1.9882(−0.0059)	1.9877(−0.0062)	1.9875(−0.0063)
Count process					
Intercept	ξ_0	2	1.1335(−0.4333)	2.0159(0.0080)	−0.4531(-1.2266)
Age	ξ_1	0.02	0.0254(0.2700)	0.0258(0.2900)	0.1617(7.0850)
Std. dev. random effect	$\sqrt{d_2}$	1.5	1.5664(0.0443)	1.4261(−0.0493)	1.7943(0.1962)
Negative-binomial parameter	α	1	–	–	–
Inflation intercept	γ_0	2	1.9724(−0.0138)	2.2087(0.1044)	–
Inflation age	γ_1	−0.2	−0.1946(−0.0270)	−0.1876(−0.0620)	–
Common parameter					
Corr. random effect	ρ	−0.5	−0.4461(0.1078)	−0.4361(0.1278)	−0.4058(−0.1884)
Frequency of convergence			233	250	250

Mean and relative bias (RB) of the parameter estimates in the ZI(NN-)(PN-), H(NN-)(PN-), and (NN-)(PN-).

Scenario 1 leads to about 75% of zeros with a similar fraction (72%) in Scenario 2. Scenario 3 is qualitatively different, with roughly 18% of zeros. Comparing mean and standard deviation shows that all three are overdispersed. Under Scenarios 1 and 2, this stems to a large part from extra zeros, whereas in Scenario 3 this is 'pure' overdispersion. When data are overdispersed, but not subject to considerable zero-inflation as in S_3, fitting models allowing for extra zeros is less important. As given in Table 8, the (NN-)(PNG), which is the true model, performs well. In addition, we observe that the (NN-)(PN-) model is also doing well, but this probably may not be the case when data are subject to much higher levels of overdispersion than those considered here.

Table 7. Simulation study under scenario S_2.

Effect	Parameter	True	ZI(NN-)(PN-) Mean (RB)	H(NN-)(PN-) Mean (RB)	(NN-)(PNG) Mean (RB)
Continuous process					
Intercept	β_0	3.3	3.3254(0.0077)	3.2939(−0.0019)	3.2958(−0.0013)
Age	β_1	0.77	0.7697(−0.0004)	0.7711(0.0014)	0.7710(0.0013)
Age × Age	β_2	−0.03	−0.0300(0.0000)	−0.0301(0.0033)	−0.0301(0.0033)
Std. dev. error	σ	0.6	0.5994(−0.0010)	0.5996(−0.0007)	0.5996(−0.0007)
Std. dev. random effect	$\sqrt{d_1}$	2	1.9850(−0.0075)	1.9856(−0.0072)	1.9855(−0.0073)
Count process					
Intercept	ξ_0	2	1.2304(−0.3848)	2.0508(0.0254)	0.0834(−0.9583)
Age	ξ_1	0.02	0.0204(0.0200)	0.0203(0.0150)	0.1948(8.7400)
Std. dev. random effect	$\sqrt{d_2}$	1.5	1.5021(0.0014)	1.4244(−0.0504)	1.2189(−0.1874)
Negative-binomial parameter	α	0	–	–	9.9406
Inflation intercept	γ_0	2	1.8843(−0.0579)	2.0889(0.0445)	–
Inflation age	γ_1	−0.2	−0.2038(−0.0190)	−0.1955(−0.0225)	–
Common parameter					
Corr. random effect	ρ	−0.5	−0.4954(−0.0092)	−0.4890(0.0220)	−0.6126(−0.2252)
Frequency of convergence			241	250	250

Mean and relative bias (RB) of the parameter estimates in the ZI(NN-)(PN-), H(NN-)(PN-), and (NN-)(PNG).

Table 8. Simulation study under scenario S_3.

Effect	Parameter	True	H(NN-)(PNG) Mean (RB)	(NN-)(PNG) Mean (RB)	(NN-)(PN-) Mean (RB)
Continuous process					
Intercept	β_0	3.3	3.2938(−0.0019)	3.2946(−0.0016)	3.2948(−0.0016)
Age	β_1	0.77	0.7721(0.0027)	0.7721(0.0027)	0.7721(0.0027)
Age × Age	β_2	−0.03	−0.0302(0.0067)	−0.0302(0.0067)	−0.0302(0.0067)
Std. dev. error	σ	0.6	0.5993(−0.0012)	0.5993(−0.0012)	0.5992(−0.0013)
Std. dev. random effect	$\sqrt{d_1}$	2	1.9872(−0.0064)	1.9878(−0.0061)	1.9875(−0.0063)
Count process					
Intercept	ξ_0	2	2.0627(0.0314)	1.9954(−0.0023)	1.9537(−0.0232)
Age	ξ_1	0.02	0.0189(−0.0550)	0.0194(−0.0300)	0.0175(−0.1250)
Std. dev. random effect	$\sqrt{d_2}$	1.5	1.4143(−0.0571)	1.4852(−0.0099)	1.5223(0.0149)
Negative-binomial parameter	α	1	1.0082(0.0082)	1.0012(0.0012)	–
Inflation intercept	γ_0	0	−1.4481	–	–
Inflation age	γ_1	0	−0.0157	–	–
Common parameter					
Corr. random effect	ρ	−0.5	−0.4963(−0.0074)	−0.4986(−0.0028)	−0.4864(−0.0272)
Frequency of convergence			250	250	249

Mean and relative bias (RB) of the parameter estimates in the H(NN-)(PNG), (NN-)(PNG), and (NN-)(PN-).

In addition, across our simulation study, we learned that the ZI models are relatively harder to fit when compared to the hurdle models, where convergence of models is never an issue, with convergence guaranteed for the latter. Further, in scenarios S_1 and S_2, though data are generated from the ZI(NN-)(PNG), the H(NN-)(PNG) is also performing very well.

7. Concluding remarks

In this paper, we have described a joint modelling strategy for a hierarchical continuous and count outcome, where the latter is subject to zero-inflation as well as overdispersion. Our work

builds upon that of Molenberghs et al. [17] and Molenberghs et al. [6] who combined normal and gamma random effects in the generalized linear model family to induce association and adjust for overdispersion. We show that these ideas can be extended to a joint modelling framework for a continuous and count sequence simultaneously. This allows for an improvement in model fit, and for enhanced statistical inferences. However, any failure to appropriately account for such features may result in a substantial impact on the parameter and precision estimates. When zero-inflation is omitted from the model, the overdispersion component will try to recover from this mis-specification, though both are eventually needed.

Fitting a ZI(NN-)(PNG) model, even when correctly specified, is relatively more complex than a H(NN-)(PNG). The latter has several additional advantages, in particular the possibility to test for zero inflation. In the real data analysis, there were no model convergence issues, which is reassuring. Of course, as we learned from analysing these data, both overdispersion as well as additional zero inflation were convincingly present. Furthermore, the set of data was very large. These are comfortable conditions to reach convergence. In the simulation study, the hurdle model was at an advantage when it came to model fit. Practically, readers may consider both approaches, hurdle and ZI, and perhaps use the relevant parameters from the hurdle model as starting values for the ZI fit.

In terms of estimation, we have focused on maximum likelihood estimation. This can be done by integrating over the random effects, using a combination of analytical and numerical techniques. Precisely, the likelihood was integrated analytically over the conjugate (gamma) random effect, using techniques outlined in Molenberghs et al.[6] The so-resulting likelihood, still conditional on the normal random effect, is integrated numerically over the said random effect, using the SAS procedure NLMIXED.

In conclusion, we note that our approach corrects for overdispersion and/or allows for joint modelling. In our example, both phenomena were present, although overdispersion results in a larger deviance reduction than joint modelling. One lesson to be drawn from this is that the user should carefully assess whether one or the other correction, both of them, or perhaps none of the two is necessary. Note also that joint modelling may be of interest in its own right. For example, one may be interested in measuring the strength of the association between both processes (estimating one or more correlation parameters) or in assessing its significance. Also, it is possible to derive prediction equations for an outcome or set of outcomes in one sequence, based on the outcomes in the other sequence and/or earlier measurements of the same sequence.

Acknowledgements

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

[1] Laird NM, Ware JH. Random effects models for longitudinal data. Biometrics. 1982;38:963–974.
[2] Verbeke G, Molenberghs G. Linear mixed models for longitudinal data. New York: Springer; 2000.
[3] Tsiatis A, Davidian M. Joint modeling of longitudinal and time-to-event data: an overview. Statist Sinica. 2004;14:809–834.
[4] Horrocks J, van den Heuvel MJ. Prediction of pregnancy: a joint model for longitudinal and binary data. Bayesian Anal. 2009;4:523–538.
[5] Molenberghs G, Verbeke G. Models for discrete longitudinal data. New York: Springer; 2005.
[6] Molenberghs G, Verbeke G, Demétrio C, Vieira A. A family of generalized linear models for repeated measures with normal and conjugate random effects. Stat Sci. 2010;25:325–347.
[7] Booth J, Casella G, Friedl H, Hobert J. Negative binomial loglinear mixed models. Stat Model. 2003;3:179–191.

[8] Lambert D. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics. 1992;34:1–14.
[9] Greene W. 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; 1994.
[10] Lee AH, Wang K, Scott J, Yau KKW, McLachlan GJ. Multi-level zero-inflated Poisson regression modelling of correlated count data with excess zeros. Stat Methods Med Res. 2006;15:47–61.
[11] Mullahy J. Specification and testing of some modified count data models. J Econom. 1986;33:341–365.
[12] Min Y, Agresti A. Random effect models for repeated measures of zero-inflated count data. Stat Model. 2005;5:1–19.
[13] Asefa M, Tessema F. Infant survivorship and occurrence of multiple-births: a longitudinal community-based study, South West Ethiopia. Ethiopian J Health Dev. 2002;16:5–11.
[14] Hinde J, Demétrio CGB. Overdispersion: models and estimation. Comput Stat Data Anal. 1998;27:151–170.
[15] Hinde J, Demétrio CGB. Overdispersion: models and estimation. São Paulo: XIII Sinape; 1998.
[16] Neelon BH, Malley AJ, Normand SL. A Bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use. Stat Model. 2010;4:421–439.
[17] Molenberghs G, Verbeke G, Demétrio C. An extended random-effects approach to modeling repeated, overdispersed count data. Lifetime Data Anal. 2007;13:513–531.

Appendix 1. SAS code

A.1. *SAS code for analysis of the Jimma Infant Growth Study*

The following SAS code can be used to fit the models to the Jimma Infant Growth Study, as described in Section 5.

```
/*
   resp is the response variable
   name=1 is an indicator for the count sequence and name=2
   is for the continuous sequence
*/

/* (NN-) & (PNG) */
proc nlmixed data=joint qpoints=20;
parms beta11=3.2818 beta12=0.7695 beta13=-0.03359 sigma=0.7202
      beta21=-1.1599 beta22=0.2250  tau1=0.6845 tau2=0.3699 alpha=17;
if name = "2" then do;
mean = beta11 + beta12*age +beta13*age*age + b1;
dens = -0.5*log(3.14159265358) - log(sigma)
       -0.5*(resp-mean)**2/(sigma**2);
ll = dens;
end;
if name = "1" then do;
eta = beta21 + beta22*age+b2;
expeta=exp(eta);
m = 1/alpha;
p = 1/(1+alpha*expeta);
ll = lgamma(resp+m)-lgamma(resp+1)-lgamma(m) +
     resp*log(alpha*expeta)-(resp+m)*log(1/p);
end;
model resp ~ general(ll);
random b1 b2 ~normal([0,0],[tau1**2, 0,tau2**2]) subject=id;
run;


/* (NN-)(PNG) */
proc nlmixed data=joint qpoints=20;
parms beta11=3.2818 beta12=0.7695 beta13=-0.03359 sigma=0.7202
beta21=-1.1599 beta22=0.2250  tau1=0.6845 tau2=0.3699  rho=-0.1 alpha=17;
if name = "2" then do;
mean = beta11 + beta12*age +beta13*age*age+b1;
dens = -0.5*log(3.14159265358) - log(sigma)
       -0.5*(resp-mean)**2/(sigma**2);
ll = dens;
end;
if name = "1" then do;
```

```
eta = beta21 + beta22*age+b2;
expeta=exp(eta);
m = 1/alpha;
p = 1/(1+alpha*expeta);
ll = lgamma(resp+m)-lgamma(resp+1)-lgamma(m) +
     resp*log(alpha*expeta)-(resp+m)*log(1+alpha*expeta);
end;
model resp ~ general(ll);
random b1 b2 ~normal([0,0],[tau1**2,rho*tau1*tau2,tau2**2]) subject=id;
run;

/*(NN-) & H(PNG) */
proc nlmixed data=joint qpoints=20;
parms beta11=3.2818 beta12=0.7695 beta13=-0.033 sigma=0.76
      beta21=2.0484 beta22=0.01788 tau1=0.688 tau2=0.48
      alpha=0.333 a0=2.04 a1=-0.037;
if name = "2" then do;
mean = beta11 + beta12*age +beta13*age*age + b1;
dens = -0.5*log(3.14159265358) - log(sigma)
       -0.5*(resp-mean)**2/(sigma**2);
ll = dens;
end;
if name = "1" then do;
eta = beta21 + beta22*age + b2;
expeta=exp(eta);
eta_prob=a0+a1*age;
expeta_prob=exp(eta_prob);
p_0=exp(eta_prob)/(1+exp(eta_prob));
m = 1/alpha;
p = 1/(1+alpha*expeta);
if  resp=0 then do;
ll = eta_prob-log(1+expeta_prob);
end;
else do;
ll = log(1-p_0)+lgamma(resp+m)-lgamma(resp+1)-lgamma(m) +
     resp*log(alpha*expeta)-(resp+m)*log(1+alpha*expeta)
     - log(1 -( 1 + alpha*expeta)**(-m));
end;
end;
model resp ~ general(ll);
random b1 b2 ~normal([0,0],[tau1**2,0,tau2**2]) subject=id;
run;

/*H(NN-)(PNG)*/
proc nlmixed data=joint  qpoints=20;
parms beta11=3.2818 beta12=0.7695 beta13=-0.033 sigma=0.76
      beta21=2.0484 beta22=0.01788 tau1=0.688 tau2=0.48
      rho=-0.1 alpha=0.333 a0=2.04 a1=-0.037;
if name = "2" then do;
mean = beta11 + beta12*age +beta13*age*age + b1;
dens = -0.5*log(3.14159265358) - log(sigma)
       -0.5*(resp-mean)**2/(sigma**2);
ll = dens;
end;
if name = "1" then do;
eta = beta21 + beta22*age + b2;
expeta=exp(eta);
eta_prob=a0+a1*age;
expeta_prob=exp(eta_prob);
p_0=exp(eta_prob)/(1+exp(eta_prob));
m = 1/alpha;
p = 1/(1+alpha*expeta);
if resp=0 then do;
ll = eta_prob-log(1+expeta_prob);
end;
else do;
```

```
ll = log(1-p_0)+lgamma(resp+m)-lgamma(resp+1)-lgamma(m) +
     resp*log(alpha*expeta)-(resp+m)*log(1+alpha*expeta)
     - log(1 -( 1 + alpha*expeta)**(-m));
end;
end;
model resp ~ general(ll);
random b1 b2 ~normal([0,0],[tau1**2,rho*tau1*tau2,tau2**2]) subject=id;
run;

/*(NN-) & ZI(PNG)*/
proc nlmixed data=joint qpoints=20;
parms beta11=3.2818 beta12=0.7695 beta13=-0.033 sigma=0.76
      beta21=2.0484 beta22=0.01788 tau1=0.688 tau2=0.48
      alpha=0.333 a0=2.04 a1=-0.037;
if name = "2" then do;
mean = beta11 + beta12*age +beta13*age*age+ b1;
dens = -0.5*log(3.14159265358) - log(sigma)
       -0.5*(resp-mean)**2/(sigma**2);
ll = dens;
end;
if name = "1" then do;
eta = beta21 + beta22*age+ b2;
expeta=exp(eta);
eta_prob=a0+a1*age;
expeta_prob=exp(eta_prob);
p_0=exp(eta_prob)/(1+exp(eta_prob));
m = 1/alpha;
p = 1/(1+alpha*expeta);
if  resp=0 then do;
ll = log(p_0 + (1-p_0)*(p**m));end;
else do;
ll = log(1-p_0)+lgamma(resp+m)-lgamma(resp+1)-lgamma(m) +
     resp*log(alpha*expeta)-(resp+m)*log(1+alpha*expeta);
end;
end;
model resp ~ general(ll);
random b1 b2 ~normal([0,0],[tau1**2,0,tau2**2]) subject=id;
run;

/*ZI(NN-)(PNG)*/
proc nlmixed data=joint qpoints=20;
parms beta11=3.2818 beta12=0.7695 beta13=-0.033 sigma=0.76
      beta21=2.0484 beta22=0.01788 tau1=0.688 tau2=0.48
      rho=-0.1 alpha=0.333 a0=2.04 a1=-0.037;
if name = "2" then do;
mean = beta11 + beta12*age +beta13*age*age + b1;
dens = -0.5*log(3.14159265358) - log(sigma)
       -0.5*(resp-mean)**2/(sigma**2);
ll = dens;
end;
if name = "1" then do;
eta = beta21 + beta22*age + b2;
expeta=exp(eta);
eta_prob=a0+a1*age;
expeta_prob=exp(eta_prob);
p_0=exp(eta_prob)/(1+exp(eta_prob));
m = 1/alpha;
p = 1/(1+alpha*expeta);
if  resp=0 then do;
ll = log(p_0 + (1-p_0)*(p**m));
end;
else do;
ll = log(1-p_0)+lgamma(resp+m)-lgamma(resp+1)-lgamma(m) +
     resp*log(alpha*expeta)-(resp+m)*log(1+alpha*expeta);
end;
end;
```

```
model resp ~ general(ll);
random b1 b2 ~normal([0,0],[tau1**2,rho*tau1*tau2,tau2**2]) subject=id;
run;


/*Test for zero-inflation (zero-altered model):
  comparing H(NN-)(PNG) with a1  versus  H(NN-)(PNG) without a1*/

/*H(NN-)(PNG) without a1*/
proc nlmixed data=joint qpoints=20;
parms beta11=3.2818 beta12=0.7695 beta13=-0.033  sigma=0.76
      beta21=2.0484 beta22=0.01788 tau1=0.688 tau2=0.48
      rho=-0.1 alpha=0.333;
if name = "2" then do;
mean = beta11 + beta12*age +beta13*age*age + b1;
dens = -0.5*log(3.14159265358) - log(sigma)
       -0.5*(resp-mean)**2/(sigma**2);
ll = dens;
end;
if name = "1" then do;
eta = beta21 + beta22*age + b2;
expeta=exp(eta);
eta_prob=beta21+beta22*age+b2;
expeta_prob=exp(eta_prob);
p_0=exp(-exp(eta_prob));
m = 1/alpha;
p = 1/(1+alpha*expeta);
if   resp=0 then do;
ll =  eta_prob-log(1+expeta_prob);end;
else do;
ll = log(1-p_0)+lgamma(resp+m)-lgamma(resp+1)-lgamma(m) +
     resp*log(alpha*expeta)-(resp+m)*log(1+alpha*expeta)
     - log(1 -( 1 + alpha*expeta)**(-m));end;
end;
model resp ~ general(ll);
random b1 b2 ~normal([0,0],[tau1**2,rho*tau1*tau2,tau2**2]) subject=id;
run;

/*H(NN-)(PNG) with a1*/
proc nlmixed data=joint qpoints=20;
parms beta11=3.2818 beta12=0.7695 beta13=-0.033  sigma=0.76
      beta21=2.0484 beta22=0.01788 tau1=0.688 tau2=0.48
      rho=-0.1 alpha=0.333 a1=0;
if name = "2" then do;
mean = beta11 + beta12*age +beta13*age*age + b1;
dens = -0.5*log(3.14159265358) - log(sigma)
       -0.5*(resp-mean)**2/(sigma**2);
ll = dens;
end;
if name = "1" then do;
eta = beta21 + beta22*age + b2;
expeta=exp(eta);
eta_prob=a1+beta21+beta22*age+b2;
expeta_prob=exp(eta_prob);
p_0=1-exp(-exp(eta_prob));
m = 1/alpha;
p = 1/(1+alpha*expeta);
if   resp=0 then do;
ll =  eta_prob-log(1+expeta_prob);end;
else do;
ll = log(1-p_0)+lgamma(resp+m)-lgamma(resp+1)-lgamma(m) +
     resp*log(alpha*expeta)-(resp+m)*log(1+alpha*expeta)
     - log(1 -( 1 + alpha*expeta)**(-m));
end;
end;
model resp ~ general(ll);
```

```
random b1 b2 ~normal([0,0],[tau1**2,rho*tau1*tau2,tau2**2]) subject=id;
run;
```

A.2. *SAS code for the simulation study*

```
/* Scenario one: generate data from ZI(NN-)(PNG)model */

data jointsim1;
call streaminit(1234);
do ss=1 to 250 ;
mean1=0; /*mean for b1*/
mean2=0; /*mean for b2*/
sig1=2; /*SD for b1*/
sig2=1.5; /*SD for b2*/
rho=-0.5; /*Correlation between b1 and b2*/
do kk=1 to 200;
r1 = rannor(1245);
r2 = rannor(2923);
b1 = mean1 + sig1*r1;/*random effect for continuous part*/
b2 = mean2 + rho*sig2*r1+sqrt(sig2**2-sig2**2*rho**2)*r2;
     /*random effect for count part*/
do TT=1 to 10; /*10 time points*/
sim=ss;
e=rand('normal',0,0.6);
id=kk;
age=TT;
mu=3.3+0.77*age-0.03*age*age+b1+e; /* continuous part*/
kappa = exp(2 + 0.02*age+b2);    /* count part*/
theta = 1;
parm1 = 1/(1+kappa/theta);
yneg = rand('NEGB',parm1,theta);
p1=2-0.2*age;   /* zero-inflation part*/
p2=exp(p1);
p=p2/(1+p2);
inf=rand('bern',p);
if inf=1 then do;
ynegzim=0;
end;
else do;
ynegzim=yneg;
end;
numdays=ynegzim;
output ;
end;
end ;
end;
data jointsim1;
set jointsim1;
run;

/* Scenario two:generate data from ZI(NN-)(PN-)model*/

data jointsim2;
call streaminit(1234);
do ss=1 to 250 ;
mean1=0; *mean for b1;
mean2=0; *mean for b2;
sig1=2; *SD for b1;
sig2=1.5; *SD for b2;
rho=-0.5; *Correlation between b1 and b2;
do kk=1 to 200;
r1 = rannor(1245);
r2 = rannor(2923);
b1 = mean1 + sig1*r1;
b2 = mean2 + rho*sig2*r1+sqrt(sig2**2-sig2**2*rho**2)*r2;
```

```
do TT=1 to 10; /*10 time points*/
sim=ss;
e=rand('normal',0,0.6);
id=kk;
age=TT;
mu=3.3+0.77*age-0.03*age*age+b1+e;
kappa = exp(2 + 0.02*age+b2);
ypois = rand('POISSON',kappa);
p1=2-0.2*age;
p2=exp(p1);
p=p2/(1+p2);
inf=rand('bern',p);
if inf=1 then do;
ypoiszim=0;
end;
else do;
ypoiszim=ypois;
end;
numdays=ypoiszim;
output ;
end;
end ;
end;
data jointsim2;
set jointsim2;
run;

/* Scenario three: generate data from (NN-)(PNG)model*/

data jointsim3;
call streaminit(1234);
do ss=1 to 250 ;
mean1=0; *mean for b1;
mean2=0; *mean for b2;
sig1=2; *SD for b1;
sig2=1.5; *SD for b2;
rho=-0.5; *Correlation between b1 and b2;
do kk=1 to 200;
r1 = rannor(1245);
r2 = rannor(2923);
b1 = mean1 + sig1*r1;
b2 = mean2 + rho*sig2*r1+sqrt(sig2**2-sig2**2*rho**2)*r2;
do TT=1 to 10; /*10 time points*/
sim=ss;
e=rand('normal',0,0.6);
id=kk;
age=TT;
mu=3.3+0.77*age-0.03*age*age+b1+e;
kappa = exp(2 + 0.02*age+b2);
theta = 1;
parm1 = 1/(1+kappa/theta);
yneg = rand('NEGB',parm1,theta);
numdays=yneg;
output ;
end;
end ;
end;
data jointsim3;
set jointsim3;
run;

/* Calculation of overall mean, std.dev. and percentage of zeros */

/* Scenario one */

proc means data=jointsim1;
```

```
var numdays; run;
proc freq data=jointsim1;
tables numdays;
run;

/*Scenario two*/

proc means data=jointsim2;
var numdays; run;
proc freq data=jointsim2;
tables numdays;
run;

/*Scenario three*/

proc means data=jointsim3;
var numdays; run;
proc freq data=jointsim3;
tables numdays;
run;
```