SORT 41 (2) July-December 2017, 191-254 DOI: 10.2436/20.8080.02.58

Hierarchical models with normal and conjugate random effects: a review

Geert Molenberghs[1,2,*], Geert Verbeke[2,1] and Clarice G.B. Demétrio[3]

Abstract

Molenberghs, Verbeke, and Demétrio (2007) and Molenberghs et al. (2010) proposed a general framework to model hierarchical data subject to within-unit correlation and/or overdispersion. The framework extends classical overdispersion models as well as generalized linear mixed models. Subsequent work has examined various aspects that lead to the formulation of several extensions. A unified treatment of the model framework and key extensions is provided. Particular extensions discussed are: explicit calculation of correlation and other moment-based functions, joint modelling of several hierarchical sequences, versions with direct marginally interpretable parameters, zero-inflation in the count case, and influence diagnostics. The basic models and several extensions are illustrated using a set of key examples, one per data type (count, binary, multinomial, ordinal, and time-to-event).

MSC: 62P10, 97K80.

Keywords: Conjugacy, frailty, joint modelling, marginalized multilevel model, mixed model, overdispersion, underdispersion, variance component, zero-inflation.

1. Introduction

Parametric or semi-parametric modelling of univariate non-Gaussian outcomes is often done within the generalized linear model (GLM) framework (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989; Agresti, 2002), which rests on the exponential family. Commonly encountered outcome types include categorical (binary, binomial, ordinal, etc.), count, and time-to-event outcomes, for which modelling typically, though not always, rests upon the Bernoulli, Poisson, and exponential/Weibull distributions, respectively. A key feature of exponential family distributions is the so-called *mean-*

[*] Corresponding author: Geert Molenberghs, I-BioStat, Universiteit Hasselt, Martelarenlaan 42, B-3500 Hasselt, Belgium. geert.molenberghs@uhasselt.be

[1] I-BioStat, Universiteit Hasselt, B-3500 Hasselt, Belgium.

[2] I-BioStat, KU Leuven, B-3000 Leuven, Belgium.

[3] ESALQ, Universidade de Saõ Paulo, Piracicaba, Brazil.

Received: May 2017.

variance relationship, i.e., the fact that the variance is a deterministic function of the mean. For example, for Bernoulli outcomes with success probability $\mu = \pi$, the variance is $v(\mu) = \pi(1-\pi)$, for counts using Poisson assumptions $v(\mu) = \mu$ and for the exponential model $v(\mu) = \mu^2$. However, for many outcome types, empirically observed data can contradict this relationship, in the sense that the observed variance may be higher or lower than what follows from the model formulation; these are referred to as overdispersion and underdispersion, respectively. The two phenomena combined are sometimes referred to as extra-model-dispersion. Especially in the somewhat older literature, more attention was given to overdispersion than to underdispersion. Hinde and Demétrio (1998ab) provide early overviews of (semi-)parametric approaches for dealing with overdispersion. Well-known models include the beta-binomial (Skellam, 1948; Kleinman, 1973) for binary and binomial data, and the negative binomial model (Breslow, 1984; Lawless, 1987) for counts. These models can be generated by assuming the so-called natural parameter to follow a carefully chosen distribution. For example, the beta-binomial models follow from assuming the outcomes follow a binomial distribution with parameter drawn from a beta distribution; the negative binomial model follows from a Poisson model with gamma distributed parameter. The resulting models have elegant parametric expressions and are relatively easy to interpret, because the outcome and random-effects distributions are *conjugate*, a precise definition of which is given in Section 4.2. Other solutions to accommodating overdispersion include mixture modelling and specific models for zero-inflated Poisson models (Ridout, Demétrio and Hinde, 1998; Böhning, 2000; McLachlan and Peel, 2000).

Nowadays, it is very common to encounter aforementioned data types in a hierarchical context, such as resulting from multivariate, longitudinal, spatial, and clustered designs. We will generically refer to these settings as repeated measures. The data hierarchies induce association among the repeated measures, which can be captured, among others, by random effects. Especially the generalized linear mixed model (GLMM; Engel and Keen, 1994; Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993) has beÃ§come a popular and widespread tool, routinely implemented in a suite of standard software packages. Reviews are given in Verbeke and Molenberghs (2000) and Molenberghs and Verbeke (2005). A key ingredient is a linear predictor that also incorporates normally distributed random effects. These random effects engender not only correlation among the repeated measures, but also some overdispersion. However, the empirical correlation and overdispersion present in the data may be hard to model with only a limited number of normal random effects. This is why Molenberghs et al. (2007; henceforth referred to as MVD) and Molenberghs et al. (2010; henceforth referred to as MVDV) have proposed a model family, the so-called *combined model* (CM) that combines conjugate and normal random effects, leading to highly increased flexibility for the triple of functions made up of the mean, variance, and correlation functions. Note that, for time-to-event data, not only GLMM but also the so-called frailty models (Duchateau and Janssen, 2007) have been used. These start from gamma rather than normally distributed random effects, which are conjugate to the exponential distribution, and lead to elegant expressions when combined with the Weibull distribution as well (see Section 7).

After introducing a set of key examples (Section 2) and reviewing several key ingredients in Section 3, the CM is introduced in Section 4. Sections 5-7 are devoted to the count, categorical, and time-to-event cases, respectively. In the count case, specific attention is given to the occurrence of extra-model zeroes, i.e., zero-inflated versions of the model. In the categorical case, we further distinguish between binary, binomial, and ordinal data. Of note is the rather different algebraic nature of the model with logit and that with probit link. In the time-to-event case, we also allow for censoring, and discuss some issues with the moment functions of the so-called Weibull-gamma-normal model and its sub-models. In Section 8 we describe maximum likelihood and some related estimation strategies.

In Section 9, we show how the CM and its sub-models can be used, in most cases, to derive explicit expressions for so-called manifest correlations, whereas often, for convenience, the latent correlation is considered. Usually, though, the manifest correlation is considerably smaller than its latent counterpart; hence, using the latter may lead to overly optimistic conclusions.

A typical problem arising with the GLMM, in contrast to the GLM, the linear mixed model (LMM) for Gaussian outcomes, and models with conjugate random effects is that deriving marginal expressions is not so straightforward and, related to this, that the model parameters have a hierarchical (i.e., conditional on the random effects) but not a marginal (i.e., averaged over a suitable population) interpretation. The CM evidently inherits this problem. While some progress is made for the specific cases discussed in Sections 5-7, it is still useful to take a different route: that of a so-called marginalized multilevel model, based on work of Heagerty (1999) and Heagerty and Zeger (2000). It will be referred to as the *combined marginalized multilevel model*, or COMMM.

Evidently, in line with a lot of contemporary work, it is perfectly possible to observe, for example, several longitudinal sequences simultaneously. The resulting designs are referred to as multivariate longitudinal or, more generically, joint modelling. The use of the CM in this context is reviewed in Section 11. Finally, Section 12 describes diagnostic measures based on local influence.

The review in this paper is based on work by MVD and MVDV, which is also based on Booth et al. (2003), and various extensions of all of these. Evidently, also different strands of research exist that extend the GLMM and increase its flexibility. In particular, we refer to Lee and Nelder (1996, 2001ab, 2003), Lee, Nelder, and Pawitan (2006), who proposed so-called *hierarchical generalized linear models*, accommodating many outcome and random-effects distributions, while being efficient in computational terms. In the particular case of count data, our model relates to theirs by considering log-gamma and log-normal random effects together. Regarding estimation, we focus primarily on marginal maximum likelihood estimation and Bayesian estimation, whereas Lee and Nelder employ so-called h-likelihood. In particular, we analytically integrate over the conjugate random effects and use numerical integration for the normal random effects.

Skrondal and Rabe-Hesketh (2004) brought together in a single model framework, multilevel modelling, structural equations modelling, latent variables, latent classes, and random-effects models for hierarchical data.

2. Case studies

We will describe five case studies. The outcomes are of a count, binary, binomial, ordinal, and time-to-event nature, respectively.

2.1. A clinical trial in epileptic patients

The data considered here are obtained from a randomized, double-blind, parallel group multicentre study for the comparison of placebo with a new anti-epileptic drug (AED), in combination with one or two other AEDs. The study is described in full detail in Faught et al. (1996). The randomization of epilepsy patients took place after a 12-week baseline period that served as a stabilization period for the use of AEDs, and during which the number of seizures were counted. After that period, 45 patients were assigned to the placebo group, 44 to the active (new) treatment group. Patients were then measured weekly. Patients were followed (double-blind) during 16 weeks, after which they were entered into a long-term open-extension study. Some patients were followed for up to 27 weeks. The outcome of interest is the number of epileptic seizures experienced during the most recent week. The research question is whether or not the additional new treatment reduces the number of epileptic seizures.

2.2. A clinical trial in onychomycosis

These data come from a randomized, double-blind, parallel group, multicentre study for the comparison of two oral treatments (coded as A and B) for toenail dermatophyte onychomycosis (TDO), described in full detail by De Backer et al. (1996). TDO is a common toenail infection, difficult to treat, affecting more than 2 out of 100 persons (Roberts, 1992). Anti-fungal compounds, classically used for treatment of TDO, need to be taken until the whole nail has grown out healthy. The development of new such compounds, however, has reduced the treatment duration to 3 months. The aim of the present study was to compare the efficacy and safety of 12 weeks of continuous therapy with treatment A or with treatment B. In total, 2×189 patients, distributed over 36 centres, were randomized. Subjects were followed during 12 weeks (3 months) of treatment and followed further, up to a total of 48 weeks (12 months). Measurements were taken at baseline, every month during treatment, and every 3 months afterwards, resulting in a maximum of 7 measurements per subject. At the first occasion, the treating physician indicates one of the affected toenails as the target nail, the nail which will be followed over time. We will restrict our analyses to only those patients for which the target nail was one of the two big toenails (146 and 148 subjects, in group A and group B, respectively). One of the responses of interest was the unaffected nail length, measured from the nail bed to the infected part of the nail, which is always at the free end of the nail, expressed in *mm*. This outcome has been studied extensively in Verbeke and Molenberghs (2000). Another important outcome in this study was the severity of the infection, coded as 0 (not severe) or 1 (severe). The question of interest was whether the percentage of severe infections decreased over time, and whether that evolution was different for the two treatment groups.

2.3. Iron-deficient diets in rats

These data result from an experiment where female rats were put on iron-deficient diets (Shepard, Mackler, and Finch, 1980). This dataset has been analysed by Liang and McCullagh (1993) and Moore and Tsiatis (1991). In Agresti (2002), the data were used to estimate several logit models. Experimental rats were divided into 4 groups, one of which is a control group. The number of female rats per group (total number of fetuses per group) are: 31 (327) for placebo, 12 (118) for low dose, 5 (58) for medium dose, and 10 (104) for high dose. Weekly injections of iron supplement were to bring the rats' iron intake to normal levels. Rats in the placebo group were given placebo injection, the others got three different doses of the iron supplements. Rats were made pregnant and sacrificed 3 weeks later and the total number of fetuses and the number of dead fetuses in each litter were counted. Hemoglobin levels of the mothers were also measured.

2.4. Diabetes study

In Belgium, the diabetes project was conducted from January 2005 until December 2006, with the aim to study the effect of implementing a structured model for chronic diabetes care on the patients' clinical outcomes. General practitioners (GPs) were offered assistance and could redirect patients to the diabetes care team, consisting of a nurse educator, a dietician, an ophthalmologist, and an internal medicine doctor. For the project, two programs were implemented and GPs were randomized to one of two groups: UQIP: Usual Quality Improvement Program and AQIP: Advanced Quality Improvement Program. A total of 120 GPs took part in the study, 53 in the UQIP group and 67 in the AQIP group, including 918 and 1577 patients, respectively.

During the project, several outcomes useful to evaluate how well diabetes is controlled were measured, at the moment the program was initiated (time T_0) and one year later (T_1). The most important outcomes were HbA1c (glycosylated hemoglobin), LDL-cholesterol (low-density lipoprotein cholesterol) and SBD (systolic blood pressure). Furthermore, experts specified cut off values defining a so-called *clinical target* for each outcome: HBA1C$<$7%, LDL-cholesterol $<$ 100 mg/dl and SBD $\leq$ 130 mmHg. As a result, for a particular time point, every patient could reach between 0 and 3 clinical targets. This number was reflected in the variable *number of clinical targets*. If at least one measurement per patient was missing, the value for the number of clinical targets was set to missing as well. The data are discussed in Borgermans et al. (2009).

2.5. Recurrent asthma attacks in children

These data have been studied in Duchateau and Janssen (2007). Asthma is occurring more and more frequently in very young children (between 6 and 24 months). Therefore, a new application of an existing anti-allergic drug is administered to children who are at higher risk to develop asthma in order to prevent it. A prevention trial is set up with such children randomized to placebo or drug, and the asthma events that developed over time are recorded in a diary. Typically, a patient has more than one asthma event. The different events are thus clustered within a patient and ordered in time. This ordering can be taken into account in the model. The data are presented in calendar time format, where the time at risk for a particular event is the time from the end of the previous event (asthma attack) to the start of the next event (start of the next asthma attack). A particular patient has different periods at risk during the total observation period which are separated either by an asthmatic event that lasts one or more days or by a period in which the patient was not under observation. The start and end of each such risk period is required, together with the status indicator to denote whether the end of the risk period corresponds to an asthma attack or not.

3. Some background

We briefly review some background on the exponential family and generalized linear models (Section 3.1), overdispersion (Section 3.2), and models with normal random effects (Section 3.3).

3.1. Generalized linear models

A random variable Y follows an exponential family distribution if the density is of the form

$$f(y) \equiv f(y|\eta,\phi) \;=\; \exp\left\{\phi^{-1}[y\eta-\psi(\eta)]+c(y,\phi)\right\}, \tag{1}$$

for a specific set of unknown parameters η ('natural parameter' or 'canonical parameter') and ϕ ('dispersion parameter'), and for known functions $\psi(\cdot)$ and $c(\cdot,\cdot)$. It follows that $\mathrm{E}(Y)=\mu=\psi'(\eta)$ and $\mathrm{Var}(Y)=\sigma^2=\phi\psi''(\eta)$, with ensuing mean-variance relationship $\sigma^2=\phi\psi''[\psi^{'-1}(\mu)]=\phi v(\mu)$, with $v(\cdot)$ the *variance function*. Commonly encountered examples and their model elements are presented in Table 1. Note that, in the normal case, there is no mean-variance relationship. In the binary case, also the probit link is commonly encountered, whence $\eta = \Phi^{-1}(\pi)$ and $\Phi(\cdot)$ is the standard normal cumulative distribution function. As explained in Section 6.1, the probit link has appealing properties when normal random effects are introduced into the model.

In the Weibull and exponential model, the decomposition $\varphi = \lambda e^{\mu}$ is often used, allowing μ to be written as a function of covariates. Note that μ is a component of the mean function, not the mean itself. The Weibull model does not belong to the exponential family in a conventional sense, unless when y is replaced by y^{ρ}. In Table 1, $\Gamma(\cdot)$ represents the gamma function.

When not the full joint distribution but, say, the first and second moments only are specified, a semi-parametric version of the model results, for which quasi-likelihood estimation has been devised (McCullagh and Nelder, 1989; Molenberghs and Verbeke, 2005).

The generalized linear model (GLM) follows from the exponential family by assuming that a set of independent replicates Y_i with p-dimensional covariate vectors $\boldsymbol{x}_i$ $(i = 1,\dots,N)$, follow exponential-family densities $f(y_i|\eta_i,\phi)$. Specification of the GLM is completed by modelling the means μ_i as functions of the covariate values: $\mu_i = h(\eta_i) = h(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\xi})$, for a known function $h(\cdot)$, and with $\boldsymbol{\xi}$ a vector of p fixed, unknown regression coefficients. Here, $h^{-1}(\cdot)$ is called the link function. In most applications, the so-called natural link function is used, i.e., $h(\cdot) = \psi'(\cdot)$, which is equivalent to assuming $\eta_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\xi}$. In other words, it is assumed that the natural parameter satisfies a linear regression model.

3.2. Overdispersion

As stated in the introduction, and as is clear from Table 1, many standard exponential family models enforce a mean-variance relationship that may be contradicted by the data, especially for count, binomial, and time-to-event data. For binary data, such a violation can only occur when the outcomes are correlated (see Section 6).

As reviewed by Hinde and Demétrio (1998ab), an obvious way to incorporate overdispersion is by allowing $\phi \neq 1$, so that the variance becomes $\mathrm{Var}(Y) = \phi v(\mu)$. An elegant way forward is through a two-stage approach. For binary data, one would assume that $Y_i|\pi_i \sim \mathrm{Bernoulli}(\pi_i)$ and further that π_i is a random variable with $\mathrm{E}(\pi_i) = \mu_i$ and $\mathrm{Var}(\pi_i) = \sigma_i^2$. Using iterated expectations, it follows that $\mathrm{E}(Y_i) = \mu_i$ and $\mathrm{var}(Y_i) = \mu_i(1-\mu_i)$, underscoring that purely Bernoulli data are unable to exhibit overdispersion. The situation is different for counts. In the Poisson case, we assume that $Y_i|\zeta_i \sim \mathrm{Poi}(\zeta_i)$ and then that ζ_i is a random variable with $\mathrm{E}(\zeta_i) = \mu_i$ and $\mathrm{Var}(\zeta_i) = \sigma_i^2$. Then, it follows that $\mathrm{E}(Y_i) = \mu_i$ and $\mathrm{var}(Y_i) = \mu_i + \sigma_i^2$. We have not assumed a particular distributional form for the random effects π_i and ζ_i, respectively. Hence, this gives rise to a semi-parametric specification. In case it is considered advantageous to make full distributional assumptions about the random effects, common choices are the beta distribution for π_i and the gamma distribution for ζ_i; of course, these are not the only ones.

The two-stage approach is made up of considering a distribution for the outcome variable, given a random effect $f(y_i|\theta_i)$ which, combined with a model for the random effect, $f(\theta_i)$, produces the marginal model:

$$f(y_i) = \int f(y_i|\theta_i) f(\theta_i) d\theta_i. \tag{2}$$

It is easy to extend this model to the case of repeated measurements by assuming a hierarchical data structure, where now Y_{ij} denotes the jth outcome measured for cluster (subject) i, $i = 1,\ldots,N$, $j = 1,\ldots,n_i$ and $\boldsymbol{Y}_i$ is the n_i-dimensional vector of all measurements available for cluster i. In the repeated-measures case, the scalar ζ_i becomes a vector $\boldsymbol{\zeta}_i = (\zeta_{i1},\ldots,\zeta_{in_i})^{\mathrm{T}}$, with $\mathrm{E}(\boldsymbol{\zeta}_i) = \boldsymbol{\mu}_i$ and $\mathrm{var}(\boldsymbol{\zeta}_i) = \boldsymbol{\Sigma}_i$. For example, for the Poisson case, similar logic as in the univariate case produces $\mathrm{E}(\boldsymbol{Y}_i) = \boldsymbol{\mu}_i$ and $\mathrm{var}(\boldsymbol{Y}_i) = \boldsymbol{M}_i + \boldsymbol{\Sigma}_i$, where $\boldsymbol{M}_i$ is a diagonal matrix with the vector $\boldsymbol{\mu}_i$ along the diagonal. Note that a diagonal structure of $\boldsymbol{M}_i$ reflects the conditional independence assumption: all dependence between measurements on the same unit stems from the random effects. Generally, a versatile class of models results. For example, assuming that the components of $\boldsymbol{\zeta}_i$ are independent, a pure overdispersion model follows, without correlation between the repeated measures. On the other hand, assuming $\zeta_{ij} = \zeta_i$, i.e., that all components are equal, then $\mathrm{var}(\boldsymbol{Y}_i) = \boldsymbol{M}_i + \sigma_i^2 \boldsymbol{J}_{n_i}$, where $\boldsymbol{J}_{n_i}$ is an $n_i \times n_i$ dimensional matrix of ones. Such a structure can be seen as a general version of compound symmetry.

Alternatively, this repeated version of the overdispersion model can be combined with normal random effects in the linear predictor. This very specific choice was also proposed by Thall and Vail (1990) and Dean (1991) for the count case.

Marginalization (2) is general and elegant, but one has to reflect on which parameter to become random, in particular when full distributional assumptions are requested. As always, this is easy for the linear mixed model, by combining a normal hierarchical model with a normal random effect, and provided θ_i is used to express the conditional mean as a linear function of covariates. It forms the basis of the two strands of random-effects models that are potentially brought together in the combined models of Section 4: on the one hand, normal random effects can be considered with non-normal outcomes, producing the GLMM; on the other hand, gamma random effects for the Poisson model, beta random effects with binomial data, and gamma random effects for the Weibull model can be considered. This is, seemingly, a disparate collection. However, they are unified through so-called *conjugacy*, in the sense of Cox and Hinkley (1974, p. 370) and Lee et al. (2006, p. 178). The topic is also discussed by Agresti (2002). Informally, conjugacy refers to the fact that the hierarchical and random-effects densities have similar algebraic forms. Conjugate distributions produce a general and closed-form solution for the corresponding marginal distribution.

Table 1: *Conventional exponential family members and extensions with conjugate random effects.*

Element	Notation	Continuous	Binary	Count	Time to event	
Standard univariate exponential family						
Model		Normal	Bernoulli	Poisson	Exponential	Weibull
Model	$f(y)$	$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(y-\mu)^2}{2\sigma^2}}$	$\pi^y(1-\pi)^{1-y}$	$\frac{e^{-\lambda}\lambda^y}{y!}$	$\varphi e^{-\varphi y}$	$\varphi\rho y^{\rho-1}e^{-\varphi y^\rho}$
Nat. param	η	μ	$\ln[\pi/(1-\pi)]$	$\ln\lambda$	$-\varphi$	
Mean function	$\psi(\eta)$	$\eta^2/2$	$\ln[1+\exp(\eta)]$	$\lambda=\exp(\eta)$	$-\ln(-\eta)$	
Norm. constant	$c(y,\phi)$	$-\frac{\ln(2\pi\phi)}{2}-\frac{y^2}{2\phi}$	0	$-\ln y!$	0	
(Over)dispersion	ϕ	σ^2	1	1	1	
Mean	μ	μ	π	λ	$-\varphi^{-1}$	$\varphi^{-1/\rho}\Gamma(\rho^{-1}+1)$
Variance	$\phi v(\mu)$	σ^2	$\pi(1-\pi)$	λ	φ^{-2}	$\varphi^{-2/\rho}\left[\Gamma(2\rho^{-1}+1)-\Gamma(\rho^{-1}+1)^2\right]$
Exponential family with conjugate random effects						
Model		Normal-normal	Beta-binomial	Negative binomial	Exponential-gamma	Weibull-gamma
Hier. model	$f(y\|\theta)$	$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(y-\theta)^2}{2\sigma^2}}$	$\theta^y(1-\theta)^{1-y}$	$\frac{e^{-\theta}\theta^y}{y!}$	$\varphi\theta e^{-\varphi\theta y}$	$\varphi\theta\rho y^{\rho-1}e^{-\varphi\theta y^\rho}$
RE model	$f(\theta)$	$\frac{1}{\sqrt{d}\sqrt{2\pi}}e^{-\frac{(\theta-\mu)^2}{2d}}$	$\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}$	$\frac{\theta^{\alpha-1}e^{-\theta/\beta}}{\beta^\alpha\Gamma(\alpha)}$	$\frac{\theta^{\alpha-1}e^{-\theta/\beta}}{\beta^\alpha\Gamma(\alpha)}$	$\frac{\theta^{\alpha-1}e^{-\theta/\beta}}{\beta^\alpha\Gamma(\alpha)}$
Marg. model	$f(y)$	$\frac{1}{\sqrt{\sigma^2+d}\sqrt{2\pi}}e^{-\frac{(y-\mu)^2}{2(\sigma^2+d)}}$	$(\alpha+\beta)\frac{\Gamma(\alpha)}{\Gamma(\alpha+y)}\frac{\Gamma(\beta)}{\Gamma(\beta+1-y)}$	$\frac{\Gamma(\alpha+y)}{y!\Gamma(\alpha)}\left(\frac{\beta}{\beta+1}\right)^y\left(\frac{1}{\beta+1}\right)^\alpha$	$\frac{\varphi\alpha\beta}{(1+\varphi\beta y)^{\alpha+1}}$	$\frac{\varphi\rho y^{\rho-1}\alpha\beta}{(1+\varphi\beta y^\rho)^{\alpha+1}}$
	$h(\theta)$	θ	$\ln[\theta/(1-\theta)]$	$\ln(\theta)$	$-\theta$	$-\theta$
	$g(\theta)$	$-\frac{1}{2}\theta^2$	$-\ln(1-\theta)$	θ	$-\ln(\theta)/\varphi$	$-\ln(\theta)/\varphi$
	ϕ	σ^2	1	1	$1/\varphi$	$1/\varphi$
	γ	$1/d$	$\alpha+\beta-2$	$1/\beta$	$\varphi(\alpha-1)$	$\varphi(\alpha-1)$
	ψ	μ	$\frac{\alpha-1}{\alpha+\beta-2}$	$\beta(\alpha-1)$	$[\beta\varphi(\alpha-1)]^{-1}$	$[\beta\varphi(\alpha-1)]^{-1}$
	$c(y,\phi)$	$-\frac{1}{2}\phi y^2-\frac{1}{2}\ln\left(\frac{2\pi}{\phi}\right)$	0	$-\ln(y!)$	$\ln(\varphi)$	$\ln\left(\varphi\rho y^{\rho-1}\right)$
	$c^*(\gamma,\psi)$	$-\frac{1}{2}\gamma\psi^2-\frac{1}{2}\ln\left(\frac{2\pi}{\gamma}\right)$	$-\ln B(\gamma\psi+1,\gamma-\psi\gamma+1)$	$(1+\gamma\psi)\ln\gamma-\ln\Gamma(1+\gamma\psi)$	$\frac{\gamma+\varphi}{\varphi}\ln(\gamma\psi)-\ln\Gamma\left(\frac{\gamma+\varphi}{\varphi}\right)$	$\frac{\gamma+\varphi}{\varphi}\ln(\gamma\psi)-\ln\Gamma\left(\frac{\gamma+\varphi}{\varphi}\right)$
Mean	$\mathrm{E}(Y)$	μ	$\frac{\alpha}{\alpha+\beta}$	$\alpha\beta$	$[\varphi(\alpha-1)\beta]^{-1}$	$\frac{\Gamma(\alpha-\rho^{-1})\Gamma(\rho^{-1}+1)}{(\varphi\beta)^{1/\rho}\Gamma(\alpha)}$
Variance	$\mathrm{Var}(Y)$	σ^2+d	$\frac{\alpha\beta}{(\alpha+\beta)^2}$	$\alpha\beta(\beta+1)$	$\alpha[\varphi^2(\alpha-1)^2(\alpha-2)\beta^2]^{-1}$	$\frac{1}{\rho(\varphi\beta)^{2/\rho}\Gamma(\alpha)}\left[2\Gamma(\alpha-2\rho^{-1})\Gamma(2\rho^{-1}) - \frac{\Gamma(\alpha-\rho^{-1})^2\Gamma(\rho^{-1})^2}{\rho\Gamma(\alpha)}\right]$

We will first define standard conjugacy, i.e., in models without the normal random effects and then, in Section 4, introduce a further property, *strong conjugacy*, necessary for situations where both normal and conventional conjugate random effects are present. To simplify notation, we will provide the definition at a general distribution level, with neither subject- nor measurement-specific subscripts, so that it can be applied to both univariate and longitudinal data. The hierarchical and random-effects densities are said to be conjugate if and only if they can be written in the generic forms:

$$f(y|\theta) = \exp\left\{\phi^{-1}[yh(\theta) - g(\theta)] + c(y,\phi)\right\}, \tag{3}$$

$$f(\theta) = \exp\left\{\gamma[\psi h(\theta) - g(\theta)] + c^*(\gamma,\psi)\right\}, \tag{4}$$

where $g(\theta)$ and $h(\theta)$ are functions, ϕ, γ, and ψ are parameters, and the additional functions $c(y,\phi)$ and $c^*(\gamma,\psi)$ are so-called normalizing constants. It can then be shown, upon constructing the joint distribution and then integrating over the random effect, that the marginal model resulting from (3) and (4) equals:

$$f(y) = \exp\left[c(y,\phi) + c^*(\gamma,\psi) - c^*\left(\phi^{-1} + \gamma, \frac{\phi^{-1}y + \gamma\psi}{\phi^{-1} + \gamma}\right)\right]. \tag{5}$$

Table 1 gives model elements, such as density or probability mass functions, conditional on random effects and marginalized over these, as well as the random effects distributions. For all models considered, the constants and functions featuring in (3)–(4) are listed, and finally marginal means and variances are provided. For some models, these are well known (Hinde and Demétrio, 1998ab) and/or easy to derive.

In the case of binary data, the model in Table 1 is the familiar beta-binomial model. Note that the variance still obeys the usual Bernoulli variance structure. This is entirely natural, given that we still focus on a single binary outcome, in contrast to the more conventional binomial basis model, where data of the format 'z_i successes out n_i trials' are considered. We do not consider this situation in this section, but rather leave it to Section 6. In such a case, the variance structure becomes $\pi_i(1-\pi_i)[1+\rho_i(n_i-1)]$, where ρ_i is a measure for correlation. All parameters, p_i and ρ_i, can be expressed in terms of α_i and β_i, 'cluster-specific' versions of the beta parameters.

For count data, the familiar negative-binomial model results. Unlike in the binary case, univariate counts are able to violate the mean-variance relationship inherent in the Poisson distribution, hence the great popularity of this and other types of models for overdispersion. The same applies to the exponential distribution. Of course, already the Weibull model, with its extra parameter ρ, alleviates the constraint.

The normal distribution case is a special one. Not only is it self-conjugate, also the model is not identified, unlike all others. This is because both random terms, seen from writing $Y_i = \mu_i + b_i + \varepsilon_i$, are in direct, linear relationship. In the generalized linear context, the various random terms have no direct linear alliance. The normal case will continue to be 'the odd one out' in models to come (Sections 3.3 and 5-7).

The parameters α and β in the beta and gamma distributions are not always jointly identified. It is therefore customary to impose restrictions, such as setting one of them equal to a fixed value, e.g., $\alpha = 1$, or constraining their mean or variance, etc. Such constraints operate differently, depending on other elements present in the models. For example, the presence of additional random effects in a model for repeated measures, such as in Section 4, alters the meaning and restrictiveness of such constraints.

3.3. Models with normal random effects

The generalized linear mixed model (GLMM; Engel and Keen, 1994; Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993) is a straightforward extension of the linear mixed model (Verbeke and Molenberghs, 2000) to non-Gaussian hierarchical data. It is implemented in many standard software tools.

Let Y_{ij} be the jth outcome measured for cluster (subject) $i = 1,\ldots,N$, $j = 1,\ldots,n_i$ and group the n_i measurements into a vector $\boldsymbol{Y}_i$. Assume that, in analogy with Section 3.1, conditionally upon q-dimensional random effects $\boldsymbol{b}_i \sim N(\mathbf{0},\boldsymbol{D})$, the outcomes Y_{ij} are independent with densities:

$$f_i(y_{ij}|\boldsymbol{b}_i,\boldsymbol{\xi},\phi) = \exp\left\{\phi^{-1}[y_{ij}\lambda_{ij} - \psi(\lambda_{ij})] + c(y_{ij},\phi)\right\}, \tag{6}$$

where

$$\eta[\psi'(\lambda_{ij})] = \eta(\mu_{ij}) = \eta[\mathrm{E}(Y_{ij}|\boldsymbol{b}_i,\boldsymbol{\xi})] = \boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\xi} + \boldsymbol{z}_{ij}^{\mathsf{T}}\boldsymbol{b}_i \tag{7}$$

for a known link function $\eta(\cdot)$, with $\boldsymbol{x}_{ij}^{\mathsf{T}}$ and $\boldsymbol{z}_{ij}$ p-dimensional and q-dimensional vectors of known covariate values, with $\boldsymbol{\xi}$ a p-dimensional vector of unknown fixed regression coefficients, and with ϕ a scale (overdispersion) parameter. Finally, let $f(\boldsymbol{b}_i|\boldsymbol{D})$ be the density of the $N(\mathbf{0},\boldsymbol{D})$ distribution for the random effects $\boldsymbol{b}_i$. These models closely follow the ones formulated in the top part of Table 1, with key differences that now: (a) data hierarchies are allowed for; (b) the natural parameter is written as a linear predictor, a function of both fixed and random effects.

4. Models combining conjugate and normal random effects

4.1. General model formulation

Combining overdispersion (Section 3.2) and normal random effects (Section 3.3) into the generalized linear model framework, produces the following general family:

$$f_i(y_{ij}|\boldsymbol{b}_i,\boldsymbol{\xi},\theta_{ij},\phi)=\exp\left\{\phi^{-1}[y_{ij}\lambda_{ij}-\psi(\lambda_{ij})]+c(y_{ij},\phi)\right\}, \tag{8}$$

The conditional mean follows as the product:

$$E\left(Y_{ij}|\boldsymbol{b_i},\boldsymbol{\xi},\theta_{ij}\right)=\mu^c_{ij}=\psi'(\lambda_{ij})=\theta_{ij}\kappa_{ij}, \tag{9}$$

where the random variable

$$\theta_{ij}\sim\Theta_{ij}\left(\upsilon_{ij},\sigma^2_{ij}\right), \tag{10}$$

with mean υ_{ij}, variance σ^2_{ij}, and the mean component

$$g(\kappa_{ij})=\boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\xi}+\boldsymbol{z}_{ij}^{\mathsf{T}}\boldsymbol{b_i} \tag{11}$$

depends on an $n_i\times p$ fixed-effects design $\boldsymbol{X}_i$ and a $n_i\times q$ random-effects design $\boldsymbol{Z}_i$ through a link function $g(\cdot)$; $\boldsymbol{\xi}$ and $\boldsymbol{b_i}\sim N(\boldsymbol{0},\boldsymbol{D})$ are fixed and random effects, respectively. The relationship between mean and natural parameter is

$$\lambda_{ij}=h(\mu^c_{ij})=h(\theta_{ij}\kappa_{ij}). \tag{12}$$

The mean satisfies:

$$\mathrm{E}(Y_{ij})=\mathrm{E}(\theta_{ij})\,\mathrm{E}(\kappa_{ij})=\mathrm{E}[h^{-1}(\lambda_{ij})]. \tag{13}$$

Depending of the type of outcome under investigation, the distribution of θ_{ij} can be chosen appropriately.

It is computationally convenient, but not strictly necessary, to assume that the sets of random effects, $\boldsymbol{\theta}_i$ and $\boldsymbol{b}_i$, are independent. Kalema and Molenberghs (2015) and Kalema, Iddi, and Molenberghs (2016) relaxed this assumption. Regarding the components θ_{ij} of $\boldsymbol{\theta}_i$, three special cases are: (1) independence; (2) correlated, implying that the univariate distributions $\mathscr{G}_{ij}(\vartheta_{ij},\sigma^2_{ij})$ must be replaced with a multivariate one; and (3) equal (useful in applications with exchangeable outcomes Y_{ij}).

4.2. Strong conjugacy

It is of interest to explore under what conditions Model (8) still allows for conjugacy, now that normal random effects have been introduced into the linear predictor, leading to the multiplicative factor κ_{ij} in the mean structure. To this end, MVDV considered conjugacy conditional upon the normally-distributed random effect $\boldsymbol{b}_i$. Write in simplified notation:

$$f(y|\kappa\theta) = \exp\left\{\phi^{-1}[yh(\kappa\theta) - g(\kappa\theta)] + c(y,\phi)\right\}, \tag{14}$$

generalizing (3), and retain (4). Applying the transformation theorem to (4) leads to

$$f(\theta|\gamma,\psi) = \kappa \cdot f(\kappa\theta|\widetilde{\gamma},\widetilde{\psi}),$$

where $\widetilde{\gamma}$ and $\widetilde{\psi}$ are appropriate parameters. Next, we request that the parametric form (4) be maintained:

$$f(\kappa\theta) = \exp\left\{\gamma^*[\psi^* h(\kappa\theta) - g(\kappa\theta)] + c^{**}(\gamma^*,\psi^*)\right\}, \tag{15}$$

where the parameters γ^* and ψ^* follow from $\widetilde{\gamma}$ and $\widetilde{\psi}$ upon absorption of κ. Then, the marginal model, in analogy with (5), equals:

$$f(y|\kappa) = \exp\left\{c(y,\phi) + c^{**}(\gamma^*,\psi^*) + c^{**}\left(\phi^{-1}+\gamma^*, \frac{\phi^{-1}y+\gamma^*\psi^*}{\phi^{-1}+\gamma^*}\right)\right\}. \tag{16}$$

Not every model satisfying conjugacy in the sense of Section 3.2 allows for this form of conjugacy, referred to as *strong conjugacy*. Examples include the normal, Poisson, and Weibull (and hence exponential) models with normal, gamma, and gamma random effects, respectively. A counterexample is provided by the Bernoulli, and hence also binomial, model. Because the probit model does not allow for conjugacy, it is out of the picture here, too. The latter does not preclude the existence of closed forms in the probit case, as was shown by MVDV. These authors noted that strong conjugacy stems from the random-effects distribution, not from the data model. For example, they showed, for a gamma random effect:

$$\frac{1}{\kappa} f(\theta|\alpha,\beta) = f(\kappa\beta|\alpha,\kappa\beta), \tag{17}$$

and hence a scaled version of a gamma random effect is still a gamma random effect, with invariant α and re-scaled β.

Strong conjugacy facilitates the use of standard software, which does not imply that such software cannot be used once strong conjugacy does not hold. Arguably, the derivation of analytic quantities, such as moments, and hence means, variances, and covariances, is simplified when the property holds.

All CM can be formulated using the same general principles. One simply has to combine the models formulated in Table 1 with the GLMM (6) and corresponding linear predictor (7). The effect θ is then replaced by $\theta_{ij}\kappa_{ij}$, where κ_{ij} is defined by setting $\eta = \eta_{ij}$ equal to the linear predictor whence κ_{ij} is expressed, for the respective models, as μ, π, λ, and ϕ.

5. Count data

The model elements in this case are:

$$
\begin{aligned}
Y_{ij} &\sim \text{Poi}(\theta_{ij}\kappa_{ij}), & (18)\\
\kappa_{ij} &= \exp\left(\boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\xi} + \boldsymbol{z}_{ij}^{\mathsf{T}}\boldsymbol{b}_i\right), & (19)\\
\boldsymbol{b}_i &\sim N(\boldsymbol{0},\boldsymbol{D}), & (20)\\
\mathrm{E}(\boldsymbol{\theta}_i) = \mathrm{E}[(\theta_{i1},\ldots,\theta_{in_i})^{\mathsf{T}}] &= \boldsymbol{\vartheta}_i, & (21)\\
\mathrm{var}(\boldsymbol{\theta}_i) &= \boldsymbol{\Sigma}_i. & (22)
\end{aligned}
$$

This model has the same structure as the one by Booth et al. (2003). In the spirit of Table 1, the θ_{ij} can be assumed to follow a gamma model, producing, what we could term, a Poisson-gamma-normal model (PGN). Recall that $\boldsymbol{b}_i$ accommodates correlation and some overdispersion, while residual overdispersion is captured by the components θ_{ij} of $\boldsymbol{\theta}_i$. Should these components be assumed dependent, then both sets of random effects capture some correlation as well as some overdispersion. In the correlated case, a multivariate extension of the gamma distribution would be needed (see, for example, Gentle, 2003).

This model enjoys strong conjugacy, as shown by MVDV. Continuing on the work of Zeger, Liang, and Albert (1988), and using expressions for the standard Poisson moments (Johnson, Kemp, and Kotz, 2005, p. 162), MVD derived the moments; conditional upon the random effects are:

$$
\mathrm{E}(Y_{ij}^k) = \sum_{\ell=0}^{k} S(k,\ell)(\theta_{ij}\kappa_{ij})^{\ell}, \qquad (23)
$$

where $S(k,\ell)$ is the so-called Stirling number of the second kind. Integrating (23) over the random effects produces:

$$
\mathrm{E}(Y_{ij}^k) = \sum_{\ell=0}^{k} S(k,\ell)\frac{\beta^{\ell}\Gamma(\alpha+\ell)}{\Gamma(\alpha)}\exp\left[\ell\boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\xi} + \tfrac{1}{2}\ell^2\boldsymbol{z}_{ij}^{\mathsf{T}}\boldsymbol{D}\boldsymbol{z}_{ij}\right]. \qquad (24)
$$

The mean components are:

$$
\mu_{ij} = \phi_{ij}\exp\left(\boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\xi} + \tfrac{1}{2}\boldsymbol{z}_{ij}^{\mathsf{T}}\boldsymbol{D}\boldsymbol{z}_{ij}\right), \qquad (25)
$$

with the variance-covariance matrix

$$
\mathrm{var}(\boldsymbol{Y}_i) = \boldsymbol{M}_i + \boldsymbol{M}_i(\boldsymbol{P}_i - \boldsymbol{J}_{n_i})\boldsymbol{M}_i, \qquad (26)
$$

where $\boldsymbol{M}_i$ is a diagonal matrix with the μ_{ij} along the main diagonal, and the $(j,k)^{\text{th}}$ element of $\boldsymbol{P}_i$ equals

$$p_{i,jk} = \exp\left(\tfrac{1}{2}\boldsymbol{z}_{ij}^{\mathsf{T}}\boldsymbol{D}\boldsymbol{z}_{ik}\right) \cdot \frac{\sigma_{i,jk} + \phi_{ij}\phi_{ik}}{\phi_{ij}\phi_{ik}} \cdot \exp\left(\tfrac{1}{2}\boldsymbol{z}_{ik}^{\mathsf{T}}\boldsymbol{D}\boldsymbol{z}_{ij}\right). \tag{27}$$

MVD also derived a series-based expression for the marginal joint distribution:

$$\begin{aligned} P(\boldsymbol{Y}_i = \boldsymbol{y}_i) = \sum_{\boldsymbol{t}} & \left[\prod_{j=1}^{n_i} \begin{pmatrix} y_{ij}+t_j \\ y_{ij} \end{pmatrix} \cdot \begin{pmatrix} \alpha_j + y_{ij} + t_j - 1 \\ \alpha_j - 1 \end{pmatrix} \cdot (-1)^{t_j} \cdot \beta_j^{y_{ij}+t_j}\right] \\ & \times \exp\left(\sum_{j=1}^{n_i} (y_{ij}+t_j)\boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\xi}\right) \\ & \times \exp\left(\frac{1}{2}\left[\sum_{j=1}^{n_i}(y_{ij}+t_j)\boldsymbol{z}_{ij}^{\mathsf{T}}\right]\boldsymbol{D}\left[\sum_{j=1}^{n_i}(y_{ij}+t_j)\boldsymbol{z}_{ij}\right]\right). \end{aligned} \tag{28}$$

In the above equation, the vector-valued index $\boldsymbol{t} = (t_1, \ldots, t_{n_i})^{\mathsf{T}}$ ranges over all non-negative integer vectors.

In Section 9, the benefit of having closed-form expressions will show when deriving quantities such as marginal correlations.

Kalema and Molenberghs (2015) and Kalema, Iddi, and Molenberghs (2016) showed how the combined model formulation can be used to generate correlated count data. Neyens, Faes, and Molenberghs (2012) adapted the framework to accommodate overdispersion in counts that arise in a spatial context.

5.1. A clinical trial in epileptic patients

We will analyse the epilepsy data, introduced in Section 2.1. Let Y_{ij} represent the number of epileptic seizures 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 measured, $t_{ij} = 1, 2, \ldots$ until at most 27. Consider the combined model (18)–(22), with specific choices

$$\ln(\kappa_{ij}) = \begin{cases} (\xi_{00} + b_i) + \xi_{01} t_{ij} & \text{if placebo} \\ (\xi_{10} + b_i) + \xi_{11} t_{ij} & \text{if treated,} \end{cases} \tag{29}$$

where the random intercept b_i is assumed to be zero-mean normally distributed with variance d. We consider special cases (a) the ordinary Poisson model (P--), (b) the negative-binomial model (PG-), (c) the Poisson-normal model (P-N), together with (d)

***Table 2:** Epilepsy study. Parameter estimates (standard error) in (1) Poisson model (P--), (2) negative-binomial model (PG-), (3) Poisson-normal model P-N), and (4) combined model (PGN), as well as their zero-inflated counterparts ZI(P--), ZI(PG-), ZI(P-N), ZI(PGN).*

		Combined models		Negative-binomial models	
Effect	Par.	ZI(PGN)	(PGN)	ZI(PG-)	(PG-)
Interc. plac.	ξ_{00}	0.947(0.167)	0.911(0.176)	1.236(0.110)	1.259(0.0.112)
Slope plac.	ξ_{01}	−0.016(0.008)	−0.025(0.008)	−0.007(0.011)	−0.013(0.011)
Interc. treatm.	ξ_{10}	0.836(0.172)	0.656(0.178)	1.397(0.110)	1.475(0.109)
Slope treatm.	ξ_{11}	−0.006(0.007)	−0.012(0.008)	−0.022(0.011)	−0.035(0.010)
Neg.-bin. par.	α_1	0.245(0.025)	2.464(0.211)	1.787(0.100)	0.527(0.026)
SD non-zero part RE	$\sqrt{d_1}$	0.997(0.085)	1.063(0.087)	–	–
Infl. Interc.	γ_0	−4.581(0.641)	–	−7.106(1.334)	–
Infl. slope	γ_1	0.092(0.034)	–	0.292(0.066)	–
SD zero part RE	$\sqrt{d_2}$	2.533(0.440)	-	–	–
Corr. RE	ρ	−0.096(0.153)	–	–	–
Pred. prob. zeros		0.352	0.321	0.185	0.158
−2log-likelihood		5317.9	5417.0	6318.9	6326.1

		Poisson-normal models		Poisson models	
Effect	Par.	ZI(P-N)	(P-N)	ZI(P--)	(P--)
Interc. plac.	ξ_{00}	0.903(0.155)	0.818(0.168)	1.485(0.043)	1.266(0.0.042)
Slope plac.	ξ_{01}	−0.004(0.005)	−0.014(0.004)	−0.007(0.005)	−0.0.013(0.004)
Interc. treatm.	ξ_{10}	0.908(0.159)	0.648(0.170)	1.806(0.040)	1.453(0.038)
Slope treatm.	ξ_{11}	−0.007(0.005)	−0.012(0.004)	−0.025(0.014)	−0.033(0.004)
SD non-zero part RE	$\sqrt{d_1}$	0.971(0.082)	1.076(0.086)	–	–
Infl. Interc.	γ_0	−3.712(0.500)	–	−0.659(4.699)	–
Infl. slope	γ_1	0.095(0.025)	–	−3.291(4.444)	–
SD zero part RE	$\sqrt{d_2}$	2.222(0.343)	–	–	–
Corr. RE	ρ	−0.154(0.157)	–	–	–
Pred. prob. zeros		0.338	0.263	0.014	0.046
−2log-likelihood		5845.1	6271.9	10912	11590

the combined model (PGN). Estimates (standard errors) are presented in Table 2. The table also contains zero-inflated versions, that will be discussed in Section 5.2. Clearly, both the negative-binomial model and the Poisson-normal model are important improvements, in terms of the likelihood, relative to the ordinary Poisson model. This should come as no surprise since the latter unrealistically assumes there is neither overdispersion nor correlation within the outcomes, while clearly both are present. In addition, when considering the combined model, there is a very strong improvement in fit when gamma and normal random effects are simultaneously allowed for. This strongly affects the point and precision estimates of such key parameters as the slope difference and the slope ratio. There is also an impact on hypothesis testing. The Poisson model leads to unequivocal significance for both the difference ($p = 0.0008$) and ratio ($p = 0.0038$), whereas for the Poisson normal this is not the case for the difference of the slopes ($p = 0.7115$), while some significance is maintained for the ratio ($p = 0.0376$). Because the Poisson-normal is commonly used, it is likely that in practice one would decide in favor of a treatment effect when considering the slope ratio. This is no longer true with the negative-binomial model, where the p-values change to $p = 0.01310$ and $p = 0.2815$, respectively. Of course, one must not forget that, while the negative-binomial model accommodates overdispersion, the θ_{ij} random effects are assumed independent, implying independence between repeated measures. Again, this is not realistic and therefore the combined model is a more viable candidate, corroborated further by the aforementioned likelihood comparison. This model produces non-significant p-values of $p = 0.2260$ and $p = 0.1591$, respectively.

Thus, in conclusion, whereas the conventionally used and broadly implemented Poisson-normal model would suggest a significant effect of treatment, our combined model issues a message of caution, because there is no evidence whatsoever regarding a treatment difference.

Molenberghs and Verbeke (2005, Ch. 19), considered a (P-N) model with random intercepts as well as random slopes in time. It is interesting to note that, when allowing for such an extension in our models, the random slopes improve the fit of the (P-N) model with random intercept, but not of the combined one with random intercept (details not shown). As a consequence, the combined model with random intercept is the best fitting one. At the same time, note that fitting such a model establishes that the presence of a conjugate random effect does not preclude the consideration of normal random effects beyond random intercepts. The data were analysed by Booth et al. (2003), too.

Let us now turn to the correlation functions. Given that the gamma random effects are assumed independent, we only need to consider the Poisson-normal and combined cases; the versions with and without random slopes are considered. Because the fixed-effects structure is not constant but rather depends on time, MVD formulated a correlation function. In the (P-N) case with random intercepts only, and for the placebo group, based on the parameter estimates in Table 2, they obtained:

$$\mathrm{Corr}(Y(t),Y(s)) = \frac{35.58 \cdot 0.99^{t+s}}{\sqrt{(4.04 \cdot 0.99^{t} + 35.58 \cdot 0.97^{t}) \cdot (4.04 \cdot 0.99^{s} + 35.58 \cdot 0.97^{s})}},$$

where $Y(t)$ represents the outcome for an arbitrary subject at time t. Calculations in all other cases are similar. The smallest and largest values for the correlation functions, for both arms, for both the Poisson-normal and combined models, and for both choices of the random-effects structure are given in Table 3. When only random intercepts are considered, the correlations range over a narrow interval; they are rather high and there is little difference between the Poisson-normal and combined models. However, turning to the models with random intercepts and random slopes, several differences become apparent. First, the values exhibit a much broader range between their smallest and largest values. Second, the range is somewhat over-estimated by the Poisson-normal model, which then narrows when we switch to the combined model, thereby incorporating overdispersion effects, random intercepts, and random slopes. Thus, the random slope allows for the correlation to range over a considerable interval, while the overdispersion effect prevents the range from becoming overly wide.

Table 3: *Epilepsy study. Observed smallest and largest values for the correlation function, for the Poisson-normal and combined models, and for both treatment arms. The time pair for which the values are observed is shown too. (RI: random intercept; RS: random slope.)*

		Smallest value		Largest value	
Model	Arm	ρ	time pair	ρ	time pair
Poisson-normal, RI	placebo	0.8577	26 & 27	0.8960	1 & 2
Poisson-normal, RI	treatment	0.8438	26 & 27	0.8794	1 & 2
Combined, RI	placebo	0.8259	26 & 27	0.8981	1 & 2
Combined, RI	treatment	0.8383	26 & 27	0.8744	1 & 2
Poisson-normal, RI+RS	placebo	0.2966	1 & 27	0.9512	26 & 27
Poisson-normal, RI+RS	treatment	0.2936	1 & 27	0.9530	26 & 27
Combined, RI+RS	placebo	0.4268	1 & 27	0.9281	26 & 27
Combined, RI+RS	treatment	0.4225	1 & 27	0.9329	26 & 27

Within each model, there is relatively little difference between the placebo and treated groups, although the difference is a bit more pronounced in the combined model. Further, the correlation range within every group is relatively narrow. The most noteworthy feature, unquestionably, is the large discrepancy between both models. This is because the (P-N) model forces the correlation and overdispersion effects to stem from a single additional parameter, the random-intercept variance d. Thus, considerable overdispersion also forces the correlation to increase, arguably beyond what is consistent with the data. In the combined model, in contrast, there are *two* additional parameters, giving proper justice to both correlation and overdispersion effects. It was already clear from the above discussion and that in MVD that the combined model is an important improvement. This now clearly manifests itself in the correlation function, too.

The above underscores the need for the combined model. Some indication came, for example, from the correlation functions in the epilepsy case. It is useful to perform formal comparison of all nested models, using Wald statistics, for each of the three cases. A summary is given in Table 4. Note that, owing to the familiar boundary problem that occurs when testing for variance components, mixtures of a χ^2_0 and χ^2_1 were used, instead of the conventional χ^2_1 (Molenberghs and Verbeke, 2007).

Table 4: *Epilepsy, onychomycosis, and asthma studies. Wald test results for comparison of nested models.*

Null model	Alternative model	Z-value	p-value
	Epilepsy study		
Poisson	Negative-binomial	20.68	<0.0001
Poisson	Poisson-normal	6.27	<0.0001
Negative-binomial	Combined	6.10	<0.0001
Poisson-normal	Combined	11.66	<0.0001
	Onychomycosis study		
Logistic	Beta-binomial	17.91	<0.0001
Logistic	Logistic-normal	10.53	<0.0001
Beta-binomial	Combined	4.28	<0.0001
Logistic-normal	Combined	8.01	<0.0001
	Asthma study		
Exponential	Exponential-gamma	8.54	<0.0001
Exponential	Exponential-normal	10.63	<0.0001
Exponential-gamma	Combined	8.54	<0.0001
Exponential-normal	Combined	3.99	<0.0001

For our case study, it is clear that: (a) independence is strongly rejected in favour of both a model with normal random effects or a model with conjugate random effects; (b) on top of one set of random effects, there is a clear need for the other set as well, hence providing very strong evidence for the proposed combined model. The evidence is extremely convincing. The table also contains results for two more case studies that will be discussed in detail in subsequent sections.

These findings, taken together, imply that the data exhibit, at the same time, within-subject correlation and overdispersion, in such a way that a single model feature cannot capture both simultaneously.

5.2. Additional zeroes

It is not uncommon when count data are collected to observe more zeroes than predicted by the model assumed, whether of a simple Poisson nature, or more elaborate, such as the combined model considered here. This feature, often referred to as zero inflation, then needs to be accommodated, in addition to correlation and/or overdispersion. Such data are often fitted by using either hurdle (Mullahy, 1986; Greene, 1994) or zero-inflated models (ZI; Lambert, 1992). In the context of the CM, additional zeroes were studied by Kassahun et al. (2014a) and Iddi and Molenberghs (2013).

We will first describe the hurdle (H) and zero-inflation (ZI) approaches for univariate data, and then turn to hierarchical versions. The hurdle model is a two-part model, whereby the first part is a binary model for the count value to be either zero or positive. Given that the value is positive, a count distribution, say f_i, is truncated at zero and fitted to the second part. Suppose Y_i is a univariate count outcome, and π_i is the probability of the i^{th} observation to be in the zero state. The hurdle model then takes the form:

$$p(\boldsymbol{Y}_i=\boldsymbol{y}_i)=\begin{cases}\pi_i & \text{if } y_i=0,\\ (1-\pi_i)\frac{f_i(y_i|\lambda_i)}{1-f_i(0|\lambda_i)} & \text{if } y_i>0.\end{cases} \tag{30}$$

An alternative approach is a zero-inflated model, which assumes zeros to come from two processes. The first process generates only zeros with probability π_i for observation i, say, while the second process generates counts with probability $1-\pi_i$. The ZI model is:

$$p(\boldsymbol{Y}_i=\boldsymbol{y}_i)=\begin{cases}\pi_i+(1-\pi_i)f_i(0|\lambda_i) & \text{if } y_i=0,\\ (1-\pi_i)f_i(y_i|\lambda_i) & \text{if } y_i>0.\end{cases} \tag{31}$$

Here, π_i and λ_i are functions of covariates. Link functions, such as the logit or probit, can be used for π_i, with the log link commonly used for λ_i.

Kassahun et al. (2014a) extended the combined model to take zero-inflation into account. The ZI version of the CM (ZICOM) is given by

$$p(Y_{ij}=y_{ij}|\boldsymbol{b}_{1i},\boldsymbol{\xi},\theta_{ij},\pi_{ij})=\begin{cases}\pi_{ij}+(1-\pi_{ij})f_i(0|\boldsymbol{b}_{1i},\boldsymbol{\xi},\theta_{ij}) & \text{if } y_{ij}=0,\\ (1-\pi_{ij})f_i(y_{ij}|\boldsymbol{b}_{1i},\boldsymbol{\xi},\theta_{ij}) & \text{if } y_{ij}>0.\end{cases} \tag{32}$$

The ZI component $\pi_{ij}=\pi(\boldsymbol{x}_{2ij}^{\top}\boldsymbol{\gamma}+\boldsymbol{z}_{2ij}^{\top}\boldsymbol{b}_{2i})$ is modelled 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 4 are now replaced by $\boldsymbol{x}_{1ij}$, $\boldsymbol{z}_{1ij}$, and $\boldsymbol{b}_{1ij}$, 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 $\boldsymbol{z}_{1ij}=\boldsymbol{z}_{2ij}=1$. The variance-covariance matrix of the random effects, assumed normally distributed, is denoted by D, as before. The model is denoted as ZI(PGN), as an obvious extension with earlier notational conventions. Three obvious special cases are ZI(P-N), ZI(PG-), and ZI(P--). Also, all four models without zero inflation are special cases as well. The conditional mean and variance of the ZI(PGN) are:

$$\mathrm{E}(Y_{ij}|\boldsymbol{b}_{1i},\boldsymbol{\xi},\theta_{ij})=\theta_{ij}\kappa_{ij}(1-\pi_{ij}), \tag{33}$$

$$\mathrm{Var}(Y_{ij}|\boldsymbol{b}_{1i},\boldsymbol{\xi},\theta_{ij})=\theta_{ij}\kappa_{ij}(1-\pi_{ij})[1+\theta_{ij}\kappa_{ij}(\pi_{ij}+1/\alpha)]. \tag{34}$$

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 ZI (parameter π_{ij}), or both.

Further model developments that allow for extra zeroes are reported in Sections 10 and 11.

5.3. A clinical trial in epileptic patients

We re-analyse the epilepsy data, introduced in Section 2.1 and analysed before in Section 5.1. 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 parameterization (29), but now accounting for zero inflation, assuming that counts are generated from a (P-N) process with λ_{ij} as in (29), or from a (PGN) process with mean $\lambda_{ij} = \theta_{ij}\kappa_{ij}$, and now κ_{ij} specified as in (29). The ZI probability (π_{ij}) is modelled as $\text{logit}(\pi_{ij}) = \gamma_0 + b_{2i} + \gamma_1 t_{ij}$. The data are analysed with the ZI(PGN), ZI(PG-), ZI(P-N), ZI(P--). One can compare the results with the non-ZI counterpart. Parameter estimates and predicted probabilities of zeros are presented in Table 2, alongside the non-ZI counterparts. Clearly, in terms of likelihood comparison, the zero-inflated versions performed much better, resulting in a substantial improvement in fit.

The ZI(PG-) is an important improvement relative to the ZI(P--), while much more improvement is gained in the case of the ZI(P-N). Moreover, the ZI(PGN) 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(PGN), 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(P-N) 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 ($p = 0.004$). 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 except the ZI(P--). Evidently, models can be extended further. For example, one could consider a version with where the ZI component is specific to treatment arm.

6. Categorical data

Categorical data come in various forms, and we usefully distinguish between them. Building on MVDV, Molenberghs et al. (2012) laid out the combined-model framework and various ramifications for the binary and binomial cases. An overview will be given in Sections 6.1 and 6.2 for the binary cases with logit and probit links, respectively, and in Section 6.3 for binomial data. The iron deficiency case study is analysed in Section 6.4. An application of the binary version of the model to the Jimma Infant study was reported in Kassahun et al. (2012). A binomial application is described in Del Fava et al. (2014). Ivanova, Molenberghs, and Verbeke (2014) developed a version of the combined model to handle ordinal data, which is the basis for Section 6.5.

6.1. Bernoulli-type models for binary data with logit link

Similar to the Poisson case in Section 5, a natural binary-data counterpart to (18)–(19) is

$$Y_{ij} \sim \text{Bernoulli}(\pi_{ij} = \theta_{ij}\kappa_{ij}), \tag{35}$$

$$\kappa_{ij} = \frac{\exp\left(\boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\xi} + \boldsymbol{z}_{ij}^{\mathsf{T}}\boldsymbol{b}_i\right)}{1+\exp\left(\boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\xi} + \boldsymbol{z}_{ij}^{\mathsf{T}}\boldsymbol{b}_i\right)}, \tag{36}$$

completing the specification with (20)–(22). Unlike in the Poisson case, closed forms for neither the mean nor the variance follow when normal random effects are present. When only overdispersion random effects are included, especially when they are assumed to follow a beta distribution, as in Table 1, conjugacy applies. However, the beta distribution does not allow for the multiplicative invariance as (17), precluding strong conjugacy.

When the overdispersion random effects are assumed to be equal: $\theta_{ij} = \theta_i$, then the beta-binomial model follows if no normal random effects are present.

Explicitly considering $\theta_{ij} \sim \text{Beta}(\alpha,\beta)$, then $\phi_{ij} = \text{E}(\theta_{ij}) = \alpha/(\alpha+\beta)$, and

$$\sigma_{ij}^2 = \text{var}(\theta_{ij}) = \sigma_{i,jj} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)},$$

$$\sigma_{i,jk} = \text{cov}(\theta_{ij},\theta_{ik}) = \rho_{ijk}\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Observe that there are two correlations: ρ_{ijk}, which described the correlation between draws from the beta distribution and $(\alpha+\beta+1)^{-1}$. It is of course possible to let α and β vary with *i* and/or *j*. In such cases, the above and below expressions will change somewhat, but computations are straightforward.

Using the general expressions, the above results can be used to derive approximate expressions for means and variance-covariance elements. For the special case of no normal random effects, but maintaining the fixed effects in (36), i.e.,

$$\kappa_{ij} = \frac{\exp\left(\boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\xi}\right)}{1+\exp\left(\boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\xi}\right)}, \tag{37}$$

we obtain

$$\begin{aligned} \mathrm{E}(Y_{ij}) &= \frac{\alpha}{\alpha+\beta}\kappa_{ij}, \qquad (38) \\ \mathrm{Var}(Y_{ij}) &= \frac{\alpha}{\alpha+\beta}\kappa_{ij} - \left(\frac{\alpha}{\alpha+\beta}\right)^2 \kappa_{ij}^2, \\ \mathrm{Cov}(Y_{ij},Y_{ik}) &= \rho_{ijk}\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}\kappa_{ij}\kappa_{ik}. \end{aligned}$$

If we further make exchangeability assumptions, i.e., $\kappa_{ij} = \kappa_{ik} \equiv \kappa_i$ and $\rho_{ijk} = \rho_i$, further simplification follows. Finally, setting $\kappa_i = 1$, the conventional beta-binomial follows. It is then easy to derive the resulting binomial version by defining:

$$Z_i = \sum_{i=1}^{n_i} Y_{ij}. \tag{39}$$

Simple algebra then shows:

$$\begin{aligned} \mathrm{E}(Z_i) &= n_i\frac{\alpha}{\alpha+\beta} = n_i\pi_i, \\ \mathrm{Var}(Z_i) &= n_i\frac{\alpha\beta}{(\alpha+\beta)^2}\left\{1+(n_i-1)\frac{1}{\alpha+\beta+1}\right\} = n_i\pi_i(1-\pi_i)\left\{1+(n_i-1)\widetilde{\rho}_i\right\}, \end{aligned}$$

with $\widetilde{\rho}_i$ the beta-binomial correlation. Hence, the conventional beta-binomial model follows.

While the logit link defeats closed-form expressions when normal random effects are introduced, this is different with the probit link. The random-effects probit model has received some attention in earlier decades (Schall, 1991; Guilkey and Murphy, 1993; Hedeker and Gibbons, 1994; McCulloch, 1994; Gibbons and Hedeker, 1997; Renard, Molenberghs, and Geys 2004).

6.2. Bernoulli-type models for binary data with probit link

Introducing the probit version of the model, while at the same time assuming that the overdispersion parameters are beta distributed, comes down to:

$$\kappa_{ij} = \Phi_1(\boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\xi} + z_{ij}^{\mathsf{T}}\boldsymbol{b}_i), \tag{40}$$

$$\theta_{ij} \sim \text{Beta}(\alpha, \beta). \tag{41}$$

Like before, α and β could be allowed to vary with i and/or j.

It now follows that the joint distribution can be written as (see MVDV):

$$f_{n_i}(\boldsymbol{y}_i = \boldsymbol{1}) = \left(\frac{\alpha}{\alpha+\beta}\right)^{n_i} \cdot \Phi_{n_i}(X_i\boldsymbol{\xi}; \boldsymbol{L}_{n_i}^{-1}), \tag{42}$$

with

$$\boldsymbol{L}_{n_i} = \boldsymbol{I}_{n_i} - \boldsymbol{Z}_i\left(\boldsymbol{D}^{-1} + \boldsymbol{Z}_i^{\mathsf{T}}\boldsymbol{Z}_i\right)^{-1}\boldsymbol{Z}_i^{\mathsf{T}}. \tag{43}$$

Note that (42) is the joint probability only for the outcome $(1,\dots,1)^{\mathsf{T}}$, a so-called success probability. However, given that the dimension n_i is arbitrary, all other probabilities can be derived by appropriate contrasts of success probabilities. Precisely,

$$f_{n_i}[\boldsymbol{y}_i = \boldsymbol{m}_i = (m_{i1},\dots,m_{in_i})^{\mathsf{T}}] = \sum_{\boldsymbol{s}\supset\iota(\boldsymbol{m}_i)} \text{sgn}(\boldsymbol{s})\Phi_{\#\boldsymbol{s}}\left(\widetilde{X}_i^{(\boldsymbol{s})}\boldsymbol{\xi}; \boldsymbol{L}_{(\boldsymbol{s})}^{-1}\right) \cdot \left(\frac{\alpha}{\alpha+\beta}\right)^{\#\boldsymbol{s}}, \tag{44}$$

with $\iota(\boldsymbol{m}_i) = \lambda(m_{i1},\dots,m_{in_i})$ the set of places for which $m_{ij} = 1$,

$$\text{sgn}(\boldsymbol{s}) = \begin{cases} 1 & \text{if } \#\boldsymbol{s} - \#\iota(\boldsymbol{m}_i) \text{ is even,} \\ 0 & \text{otherwise,} \end{cases}$$

$\widetilde{X}_i^{(\boldsymbol{s})}$ contains the rows from X_i with row number in $\boldsymbol{s}$, and $\boldsymbol{L}_{(\boldsymbol{s})}$ is the $\#s$-dimensional matrix built from the appropriate sub-matrices of these used in (43). The above developments straightforwardly generalize when (41) is replaced with $\theta_{ij} \sim \text{Beta}(\alpha_j, \beta_j)$.

Next, the means, variances, and covariances can be derived from (42), by evaluating it for the one- and two-dimensional cases. We find:

$$\text{E}(Y_{ij}) = \frac{\alpha}{\alpha+\beta} \cdot \Phi_1(\boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\xi}; \boldsymbol{L}_1^{-1}) = \frac{\alpha}{\alpha+\beta} \cdot \Phi_1(|\boldsymbol{I} + \boldsymbol{D}z_{ij}z_{ij}^{\mathsf{T}}|^{-1/2}\boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\xi}), \tag{45}$$

$$\text{Var}(Y_{ij}) = \frac{\alpha}{\alpha+\beta} \cdot \Phi_1(\boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\xi}; \boldsymbol{L}_1^{-1}) \cdot \left[1 - \frac{\alpha}{\alpha+\beta}.\Phi_1(\boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\xi}; \boldsymbol{L}_1^{-1})\right], \tag{46}$$

$$\mathrm{Cov}(Y_{ij},Y_{ik}) = \left(\frac{\alpha}{\alpha+\beta}\right)^2 \cdot \left\{ \Phi_2\left[\begin{pmatrix} \boldsymbol{x}_{ij}^{\mathsf{T}} \\ \boldsymbol{x}_{ik}^{\mathsf{T}} \end{pmatrix}\boldsymbol{\xi}, \boldsymbol{L}_{2jk}^{-1}\right] - \Phi_1(\boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\xi};\boldsymbol{L}_{1j}^{-1})\Phi_1(\boldsymbol{x}_{ik}^{\mathsf{T}}\boldsymbol{\xi};\boldsymbol{L}_{1k}^{-1}) \right\}, \tag{47}$$

where

$$\boldsymbol{L}_{2jk} = \boldsymbol{I}_2 - \begin{pmatrix} z_{ij}^{\mathsf{T}} \\ z_{ik}^{\mathsf{T}} \end{pmatrix} \left[\boldsymbol{D}^{-1} + \begin{pmatrix} z_{ij}^{\mathsf{T}} \\ z_{ik}^{\mathsf{T}} \end{pmatrix} (z_{ij}\ z_{ik}) \right]^{-1} (z_{ij}\, z_{ik}).$$

The rightmost density in (45) is the standard normal one. Evidently, (42) and (44) lead, not only to the mean, variance, and covariance expressions, but also to the higher-order moments.

MVDV noted that the existence of closed-form expressions for the probit case opens a window of opportunity for the logit case. Indeed, the well-known approximation formulae, linking the normal and logistic densities, prove useful here. As shown in Johnson and Kotz (1970, p. 6) and used in Zeger et al. (1988):

$$\frac{e^y}{1+e^y} \approx \Phi_1(cy), \tag{48}$$

with $c = (16\sqrt{3})/(15\pi)$. Applied to (35)–(36), it follows that

$$\pi_{ij} \sim \theta_{ij}\frac{\exp\left(\boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\xi} + z_{ij}^{\mathsf{T}}\boldsymbol{b}_i\right)}{1+\exp\left(\boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\xi} + z_{ij}^{\mathsf{T}}\boldsymbol{b}_i\right)} \approx \theta_{ij}\Phi_1[c(\boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\xi} + z_{ij}^{\mathsf{T}}\boldsymbol{b}_i)]. \tag{49}$$

Applying (49) to (42), yields

$$f_{n_i}(\boldsymbol{y}_i = \boldsymbol{1}) \approx \left(\frac{\alpha}{\alpha+\beta}\right)^{n_i} \cdot \Phi_{n_i}\left(cX_i\boldsymbol{\xi};\widetilde{\boldsymbol{L}}_{n_i}^{-1}\right), \tag{50}$$

with

$$\widetilde{\boldsymbol{L}}_{n_i} = \boldsymbol{I}_{n_i} - c^2\boldsymbol{Z}_i\left(\boldsymbol{D}^{-1} + \boldsymbol{Z}_i^{\mathsf{T}}\boldsymbol{Z}_i\right)^{-1}\boldsymbol{Z}_i^{\mathsf{T}}.$$

For the expectation, we find, based on (49) and (45):

$$\mathrm{E}(Y_{ij}) \approx \frac{\alpha}{\alpha+\beta} \cdot \Phi_1\left(|\boldsymbol{I} + c^2\boldsymbol{D}z_{ij}z_{ij}^{\mathsf{T}}|^{-1/2}c\boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\xi}\right), \tag{51}$$

with similar expressions for the variance and covariance terms. Upon estimating the parameters within the probit approximation paradigm, back-transformation to the original logit scale is possible, using expressions such as (49) and (51). This opens perspectives for alternative estimation methods for the combined model with logit link, with the important special case of the normal-logistic GLMM.

In the Bernoulli case, calculating the moments is extremely simple. Indeed, the Bernoulli moments are all identical. The conditional moments are all $\mathrm{E}(Y_{ij}^k|\theta_{ij},\boldsymbol{b}_i) = \theta_{ij}\kappa_{ij}$ ($k = 1,2,\dots$). Hence, they all reduce to (38). In the probit case, they are equal to (45).

6.2.1. A clinical trial in onychomycosis

We present the MVDV analysis of the binary onychomycosis data, introduced in Section 2.2. For the logit, consider the model:

$$\begin{aligned} Y_{ij}|(b_i) &\sim \mathrm{Bernoulli}(\pi_{ij}), \\ \mathrm{logit}(\pi_{ij}) &= \xi_1(1-T_i) + b_i + \xi_2(1-T_i)t_{ij} + \xi_3 T_i + \xi_4 T_i t_{ij}, \end{aligned} \tag{52}$$

where T_i is the treatment indicator for subject i, t_{ij} is the time-point at which the jth measurement is taken for the ith subject, and $b_i \sim N(0,d)$. Parameter estimates for the logistic model, with and without the normal random effect on the one hand, and with and without the beta-binomial component on the other hand, as described in Section 6.1, are presented in Table 5. Observe that the model becomes hard to fit when the beta random

Table 5: *Onychomycosis study. Parameter estimates (standard errors) for the regression coefficients in (1) the logistic model, (2) the beta-binomial model, (3) the logistic-normal model, and (4) the combined model. Estimation was done by maximum likelihood using numerical integration over the normal random effect, if present.*

Effect	Par.	Logistic	Beta-binomial
Intercept treatment A	ξ_0	−0.5571 (0.1090)	17.9714 (1482.6)
Slope treatment A	ξ_1	−0.1769 (0.0246)	5.2454 (12970.0)
Intercept treatment B	ξ_2	−0.5335 (0.1122)	18.6744 (2077.13)
Slope treatment B	ξ_3	−0.2549 (0.0309)	4.7775 (12912.0)
Std. dev random effect	$\sqrt{d}$	—	—
Ratio	α/β	—	3.6739 (0.2051)
−2log-likelihood		1812	1980
Effect	Par.	Logistic-normal	Combined
Intercept treatment A	ξ_0	−1.6299 (0.4354)	−1.6042 (4.0263)
Slope treatment A	ξ_1	−0.4042 (0.0460)	−6.4783 (1.4386)
Intercept treatment B	ξ_2	−1.7486 (0.4478)	−16.2079 (3.5830)
Slope treatment B	ξ_3	−0.5634 (0.0602)	−8.0745 (1.5997)
Std. dev random effect	$\sqrt{d}$	4.0150 (0.3812)	60.8835 (14.2237)
Ratio	α/β	—	0.2805 (0.0350)
−2log-likelihood		1248	1240

effects are present, which is seen from estimates and standard errors in both the beta-binomial model as well as the combined model. To understand this, we must observe that the conjugate random effects in the Bernoulli case, unlike in the Poisson, binomial, and Weibull cases, cannot add to the variability, only to the correlation structure. This means that there is considerably less information available than in the other cases. This does not mean that the beta random effects are unnecessary, but rather that they challenge the stable estimation of other model parameters.

6.3. Models for binomial data with logit and probit link

Molenberghs et al. (2012) supplemented the study of the binary case with the binomial one. Starting from the Bernoulli expressions (35) and (36) but now for three rather than two levels, they got:

$$Y_{ijk} \sim \text{Bernoulli}(\pi_{ijk} = \theta_{ijk}\kappa_{ijk}), \tag{53}$$

$$\kappa_{ijk} = \frac{\exp\left(\boldsymbol{x}_{ijk}^{\mathsf{T}}\boldsymbol{\xi} + z_{ijk}^{\mathsf{T}}\boldsymbol{b}_i\right)}{1+\exp\left(\boldsymbol{x}_{ijk}^{\mathsf{T}}\boldsymbol{\xi} + \boldsymbol{z}_{ijk}^{\mathsf{T}}\boldsymbol{b}_i\right)}, \tag{54}$$

where i stands for the independent block, as before, j for occasion, and k for the repeats of the Bernoulli trials. It is natural to define $Z_{ij} = \sum_{k=1}^{m_{ij}} Y_{ijk}$. Also here, there are no closed-form expressions for the moments when a logit link is used, but they do exist for the probit case. The data consists of an array of successes $z_i = (z_{i1},\ldots,z_{in_i})^{\mathsf{T}}$ out of $\boldsymbol{m}_i = (m_{i1},\ldots,m_{in_i})^{\mathsf{T}}$ trials. It is also convenient to provide for multi-indices $\boldsymbol{t} = (t_1,\ldots,t_{n_i})^{\mathsf{T}}$ and for vectors of the parameters $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_{n_i})^{\mathsf{T}}$ and $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_{n_i})^{\mathsf{T}}$. The joint distribution can then be written as:

$$f(z_i|\boldsymbol{m}_i,\boldsymbol{\xi},\boldsymbol{D},\boldsymbol{\alpha},\boldsymbol{\beta}) = \sum_{\boldsymbol{t}=\boldsymbol{0}}^{\boldsymbol{m}_i - z_i}\left[\prod_{j=1}^{n_i}\frac{(-1)^{t_j}}{B(\alpha_j,\beta_j)}\begin{pmatrix} m_{ij} \\ z_{ij}\end{pmatrix}\begin{pmatrix} m_{ij}-z_{ij} \\ t_j\end{pmatrix}B(z_{ij}+\alpha_j+t_j,\beta_j)\right]\times$$

$$\times\,\Phi_{\sum_j t_j}\left[(X_i(t)\boldsymbol{\xi};\boldsymbol{L}(t)^{-1}\right]. \tag{55}$$

Here, $\boldsymbol{X}_i(t)$ is the design matrix, built from X_i, with row j in X_i replicated t_j times. The design matrix X_i is built similarly, and then, in analogy with (43),

$$\boldsymbol{L}(t) = \boldsymbol{I}_{\sum_j t_j} - \boldsymbol{Z}_i(t)\left[\boldsymbol{D}^{-1} + \boldsymbol{Z}_i(t)^{\mathsf{T}}\boldsymbol{Z}_i(t)\right]^{-1}\boldsymbol{Z}_i(t)^{\mathsf{T}}. \tag{56}$$

6.4. Iron-deficient diets in rats

We turn to the data in Section 2.3. Because the probability of a fetus dying varies from litter to litter, the total variance of the proportions will be greater than that predicted by a binomial model, even when covariates are accounted for. Hence, overdispersion and correlation need to be accommodated.

Construct predictor function $\eta_i = \xi_0 + \xi_2 x_{2i} + \xi_3 x_{3i} + \xi_4 x_{4i}$ with $x_{g_i} = 1$ if litter i belongs to group g and 0 otherwise. The placebo group figures as a reference category. Further, let $Z_i = \sum_{j=1}^{n_i} Y_{ij} \sim \text{Binomial}(n_i, \pi_i)$ be the number of dead fetuses out of n_i in litter i. Five models are considered: (a) the binomial model, $\text{logit}(\pi_i) = \eta_i$; (b) the GLMM: $\text{logit}(\pi_i) = \eta_i + b_i$, where $b_i \sim N(0,d)$; (c) the beta-binomial model, $\text{logit}(\mu_i) = \eta_i$, where $\pi_i \sim \text{Beta}(\alpha,\beta)$, and $\mu_i = \text{E}(\pi_i)$; (d) the beta-binomial model with normal random effects: for $b_i \sim N(0,d)$, $\text{logit}(\mu_i) = \eta_i$, and π_i and μ_i as in the beta-binomial; (e) in the combined model: $\text{logit}(\kappa_i) = \eta_i + b_i$ where $\pi_i = \theta_i \kappa_i$, $\theta_i \sim \text{Beta}(\alpha,\beta)$, and $b_i \sim N(0,d)$. The constraint $\alpha\beta \equiv 1$ is imposed in the latter case.

The results of the various models are presented in Table 6. We observe that the two models that simultaneously account for overdispersion and correlation perform better than the others. The classical beta-binomial model with normal random effects has the same double negative log-likelihood as the combined model. This is the case only for cross-sectional data; even though their hierarchical formulations are different, they marginally coincide in this case. That said, the parameters have a different meaning, as they are to be interpreted conditionally on the assumed random-effects structure. Differences may be very noticeable when binomial measurements are collected repeatedly over time or in an otherwise hierarchical fashion.

Between these two, the estimates' precision is best in the combined model. Owing to conjugacy, the mean model and overdispersion parameter estimators are less correlated, leading to increased precision, even though the effect is modest.

Table 6: *Iron-deficiency study. Parameter estimates (standard errors) for (1) the binomial model, (2) the GLMM, (3) the beta-binomial model, (4) the conventional beta-binomial model with random effect in the linear predictor, and (5) the combined model.*

Effect	Par.	Binomial	GLMM	BB	BB-normal	Combined
Intercept	ξ_0	1.14(0.13)	1.80(0.36)	1.35(0.25)	1.79(0.38)	1.80(0.36)
Group2	ξ_2	−3.32(0.33)	−4.52(0.74)	−3.11(0.50)	−4.49(0.80)	−4.51(0.74)
Group3	ξ_3	−4.48(0.73)	−5.86(1.19)	−3.87(0.81)	−5.81(1.30)	−5.85(1.19)
Group4	ξ_4	−4.13(0.48)	−5.60(0.92)	−3.93(0.67)	−5.57(0.97)	−5.59(0.92)
Std. dev. RE	$\sqrt{d}$	—	1.54(0.29)	—	1.52(0.37)	1.53(0.29)
Overdispersion		—	—	0.24(0.06)	0.005(0.051)	0.0005(0.0018)
−2log-likelihood		244.9	183.9	186.9	183.8	183.8

6.5. Ordinal data: a combined proportional odds-beta-normal model

The ordinal case was studied by Ivanova et al. (2014). Assume the ordinal outcome Y_{ij} can take values $r = 1,\dots,R$, and replace it by a set of R dummies:

$$Z_{r,ij} = \begin{cases} 1 & \text{if } Y_{r,ij} = r, \\ 0 & \text{otherwise,} \end{cases}$$

for $r = 1,\dots,R$. Evidently, there are redundant dummies, but any subset of $R-1$ components is not. Group the dummies into vectors $\boldsymbol{Z}_{ij}$ and $\boldsymbol{Z}_i$ for a specific subject i and occasion j, and for a specific subject i, respectively. We assume a multinomial distribution $\boldsymbol{Z}_{ij} \sim \text{multinomial}(\boldsymbol{\pi}_{ij})$, with $\boldsymbol{\pi}_{ij} = (\pi_{1,ij},\dots,\pi_{r,ij},\dots,\pi_{R,ij})$. The multinomial distribution at a given occasion is determined by the modelling choice for the ordinal outcome. Under a proportional odds assumption, using normal random effects $\boldsymbol{b}_i \sim N(0,D)$ in the linear predictor, and beta random effects $\theta_{ij} \sim \text{Beta}(\alpha_j,\beta_j)$ to capture further overdispersion, the probabilities can be written as:

$$\pi_{r,ij} = \begin{cases} \theta_{ij}\kappa_{1,ij} & \text{if } r = 1, \\ \theta_{ij}(\kappa_{r,ij} - \kappa_{r-1,ij}) & \text{if } 1 < r < R, \\ 1 - \theta_{ij}\kappa_{R-1,ij} & \text{if } r = R. \end{cases} \tag{57}$$

where

$$\kappa_{r,ij} = \frac{\exp\left(\xi_{0r} + \boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\xi} + \boldsymbol{z}_{ij}^{\mathsf{T}}\boldsymbol{b}_i\right)}{1 + \exp\left(\xi_{0r} + \boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\xi} + \boldsymbol{z}_{ij}^{\mathsf{T}}\boldsymbol{b}_i\right)}. \tag{58}$$

Here, $\xi_{01} \le \cdots \le \xi_{0,R-1}$ are intercepts, $\boldsymbol{\xi}$ are fixed regression coefficients, and $\boldsymbol{x}_{ij}$ ($\boldsymbol{z}_{ij}$) is the design vector for the fixed (random) effects at occasion j. Also here, some choices in the above can be relaxed and/or altered. For example, like before, the α_j and β_j parameters, describing the beta distribution, need not be dependent on j. To ensure identifiability, a constraint needs to be applied to it, e.g., $\alpha_j\beta_j = 1$, but it is mathematically convenient to retain them as two separate parameters, with the understanding that the constraint does apply. Finally, the θ_{ij} within a subject are assumed different from each other and independent. One could allow them to be correlated, or even constant across subjects. This will not be considered here.

As argued in MVDV, MVID, and Molenberghs et al. (2012), closed-form expressions for marginal means, variances, covariances, and even the entire marginal distribution, i.e., integrated over both sets of random effects, cannot be derived in the binary case with logit link and normal random effects (regardless of the overdispersion random effects). Evidently, the same will be true for the ordinal case. If necessary, numerical integration or other Monte Carlo methods can be used to derive such marginal quantities.

6.6. Diabetes study

We describe the analysis of the diabetes study (Section 2.4), reported in Ivanova et al. (2014). Let $Y_{ij} = 0, \ldots, 3$ be the number of clinical targets patient i reached at occasion j. Also, let $t_{ij} = 0, 1$ be the time point at which the jth measurement was taken. Consider the combined proportional odds logistic regression model:

$$\text{logit}[P(Y_{ij} \leq r | t_{ij}, X_i)] = \xi_{0r} + b_i + \xi_1 t_{ij} + \xi_2 X_i,$$

$(r = 0, \ldots, 3)$, where the random intercept b_i is assumed $N(0, d)$ distributed, and X_i is an indicator for group. The beta random effect is re-parameterized such that

$$\nu = \frac{e^{\delta}}{1 + e^{\delta}} = \frac{\alpha}{\alpha + \beta},$$

thus simultaneously avoiding identifiability and range violation issues. The parameter δ is the one entered into the likelihood function. We consider (1) the ordinary proportional odds model, (2) the proportional odds model with beta overdispersion effect, (3) the proportional odds model with random normal effect, and (4) the combined model. Estimates (standard errors) are presented in Table 7. Clearly, there is no significant im-

Table 7: *Diabetes study. Parameter estimates (standard errors) from the regression coefficients in (1) the ordinary proportional odds model, (2) the proportional odds model with beta overdispersion effect, (3) the proportional odds model with random normal effect, together with (4) the combined model. Estimation was done by maximum likelihood using numerical integration over the normal random effect, if present.*

Effect	Par.	PO	PO-Beta
Intercept 0	ξ_{00}	–0.7130 (0.0662)	−1.7129 (0.0662)
Intercept 1	ξ_{01}	0.2668 (0.0560)	0.2667 (0.0560)
Intercept 2	ξ_{02}	2.0279 (0.0648)	2.0277 (0.0650)
Slope time	ξ_1	–0.7614 (0.0575)	–0.7610 (0.0575)
Slope group	ξ_2	–0.2053 (0.0587)	–0.2053 (0.0587)
Std. dev. RE	$\sqrt{d}$	—	—
Beta parameter	δ	—	13.1622 (390.44)
−2 log-likelihood		10588.18	10588.18
Effect	Par.	PO-Normal	PO-Beta-Normal
Intercept 0	ξ_{00}	−2.3201 (0.0100)	−2.3201 (0.0999)
Intercept 1	ξ_{01}	0.3336 (0.0818)	0.3335 (0.0818)
Intercept 2	ξ_{02}	2.7727 (0.1035)	2.7728 (0.1035)
Slope time	ξ_1	−1.0268 (0.0659)	−1.0268 (0.0659)
Slope group	ξ_2	–0.2605 (0.0912)	–0.2605 (0.0912)
Std. dev. RE	$\sqrt{d}$	1.5105 (0.0729)	1.5205 (0.0729)
Beta parameter	δ	—	15.4925 (246.55)
−2 log-likelihood		10320.39	10320.39

provement, neither when we switch from model (1) to model (2), nor when we move from (3) to (4). The estimate for the beta-parameter δ is large and has a very large standard error. This indicates that there is probably no overdispersion in the data.

7. Time-to-event data

MVDV, using their general framework, also focused on time-to-event data, combining the Weibull model with normal and gamma random effects. The model extends both the GLMM and the gamma frailty model. Molenberghs et al. (2015) extended the approach to allow for censoring. In what follows, we will give an overview of these developments. Efendi and Molenberghs (2013) paid particular attention to various estimation strategies. Abrams et al. (2017) integrated this framework in the modelling of current-status data, in the context of infectious diseases modelling.

Molenberghs and Verbeke (2011a), using closed-form expressions for the model's moments, pointed to both probabilistic as well as data-analytic implications of using (gamma) frailty models. We give a brief summary of these in Section 7.2.

The general Weibull model for repeated measures, with both gamma and normal random effects can be expressed as

$$f(\boldsymbol{y}_i|\boldsymbol{\theta}_i,\boldsymbol{b}_i) = \prod_{j=1}^{n_i} \lambda\rho\theta_{ij}y_{ij}^{\rho-1}e^{\boldsymbol{x}_{ij}^{\top}\boldsymbol{\xi}+z_{ij}^{\top}\boldsymbol{b}_i}e^{-\lambda y_{ij}^{\rho}\theta_{ij}e^{\boldsymbol{x}_{ij}^{\top}\boldsymbol{\xi}+z_{ij}^{\top}\boldsymbol{b}_i}}, \tag{59}$$

$$f(\boldsymbol{\theta}_i) = \prod_{j=1}^{n_i} \frac{1}{\beta_j^{\alpha_j}\Gamma(\alpha_j)}\theta_{ij}^{\alpha_j-1}e^{-\theta_{ij}/\beta_j}, \tag{60}$$

$$f(\boldsymbol{b}_i) = \frac{1}{(2\pi)^{q/2}|\boldsymbol{D}|^{1/2}}e^{-\frac{1}{2}\boldsymbol{b}_i^{\top}\boldsymbol{D}^{-1}\boldsymbol{b}_i}. \tag{61}$$

A few observations are in place. First, setting $\rho = 1$ leads to the special case of an exponential time-to-event distribution. Second, the classical gamma frailty model (i.e., no normal random effects) and the Weibull-based GLMM (i.e., no gamma random effects) follow as special cases. Third, strong conjugacy applies. This is definitely true for the exponential model, but carries over to the Weibull model, using the transformation Y_{ij}^{ρ}. It is equally possible to derive this result by merely re-writing the factor $\phi = \lambda\kappa$. Fourth, the above expressions are derived for a two-parameter gamma density. It is customary in a gamma frailty context (Duchateau and Janssen, 2007) to set $\alpha_j\beta_j = 1$, for reasons of identifiability. In this case, (60) is replaced by

$$f(\boldsymbol{\theta}_i) = \prod_{j=1}^{n_i} \frac{1}{\left(\frac{1}{\alpha_j}\right)^{\alpha_j}\Gamma(\alpha_j)}\theta_{ij}^{\alpha_j-1}e^{-\alpha_j\theta_{ij}}, \tag{62}$$

Alternatively, assuming $\alpha_j = 1$ and $\beta_j = 1/\delta_j$, one could write

$$f(\boldsymbol{\theta}_i) = \prod_{j=1}^{n_i} \delta_j e^{-\delta_j \theta_{ij}}, \tag{63}$$

implying that the gamma density is reduced to an exponential one.

MVDV derived a multi-index series formulation of the marginal joint distribution:

$$\begin{aligned} f(\boldsymbol{y}_i) = \sum_{(m_1,\ldots,m_{n_i})} \prod_{j=1}^{n_i} \frac{(-1)^{m_j}}{m_j!} \frac{\Gamma(\alpha_j + m_j + 1)\beta_j^{m_j+1}}{\Gamma(\alpha_j)} \lambda^{m_j+1} \rho y_{ij}^{(m_j+1)\rho-1} \\ \times \exp\left\{(m_j+1)\left[\boldsymbol{x}_{ij}^{\top}\boldsymbol{\xi} + \tfrac{1}{2}(m_j+1)\cdot \boldsymbol{z}_{ij}^{\top}\boldsymbol{D}\boldsymbol{z}_{ij}\right]\right\}. \end{aligned} \tag{64}$$

In case censorship applies, it is easy to integrate (64) over the interval $[C_{ij}, +\infty[$ or, in a multivariate fashion, over the cube $[\boldsymbol{0},\boldsymbol{C}_i]$:

$$\begin{aligned} F(\boldsymbol{C}_i) = \sum_{(m_1,\ldots,m_{n_i})} \prod_{j=1}^{n_i} \frac{(-1)^{m_j}}{(m_j+1)!} \frac{\Gamma(\alpha_j + m_j + 1)\beta_j^{m_j+1}}{\Gamma(\alpha_j)} \lambda^{m_j+1} C_{ij}^{(m_j+1)\rho} \\ \times \exp\left\{(m_j+1)\left[\boldsymbol{x}_{ij}^{\top}\boldsymbol{\xi} + \tfrac{1}{2}(m_j+1)\cdot \boldsymbol{z}_{ij}^{\top}\boldsymbol{D}\boldsymbol{z}_{ij}\right]\right\}. \end{aligned} \tag{65}$$

Evidently, if censorship applies to some but not all of the times within the vector, then the integration can be restricted to these, and the corresponding contribution will be an amalgamation of components taken from (64) and (65).

MVDV also derived the following moment expression, with mean, variance, and covariance expressions:

$$\mathrm{E}(Y_{ij}^k) = \frac{\alpha_j B(\alpha_j - k/\rho, k/\rho + 1)}{\lambda^{k/\rho}\beta_j^{k/\rho}} \exp\left(-\frac{k}{\rho}\boldsymbol{x}_{ij}^{\top}\boldsymbol{\xi} + \frac{k^2}{2\rho^2}\boldsymbol{z}_{ij}^{\top}\boldsymbol{D}\boldsymbol{z}_{ij}\right), \tag{66}$$

$$\mathrm{E}(Y_{ij}) = \frac{\alpha_j B(\alpha_j - 1/\rho, 1/\rho + 1)}{\lambda^{1/\rho}\beta_j^{1/\rho}} \exp\left(-\frac{1}{\rho}\boldsymbol{x}_{ij}^{\top}\boldsymbol{\xi} + \frac{1}{2\rho^2}\boldsymbol{z}_{ij}^{\top}\boldsymbol{D}\boldsymbol{z}_{ij}\right), \tag{67}$$

$$\begin{aligned} \mathrm{Var}(Y_{ij}) = \frac{\alpha_j}{\lambda^{2/\rho}\beta_j^{2\rho}} \exp\left(-\frac{2}{\rho}\boldsymbol{x}_{ij}^{\top}\boldsymbol{\xi} + \frac{1}{\rho^2}\boldsymbol{z}_{ij}^{\top}\boldsymbol{D}\boldsymbol{z}_{ij}\right) \\ \times \left[B(\alpha_j - 2/\rho, 2/\rho + 1)\exp\left(\frac{1}{\rho^2}\boldsymbol{z}_{ij}^{\top}\boldsymbol{D}\boldsymbol{z}_{ij}\right) - \alpha_j B\left(\alpha_j - \frac{1}{\rho}, \frac{1}{\rho} + 1\right)^2\right], \end{aligned} \tag{68}$$

$$\begin{aligned}\mathrm{Cov}(Y_{ij},Y_{ik}) &= \frac{\alpha_j\alpha_k}{\lambda^{2/\rho}\beta_j^{1/\rho}\beta_k^{1/\rho}}\exp\left[-\frac{1}{\rho}(\boldsymbol{x}_{ij}^{\top}\boldsymbol{\xi}+\boldsymbol{x}_{ik}^{\top}\boldsymbol{\xi})\right]\\ &\times B\left(\alpha_j-\frac{1}{\rho},\frac{1}{\rho}+1\right)B\left(\alpha_k-\frac{1}{\rho},\frac{1}{\rho}+1\right)\\ &\times\exp\left[\frac{1}{2\rho^2}(\boldsymbol{z}_{ij}^{\top}\boldsymbol{D}\boldsymbol{z}_{ij}+\boldsymbol{z}_{ik}^{\top}\boldsymbol{D}\boldsymbol{z}_{ik})\right]\left[\exp\left(\frac{1}{\rho^2}\boldsymbol{z}_{ij}^{\top}\boldsymbol{D}\boldsymbol{z}_{ik}\right)-1\right].\end{aligned} \tag{69}$$

7.1. Recurrent asthma attacks in children

MVDV analysed the times-to-event, introduced in Section 2.5. They considered an exponential model, i.e., a model of the form (59) with $\rho = 1$, and further a predictor of the form:

$$\kappa_{ij} = \xi_0 + b_i + \xi_1 T_i,$$

where T_i is an indicator for treatment and $b_i \sim N(0,d)$. Results from fitting all four models (with/without normal random effect; with/without gamma random effect) can be found in Table 8. A formal assessment of the treatment effect from all four models is given in Table 9. The treatment effect ξ_1 is stably identifiable in all four models. As can be seen from Table 9, the treatment effects are similar in strengths, but including both random effects reduces the evidence, relative to the exponential model. Needless to say that too parsimonious an association structure might lead to liberal test behaviour.

Table 8: *Asthma study. Parameter estimates (standard errors) for the regression coefficients in (1) the exponential model, (2) the exponential-gamma model, (3) the exponential-normal model, and (4) the combined model. Estimation was done by maximum likelihood using numerical integration over the normal random effect, if present.*

Effect	Par.	Exponential	Exponential-gamma
Intercept	ξ_0	−3.3709(0.0772)	−3.9782(15.354)
Treatment effect	ξ_1	−0.0726(0.0475)	−0.0755(0.0605)
Shape parameter	λ	0.8140(0.0149)	1.0490(16.106)
Std. dev. random effect	$\sqrt{d}$	—	—
Gamma parameter	γ	—	3.3192(0.3885)
−2log-likelihood		18,693	18,715
Effect	**Par.**	**Exponential-normal**	**Combined**
Intercept	ξ_0	−3.8095(0.1028)	3.9923(20.337)
Treatment effect	ξ_1	−0.0825(0.0731)	−0.0887(0.0842)
Shape parameter	λ	0.8882(0.0180)	0.8130(16.535)
Std. dev. random effect	$\sqrt{d}$	0.4097(0.0386)	0.4720(0.0416)
Gamma parameter	γ	—	6.8414(1.7146)
−2log-likelihood		18,611	18,629

Table 9: *Asthma study. Wald test results for the assessment of treatment effect.*

Model	Z value	p-value
Exponential	-1.5283	0.1264
Exponential-gamma	-1.1293	0.2588
Exponential-normal	-1.2480	0.2120
Combined	-1.0534	0.2921

7.2. Probabilistic and data-analytic issues with frailty models and their combined-model extensions

Based on moment expression (66), Molenberghs and Verbeke (2011a) observed that there can be a problem with models combining Weibull outcomes with gamma random effects, as well as with several extensions and sub-models. In particular, they established a connection with the so-called log-logistic distribution (Shoukri, Mian and Tracy, 1988), a transformation of the logistic distribution to the half line with only a finite number of finite moments.

To make their point, they started from a univariate Weibull distribution with gamma random effects (adding the normal random effects to the linear predictor does not substantially change anything), for which all expressions are given in the last column of Table 1. Like before, setting $\alpha\beta = 1$, and using formulation (62), the gamma and marginal distributions are written as:

$$f(\theta) = \frac{1}{\left(\frac{1}{\alpha}\right)^{\alpha}\Gamma(\alpha)}\theta^{\alpha-1}e^{-\alpha\theta}, \tag{70}$$

$$f(y) = \frac{\varphi\rho y^{\rho-1}\alpha^{\alpha+1}}{(\alpha+\varphi y^{\rho})^{\alpha+1}}. \tag{71}$$

Molenberghs and Verbeke (2011a) term this Case I. They also considered Case II, obtained by setting $\alpha = 1$ and $\beta = 1/\delta$, line in (63), henceforth, Case II:

$$f(\theta) = \delta e^{-\delta\theta}, \tag{72}$$

$$f(y) = \frac{\varphi\rho y^{\rho-1}\delta}{(\delta+\varphi y^{\rho})^{2}}. \tag{73}$$

Here, the gamma distribution has been replaced by its exponential special case and (73) is the log-logistic distribution (Bennett, 1983; Collett, 2003).

The moments follow from (66). For the general case, with α and β free parameters, for Case I, and for Case II, they are, respectively:

$$\text{General}: \mathrm{E}(Y^k) = \frac{\alpha B(\alpha - k/\rho, k/\rho + 1)}{(\beta\varphi)^{k/\rho}}, \tag{74}$$

$$\text{Case I}: \mathrm{E}(Y^k) = \left(\frac{\alpha}{\varphi}\right)^{k/\rho} \frac{k}{\rho} B(\alpha - k/\rho, k/\rho), \tag{75}$$

$$\text{Case II (log-logistic)}: \mathrm{E}(Y^k) = \frac{k}{\rho}\left(\frac{\delta}{\varphi}\right)^{k\rho} \cdot \Gamma(1-k/\rho)\cdot\Gamma(k/\rho). \tag{76}$$

The moments (74) are finite if and only if $k < \alpha\rho$. Hence, if $\alpha\rho$ is small, there is a risk that even lower-order moments do not exist, which evidently is problematic. Molenberghs and Verbeke (2011a) gave an example, using data from Duchateau and Janssen (2007). For certain methods of estimation in the context of the Weibull-Gamma frailty model, this would imply that regularity conditions are not satisfied. For the log-logistic case, this becomes $k < \rho$. The moments have been presented by Rinne (2009, p. 157) as well, though without reference to the irregularity issue.

8. Estimation

MVD and MVDV showed that fitting the combined model is relatively easy, and that standard software tools, such as the SAS procedure NLMIXED, can be used for maximum likelihood estimation in this case. More generically, any sufficiently flexible likelihood maximization tool that allows for normally distributed random effects can be used to this effect. This can typically be done with relatively little programming effort. Efendi and Molenberghs (2013) expanded upon this for the specific case of time-to-event data, and supplemented maximum likelihood with pairwise likelihood and Bayesian estimation. Their simulations indicated that, while maximum likelihood can be faster than pairwise likelihood, the latter has somewhat better convergence properties.

A priori, fitting a combined model of the type described in Section 4, proceeds by integrating over the random effects. The likelihood contribution of subject i is

$$f_i(\boldsymbol{y}_i|\boldsymbol{\vartheta},\boldsymbol{D},\boldsymbol{\vartheta}_i,\boldsymbol{\Sigma}_i) = \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{\vartheta},\boldsymbol{b}_i,\boldsymbol{\theta}_i)\; f(\boldsymbol{b}_i|\boldsymbol{D})\; f(\boldsymbol{\theta}_i|\boldsymbol{\vartheta}_i,\boldsymbol{\Sigma}_i)\; d\boldsymbol{b}_i\, d\boldsymbol{\theta}_i. \tag{77}$$

Here, $\boldsymbol{\vartheta}$ groups all parameters in the conditional model for Y_i. From (77) the likelihood derives as:

$$\begin{aligned} \boldsymbol{L}(\boldsymbol{\vartheta},\boldsymbol{D},\boldsymbol{\vartheta},\boldsymbol{\Sigma}) &= \prod_{i=1}^{N} f_i(\boldsymbol{y}_i|\boldsymbol{\vartheta},\boldsymbol{D},\boldsymbol{\vartheta}_i,\boldsymbol{\Sigma}_i) \\ &= \prod_{i=1}^{N}\int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\boldsymbol{\vartheta},\boldsymbol{b}_i,\boldsymbol{\theta}_i)\; f(\boldsymbol{b}_i|\boldsymbol{D})\; f(\boldsymbol{\theta}_i|\boldsymbol{\vartheta}_i,\boldsymbol{\Sigma}_i)\; d\boldsymbol{b}_i\, d\boldsymbol{\theta}_i. \end{aligned} \tag{78}$$

The key problem in maximizing (78) is the presence of N integrals over the random effects $\boldsymbol{b}_i$ and $\boldsymbol{\theta}$. It is widely claimed that the absence of a closed-form solution precludes an analytical-integration based solution (Molenberghs and Verbeke, 2005), explaining the popularity of Taylor-series expansion based methods, such as PQL and MQL, Laplace approximation, and numerical-integration based methods. These have been implemented in, for example, the SAS procedures GLIMMIX and NLMIXED. Several of the series expansion methods tend to exhibit bias, an issue taken up in Breslow and Lin (1995), and suggesting the use of alternative methods.

However, thanks to our results in Section 4, further progress can be made. Closed-form integration, apart from the normal case, is within reach for the Poisson, probit, and Weibull cases. Now, some closed forms involve series expansions, and may be either time consuming or cumbersome to implement. This notwithstanding, a variety of alternative approaches are possible.

Let us turn to the Poisson case. While closed-form expressions can be used to implement maximum likelihood estimation, with numerical accuracy governed by the number of terms included in the series, one can also proceed by what we will term partial marginalization. By this we refer to integrating (18)–(22) over the gamma random effects only, leaving the normal random effects untouched. The corresponding probability is:

$$f(y_{ij}|\boldsymbol{b}_i) = \begin{pmatrix} \alpha_j + y_{ij} - 1 \\ \alpha_j - 1 \end{pmatrix} \cdot \left(\frac{\beta_j}{1+\kappa_{ij}\beta_j} \right)^{y_{ij}} \cdot \left(\frac{1}{1+\kappa_{ij}\beta_j} \right)^{\alpha_j} \kappa_{ij}^{y_{ij}}, \tag{79}$$

where $\kappa_{ij} = \exp[\boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\xi} + \boldsymbol{z}_{ij}^{\mathsf{T}}\boldsymbol{b}_i]$. Note that, with this approach, we assume that the gamma random effects are independent within a subject. This is fine, given the correlation is induced by the normal random effects.

Similarly, for the Weibull case we obtain

$$f(y_{ij}|\boldsymbol{b}_i) = \frac{\lambda\kappa_{ij}e^{\mu_{ij}}\rho y_{ij}^{\rho-1}\alpha_j\beta_j}{(1+\lambda\kappa_{ij}e^{\mu_{ij}}\beta_j y_{ij}^{\rho})^{\alpha_j+1}}. \tag{80}$$

Now, in the survival case it is evidently very likely that censoring occurs. Focusing on right-censored data, it is then necessary to integrate the marginal density over the survival time within the interval $[0, C_i]$. The corresponding cumulative distribution is given in (65). In the spirit of (80), the partial marginalization of a censored component takes the form:

$$f(C_{ij}|\boldsymbol{b}_i) = \int_{C_{ij}}^{+\infty} f(y_{ij}|\boldsymbol{b}_i)dy_{ij} = \frac{1}{(1+\lambda\kappa_{ij}e^{\mu_{ij}}C_{ij}^{\rho})^{\alpha_j}}. \tag{81}$$

The concept of partial integration always applies whenever strong conjugacy holds. Indeed, an expression of the form (16) corresponds to integrating over the conjugate random effect θ, while leaving the normally distributed random effect embedded in the predictor, κ in this notation. Recall that, while expressions of the type (16) appear to be for the univariate case, they extend without problem to the longitudinal setting as well.

Because there is lack of strong conjugacy, the logit case defies the mere exploitation of conjugacy, such as the negative binomial form (79) and the Weibull-gamma frailty form (80). Nevertheless, it is easy to derive, for this case:

$$f(y_{ij}|\boldsymbol{b}_i) = \frac{1}{\alpha_j+\beta_j} \cdot (\kappa_{ij}\alpha_j)^{y_{ij}} \cdot [(1-\kappa_{ij})\alpha_j+\beta_j]^{1-y_{ij}}. \tag{82}$$

For all of these, it is straightforward to obtain the fully marginalized probability by numerically integrating the normal random effects out of (79), (80), and (82), using a tool such as the SAS procedure NLMIXED that allows for normal random effects in arbitrary, user-specified models.

For the specific case of the marginalized probit model, the computational challenge stems from the presence of a multivariate normal integral of the form (42), a phenomenon also known from the fully marginally specified multivariate probit model (Ashford and Sowden, 1970; Lesaffre and Molenberghs, 1991; Molenberghs and Verbeke, 2005). Specific to the context of the probit models with random effects, Zeger et al. (1988) derived the marginal mean function, needed for their application of generalized estimating equations as a fitting algorithm for the marginalized probit model.

In the ordinal case, the partially marginalized density at occasion j for subject i takes the form:

$$f(y_{ij}|\boldsymbol{b}_i) = \frac{\alpha_j}{\alpha_j+\beta_j} \cdot (\kappa_{1,ij})^{z_{1,ij}} \cdot \prod_{r=2}^{R-1}(\kappa_{r,ij}-\kappa_{r-1,ij})^{z_{r,ij}} \cdot \left(\frac{\alpha_j+\beta_j}{\alpha_j}-\kappa_{R-1,ij}\right)^{z_{R,ij}}.$$

From these, the likelihood can be constructed by assembling all contributions over subjects and repeated measurements within subjects.

MVDV discussed a number of alternative estimation strategies. These include pseudo-likelihood (or: pairwise likelihood; Aerts et al., 2002; Molenberghs and Verbeke, 2005), Bayesian inferences, non-parametric maximum likelihood (Booth et al., 2003: Aitkin, 1999; Alfò and Aitkin, 2000). Also, hierarchical generalized linear models (Lee and Nelder, 1996; Lee et al., 2006) can be used. They also referred to transformation-based methods, whereby non-normal random effects are transformed to normal ones, or vice versa (Liu and Yu, 2008; Nelson et al., 2006).

An important point is that not all parameters may be simultaneously identifiable. For example, the gamma-distribution parameters in the Poisson case, α and β, are not simultaneously identifiable when the linear-predictor part is also present, because there is aliasing with the intercept term. Therefore, one can set, for example, β equal to a constant, removing the identifiability problem. It is then clear that α, in the univariate case, or the set of α_j in the repeated-measures case, describe the additional overdispersion, in addition to what stems from the normal random effect(s). A similar phenomenon also plays in the binary case, where both beta-distribution parameters are not simultaneously estimable.

In addition, also Bayesian estimation and inference can be considered. Ghebretinsae et al. (2013) considered a Bayesian version in the time-to-event case. Ghebretinsae et al. (2012) presented a Bayesian joint CM. Efendi and Molenberghs (2013) juxtaposed likelihood-based and Bayesian estimation. The performance of the Bayesian method for the count case was assessed, using simulations, by Aregay, Shkedy, and Molenberghs (2013) and Rizzato et al. (2016). Aregay, Shkedy, and Molenberghs (2015) compared model versions with additive and multiplicative random effects. On a related note, Iddi et al. (2014) examined empirical Bayes estimation for the combined model.

9. Implication for computation of correlation and derived quantities

As we have seen, the combined model allows for closed-form expressions for moments, and hence for means and variances, for the normal, Poisson, probit, and Weibull cases, with a combination of normal random effects on the one hand, supplemented on the other hand with conjugate random effects, taking a normal, gamma, beta, and gamma form, respectively. The obvious one missing from the list is the logit model, but then the logit-probit connection, as discussed in Section 6.2, comes to the rescue.

These closed-form moments enable easy calculations of such derived quantities as correlations. For the count case, this was done by Vangeneugden et al. (2011), while Vangeneugden et al. (2014) focused on the binary setting.

For the count combined model, Vangeneugden et al. (2011) used the following derivation. The mean vector $\boldsymbol{\mu}_i = \mathrm{E}(\boldsymbol{Y}_i)$ has components:

$$\mu_{ij} = \phi_{ij} \exp\left(\boldsymbol{x}_{ij}^{\top}\boldsymbol{\xi} + \tfrac{1}{2}\boldsymbol{z}_{ij}^{\top}\boldsymbol{D}\boldsymbol{z}_{ij}\right), \tag{83}$$

and the variance-covariance matrix is given by

$$\mathrm{var}(\boldsymbol{Y}_i) = \boldsymbol{M}_i + \boldsymbol{M}_i\left(\boldsymbol{P}_i - \boldsymbol{J}_{n_i}\right)\boldsymbol{M}_i, \tag{84}$$

where $\boldsymbol{\phi}_i$ is the mean vector of the overdispersion random effects, with components ϕ_{ij}, $\boldsymbol{\Sigma}_i$ is the variance-covariance matrix of the overdispersion random effects, with components σ_{ij}, and $\boldsymbol{M}_i$ is a diagonal matrix with elements μ_{ij}. Further, the $(j,k)^{\text{th}}$ element of $\boldsymbol{P}_i$ equals

$$p_{i,jk} = \exp\left(\tfrac{1}{2}\boldsymbol{z}_{ij}^{\top}\boldsymbol{D}\boldsymbol{z}_{ik}\right)\cdot\frac{\sigma_{i,jk}+\phi_{ij}\phi_{ik}}{\phi_{ij}\phi_{ik}}\cdot\exp\left(\tfrac{1}{2}\boldsymbol{z}_{ik}^{\top}\boldsymbol{D}\boldsymbol{z}_{ij}\right). \tag{85}$$

Evidently, from this variance-covariance structure, the correlations immediately follow.

For the binary combined model, with probit link, the means, variances, and covariances were given in (45)–(47). When the logit link is used, no similar closed form exist. One can proceed by approximating the logit function via the probit function, or by using Taylor-series-based expressions. Details on these can be found in Vangeneugden et al. (2014).

The availability of closed-form correlation and other moment-based functions is useful in a number of contexts. For example, when studying psychometric reliability and generalizability (Vangeneugden et al., 2008; 2010), the correlation function is the basic building block. Correlation functions are also used in the context of surrogate marker evaluation from clinical-trial data (Alonso et al., 2017). Milanzi et al. (2015) used developments of this type to underscore the difference between manifest and latent correlations, for example when reliability measures are calculated in item response theory.

10. Marginalized versions of the combined model

As is clear from Sections 4–7, for many though not all versions of the CM there are explicit moment expressions and quantities derived there from. Nevertheless, they are algebraically involved, chiefly due to the non-conjugate nature of the normal random effects. To simplify the derivation of marginal quantities, such as effect measures, mean functions, etc., it is sensible to turn to the methodology of Heagerty (1999) and Heagerty and Zeger (2000), who modified the GLMM so that the first-order moments, i.e., the mean functions, are directly marginally interpretable. They originally focused on the logistic-normal model for binary longitudinal data, but they and others then extended the framework to other data types and link functions. The method specifies, at first sight contrary to intuition, a separate model for the marginal and conditional means. But this works thanks to a connector function that depends on covariates, marginal parameters, and the random-effects specification. Hence, both a marginal and conditional interpretation of the parameters can be maintained. The model, called the marginalized multilevel model (MMM), also allows for the use of maximum likelihood and Bayesian inferences, which is useful when data are incomplete.

To bring together the flexibility of the CM and the marginal interpretability of the MMM, Iddi and Molenberghs (2012ab) developed the *combined overdispersed and marginalized multilevel model* (COMMM). They focused on binary data and to some extent on counts. Kassahun et al. (2014b) studied further the count data case. The time-to-event case was studied by Efendi, Molenberghs, and Iddi (2014). Molenberghs et al. (2013) and Kenward and Molenberghs (2016) established connections between various ways of deriving marginally interpretable random-effects models, of which the MMM idea is one. Iddi and Molenberghs (2013) and Kassahun et al. (2014b) combined the MMM idea, for counts, with the occurrence of zero-inflation.

The rest of this section is organized in the following way. In Section 10.1, the general MMM and COMMM methodology is given. The analysis of the epilepsy, onychomycosis, and asthma cases studies is presented in Sections 10.2–10.4. In Section 10.5, we show how further zero inflation in the count case can be added.

10.1. Methodology

The general formulation of the CM was given in Section 4. The other building block that we need is the general marginalized multilevel model (MMM), after which both will be merged.

The general marginalized multilevel model due to Heagerty (1999) can be written as:

$$g_1(\mu_{ij}^m) = \boldsymbol{x}_{ij}^{\mathsf{T}} \boldsymbol{\xi}^m, \tag{86}$$

$$g_2(\mu_{ij}^c) = \Delta_{ij} + \boldsymbol{z}_{ij}^{\mathsf{T}} \boldsymbol{b_i}, \tag{87}$$

$$\boldsymbol{b_i} \sim F_b(\boldsymbol{0}, \boldsymbol{D}), \tag{88}$$

$$Y_{ij}^c = Y_{ij}|\boldsymbol{b_i} \sim F_{Y^c}\left(\mu_{ij}^c, \upsilon\right). \tag{89}$$

The two link functions g_1 and g_2 can be different, although frequently they will be identical and then denoted by g. Further F_b is an arbitrary distribution. Here, υ is a dispersion parameter, similar to the overdispersion parameter ϕ in the exponential family. The marginal mean $\mu_{ij}^m = \mathrm{E}(Y_{ij})$ is made to depend on an $n_i \times p$ matrix of p linear predictors $\boldsymbol{X}_i$ through a link function $g(\cdot)$. Further, the conditional mean $\mu_{ij}^c = \mathrm{E}(Y_{ij}|\boldsymbol{b_i})$ relates to the random variable $\boldsymbol{b_i}$ with distribution (88) and the function Δ_{ij} connects the marginal and conditional means through the same link function; the latter aspect could be relaxed if desired. The conditional response distribution is given by F_{Y^c}. The function Δ_{ij} is obtained from the solution to the integral equation

$$\mu_{ij}^m = g^{-1}(\boldsymbol{x}_{ij}^{\mathsf{T}} \boldsymbol{\xi}^m) = \int_b g^{-1}(\Delta_{ij} + \boldsymbol{z}_{ij}^{\mathsf{T}} \boldsymbol{b_i}) dF_b. \tag{90}$$

For example, when the link function is logit and the distribution of the random effect is normal, the expression of Δ_{ij} is obtained from:

$$\mathrm{expit}(\boldsymbol{x}_{ij}^{\mathsf{T}} \boldsymbol{\xi}^m) = \int_b \mathrm{expit}(\Delta_{ij} + \boldsymbol{z}_{ij}^{\mathsf{T}} \boldsymbol{b_i}) \varphi(\boldsymbol{b_i}|\boldsymbol{0}, \boldsymbol{D}) d\boldsymbol{b_i}.$$

Here, $\mathrm{expit}(\eta) = e^{\eta}/(1+e^{\eta})$. Griswold and Zeger (2004) expanded the model by relaxing the common link function assumed for both the marginal and conditional model specification. For example, using a logistic-probit-normal model:

$$\begin{aligned}\text{logit}(\mu_{ij}^m) &= \boldsymbol{x}_{ij}^{\top}\boldsymbol{\xi}^m,\\ \Phi^{-1}(\mu_{ij}^c) &= \Delta_{ij} + \boldsymbol{z}_{ij}^{\top}\boldsymbol{b_i},\\ \boldsymbol{b_i} &\sim F_b(\boldsymbol{0},\boldsymbol{D}),\\ Y_{ij}^c|\boldsymbol{b_i} = Y_{ij} &\sim F_{Y^c}\left(\mu_{ij}^c,\upsilon\right).\end{aligned}$$

(90) becomes:

$$\Delta_{ij} = \left(\sqrt{1+\boldsymbol{z}_{ij}^{\top}\boldsymbol{D}\boldsymbol{z}_{ij}}\right)\cdot\Phi^{-1}\{\text{expit}(\boldsymbol{x}_{ij}^{\top}\boldsymbol{\xi}^m)\}. \tag{91}$$

The logit-probit-normal is more attractive than the logit-logit normal version in the sense that, for example, the marginal parameters will enjoy the odds ratio interpretation while at the same time retaining the computational advantage associated with the probit-normal relationship. Of course, when both link functions are of probit form, (91) becomes:

$$\Delta_{ij} = \left(\sqrt{1+\boldsymbol{z}_{ij}^{\top}\boldsymbol{D}\boldsymbol{z}_{ij}}\right)\cdot\boldsymbol{x}_{ij}^{\top}\boldsymbol{\xi}^m. \tag{92}$$

For count data, a log-log-normal specification leads to

$$\Delta_{ij} = \boldsymbol{x}_{ij}^{\top}\boldsymbol{\xi}^m - \boldsymbol{z}_{ij}^{\top}\boldsymbol{D}\boldsymbol{z}_{ij}/2. \tag{93}$$

Note from this expression that, in particular for a random intercept model, i.e., one where $\boldsymbol{z}_{ij}^{\top}\boldsymbol{b_i} = b_i$ with $b_i \sim N\left(0,\tau^2\right)$, then $\boldsymbol{z}_{ij}^{\top}\boldsymbol{D}\boldsymbol{z}_{ij} = \sqrt{1+\tau^2}$, which implies that only fixed intercept parameters will be affected in the MMM model compared to their counterparts in the conditional GLMM model. For a general random-effects design $\boldsymbol{z}_{ij}^{\top}\boldsymbol{b_i}$, this will not be the case. The expression for Δ_{ij}, in the case of probit-probit-normal, log-log-gamma model and the logistic-logistic-Bridge MMM can be found in Griswold and Zeger (2004).

Iddi and Molenberghs (2012), and Efendi et al. (2014) combined the MMM with the CM, by combining (9), (10), and (11) from the CM with (86), (88), and (89) from the MMM in the following way:

$$\begin{aligned}g(\mu_{ij}^m) &= \boldsymbol{x}_{ij}^{\top}\boldsymbol{\xi}^m\\ g(\kappa_{ij}) &= \Delta_{ij} + \boldsymbol{z}_{ij}^{\top}\boldsymbol{b_i}\\ \mu_{ij}^c &= \theta_{ij}\kappa_{ij}\\ \theta_{ij} &\sim \Theta_{ij}\left(\tau_{ij},\sigma_{ij}^2\right)\\ \boldsymbol{b_i} &\sim F_b(\boldsymbol{0},\boldsymbol{D})\\ Y_{ij}^c = (Y_{ij}|\theta_{ij},\boldsymbol{b_i}) &\sim F_{Y^c}\left(\mu_{ij}^c,\upsilon\right).\end{aligned}$$

Note that the response distribution is now conditioned on two sets of random effects, namely the overdispersion and longitudinal ones. This implies that the expression for Δ_{ij} will change slightly. Because $\mu_{ij}^c = \mathrm{E}(Y_{ij}|\theta_{ij}, \boldsymbol{b_i})$, the function Δ_{ij} will then be obtained from the integral equation

$$\begin{aligned}\mu_{ij}^m = g^{-1}(\boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\xi}^m) &= \int_b \int_\theta \theta_{ij} g^{-1}(\Delta_{ij} + \boldsymbol{z}_{ij}^{\mathsf{T}}\boldsymbol{b_i}) d\Theta_\theta dF_b \\ &= \int_b \mathrm{E}(\theta_{ij}) g^{-1}(\Delta_{ij} + \boldsymbol{z}_{ij}^{\mathsf{T}}\boldsymbol{b_i}) dF_b. \qquad (94)\end{aligned}$$

These authors showed that for the logistic-probit-normal model with beta distribution for the overdispersion parameter, i.e., $\theta_{ij} \sim \mathrm{Beta}(\alpha_{1j}, \beta_{2j})$, (94) becomes

$$\Delta_{ij} = \left(\sqrt{1 + \boldsymbol{z}_{ij}^{\mathsf{T}}\boldsymbol{D}\boldsymbol{z}_{ij}}\right) \cdot \Phi^{-1}\{(1 + c_j) \cdot \mathrm{expit}(\boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\xi}^m)\},$$

where $c_j = \beta_{2j}/\alpha_{1j}$, which can serve as one of several possible constraints, given that the model is now over-parameterized. For the log-log-normal MMM model with $\theta_{ij} \sim \mathrm{Gamma}(\alpha_{1j}, \alpha_{2j})$,

$$\Delta_{ij} = -\log(\alpha_{1j}\alpha_{2j}) + \boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\xi}^m - \boldsymbol{z}_{ij}^{\mathsf{T}}\boldsymbol{D}\boldsymbol{z}_{ij}/2.$$

The fully marginalized joint distribution can be obtained from integrating out the two random effects. Less effort is needed here because the expressions for the marginal distribution are similar to those found in Molenberghs et al. (2010), except for replacing κ_{ij} with $\kappa_{ij} = g^{-1}(\Delta_{ij} + \boldsymbol{z}_{ij}^{\mathsf{T}}\boldsymbol{b_i})$.

Efendi et al. (2014) showed that, in the particular case of the Weibull-gamma-normal model, the integral equation leads to:

$$\Delta_{ij} = -\log(\alpha_j\beta_j) + \boldsymbol{x}_{ij}^{\mathsf{T}}\boldsymbol{\xi}^m - \boldsymbol{z}_{ij}^{\mathsf{T}}\boldsymbol{D}\boldsymbol{z}_{ij}/2. \qquad (95)$$

Should there be no gamma random effects, then the first term on the right hand side of (95) simply drops.

Parameter estimation conveniently proceeds by using the partially marginalized distribution method, explained in Section 8. Only here, the conditional distribution is partly specified through the marginal mean function, which is passed on to the conditional mean function via the connector function.

10.2. A clinical trial in epileptic patients

Building further on the models fitted in Sections 5.1 and 5.3, assume Y_{ij} to follow a Poisson distribution with marginal mean

$$\log(\pi_{ij}^m) = \begin{cases} \beta_{00} + \beta_{01} t_{ij} & \text{if placebo,} \\ \beta_{10} + \beta_{11} t_{ij} & \text{if treatment.} \end{cases} \tag{96}$$

Write the conditional model $\log(\pi_{ij}^c) = \Delta_{ij} + b_i$, with $b_i \sim N(0,d)$ and Δ_{ij} the connector. If also overdispersion is present, consider the COMMM version with then $\pi_{ij}^c = \theta_{ij}\exp(\Delta_{ij} + b_i)$ where $\theta_{ij} \sim \text{Gamma}(\alpha_1, \alpha_2)$ and impose constraint $\alpha_2 = 1/\alpha_1$.

Table 10: *Epilepsy study. Comparison of the log-log-normal MMM with the combined gamma and log-log-normal MMM.*

		CM	MMM	COMMM
		Gamma and	Log-Log-	Gamma and
Effect	Par.	log-normal	normal	Log-Log normal
Interc. plac.	β_{00}	0.9112(0.1755)	1.3960 (0.1887)	1.4757 (0.1962)
Slope plac.	β_{01}	−0.0248(0.0077)	−0.0143 (0.0044)	−0.0248 (0.0077)
Interc. treatm.	β_{10}	0.6555(0.1782)	1.2256 (0.1901)	1.2200 (0.1970)
Slope treatm.	β_{11}	−0.0118(0.0075)	−0.0120 (0.0043)	−0.0118 (0.0075)
SD RE	$\sqrt{d}$	1.0625(0.0871)	1.0755 (0.0857)	1.0625 (0.0871)
Neg.-bin. par.	α_1	2.4640(0.2113)	—	2.4640 (0.2113)
Neg.-bin. par.	$\alpha_2 = \frac{1}{\alpha_1}$	0.4059(0.0348)	—	0.4059 (0.0348)
−2 log-likelihood		−7664	−6810	−7664

Parameter estimates and standard errors for the log-log-normal MMM and the gamma-log-log-normal COMMM model are presented in Table 10. Observe that the parameter estimates for the two models are very similar, with the same holding for the standard errors. The log-log-normal model improves when the gamma random effect is introduced, as seen from a likelihood ratio comparison. This crucially affects inferences about the difference between the slopes as well as the ratio of the slopes. For the log-log-normal model, the difference of the slopes $\beta_{11} - \beta_{01}$ was found not to be significantly different form zero while the ratio of the slopes β_{11}/β_{01} showed a significant difference from one ($p = 0.7111$ and $p = 0.0376$, respectively). On the other hand both the slope difference ($p = 0.2260$) and ratio ($p = 0.1591$) showed non-significance in the combined model. To understand this, two things need to be borne in mind. First, the above demonstrates that, due to more careful modelling of the association and dispersion structures, inferences about functions of the model parameters may be erroneous in the simpler model, underscoring that care must be taken regarding conclusions based on the simpler model. Indeed, it would lead to a significant treatment difference, whereas the more general combined model showed no evidence for treatment difference. Similar observations were also made by MVD, where the combined Poisson-Gamma-normal showed a strong improvement of the Poisson GLMM model, underscoring the importance of introducing the gamma random effect. Second, and very important, one should not directly compare the estimates in the marginalized and the conditional version. Indeed, in the MMM model, treatment effects, slopes, etc. have a marginal interpretation. In addition, we can examine the results of fitting a combined beta and log-normal model, which is purely conditionally specified. The interpretation of the latter should be considered at the individual level, or at least for a change between two patients with different covariate profile (e.g., treated versus non-treated), but with the same level of the random effect.

We note from these results that for a random intercept model, only the intercepts parameters are affected but all other parameters remain the same compared to the combined Gamma and log-log-normal model. These would, however, not be the same, for example, for a random intercept and slope model. Given that the log link was used for both marginal and conditional models, we see further that the log-likelihood remains the same across both combined models.

10.3. A clinical trial in onychomycosis

Also here, both the conditional as well as the marginal mean are specified:

$$\begin{aligned}
Y_{ij}|b_i &\sim \text{Bernoulli}(\pi^c_{ij}),\\
\Phi^{-1}\left(\pi^c_{ij}\right) &= \Delta_{ij} + b_i,\\
b_i &\sim N(0,d),\\
\text{logit}(\pi^m_{ij}) &= \beta_0 + \beta_1 X_i + \beta_2 t_{ij} + \beta_3 X_i t_{ij}.
\end{aligned}$$

Recall that X_i is an indicator for the treatment applied to subject i, t_{ij} is the time at which the jth measurement is taken. For the COMMM model, the conditional mean model is specified as $\pi^c_{ij} = \theta_{ij}\Phi(\Delta_{ij} + b_i)$ where $\theta_{ij} \sim \text{Beta}(\alpha_1, \alpha_2)$ and Φ^{-1} is the probit link. The constraint $c = \alpha_2/\alpha_1$ was imposed.

From the results presented in Table 11, it is again clear that introducing the beta random effect improves significantly the model fit when comparing the log-likelihoods (smaller AIC). Parameter estimates from both models are slightly different, but a much more dramatic effect is seen in precision estimation. For many, but not all parameters, the extended model yields a higher precision. Furthermore, we observed that whereas the broader model encompassing both overdispersion and correlation concludes that there is no effect of the evolution of treatment (β_3) on the response with p-value of $p = 0.0790$, the MMM model results in a significant treatment evolution ($p = 0.0155$). Also presented in Table 11 are the results for a combined beta and probit-normal model whose parameters have a conditional interpretation. The treatment evolution was found to be significant with $p = 0.0343$. By comparing the two combined models, which both account for overdispersion and correlation simultaneously but with different interpretation of parameters, we may conclude that, while there is a significant treatment evolution given subjects, there is no evidence of population average treatment evolution.

Table 11: *A clinical trial in onychomycosis. Comparison of logistic-probit-normal MMM with the combined Beta and logistic-probit-normal MMM.*

		CM	MMM	COMMM
		Beta and	logistic-	Beta and logistic-
Effect	Par.	probit-normal	probit-normal	probit-normal
Interc.	β_0	−0.7285(0.8622)	−0.6154 (0.1493)	−0.4762 (0.0408)
Treatment	β_1	−0.7404(1.1816)	−0.0382 (0.2120)	−0.1858 (0.1240)
Time	β_2	−0.9109(0.2321)	−0.1529 (0.0190)	−0.1832 (0.0241)
Interaction	β_3	−0.3989(0.1876)	−0.0702 (0.0288)	−0.0691 (0.0392)
SD RE	$\sqrt{d}$	8.6763(1.9535)	2.1061 (0.1904)	8.8901 (0.0152)
Beta-bin. par.	α_2/α_1	0.2828(0.0372)	—	0.2769 (0.0363)
−2 log-likelihood		1259.9	1265.2	1254.0

10.4. Recurrent asthma attacks in children

We now turn to the recurrent asthma data, described in Section 2.5. For each of the 226 patients, their treatment allocation and repeated time-to-event outcomes, the time between the end of the previous to onset of the next attack, Y_{ij} is recorded; the outcome is subject to censoring. Also here, the combined model and its marginalized version are presented next to each other. Regarding the normal random-effects structure, a random intercept b_{i1} (with variance σ_i^2) and a random slope b_{i2} (with variance σ_e^2) is included. While this could be relaxed, both random effects are assumed to be independently normally distributed. Model fitting is done using both full and pairwise likelihood. Parameter estimates (standard errors) are presented in Table 12.

Full likelihood estimates between the ordinary and marginalized models are similar. Treatment effect is not significant. Because marginalization does not change the likelihood, the likelihood ratios are invariant to this operation (Griswold and Zeger, 2004). Because we now include two normally distributed random effects, the connector function (95) uses a different vector z_{ij}. This now implies that the treatment effect estimate changes upon marginalization, although the change is minor.

Table 12: *Asthma study. Original and marginalized combined model results. 'WGN' refers to the Weibull-gamma-normal model, whilst 'C' and 'CM' means censored and censored-marginalized, respectively.*

Effect	Par	WGN-C	WGN-CM	WGN-C	WGN-CM
		Full likelihood		*Pairwise likelihood*	
		Estimate(s.e.)	Estimate(s.e.)	Estimate(s.e.)	Estimate(s.e.)
Treatment	ξ	−0.113(0.106)	−0.111(0.102)	−0.127(0.105)	−0.127(0.105)
Shape	λ	0.014(0.001)	0.017(0.001)	0.025(0.002)	0.027(0.003)
Conj.RE	α	3.566(0.632)	3.566(0.632)	4.583(0.708)	4.584(0.708)
s.d. norm. R.int.	σ_i	0.560(0.068)	0.560(0.068)	0.445(0.039)	0.445(0.039)
s.d. norm. R.eff.	σ_e	0.077(0.734)	0.077(0.741)	11E-4(11E-4)	20E-6(20E-6)
−2 log-likelihood		16649	16649		

Turning attention to results using pairwise likelihood estimation, it is found that the estimates before and after marginalization are still similar. We also see that the estimate of the random slope parameter is virtually zero in all cases, although more pronounced in the pairwise-likelihood case. This does not contradict the results from full likelihood, where this component was non-significant, although the numerical behaviour is quite different.

In the four versions presented in the table, the conjugate random effect parameter is statistically significant. This is important and underscores that neither the standard GLMM nor the available marginalized model of Griswold and Zeger (2004) fit the data adequately.

10.5. Adding zero inflation to the COMMM in the case of counts

In line with Kassahun et al. (2014b), the above construction can be combined with the concepts of Section 5.2, where additional zeroes are allowed for in the CM for count data.

We have to be careful regarding the correct logic. We first specify the model fully hierarchically, derive its marginal mean function, model the former including connector functions, and the latter in the usual parametric way.

10.5.1. Zero-inflation

Dropping indices to diminish notational clutter at this point, the conditional model specification is:

$$P(\boldsymbol{Y}=\boldsymbol{y}|\theta,\boldsymbol{b}) = \begin{cases} \pi^c + (1-\pi^c)f(0|\lambda^c) & y=0, \\ (1-\pi^c)f(y|\lambda^c) & y>0, \end{cases} \tag{97}$$

$$\pi^c = \Phi(\Delta_1 + \boldsymbol{z}_1^{\mathsf{T}}\boldsymbol{b}_1), \tag{98}$$

$$\lambda^c = \theta \exp(\Delta_2 + \boldsymbol{z}_2^{\mathsf{T}}\boldsymbol{b}_2), \tag{99}$$

$$\theta \sim \text{Gamma}(\alpha,\beta), \tag{100}$$

$$\boldsymbol{b} = \begin{pmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{pmatrix}\right]. \tag{101}$$

It now follows:

$$\text{E}(Y|\theta,\boldsymbol{b}) = [\pi^c + (1-\pi^c)f(0|\lambda^c)]\cdot 0 + \sum_{y=1}^{\infty} y\frac{e^{-\lambda^c}(\lambda^c)^y}{y!} = (1-\pi^c)\lambda^c. \tag{102}$$

We then require that the marginal mean is of the form:

$$\mathrm{E}(Y) = (1-\pi^m)\lambda^m. \tag{103}$$

The fact that calculating the mean form (102) results in the form (103) does not imply that the marginal model behind (97)–(101) is equal to (103). In fact, as stated before, we know this is not true.

Focusing on the mean functions, as we should, leads to the requirement:

$$\int\int (1-\pi^c)\lambda^c f(\theta) f(\boldsymbol{b}) d\theta d\boldsymbol{b} = (1-\pi^m)\lambda^m. \tag{104}$$

It looks like this is straightforward, but there is a caveat: π^c and λ^c are connected through correlated random effects. In the special but relevant case that $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ are uncorrelated, and hence that $\boldsymbol{D}_{12} = 0$, we can solve the system:

$$\int \pi^c f(\boldsymbol{b}_1) d\boldsymbol{b}_1 = \pi^m, \tag{105}$$

$$\int\int \lambda^c f(\boldsymbol{b}_2) f(\theta) d\boldsymbol{b}_2\, d\theta = \lambda^m. \tag{106}$$

Now, (105) is the classical binary connector function integral equation; (106) is the counterpart for the Poisson case.

In case $\boldsymbol{D}_{12} \neq 0$, the integral equation takes the form:

$$\int\int\int (1-\pi^c)\lambda^c f(\theta) f(\boldsymbol{b}_1) f(\boldsymbol{b}_2|\boldsymbol{b}_1) d\theta\, d\boldsymbol{b}_1\, d\boldsymbol{b}_2 = (1-\pi^m)\lambda^m. \tag{107}$$

Given that

$$\boldsymbol{b}_2|\boldsymbol{b}_1 \sim N\left(\boldsymbol{D}_{21}\boldsymbol{D}_{11}^{-1}\boldsymbol{b}_1, \boldsymbol{D} = \boldsymbol{D}_{22} - \boldsymbol{D}_{21}\boldsymbol{D}_{11}^{-1}\boldsymbol{D}_{12}\right),$$

and with some straightforward algebra, we obtain the following intermediate step:

$$\mathrm{E}(\theta) e^{\Delta_2 + \frac{1}{2}\boldsymbol{z}_2^\mathsf{T}\boldsymbol{D}\boldsymbol{z}_2} \int (1-\pi^c) e^{\boldsymbol{z}_2^\mathsf{T}\boldsymbol{D}_{21}\boldsymbol{D}_{11}^{-1}\boldsymbol{b}_1} f(\boldsymbol{b}_1) d\boldsymbol{b}_1 = (1-\pi^m)\lambda^m.$$

This, in turn, leads to

$$\mathrm{E}(\theta) e^{\Delta_2 + \frac{1}{2}\boldsymbol{z}_2^\mathsf{T}\boldsymbol{D}_{22}\boldsymbol{z}_2} \int (1-\pi^c) f(\boldsymbol{b}_1; \mu = \boldsymbol{D}_{12}\boldsymbol{z}_2) d\boldsymbol{b}_1.$$

Upon applying a final transformation ($\widetilde{\boldsymbol{b}}_1 = \boldsymbol{b}_1 - \boldsymbol{D}_{12}z_2 \sim N(\boldsymbol{0}, \boldsymbol{D}_{11})$), we find that the Poisson connector remains the same, but for the binary connector, we need to solve:

$$\pi^c = \Phi(\Delta_1 + z_1^{\mathrm{T}}\widetilde{\boldsymbol{b}}_1 + z_1^{\mathrm{T}}\boldsymbol{D}_{12}z_2).$$

Of course, this is equal to the standard binary connector problem, but merely with a shift applied to Δ_1.

10.5.2. Hurdle models

Using the same simplified notation as before, we now have:

$$P(\boldsymbol{Y} = \boldsymbol{y}|\theta, \boldsymbol{b}) = \begin{cases} \pi^c & y = 0, \\ (1-\pi^c)\frac{f(y|\lambda^c)}{1-f(0|\lambda^c)} & y > 0, \end{cases}, \tag{108}$$

with the rest of the model specified by (98)–(101). It now follows:

$$\mathrm{E}(Y|\theta, \boldsymbol{b}) = \pi^c \cdot 0 + \frac{1-\pi^c}{1-f(0|\lambda^c)}\sum_{y=1}^{\infty} f(y|\lambda^c) = \frac{1-\pi^c}{1-f(0|\lambda^c)} \cdot \lambda^c = \frac{1-\pi^c}{1-e^{-\lambda^c}} \cdot \lambda^c. \tag{109}$$

Also here, we require conditional mean (109) to take the same form marginally:

$$\mathrm{E}(Y) = (1-\pi^m) \cdot \frac{\lambda^m}{1-e^{-\lambda^m}}.$$

When $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ are independent, we find the classical connector integral equation for the binary component:

$$\int \pi^c f(\boldsymbol{b}_1) d\boldsymbol{b}_1 = \pi^m.$$

For the count connector function, we need to solve:

$$\iint \frac{\lambda^c}{1-e^{-\lambda^c}} f(\boldsymbol{\theta}) f(\boldsymbol{b}_2) d\theta d\boldsymbol{b}_2 = \frac{\lambda^m}{1-e^{-\lambda^m}}.$$

More explicitly,

$$\iint \frac{\theta e^{\Delta_2 + z_2^{\mathrm{T}}\boldsymbol{b}_2}}{1 - e^{-\left[\theta e^{\Delta_2 + z_2^{\mathrm{T}}\boldsymbol{b}_2}\right]}} f(\boldsymbol{\theta}) f(\boldsymbol{b}_2) d\theta d\boldsymbol{b}_2 = \frac{e^{x_2^{\mathrm{T}}\boldsymbol{\xi}}}{1 - e^{-e^{x_2^{\mathrm{T}}\boldsymbol{\xi}}}}.$$

Of course, also here, a further modification is needed when the two normal random effects are correlated. In line with what we find in the zero-inflated case, we now have:

$$\iiint \Phi(\Delta_1 + \boldsymbol{z}_1^\top \boldsymbol{b}_1) \cdot \frac{\theta e^{\Delta_2 + \boldsymbol{z}_2^\top \boldsymbol{b}_2}}{1 - e^{-\left[\theta e^{\Delta_2 + \boldsymbol{z}_2^\top \boldsymbol{b}_2}\right]}} f(\boldsymbol{\theta}) f(\boldsymbol{b}_1) f(\boldsymbol{b}_2 | \boldsymbol{b}_1) d\theta d\boldsymbol{b}_1 d\boldsymbol{b}_2 = \Phi(\boldsymbol{x}_1^\top \gamma) \cdot \frac{e^{\boldsymbol{x}_2^\top \boldsymbol{\xi}}}{1 - e^{-e^{\boldsymbol{x}_2^\top \boldsymbol{\xi}}}}.$$

However now, the denominator under the integrand implies that simplification is less straightforward, and hence a Newton-Raphson approach for the pair (Δ_1, Δ_2) is an obvious way forward. Note that in the zero-inflated case, we were able to derive intuitive expressions for Δ_1 and Δ_2, but these are not unique, given that there is one integral equation with two tuning parameters. Thus, at best, one can find an algebraic expression for Δ_1, because even in the uncorrelated random-effects case, there is no closed form for the count connector. Therefore, we can simply set one of the two equal to zero, $\Delta_1 \equiv 0$, say, and then solve the reduced integral equation for Δ_2.

11. Joint modelling of several outcomes

The common recording of not one but several longitudinal sequences is common practice nowadays. The use of normal random effects in the combined model allows one to simultaneously analyse several longitudinal sequences, which do not even need to be of the same type.

Iddi and Molenberghs (2012a) made use of this possibility to jointly model a continuous and a binary longitudinal sequence. Kassahun et al. (2015) jointly modelled a continuous and a zero-inflated count sequence. Njeru Njagu et al. (2016) considered the case where repeated time-to-event outcomes are coupled with a longitudinal outcome of various types (continuous, binary, count) as well as the joint modelling of a continuous and binary outcome. Ivanova, Molenberghs, and Verbeke (2016) allow for ordinal outcomes as well. Ghebretinsae et al. (2012) used CM joint modelling to analyse comet assay data.

To give an example, let us consider Case 1 of Njeru Njagu et al. (2016), where a linear mixed model for the continuous outcome is coupled with a Weibull-gamma-normal model for the time-to-event outcome. The joint model, conditional on both the normal and gamma random effects, takes the form:

$$\begin{aligned} f(\boldsymbol{t}_i, \boldsymbol{y}_i | \boldsymbol{b}_i, \boldsymbol{\psi}_i) = & \prod_k \lambda_k \rho_k t_{ik}{}^{\rho_k - 1} \psi_{ik} e^{\mu_{ik} + d_{ik}} e^{-\lambda_k t_{ik}{}^{\rho_k} \psi_{ik} e^{\mu_{ik} + d_{ik}}} \\ & \times \frac{1}{(2\pi)^{\frac{n_i}{2}} |\boldsymbol{\Sigma}_i|^{\frac{1}{2}}} e^{\frac{-1}{2} (\boldsymbol{y}_i - X_i \boldsymbol{\xi} - Z_i \boldsymbol{b}_i)^\top \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{y}_i - X_i \boldsymbol{\xi} - Z_i \boldsymbol{b}_i)}, \end{aligned} \tag{110}$$

with $\boldsymbol{\Sigma}_i$ an n_i by n_i diagonal covariance matrix with diagonal elements σ^2. Also, $\boldsymbol{t}_i$ is the set of p_i survival times for cluster i, while $\boldsymbol{y}_i$ is the vector of n_i continuous outcomes. Moreover, $d_{ik} = \boldsymbol{w}_{ik}^{\top}\boldsymbol{b}_i$, where $\boldsymbol{w}_{ik}^{\top}$ is a vector of scale factors. Here, the index k refers to the kth survival time in cluster i. For the scale and shape parameters in the baseline hazard, we consider a more general case, where both λ and ρ are allowed to vary between members of a cluster. The continuous and survival processes are assumed independent, conditional on the shared normal random effects. Note that the shared random effect in the way considered here is generic. For example, one can choose $\boldsymbol{z}_{ij}$ and $\boldsymbol{w}_{ij}$ such that some random effects are present in the normal-outcome linear predictors, with others influencing the Weibull predictor, and a third set influencing both. As such, our paradigm encompasses both shared as well as correlated random effects.

12. Influence diagnostics

Because of the relative novelty of the CM and its extensions, development regarding model assessment and diagnostic tools has been limited. Rakhmawati et al. (2017) presented local influence diagnostic tools for the count-data CM. Rakhmawati et al. (2016ab) extended this to allow for zero inflation and incomplete data, respectively.

Local influence was presented by Cook (1986). The impact of individuals and measurements on the analysis is assessed by comparing standard maximum likelihood estimates with those resulting from slightly perturbing the contribution of an individual or measurement. The method is to be contrasted with global influence (case deletion), where impact is assessed by simply deleting an individual or measurement. While conceptually a bit technical, it is easy and fast to use in practice and in several cases it leads to interpretable components of influence. Lesaffre and Verbeke (1998) introduced influence assessment for the linear mixed model. Ouwens, Tan, and Berger (2001) applied local influence to the Poisson-normal model. Rakhmawati et al. (2017) followed their ideas, but with extensions in three directions. First, they provided closed-form expressions, based on an analytical form for the marginal likelihood function, as well as based on an integral form for the said likelihood. Second, they considered three important cases: binary, count, and time-to-event. Third, they started from the combined model, rather than merely from the GLMM.

The general theory behind so-called case-weighted likelihood is as follows. Let the log-likelihood for the generalized linear mixed model or its combined extension take the form

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} \ell_i(\boldsymbol{\theta}), \tag{111}$$

in which $\ell_i(\boldsymbol{\theta})$ is the contribution of the ith individual to the log-likelihood. Let

$$\ell(\boldsymbol{\theta}|\boldsymbol{\omega}) \;=\; \sum_{i=1}^{N} \omega_i \ell_i(\boldsymbol{\theta}), \tag{112}$$

and denote the perturbed version of $\ell(\boldsymbol{\theta})$, depending on an N-dimensional vector $\boldsymbol{\omega}$ of weights, assumed to belong to an open subset Ω of $\mathbb{R}^N$. The original log-likelihood (111) follows for $\boldsymbol{\omega} = \boldsymbol{\omega}_0 = (1,1,\ldots,1)^{\mathrm{T}}$. Let $\widehat{\boldsymbol{\theta}}$ be the maximum likelihood estimator for $\boldsymbol{\theta}$, obtained by maximizing $\ell(\boldsymbol{\theta})$, and let $\widehat{\boldsymbol{\theta}}_\omega$ denote the estimator for $\boldsymbol{\theta}$ under $\ell(\boldsymbol{\theta}|\boldsymbol{\omega})$. Cook (1986) proposed to measure the distance between $\widehat{\boldsymbol{\theta}}_\omega$ and $\widehat{\boldsymbol{\theta}}$ by the likelihood displacement: $\mathrm{LD}(\boldsymbol{\omega}) \;=\; 2\left(\ell(\hat{\boldsymbol{\theta}}) - \ell(\hat{\boldsymbol{\theta}}_{\boldsymbol{\omega}})\right)$. $\mathrm{LD}(\boldsymbol{\omega})$ will be large if $\ell(\boldsymbol{\theta})$ is strongly curved at $\widehat{\boldsymbol{\theta}}$. A graph of $\mathrm{LD}(\boldsymbol{\omega})$ versus $\boldsymbol{\omega}$ brings out information on the influence of case-weight perturbations. The graph is the geometric surface formed by the values of the $(N+1)$-dimensional vector

$$\boldsymbol{\xi}(\boldsymbol{\omega}) = \begin{pmatrix} \boldsymbol{\omega} \\ \mathrm{LD}(\boldsymbol{\omega}) \end{pmatrix}$$

as $\boldsymbol{\omega}$ varies throughout Ω. Following Cook (1986) and Verbeke and Molenberghs (2000), we will refer to $\boldsymbol{\xi}(\boldsymbol{\omega})$ as an influence graph.

Cook (1986) derived a convenient computational scheme. Let $\boldsymbol{\Delta}_i$ be the s-dimensional vector of second-order derivatives of $\ell(\boldsymbol{\theta}|\boldsymbol{\omega})$, w.r.t. ω_i and all components of $\boldsymbol{\theta}$, and evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\omega} = \boldsymbol{\omega}_0$. Also, write Δ for the $s \times r$ matrix with $\boldsymbol{\Delta}_i$ in the ith column. Let $\ddot{L}$ denote the $s \times s$ matrix of second derivatives of $\ell(\boldsymbol{\theta})$, evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$. For any unit vector $\boldsymbol{h}$ in Ω, it follows that:

$$C_h = 2\left|\, \boldsymbol{h}^{\mathrm{T}} \boldsymbol{\Delta}^{\mathrm{T}} \ddot{\boldsymbol{L}}^{-1} \boldsymbol{\Delta} \boldsymbol{h} \,\right|. \tag{113}$$

Various choices for $\boldsymbol{h}$ have received attention. First, as will be done here, one can focus on subject i only, by choosing $\boldsymbol{h} = \boldsymbol{h_i}$, the zero vector with a sole 1 in the ith position. Local influence then is

$$C_i \;\equiv\; C_{h_i} = 2\left|\, \boldsymbol{\Delta}_i^{\mathrm{T}} \ddot{\boldsymbol{L}}^{-1} \boldsymbol{\Delta}_i \,\right|. \tag{114}$$

Second, $\boldsymbol{h} = \boldsymbol{h}_{\max}$ can be considered, the direction of maximal normal curvature (Verbeke and Molenberghs 2000). Expressions can be derived when only a sub-vector of the parameter vector is of interest as well. We refer to Rakhmawati et al. (2017) for details.

These authors derived interpretable expressions for several cases. For example, for the probit-normal case they showed that

$$||\boldsymbol{\Delta}_i||^2 = \left(\sum_{j=1}^{n_i} r_{ij}\boldsymbol{x}_{ij}\right)\left(\sum_{j=1}^{n_i} r_{ij}\boldsymbol{x}_{ij}\right)^{\mathsf{T}} + \sum_{k,l}\left\{-\tfrac{1}{2}(\boldsymbol{D}^{-1})_{kl} + \frac{1}{2}(\boldsymbol{D}^{-1}\boldsymbol{D}^{-1})_{kl}\mathrm{Var}(\boldsymbol{b_i})\right\}^2 .$$

Let $C_i = C_{1i} + C_{2i}$ with:

$$C_{1i} = 2||\ddot{\boldsymbol{L}}^{-1}||\ ||\boldsymbol{r}_i^{\mathsf{T}}\boldsymbol{x}_i||^2\cos(\varphi_i), \tag{115}$$

$$C_{2i} = \tfrac{1}{2}||\ddot{\boldsymbol{L}}^{-1}||\ ||(\boldsymbol{D}^{-1})_{kl} - (\boldsymbol{D}^{-1}\boldsymbol{D}^{-1})_{kl}\mathrm{Var}(\boldsymbol{b_i})||^2\cos(\varphi_i), \tag{116}$$

where $\boldsymbol{r}_i^{\mathsf{T}}\boldsymbol{x}_i = \sum_{j=1}^{n_i} r_{ij}\boldsymbol{x}_{ij}$. Note that C_{1i} and C_{2i} are the contributions of subject i to local influence C_i from $\boldsymbol{\beta}$ and $\boldsymbol{D}$, respectively. Now, C_{1i} and C_{2i} were shown to equal:

$$C_{1i} = 2||\ddot{\boldsymbol{L}}^{-1}||\ ||\boldsymbol{x}_i\boldsymbol{x}_i^{\mathsf{T}}||\ ||\boldsymbol{r}_i||^2\cos(\alpha_i)\cos(\varphi_i), \tag{117}$$

$$\begin{aligned} C_{2i} = {} & \tfrac{1}{2}||\ddot{\boldsymbol{L}}^{-1}||\cos(\varphi_i) \times \left[\mathrm{tr}\left\{(\boldsymbol{D}^{-1})^2_{kl}\right\} - \mathrm{tr}\left\{2(\boldsymbol{D}^{-1})_{kl}(\boldsymbol{D}^{-1}\boldsymbol{D}^{-1})_{kl}\mathrm{Var}(\boldsymbol{b_i})\right\})\right. \\ & \left. + \mathrm{tr}\left\{(\boldsymbol{D}^{-1}\boldsymbol{D}^{-1})^2_{kl}\mathrm{Var}(\boldsymbol{b_i})^2\right\}\right], \end{aligned} \tag{118}$$

where $\cos(\alpha_i)$ is the angle between $\mathrm{vec}(\boldsymbol{x}_i\boldsymbol{x}_i^{\mathsf{T}})$ and $\mathrm{vec}(\boldsymbol{r}_i\boldsymbol{r}_i^{\mathsf{T}})$, and φ_i is the angle between $\mathrm{vec}(-\ddot{\boldsymbol{L}}^{-1})$ and $\mathrm{vec}(\boldsymbol{\Delta}_i\boldsymbol{\Delta}_i^{\mathsf{T}})$. Hence, the interpretable components of C_i in the case of the Poisson-normal model can be described using the 'length of the fixed effect' ($||\boldsymbol{x}_i\boldsymbol{x}_i^{\mathsf{T}}||$), the 'squared length of the residual' ($||\boldsymbol{r_i}||^2$), and the 'squared of random effect variability' ($\mathrm{Var}(\boldsymbol{b_i})^2$).

Rakhmawati et al. (2017) derived similar expressions for the probit-normal, logit-normal and Weibull-normal models.

12.1. A clinical trial in epileptic patients

We start from the Poisson-normal (P-N) and Poisson-gamma-normal (PGN) models studied before:

$$\ln(\lambda_{ij}) = \begin{cases} (\xi_{00} + b_i) + \xi_{01}t_j & \text{if placebo} \\ (\xi_{10} + b_i) + \xi_{11}t_j & \text{if treated,} \end{cases} \tag{119}$$

where Y_{ij} represent the number of epileptic seizures patient i experienced during week j, t_j is the time point at which Y_{ij} was measured, and with random intercept $b_i \sim N(0,d)$. Parameter estimates are given in Table 13. Index plots (versus patient ID) for various local influence analyses are given in Figure 2. The top row of the plot represents the total local influence, with subsequent rows representing influence for sub-vectors: fixed effects, random-intercept variance d, and, for the (PGN), the overdispersion parameter

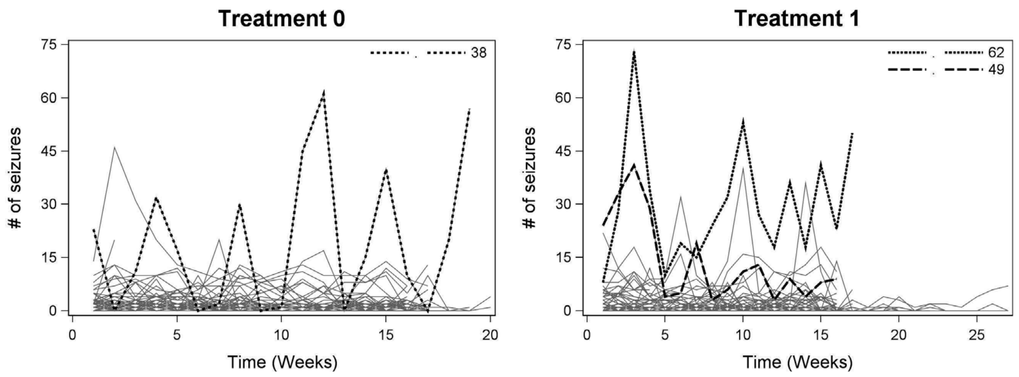
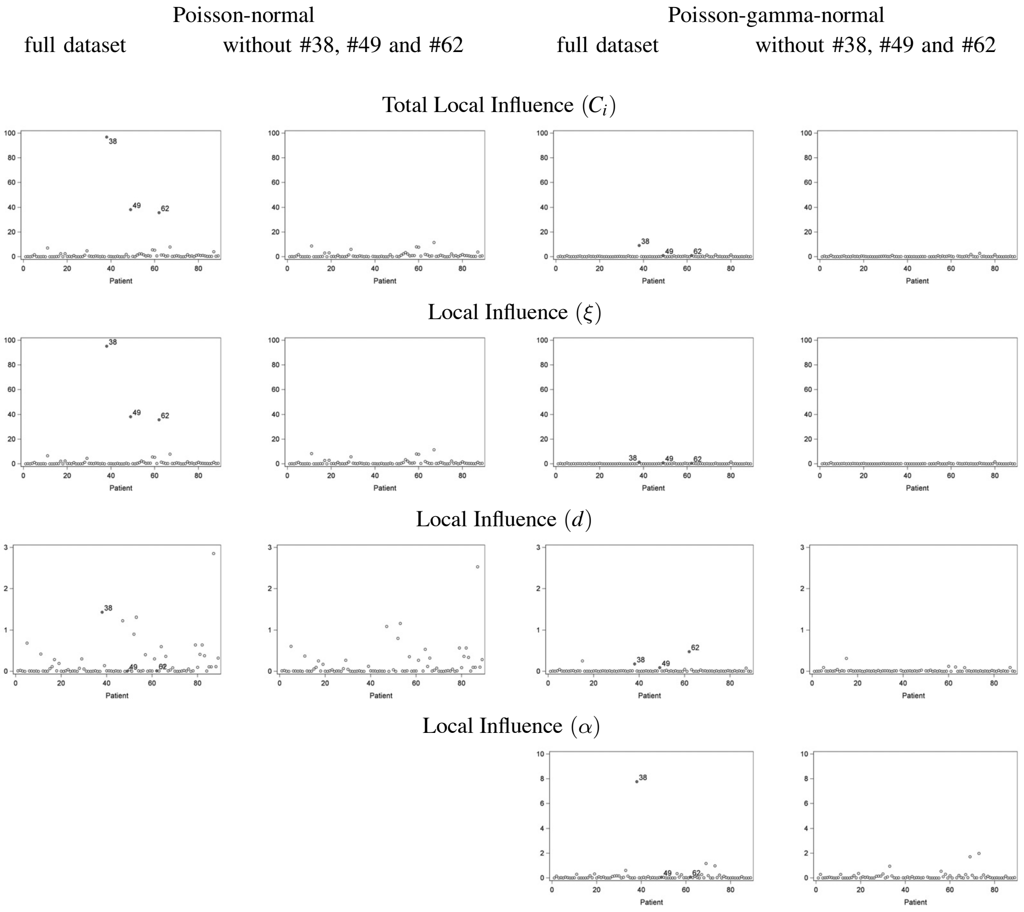


Figure 1: *Epilepsy data. Individual profiles.*



Total Local Influence (C_i)

Local Influence (ξ)

Local Influence (d)

Local Influence (α)

Figure 2: Epilepsy data. Local influence plots.

Table 13: *Local influence. Parameter estimates (standard errors) for the generalized linear mixed and combined models.*

Epilepsy		Poisson-normal		Poisson-gamma-normal	
Effect	Par.	Full	#(38,49,62)	Full	#(38,49,62)
Interc. plac.	ξ_{00}	0.818(0.168)	0.903(0.157)	0.911(0.176)	0.907(0.163)
Slope plac.	ξ_{01}	−0.014(0.004)	−0.031(0.005)	−0.025(0.008)	−0.031(0.008)
Interc. treat.	ξ_{10}	0.648(0.170)	0.492(0.162)	0.656(0.178)	0.510(0.169)
Slope treat.	ξ_{11}	−0.012(0.004)	−0.007(0.005)	−0.012(0.007)	−0.009(0.007)
Treat. eff.	$\xi_{11}-\xi_{10}$	0.002(0.006)	0.024(0.007)	0.013(0.011)	0.022(0.011)
Treat. eff.	ξ_{11}/ξ_{10}	0.840(0.398)	0.236(0.170)	0.475(0.335)	0.281(0.250)
Std. rand. int.	σ	1.076(0.086)	0.982(0.081)	1.063(0.087)	0.969(0.082)
Overdisp. par.	α			2.464(0.211)	3.109(0.329)

Onychomycosis		Logit-normal		Logit-beta-normal	
Effect	Par.	Full	#(6,30,53)	Full	#(6,30,53)
Interc. plac.	ξ_0	−1.630(0.435)	−1.940(0.523)	−1.604(4.026)	−2.420(3.089)
Slope plac.	ξ_1	−0.404(0.046)	−0.430(0.049)	−6.478(1.439)	−6.075(1.264)
Interc. treat.	ξ_2	−1.749(0.448)	−1.604(0.536)	−16.21(3.58)	−15.21(3.02)
Slope treat.	ξ_3	−0.563(0.060)	−0.872(0.100)	−8.075(1.600)	−8.755(1.437
Treat. eff.	$\xi_{11}-\xi_{10}$	−0.159(0.072)	−0.442(0.105)	−1.596(0.858)	−2.680(0.822)
Treat. eff.	ξ_{11}/ξ_{10}	1.394(0.206)	2.028(0.302)	1.246(0.148)	1.441(0.171)
Std. rand. int.	σ	4.015(0.381)	4.814(0.490)	60.88(14.22)	56.47(11.69)
Overdisp. par.	α/β			0.281(0.035)	0.231(0.031)

α, respectively. Patients #38, #49, and #62 stand out with large total influence C_i when compared to other patients. Importantly, influences show a major drop when switching from (P-N) to (PGN). This is most prominently seen for #38. For an explanation, turn to the right hand panel of Figure 1. Patient #38 (and to some extent also #62 on the left hand side) alternates periodically between very high numbers of episodes and periods virtually without. This implies that their mean, variance, and association structure are rather different from the majority of subjects. The impact on the mean structure, by way of the fixed effects, is evident in the second row. For the (P-N) it is less clear when turning to d, but we gain a lot of insight from the (PGN) results. Overall influence and influence on $\boldsymbol{\xi}$ reduce drastically, but there now is clear influence on d and α. What it means is that with these subjects present, the overdispersion parameter helps capturing their anomalous behaviour, which 'deflates' d. In other words, adding overdispersion protects the inferentially crucial fixed-effects parameter vector. When removing these subjects, and also #49, little or no influence is left.

Note that the (PGN) model fitted to the full dataset exhibits a smaller value for α, which corresponds to more overdispersion (no overdispersion corresponds to α approaching $+\infty$), while it does not vanish with removal of the three subjects. Thus, there appears to be genuine overdispersion in the data, further inflated by the influential subjects.

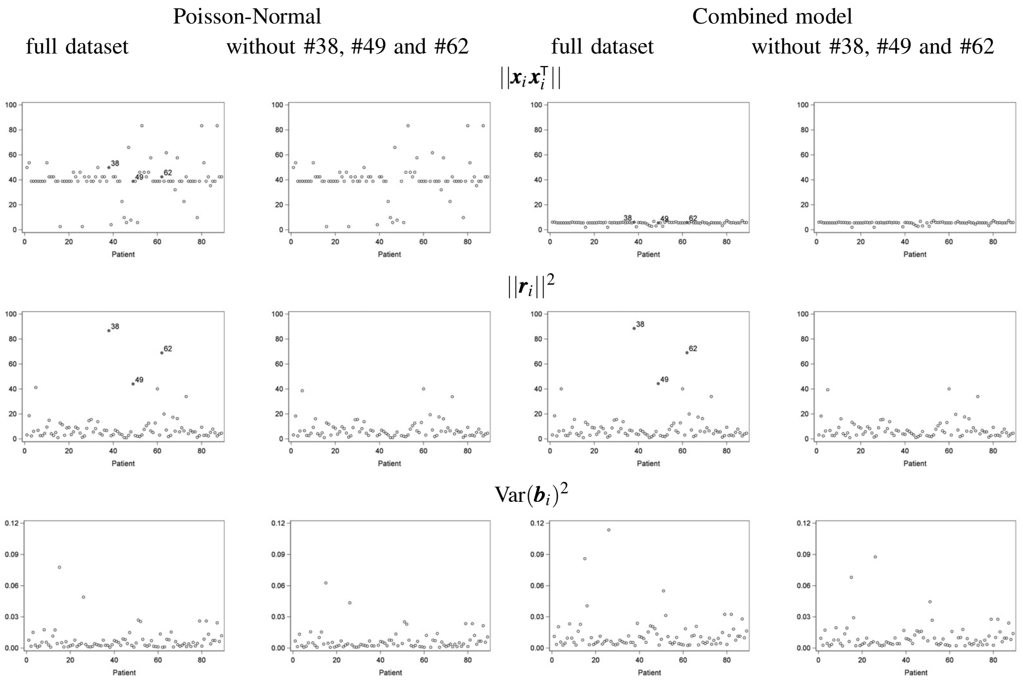
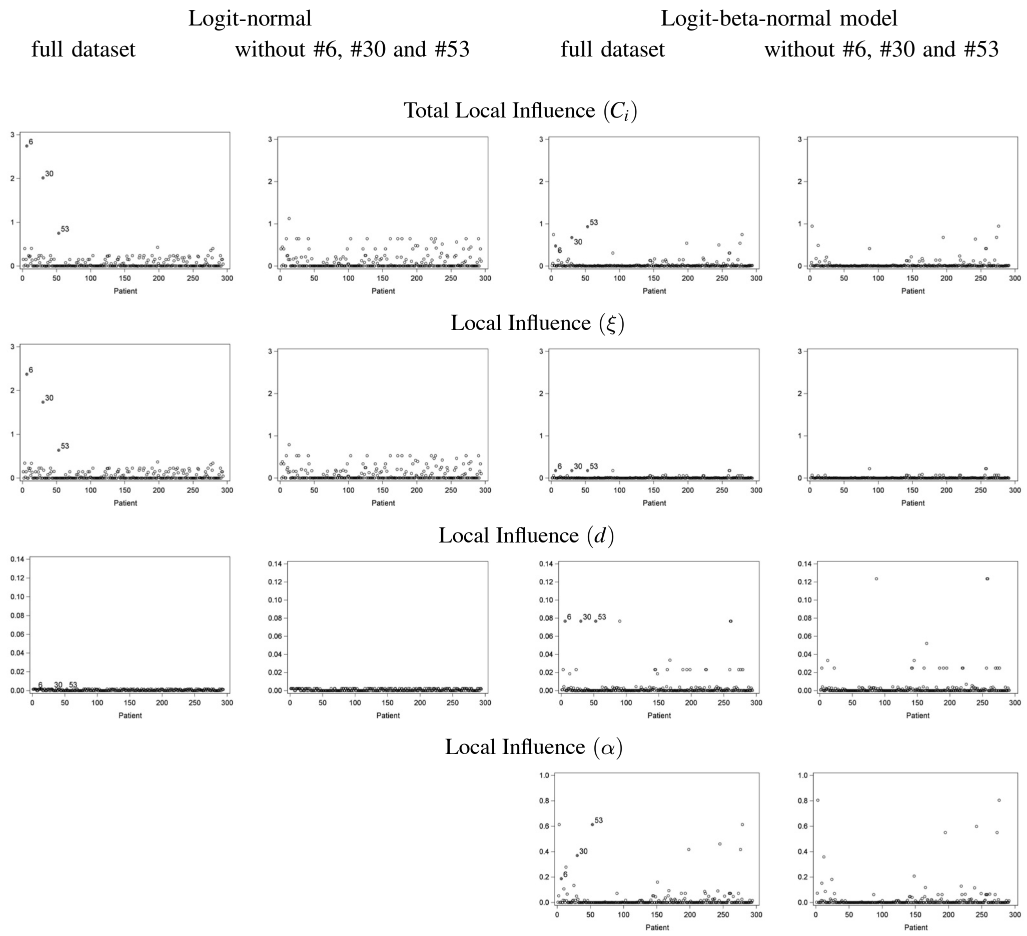


Figure 3: Epilepsy data. Plots of interpretable components of local influence.

In agreement with MVD, MVDV, and our earlier analysis, Rakhmawati et al. (2017) considered the treatment effect in additive ($\xi_{11} - \xi_{01}$) and multiplicative (ξ_{11}/ξ_{01}) form. Important differences are seen on the additive scale. (P-N) shows no significance ($p = 0.7106$), which is sustained for (PGN), with $p = 0.2225$. Removing the influential subjects leads to a highly significant result for (P-N), with $p = 0.0009$, which changes to the still significant $p = 0.0350$ for (PGN). Hence, the influential subjects mask a treatment effect. This is logical, because the influential subjects exhibit an oscillating behaviour, introducing an important source of variability. At the multiplicative level, where the null hypothesis is for the ratio to be 1, the story is nicely confirmed, with $p = 0.6872$ and $p = 0.1166$ for (P-N) and (PGN), respectively; the counterparts after deletion are $p < 0.0001$ and $p = 0.0040$, respectively.

To get further insight as to why these subject have higher influence than others, plots with interpretable components are given in Figure 3: 'squared length of the fixed effects' $||\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}||$, 'squared length of the residual' $||\boldsymbol{r}_i||^2$, and 'random-effect variability' $\mathrm{Var}(b_i)^2$. It is hardly surprising that #38 stands out in terms of $||\boldsymbol{r}_i||^2$. Influences on #49 and #62 are less pronounced.

Our analysis has provided insight not available from earlier analysis. The influential subjects exhibit a cyclic behaviour not observed in the majority of patients, but at the same time well documented. Based on these findings, a focused clinical discussion can take place, to determine the course of action. Options include removal, retention, or


Logit-normal

Logit-beta-normal model

full dataset

without #6, #30 and #53

full dataset

without #6, #30 and #53

Total Local Influence (C_i)

Local Influence (ξ)

Local Influence (d)

Local Influence (α)

Figure 4: Onychomycosis data. Local influence plots.

even setting up a dedicated study to further scrutinize this sub-population. In this case, a small group of patients with oscillating behaviour between two poles has been identified.

12.2. A clinical trial in onychomycosis

Before, we assumed $Y_{ij}|b_i \sim \text{Bernoulli}(\pi_{ij})$, where Y_{ij} is severity of infection (1 for severe, 0 for non-severe) for patient i at occasion j, T_i is the treatment indicator (1 for experimental, 0 for standard) for subject, t_j is the time point (months) at which the jth measurement has been taken, and $b_i \sim N(0,d)$. The conditional success probability is expressed as:

$$\text{logit}(\pi_{ij}) = \xi_1(1-T_i) + \xi_2(1-T_i)t_{ij} + \xi_3 T_i + \xi_4 T_i t_{ij} + b_i.$$

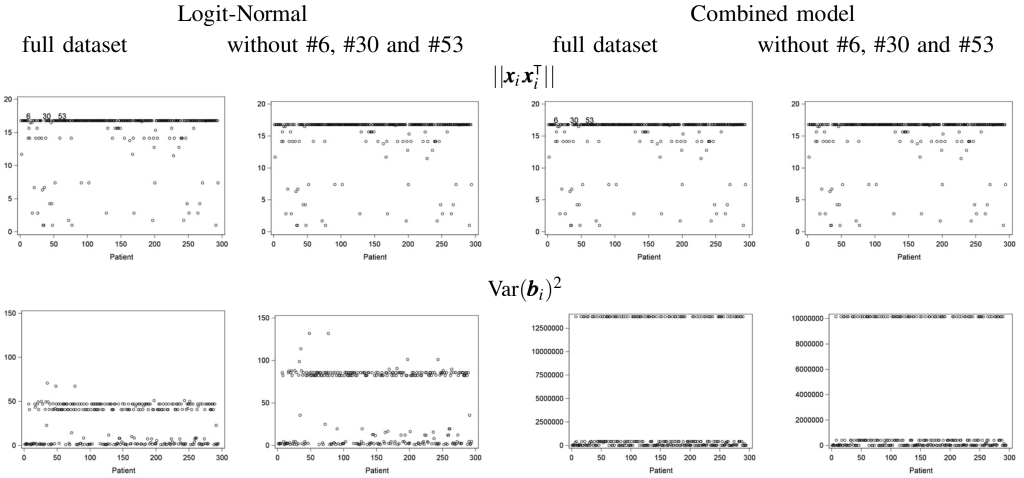


Figure 5: Onychomycosis data. Plot of interpretable components of local influence.

Both the logit-normal (L-N) and logit-beta-normal (LBN) are fitted. Parameter estimates (standard errors) are displayed in Table 13, with local influence plots in Figure 4. Subjects #6, #30, and #53 are detected as influential, overall, and with respect to the fixed effects, in the (L-N). Accommodating overdispersion, hence turning to the (LBN), deflates the magnitude of influence. Likewise, influence is drastically diminished by removing these three subjects. Thus, in case the influential subjects should remain in the analysis, the (LBN) may be the most sensible route forward. Alternatively, in case they are considered anomalous, one can remove them. To decide on which scenario is preferred in this case, we note that all three subjects are unusual: they set out with a sequence of non-severe ratings, but then switch to a severe rating ('0000111' for #6, '0000011' for #30, and '0000001' for #53). Arguably, there is no reason to remove these subjects from analysis, partly also to safeguard randomization. However, it is uncommon to switch from non-severe to severe in this particular way, so these patients must be further clinically scrutinized. Also for these data, the interpretable components do not lead to further insight (Figure 5).

The (L-N) and (LBN) lead to borderline significance when applied to the full data [$p = 0.0268$ additively and $p = 0.0560$ multiplicatively for (L-N); $p = 0.0627$ additively and $p = 0.0964$ multiplicatively for (LBN)]. When influential subjects are removed, these values all become highly significant [in the same order, $p < 0.0001$, $p = 0.0007$, $p = 0.0011$, and $p = 0.0099$]. These findings are qualitatively similar to the epilepsy cases.

13. Concluding remarks

Based on work by MVD, MVDV, and subsequent references, we have reviewed a general and flexible framework for such combinations, starting from arbitrary generalized linear models and exponential family members. Specific emphasis is placed on normally distributed, binary, binomial, count, and time-to-event outcomes. There are various reasons to do so. First, non-Gaussian hierarchical data exhibit three important features: (1) the mean structure; (2) the variance structure; and (3) the correlation structure. Our proposed framework features: (a) a mean structure; (b) overdispersion, often conjugate random-effects; (c) normal random effects. It will be clear from our case studies that model fit can be improved and hence model interpretation changed, by shifting to the extended model. Second, especially in cases where the variance and/or correlation structures are of interest (e.g., surrogate marker evaluation, psychometric evaluation, etc.) such extensions are useful. Third, even when interest remains with more conventional models, such as the GLMM, the extended model can serve as a goodness-of-fit tool. Fourth, because we can derive closed-form expressions for both standard and extended models, the accuracy of parameter estimation and resulting inferences can be improved, while obviating the need for tedious numerical integration techniques. Fifth, the analysis of the case studies corroborates this need. While the model extends the classical GLMM, it is actually easy to fit when standard non-linear mixed-model software is available, such as the SAS procedure NLMIXED.

Because for most of these combined models, and their GLMM sub-models, closed form moment expressions are available, derived quantities such as correlation are easy to obtain. Furthermore, versions with mean parameters that are directly marginally interpretable can be constructed. Also, the model lends itself naturally to the joint modelling of several hierarchical sequences simultaneously. Diagnostics based on local influence ideas have been developed as well.

While we have aimed to give an extensive overview of a modelling framework to accommodate data hierarchies and overdispersion, inevitably a number of topics have been left untouched. For example, Molenberghs and Verbeke (2011b), Pryseley et al. (2011) examined the occurrence of negative variance components in hierarchical data, which is also relevant for this context. Likewise, underdispersion has received some treatment (Oliveira et al., 2016; 2017).

Acknowledgment

Financial support from the IAP research network #P7/06 of the Belgian Government (Belgian Science Policy) is gratefully acknowledged. This work was partially supported by CNPq, a Brazilian science funding agency.

References

Abrams, S., Aerts, M., Molenberghs, G. and Hens, N. (2017). Parametric overdispersed frailty models for current status data. *Biometrics*, DOI: 10.1111/biom.12692.

Aerts, M., Geys, H., Molenberghs, G. and Ryan, L. (2002). *Topics in Modelling of Clustered Data*. London: Chapman & Hall.

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

Aitkin, M. (l999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, 55, 117–128.

Alonso, A., Bigirumurame, T., Burzykowski, T., Buyse, M., Molenberghs, G., Muchene, L., Perualila, N.J., Shkedy, Z. and Van der Elst, W. (2017). *Applied Surrogate Endpoint Evaluation with SAS and R*. Boca Raton: Chapman & Hall/CRC.

Alfò, M. and Aitkin, M. (2000). Random coefficient models for binary longitudinal responses with attrition. *Statistics and Computing*, 10, 279–288.

Aregay, M., Shkedy, Z. and Molenberghs, G. (2013). A hierarchical Bayesian approach for the analysis of longitudinal count data with overdispersion: a simulation study. *Computational Statistics and Data Analysis*, 57, 233–245.

Aregay, M., Shkedy, Z. and Molenberghs, G. (2015). Comparison of additive and multiplicative Bayesian models for longitudinal count data with overdispersion parameters: a simulation study. *Communications in Statistics, Computation and Simulation,* 44, 454–473.

Ashford, J.R. and Sowden, R.R. (1970) Multivariate probit analysis. *Biometrics*, 26, 535–546.

Bennett, S. (1983). Log-logistic regression models for survival data. *Applied Statistics*, 32, 165–171.

Böhning, D. (2000) *Computer-assisted Analysis of Mixtures and Applications. Meta-analysis, Disease Mapping and Others*. London: Chapman & Hall/CRC.

Booth, J.G., Casella, G., Friedl, H. and Hobert, J.P. (2003). Negative binomial loglinear mixed models. *Statistical Modelling*, 3, 179–181.

Borgermans, L., Goderis, G., Van Den Broeke, C., Verbeke, G., Carbonez, A., Ivanova, A., Mathieu, C., Aertgeerts, B., Heyrman, J. and Grol, R. (2009). Interdisciplinary diabetes care Teams operating on the interface between primary and specialty care are associated with improved outcomes of care: Findings from the Leuven Diabetes Project, *BMC Health Services Research*, 9, 179.

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.

Collett, D. (2003). *Modelling Survival Data in Medical Research* (2^{nd} ed.). Boca Raton: CRC Press.

Cook, R.D. (1986). Assessment of local influence. *Journal of the Royal Statistical Society, Series B*, 48, 133–169.

Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics*. London: Chapman & Hall/CRC.

Dean, C.B. (1991). Estimating equations for mixed-Poisson models. In: *Estimating Functions*, V.P. Godambe (Ed.). Oxford: Oxford University Press.

De Backer, M., De Keyser, P., De Vroey, C. and Lesaffre, E. (1996). A 12-week treatment for dermatophyte toe onychomycosis: terbinafine 250 mg/day vs. itraconazole 200 mg/day – a double-blind comparative trial. *British Journal of Dermatology*, 134, 16–17.

Del Fava, E., Shkedy, Z., Aregay, M. and Molenberghs, G. (2014). Modelling multivariate, overdispersed binomial data with additive and multiplicative random effects. *Statistical Modelling*, 14, 99–133.

Duchateau, L. and Janssen, P. (2007). *The Frailty Model*. New York: Springer.

Efendi, A. and Molenberghs, G. (2013). A multilevel model for hierarchical, repeated, and overdispersed time-to-event outcomes and its estimation strategies. *Journal of Biopharmaceutical Statistics*, 23, 1420–1434.

Efendi, A., Molenberghs, G. and Iddi, S. (2014). A Marginalized combined gamma frailty and normal random-effects model for repeated, overdispersed time-to-event outcomes. *Communications in Statistics*, 43, 4806–4828.

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.

Gentle, J.E. (2003). *Random Number Generation and Monte Carlo Methods*. New York: Springer.

Ghebretinsae, A.H., Faes, C., Molenberghs, G., De Boeck, M. and Geys, H. (2013). A Bayesian generalized frailty model for comet assays. *Journal of Biopharmaceutical Statistics*, 11, 449–455.

Ghebretinsae, A., Faes, C., Molenberghs, G., Geys, H. and Van der Leede, B.-J. (2012). Joint modelling of hierarchically clustered and overdispersed non-Gaussian continuous outcomes for comet assay data. *Pharmaceutical Statistics*, 11, 449–455.

Gibbons, R.D. and Hedeker, D. (1997). Random effects probit and logistic regression models for three-level data. *Biometrics*, 53, 1527–1537.

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.

Griswold, M.E. and Zeger, S.L. (2004). *On Marginalized Multilevel Models and their Computation*. John Hopkins University, Dept. of Biostatistics Working Papers.

Guilkey, D.K. and Murphy, J.L. (1993). Estimation and testing in the random effects probit model. *Journal of Econometrics*, 59, 301–317.

Heagerty, P.J. (1999). Marginally specified logistic-normal models for longitudinal binary data *Biometrics*, 55, 688–698.

Heagerty, P.J. and Zeger, S.L. (2000). Marginalized multilevel models and likelihood inference, *Statistical Science*, 15, 1–26.

Hedeker, D. and Gibbons, R.D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics*, 51, 933–944.

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.

Iddi, S. and Molenberghs, G. (2012a). A combined overdispersed and marginalized multilevel model. *Computational Statistics and Data Analysis*, 56, 1944–1951.

Iddi, S. and Molenberghs, G. (2012b). A joint marginalized multilevel model for continuous and binary longitudinal outcomes. *Journal of Applied Statistics*, 56, 1944–1951.

Iddi, S. and Molenberghs, G. (2013). A marginalized model for zero-inflated, overdispersed and correlated count data. *Electronic Journal of Applied Statistical Analysis*, 6, 149–165.

Iddi, S., Molenberghs, G., Aregay, M. and Kalema, G. (2014). Empirical Bayes estimates for correlated hierarchical data with overdispersion. *Pharmaceutical Statistics*, 13, 316–326.

Ivanova, A., Molenberghs, G. and Verbeke, G. (2014). A model for overdispersed hierarchical ordinal data. *Statistical Modelling*, 14, 399–415.

Ivanova, A., Molenberghs, G. and Verbeke, G. (2016). Mixed model approaches for joint modelling of different types of responses. *Journal of Biopharmaceutical Statistics*, 26, 601–618.

Johnson, N.L., Kemp, A., Kotz, S. (2005). *Univariate Discrete Distributions* (3rd ed.). Hoboken: John Wiley & Sons.

Johnson, N.L. and Kotz, S. (1970). *Distributions in Statistics, Continuous Univariate Distributions*, Vol. 2. Boston: Houghton-Mifflin.

Kalema, G., Iddi, S. and Molenberghs, G. (2016). The combined model: a tool for simulating correlated counts with overdispersion. *Communications in Statistics*, 45, 2491–2510.

Kalema, G. and Molenberghs, G. (2015). Generating correlated and/or overdispersed count data; a SAS implementation. *Journal of Statistical Software*, 00, 000–000.

Kassahun, W., Neyens, T., Faes, C., Molenberghs, G. and Verbeke, G. (2014a). A Zero-inflated overdispersed hierarchical Poisson model. *Statistical Modelling*, 14, 439–456.

Kassahun, W., Neyens, T., Molenberghs, G., Faes, C. and Verbeke, G. (2012). Modelling overdispersed longitudinal binary data from the Jimma longitudinal studies using a combined beta and normal random-effects model. *Archives of Public Health*, 70, Article 7.

Kassahun, W., Neyens, T., Molenberghs, G., Faes, C. and Verbeke, G. (2014b). Marginalized multilevel hurdle and zero-inflated models for overdispersed and correlated count data with excess zeros. *Statistics in Medicine*, 33, 4402–4419.

Kassahun, W., Neyens, T., Molenberghs, G., Faes C. and Verbeke, G. (2015). A joint model for hierarchical continuous and zero-inflated overdispersed count data. *Journal of Statistical Computation and Simulation*, 85, 552–571.

Kenward, M.G. and Molenberghs, G. (2016). A taxonomy of mixing and outcome distributions based on conjugacy and bridging. *Communications in Statistics, Theory and Methods*, 45, 1953–1968.

Kleinman, J. (1973). Proportions with extraneous variance: single and independent samples. *Journal of the American Statistical Association*, 68, 46–54.

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, Y. and Nelder, J.A. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society, Series B*, 58, 619–678.

Lee, Y. and Nelder, J.A. (2001a). Two ways of modelling overdispersion. *Applied Statistics*, 49, 591–598.

Lee, Y. and Nelder, J.A. (2001b). Hierarchical generalized linear models: a synthesis of generalized linear models, random-effect models and structured dispersions. *Biometrika*, 88, 987–1006.

Lee, Y. and Nelder, J.A. (2003). Extended-REML estimators. *Journal of Applied Statistics*, 30, 845–856.

Lee, Y., Nelder, J.A. and Pawitan, Y. (2006). *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood*. Boca Raton: Chapman & Hall/CRC.

Lesaffre, E. and Molenberghs, G. (1991). Multivariate probit analysis: a neglected procedure in medical statistics. *Statistics in Medicine*, 10, 1391–1403.

Lesaffre, E. and Verbeke, G. (1998). Local influence in linear mixed models. *Biometrics*, 54, 570–582.

Liang, K.Y. and McCullagh, P. (1993). Case studies in binary dispersion. *Biometrics*, 49, 623–630.

Liu, L. and Yu, Z. (2008) A likelihood reformulation method in non-normal random-effects models. *Statistics in Medicine*, 27, 3105–3124.

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

McCulloch, C.E. (1994). Maximum likelihood variance components estimation for binary data. *Journal of the American Statistical Association*, 89, 330–335.

McLachlan, G. and Peel, D.A. (2000). *Finite Mixture Models*. New York: John Wiley & Sons.

Milanzi, E., Molenberghs, G., Alonso, A., Verbeke, G. and De Boeck, P. (2015). Reliability measures in item response theory: Manifest versus latent correlation functions. *British Journal of Mathematical and Statistical Psychology*, 68, 43–64.

Molenberghs, G., Kenward, M.G., Verbeke, G., Efendi, A. and Iddi, S. (2013). On the connections between bridge distributions, marginalized multilevel models, and generalized linear mixed models. *International Journal of Statistics and Probability*, 2, 1–21.

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

Molenberghs, G. and Verbeke, G. (2007). Likelihood ratio, score, and Wald tests in a constrained parameter space. *The American Statistician*, 61, 1–6.

Molenberghs, G. and Verbeke, G. (2011a). On the Weibull-Gamma frailty model, its infinite moments, and its connection to generalized log-logistic, logistic, Cauchy, and extreme-value distributions. *Journal of Statistical Planning and Inference*, 141, 861–868.

Molenberghs, G. and Verbeke, G. (2011b). A note on a hierarchical interpretation for negative variance components. *Statistical Modelling*, 11, 389–408.

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

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

Molenberghs, G., Verbeke, G., Efendi, A., Braekers, R. and Demétrio, C.G.B. (2015). A combined gamma frailty and normal random-effects model for repeated, overdispersed time-to-event data. *Statistical Methods in Medical Research*, 24, 434–452.

Molenberghs, G., Verbeke, G., Iddi, S. and Demétrio, C.G.B. (2012). A combined beta and normal random-effects model for repeated, overdispersed binary and binomial data. *Journal of Multivariate Analysis*, 111, 94–109.

Moore, D.F. and Tsiatis, A.A. (1991). Robust estimation of the variance in moment methods for extra-binomial and extra-Poisson variation. *Biometrics*, 47, 383–401.

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 A*, 135, 370–384.

Nelson, K.P., Lipsitz, S.R., Fitzmaurice, G.M., Ibrahim, J., Parzen, M. and Strawderman, R. (2006). Use of the probability integral transformation to fit nonlinear mixed-effects models with non-normal random effects. *Journal of Computational and Graphical Statistics*, 15, 39–57.

Neyens, T., Faes, C., and Molenberghs, G. (2012). A generalized Poisson-gamma model for spatially overdispersed data. *Spatial and Spatio-temporal Epidemiology*, 3, 185–194.

Njeru Njagi, E., Molenberghs, G., Rizopoulos, D., Verbeke, G., Kenward, M.G., Dendale, P. and Willekens, K. (2016). A flexible joint-modelling framework for longitudinal and time-to-event data with overdispersion. *Statistical Methods in Medical Research*, 25, 1661–1676.

Oliveira, I.R.C., Molenberghs, G., Demétrio, C.G.B., Giolo, S. and Dias, C.T.S. (2016). Quantifying intraclass correlations for nonnegative traits. *Biometrical Journal*, 58, 852–867.

Oliveira, I.R.C., Molenberghs, G., Verbeke, G., Demétrio, C.G.B. and Dias, C.T.S. (2017). Negative variance components for non-negative hierarchical data with correlation, over-, and/or underdispersion. *Journal of Applied Statistics*, 44, 1047–1063.

Ouwens, M.J.N.M., Tan, F.E.S. and Berger, M.P.F. (2001). Local influence to detect influential data structures for generalized linear mixed models. *Biometrics*, 57, 1166–1172.

Pryseley, A., Tchonlafi, C., Verbeke, G. and Molenberghs, G. (2011). Estimating negative variance components from Gaussian and non-Gaussian data: a mixed models approach. *Computational Statistics and Data Analysis*, 55, 1071–1085.

Rakhmawati, T., Molenberghs, G., Verbeke, G. and Faes, C. (2016a). Local influence diagnostics for hierarchical count data models with overdispersion and excess zeros. *Biometrical Journal*, 58, 1390–1408.

Rakhmawati, T., Molenberghs, G., Verbeke, G. and Faes, C. (2016b). Local influence diagnostics for incomplete overdispersed longitudinal counts. *Journal of Applied Statistics*, 43, 1722–1737.

Rakhmawati, T., Molenberghs, G., Verbeke, G. and Faes, C. (2017). Local influence diagnostics for generalized linear mixed models with overdispersion. *Journal of Applied Statistics*, 44, 620–641.

Renard, D., Molenberghs, G. and Geys, H. (2004). A pairwise likelihood approach to estimation in multilevel probit models. *Computational Statistics and Data Analysis*, 44, 649–667.

Ridout, M., Demétrio, C.G.B. and Hinde, J. (1998). Models for count data with many zeros. In: *International Biometric Conference XIX*, Cape Town. Invited Papers, pp. 179–192.

Rinne, H. (2009). *The Weibull Distribution. A Handbook.* Boca Raton: CRC/Chapman & Hall.

Rizzato, F.B., Leandro, R.A., Demétrio, C.G.B. and Molenberghs, G. (2016). A Bayesian approach to analyse overdispersed longitudinal count data. *Journal of Applied Statistics*, 43, 2085–2109.

Roberts, D.T. (1992). Prevalence of dermatophyte onychomycosis in the United Kingdom: Results of an omnibus survey. *British Journal of Dermatology*, 126, 23–27.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, 78, 719–729.

Shepard, T.H., Mackler, B. and Finch, C.A. (1980). Reproductive studies in the iron-deficient rat. *Teratology*, 22, 329–334.

Shoukri, M.M., Mian, I.U.M. and Tracy, D.S. (1988). Sampling properties of estimators of the log-logistic distribution with application to Canadian precipitation data. *Canadian Journal of Statistics*, 16, 223–236.

Skellam, J.G. (1948). A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials, *Journal of the Royal Statistical Society, Series B*, 10, 257–261.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modelling*. London: Chapman & Hall/CRC.

Thall, P.F. and Vail, S.C. (1990). Some covariance models for longitudinal count data with overdispersion. *Biometrics*, 46, 657–671.

Vangeneugden, T., Molenberghs, G., Laenen, A., Alonso, A. and Geys, H. (2008). Generalizability in non-Gaussian longitudinal clinical trial data based on generalized linear mixed models. *Journal of Biopharmaceutical Statistics*, 18, 691–712.

Vangeneugden, T., Molenberghs, G., Laenen, A., Geys, H., Beunckens, C. and Sotto, C. (2010). Marginal correlation in longitudinal binary data based on generalized linear mixed models. *Communications in Statistics, Theory & Methods*, 39, 3540–3557.

Vangeneugden, T., Molenberghs, G., Verbeke, G. and Demétrio, C. (2011). Marginal correlation from an extended random-effects model for repeated and overdispersed counts. *Journal of Applied Statistics*, 38, 215–232.

Vangeneugden, T., Molenberghs, G., Verbeke, G. and Demétrio, C.G.B. (2014). Marginal correlation from logit- and probit-beta-normal models for hierarchical binary data. *Communications in Statistics*, 43, 4164–4178.

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

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

Zeger, S.L., Liang, K.-Y. and Albert, P.S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, 44, 1049–1060.