Made available by Hasselt University Library in https://documentserver.uhasselt.be

Modeling multivariate, overdispersed binomial data with additive and multiplicative random effects

Peer-reviewed author version

DEL FAVA, Emanuele; SHKEDY, Ziv; AREGAY, Mehreteab & MOLENBERGHS, Geert (2014) Modeling multivariate, overdispersed binomial data with additive and multiplicative random effects. In: Statistical modelling, 14 (2), p. 99-133.

DOI: 10.1177/1471082X13503450
Handle: http://hdl.handle.net/1942/16620

Modeling Multivariate, Overdispersed Binomial Data with Additive and Multiplicative Random Effects

Del Fava Emanuele[1], Shkedy Ziv[1], Aregay Mehreteab[2], and Molenberghs Geert[1,2]

[1]I-BioStat, Universiteit Hasselt, Diepenbeek, Belgium.

[2]I-BioStat, Katholieke Universiteit Leuven, Leuven, Belgium.

August 6, 2012

Abstract

Often, when modeling longitudinal binomial data, one needs to take into consideration both clustering and overdispersion. When the primary interest is in accommodating both phenomena, we can use separate sets of random effects that capture the within-cluster association and the extra variability due to overdispersion. In this paper, we propose a series of hierarchical Bayesian generalized linear mixed models that deal simultaneously with both phenomena. The proposed models are applied to a sample of multivariate data on hepatitis C virus (HCV) and human immunodeficiency virus (HIV) infection prevalence in injecting drug users in Italy from 1998 to 2007.

Keywords: Binomial data; Clustering; Generalized Linear Mixed Models; MCMC; Overdispersion.

Address for correspondence: Emanuele Del Fava, Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat), Universiteit Hasselt, Agoralaan, Gebouw D, B-3590 Diepenbeek, Belgium. Tel: +32 11 26 8297. Fax: +32 11 26 8299. E-mail: emanuele.delfava@uhasselt.be

1 Introduction

Clustering and overdispersion are major issues that must be addressed when modeling data that cannot be assumed to be normally distributed, e.g., aggregated binary data and count data.

The clustering issue refers to the hierarchical structure of data, where measurements belonging to the same cluster are assumed to be associated. This issue can be accommodated using cluster-specific random effects, usually assumed to be normally distributed, which induce the association between the repeated or multivariate measurements. Such models can be easily fitted within the framework of generalized linear mixed models (GLMM, Breslow and Clayton, 1993; Molenberghs and Verbeke, 2005).

We encounter issues of overdispersion when the data present additional variability than the that prescribed by the mean-variance relation of the distribution. The phenomenon of overdispersion has been widely considered in literature, most of all in relation to the binomial and the Poisson distributions. Ignoring to account for overdispersion can lead to the underestimation of the standard errors and therefore to a wrong inference for the regression parameters. Possible solutions to this issue can be of two types (Hinde and Demétrio, 1998). A first approach consists in generalizing the variance function by including additional parameters, such as the heterogeneity factor in overdispersed binomial data, and then estimating the regression parameters using quasi-likelihood methods (Agresti, 2002). A second approach assumes a two-stage model, where in the first stage we define for the data a distribution depending on certain parameters, whose distribution is then specified in the second stage. Examples are the beta-binomial model (Skellam, 1948) for binomial data and the negative binomial model (Breslow, 1984) for count data, but also some versions of the GLMMs. A wide review of approaches able to deal with overdispersion can be found in Hinde and Demétrio (1998).

However, often interest may lie in simultaneously combining these two phenomena, clustering and overdispersion. Both marginal and random-effects models can be used to address them. If the focus is on the estimation of the fixed effects rather than on modeling the correlation structure, marginal models can be used to adjust the variance-covariance structure in order to accommodate for clustering and overdispersion. For instance, both the GEE2 approach (Qaqish and Liang, 1992) and, better suited for binomial data, the alternat-

ing logistic regression (Carey et al., 1993) can be used to model the marginal means of the outcomes as well as the correlation between pairs of within-cluster measurements. Chen and Ahn (1997) go further in this direction and develop a marginal model for multivariate overdispersed binomial data where the mean structure depends on two multiplicative nested random effects. On the other hand, conditional models depending on random effects are a better choice if we are more interested in modeling the individual profiles rather than the population mean and in estimating the correlation among the measurements. Molenberghs et al. (2007, 2010) proposed a class of GLMMs that accommodate for clustering and overdispersion, making use of two separate sets of random effects, which then are estimated with maximum likelihood (ML) methods. These GLMMs are meant to be used for modeling normal, binomial, Poisson and time-to-event data.

In this paper, we focus on conditional models and we thus extend the work presented in Molenberghs et al. (2010) focusing on modeling multivariate, repeated and overdispersed binomial data. For this purpose, we develop a series of GLMMs that account for overdispersion through a set of random effects, either additively or multiplicatively included in the model, while dealing with clustering. To avoid the difficulties encountered with the ML estimation (Molenberghs et al., 2010), we fit the GLMMs within a Bayesian framework using Monte Carlo Markov Chain (MCMC) methods (Clayton, 1996). In such a way, it is possible to specify a prior distribution for the unknown parameters, in particular for the overdispersion random effects and for their covariance matrix, and then calculate their posterior distribution through Gibbs sampling.

We apply the proposed methodology to a sample of prevalence data of hepatitis C virus (HCV) and human immunodeficiency virus (HIV) infection of injecting drug users (IDUs) in treatment from the 20 Italian regions from 1998 to 2007. The paper is organized as follows. In Section 2 we introduce the data. In Section 3 we describe the proposed methodology, giving details about the GLMMs with additive and multiplicative overdispersion parameters and about model selection. Section 4 is dedicated to the presentation of the main results, while in Section 5 we wrap up with a discussion of the proposed methodology.

2 Data

The data analyzed in the paper were reported to the European Monitoring Centre for Drugs and Drug Addiction (EMCDDA) and consist of diagnostic testing binomial data providing information about the HCV and HIV infection prevalence from samples of IDUs in treatment from the twenty regions of Italy between 1998 and 2007. For each IDU, a serum specimen was taken and tested for antibodies against HCV and HIV. Further details about data collection can be found in Del Fava et al. (2011). Note that the data analyzed in this paper are updated to 2007.

Figure 1 (panel a) shows the observed prevalence profiles over the years for HCV and HIV infections, with a bold line representing the national prevalence profile, obtained by pooling together the regional results. We notice that the prevalence of HCV infection is much higher than the prevalence of HIV infection, reflecting the fact that HCV is reported to be about 10 times more infectious than HIV (Crofts et al., 2001). In addition, Figure 1 (panel a) reveals a pattern of large between-region and within-region variability, revealing an issue of overdispersion within regions over the years.

3 Methodology

3.1 A Joint Model for HCV and HIV Infection

The data consist of multivariate repeated binomial measurements collected in a period of 10 years, from 1998 to 2007. Let $\boldsymbol{Y}_i = (\boldsymbol{Y}_{i1}, \boldsymbol{Y}_{i2})$ be the response vector representing the number of reported cases of HCV and HIV infection, respectively, in region i. In turn, let $\boldsymbol{Y}_{ik} = (Y_{i1k}, \ldots, Y_{iJk})$ be the response vector containing the repeated measurements per infection k within the ith region in year j. Del Fava et al. (2011) discussed a joint hierarchical GLMM for HCV and HIV infection prevalence that took into account merely the regional clustering of data. In the first stage of the hierarchical model, they assumed that the distribution of Y_{ijk} is binomial, with sample size equal to n_{ijk}:

$$Y_{ijk} \sim Bin(\pi_{ijk}, n_{ijk}), \qquad i = 1, \ldots, 20 \qquad j = 1, \ldots, 10 \qquad k = 1, 2. \tag{3.1}$$

The primary interest is in the estimation of π_{ij1} and π_{ij2}, which are the prevalence of HCV and HIV infections in the ith region in year j, respectively, and in the association between the two infections at the population level, $\rho(\pi_{ij1}, \pi_{ij2})$.

In the second stage of the hierarchical model, for infection k, they specified a logistic model for the prevalence $\boldsymbol{\pi}_{ij}$. It contains unstructured fixed effects adjusting for time, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}_j$, and region-specific random intercepts, $\boldsymbol{\gamma}_i = (\gamma_{i1}, \gamma_{i2})$:

$$\begin{cases} \text{logit}(\pi_{ij1}) = \alpha_1 + \beta_{j1} + \gamma_{i1}, \\ \text{logit}(\pi_{ij2}) = \alpha_2 + \beta_{j2} + \gamma_{i2}. \end{cases} \tag{3.2}$$

The random intercepts γ_{ik}, which are assumed to follow a bivariate normal distribution, account for the association between the repeated measurements within region, independently of time. Depending on the structure of their covariance matrix $\boldsymbol{D}_\gamma$, the random intercepts might even account for the association between the two infections, $\rho(\pi_{ij1}, \pi_{ij2})$. Del Fava et al. (2011) fitted a series of joint models with different variance structures and concluded that the model with the best goodness-of-fit is the so-called correlated random-effects model. For this model, the region-specific random intercepts are assumed to follow a bivariate normal distribution with a mean vector of zeros and unstructured covariance matrix $\boldsymbol{D}_\gamma$:

$$\boldsymbol{D}_\gamma = \begin{pmatrix} \sigma^2_{\gamma_1} & \rho_{\gamma_1\gamma_2}\sigma_{\gamma_1}\sigma_{\gamma_2} \\ \rho_{\gamma_1\gamma_2}\sigma_{\gamma_1}\sigma_{\gamma_2} & \sigma^2_{\gamma_2} \end{pmatrix}. \tag{3.3}$$

The infection-specific variances of the random intercepts, $\sigma^2_{\gamma_1}$ and $\sigma^2_{\gamma_2}$, account for the within-region association between the repeated measurements, whereas the parameter $\rho_{\gamma_1\gamma_2}$ is the correlation coefficient between the two infections at the level of the linear predictor, being therefore a measure of the association between HCV and HIV infection, $\rho(\pi_{ij1}, \pi_{ij2})$. The prior distributions of the unknown parameters can be found in Del Fava et al. (2011), where this basic model is introduced.

3.2 Joint Model with Additive Overdispersion Parameters

In this section, we propose extensions to the basic joint model (3.2) to deal simultaneously with clustering and overdispersion. This is achieved by including separate sets of random

effects for overdispersion and for clustering. We opt for a set of random effects θ_{ijk}, which, for convenience, are assumed to be independent of the random intercepts γ_{ik}. In this section, we focus on additive overdispersion parameters θ_{ijk} (McLachlan, 1997), which are introduced on the same scale of the linear predictor. We consider four possible situations of interest for the overdispersion random effects: (1) they are shared by the two infections; (2) they are differentiated by infection and independent; (3) they are differentiated by infection and allowed to be dependent; (4) they are differentiated by infection and by year, leading thus to time-specific covariance matrices, $\boldsymbol{D}_{\theta j}$.

3.2.1 Shared Overdispersion Parameters

The first model we consider is the shared overdispersion model, where the overdispersion parameters are shared between the infections, i.e., $\theta_{ij1} = \theta_{ij2} = \theta_{ij}$:

$$\begin{cases} \text{logit}(\pi_{ij1}) = \alpha_1 + \beta_{j1} + \gamma_{i1} + \theta_{ij}, \\ \text{logit}(\pi_{ij2}) = \alpha_2 + \beta_{j2} + \gamma_{i2} + \delta\theta_{ij}. \end{cases} \tag{3.4}$$

We assume that θ_{ij} has a normal prior distribution, $\theta_{ij} \sim N(0, \sigma^2_\theta)$, where σ^2_θ is an hyperparameter with a flat inverse Gamma (IG) prior distribution, that implies $\tau_\theta = 1/\sigma^2_\theta \sim \Gamma(0.01, 0.01)$.

The underlying assumption behind the shared overdispersion model (3.4) is that the correlation between the infections described by the overdispersion parameters is equal to one. We use the parameter δ to relax the assumption of common variance between the random slopes of HCV and HIV infections, since $\sigma^2_{\theta_{HIV}} = \delta^2\sigma^2_{\theta_{HCV}}$. Note that the case with $\sigma^2_\theta = 0$ implies the absence of regional-specific evolution patterns during the years.

3.2.2 Independent Overdispersion Parameters

Another possible model is the independent overdispersion model, which, differently from Model (3.4), includes infection-specific random effects, θ_{ijk}:

$$\begin{cases} \text{logit}(\pi_{ij1}) = \alpha_1 + \beta_{j1} + \gamma_{i1} + \theta_{ij1}, \\ \text{logit}(\pi_{ij2}) = \alpha_2 + \beta_{j2} + \gamma_{i2} + \theta_{ij2}. \end{cases} \tag{3.5}$$

For θ_{ijk} we now assume a bivariate normal prior distribution, with covariance between θ_{ij1} and θ_{ij2} equal to zero:

$$\begin{pmatrix} \theta_{ij1} \\ \theta_{ij2} \end{pmatrix} \sim MVN \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{D}_\theta = \begin{pmatrix} \sigma^2_{\theta_1} & 0 \\ 0 & \sigma^2_{\theta_2} \end{pmatrix} \right]. \tag{3.6}$$

Regarding the variances of the overdispersion parameters θ_{ijk}, we assume that $\sigma^2_{\theta_1}$ and $\sigma^2_{\theta_2}$ are independently distributed according to a flat IG prior distribution, which implies $\tau_{\theta_1} = 1/\sigma^2_{\theta_1} \sim \Gamma(0.01, 0.01)$ and $\tau_{\theta_2} = 1/\sigma^2_{\theta_2} \sim \Gamma(0.01, 0.01)$.

This model assumes that, although there is overdispersion in the time evolution of prevalence among the regions, all correlation between HCV and HIV infections at the regional level is captured fully by the random intercepts, not by the overdispersion parameters.

3.2.3 Correlated Overdispersion Parameters

As a further extension, Model (3.5) can be expressed as a correlated overdispersion model. The new model is similar to the independent overdispersion model (3.5), except for the overdispersion parameters that are now correlated between the infections:

$$\begin{pmatrix} \theta_{ij1} \\ \theta_{ij2} \end{pmatrix} \sim MVN \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{D}_\theta = \begin{pmatrix} \sigma^2_{\theta_1} & \rho_{\theta_1\theta_2}\sigma_{\theta_1}\sigma_{\theta_2} \\ \rho_{\theta_1\theta_2}\sigma_{\theta_1}\sigma_{\theta_2} & \sigma^2_{\theta_2} \end{pmatrix} \right]. \tag{3.7}$$

For the covariance matrix $\boldsymbol{D}_\theta$, we specify an inverse-Wishart (IW) prior distribution, corresponding to a Wishart distribution for its inverse, $\boldsymbol{D}_\theta^{-1} \sim W_2(\boldsymbol{\Psi}, 2)$, where $\boldsymbol{\Psi}$ is a 2×2 identity matrix.

Assuming an unstructured covariance matrix for $\boldsymbol{D}_\theta$, it is possible to estimate an additional correlation between the infections, $\rho_{\theta_1\theta_2}$. This implies that, after having accounted for the correlation between the infections using the region-specific random intercepts γ_{ik}, there is still correlation between HCV and HIV infections in the time evolution of prevalence that is captured by the overdispersion parameters.

3.2.4 Correlated Overdispersion Parameters with Time-Dependent Correlation

The last additive model that we consider extends Model (3.7) relaxing the hypothesis of a constant correlation between HCV and HIV infections captured by the overdispersion parameters. We now let the covariance matrix $\boldsymbol{D}_\theta$ change in each year:

$$\begin{pmatrix} \theta_{ij1} \\ \theta_{ij2} \end{pmatrix} \sim MVN \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{D}_{\theta j} = \begin{pmatrix} \sigma^2_{\theta_1|j} & \rho_{\theta_1\theta_2|j}\sigma_{\theta_1|j}\sigma_{\theta_2|j} \\ \rho_{\theta_1\theta_2|j}\sigma_{\theta_1|j}\sigma_{\theta_2|j} & \sigma^2_{\theta_2|j} \end{pmatrix} \right]. \quad (3.8)$$

For the covariance matrix $\boldsymbol{D}_{\theta j}$, we specify an inverse-Wishart (IW) prior distribution different from each year j, $\boldsymbol{D}^{-1}_{\theta j} \sim W_2(\boldsymbol{\Psi}, 2)$, where $\boldsymbol{\Psi}$ is a 2×2 identity matrix and $j = 1, \ldots, 10$.

3.3 Joint GLMM with Multiplicative Overdispersion Parameters

We consider a setting in which we account for overdispersion using multiplicative effects (McLachlan, 1997; Molenberghs et al., 2010). While the random intercepts γ_{ik} induce association between the clustered measurements, the parameters θ_{ijk} account for additional overdispersion. Hence, in this section, we assume that

$$\begin{cases} Y_{ijk} \sim Bin(\pi_{ijk} = \theta_{ijk} \cdot \kappa_{ijk}, n_{ijk}), \\ \text{logit}(\kappa_{ij1}) = \alpha_1 + \beta_{j1} + \gamma_{i1}, \\ \text{logit}(\kappa_{ij2}) = \alpha_2 + \beta_{j2} + \gamma_{i2}. \end{cases} \quad (3.9)$$

Note that $0 \leq \theta_{ijk} \leq 1$ must hold to ensure that $0 \leq \theta_{ijk}\kappa_{ijk} \leq 1$.

We specify a Beta prior distribution for θ_{ijk}. As a special case, we use a Beta distribution with parameters equal to 1, equivalent to a Uniform distribution over the range (0,1), which implies a noninformative prior for the overdispersion parameters:

$$\begin{aligned} \theta_{ij1} &\sim Be(1,1), \\ \theta_{ij2} &\sim Be(1,1). \end{aligned} \quad (3.10)$$

However, in general, we can assume that the parameters of the beta prior distributions are

hyperparameters to be estimated:

$$\begin{aligned} \theta_{ij1} &\sim Be(a,b), \\ \theta_{ij2} &\sim Be(a,b). \end{aligned} \tag{3.11}$$

Then, the infection-specific variance of the overdispersion random effects can be calculated as

$$\sigma^2 = \frac{ab}{(a+b)^2(a+b+1)}. \tag{3.12}$$

As concerns the prior distribution of the hyperparameters a and b in (3.12), we choose independent diffuse Uniform distributions within the range [0,100]:

$$\begin{aligned} a &\sim U(0,100), \\ b &\sim U(0,100). \end{aligned} \tag{3.13}$$

3.4 Model Selection

A hierarchical mixed-effects model may be seen as a missing data problem, where the random effects are regarded as the missing information. When dealing with missing data problems, Celeux et al. (2006) showed that the deviance information criterion (DIC, Spiegelhalter et al., 2002), which is the typical selection criterion used for Bayesian models, does not work properly with distributions outside the exponential family. Moreover, Plummer (2008) suggested that the approximation used to compute the DIC is valid only when the effective number of parameters (the penalty pD) is much smaller than the number of independent observations n.

To overcome these issues, we use two different criteria to select the best model: the penalized expected deviance (PED, Plummer, 2008), and the difference in posterior deviance (Aitkin et al., 2009; Aitkin, 2010).

The PED can be considered as a loss function when predicting the data Y using the same data Y. The issue of using the data twice (for estimation as well as for prediction) makes the expected deviance $\overline{D(\theta)}$ too optimistic (Plummer, 2008). Thus, the PED penalizes it with a measure for model complexity, p_{opt}:

$$PED = \overline{D(\theta)} + p_{opt}. \tag{3.14}$$

Even though outside the exponential family it is hard to calculate the optimism parameter p_{opt}, it can be estimated for general models using MCMC methods. According to Plummer (2011), the software JAGS uses importance sampling to estimate the parameter p_{opt}. However, the author warns that the estimates may result numerically unstable when the effective number of parameters is high, as it typically occurs with random-effects models. Similarly to the DIC, the smaller the PED, the better.

The difference in posterior deviances (Aitkin et al., 2009; Aitkin, 2010) permits to compare pairs of models to select the best one. This approach is based on the observation that models with growing numbers of parameters are automatically penalized by the increasing diffuseness of the posterior distributions in their parameters. Thus, we can base the selection between two models on the difference between the whole posterior distributions of their deviances,

$$\left\{ D_{1,2}^{(m)} = D_1^{(m)} - D_2^{(m)} : m = 1, \ldots, M \right\}, \tag{3.15}$$

where M is the length of the MCMC chain. We can derive the posterior probability that Model 1 is better than Model 2,

$$P(D_{1,2}^{(m)} < 0) = \frac{1}{M} \sum_{m=1}^{M} I(D_{1,2}^{(m)} < -2\log 9 = 4.39), \tag{3.16}$$

where the value $-2\log 9$ is calibrated in order to correspond to a likelihood ratio test favoring Model 1 with a posterior probability of 0.9 (Aitkin, 2010). It is also possible to derive 95% credible intervals (CI) for the difference in deviances: a 95% CI totally negative implies that we favor Model 1 over Model 2.

4 Results

The hierarchical Bayesian models presented in the previous section are fitted to data using MCMC methods, specifically Gibbs sampling implemented through JAGS software (Plummer, 2003). For each model, we used three chains of 250000 iterations each, burn-in of 125000 and thinning of 125. Convergence for all parameters was assessed with the potential scale reduction factor (Gelman and Rubin, 1992), for which approximate convergence is diagnosed when the factor approaches one. For each model the DIC and the PED are

computed, based on further 20000 iterations; furthermore, we compute the difference in posterior deviances for each pair of models. We refer to Table 1 for a summary of the main results. For each model, we give the values of each selection criterion. We notice that the PED and the difference in posterior deviances lead us to favor different models. As expected, the worst model is the basic joint GLMM (Model (3.2)). Among the models with additive overdispersion parameters, the PED indicates that the shared random-effects model is the best model (Model (3.4)), whereas the difference in posterior deviances favors the model with correlated overdispersion parameters and time-dependent correlation coefficients $\rho_{\theta_1\theta_2}$ between HCV and HIV infection (Model (3.8)). We discard the model with correlated overdispersion parameters (Model (3.7)), implying that the assumption of a constant correlation between HCV and HIV infection over the years on the scale of the overdispersion parameters is not tenable. The same results given by the difference in posterior deviances are obtained when the DIC is used for model selection. Indeed, from Figure 2 (panel d), showing the posterior means of the time-dependent correlation coefficients with their respective 95% CI, we observe that the correlation is never significantly different from zero, except for 2006, when it becomes significantly positive. For the multiplicative models, according to the difference in posterior deviances and DIC, Model (3.10) outperforms Model (3.11), while the PED ranks the models in the other way around.

Finally, when comparing the best additive model and the best multiplicative model, the PED favors the additive model with shared random effects, while, according to the difference in posterior deviance we do not have enough confidence to choose between Model (3.8) and Model (3.10). The DIC criterion favors Model (3.8). Figure 2, (panels a–c), for each fitted model, presents the posterior means of the variance components (with 95% CI), for HCV and HIV infection, respectively, and of the correlation for the clustering random effects γ_{ik}. For the models with additive random effects only, (Models 3.2 – 3.8), we notice that misspecifying the overdispersion random effects or even ignoring them affects neither the estimates of the variances components for the clustering random effects, nor the length of their credible intervals (panels a and b). This may be in relation to the fact the γ_{ik} and θ_{ik} are assumed to be independent and are both introduced on the same scale of the linear predictor, therefore the former accommodate the within-region association all over the years, while the latter capture the unexplained additional variability. However, the same

argument does not hold for the models with multiplicative random effects. For instance, for Model (3.10) and, to a lesser extent, for Model (3.11), we notice that $\hat{\sigma}^2_{\gamma_1}$ (but not $\hat{\sigma}^2_{\gamma_2}$) is larger with a wider CI. This is reflected in smaller values of the correlation $\rho_{\gamma_1\gamma_2}$ and longer CI, as can be observed in Figure 2 (panel c).

We refer to Figures 1 – 4 for a graphical representation of the results. Figure 1 displays the observed regional profiles and the fitted profiles from the basic model. We notice that Model (3.2) in Figure 1 (panel b) shows parallel regional profiles, because only the cluster-specific random effects are specified, thus failing to describe all the excess variability within the data. This is instead accomplished by the best additive and multiplicative models (according to PED and the difference in posterior deviances), plotted in Figure 3. What is most striking from the graphical representation is that the fitted regional profiles from the pair of additive models look very similar, as well as it happens with the pair of multiplicative models. Finally, Figure 4 displays the marginal prevalence of HCV and HIV infection with the respective 95% CI for the basic model as well as the overdispersion models with additive and multiplicative random effects. For comparison, we plotted also the national prevalence per year, obtained by pooling together all the regional results. We notice that the five estimated prevalence profiles are fairly close to the national observed prevalence. What really distinguishes the marginal prevalence estimates from each other is their credible intervals, which account for all the variability shown by the regional profiles.

5 Discussion

In this paper, we discussed a set of hierarchical models that can account simultaneously with clustering and overdispersion for binomial data. This objective has been achieved using two separate sets of random effects, one to induce association among the within-region measurements and between the infections, the other to account for the extra binomial variability in the data. For the latter set of random effects, two different settings have been considered: (1) the overdispersion parameters are included additively into the model, on the same scale of the linear predictor; (2) the overdispersion parameters are included multiplicatively in the model, correcting the prevalence in order to encompass the additional variability. All fitted combined models outperform the so-called basic joint model (Del Fava

et al., 2011), which accommodates only the clustered nature of the data. This result was expected since the random-intercept model (3.2) cannot capture the complete association structure in the data.

Under the additive approach, the overdispersion parameters can be seen as random slopes adjusting for clustering, as it happens with normal outcomes (Molenberghs et al., 2010), where overdispersion and cluster-specific random effects coincide. Indeed, the region-specific random intercepts γ_{ik} induce association averaging out the yearly measurements, while the overdispersion random slopes θ_{ijk} further adjust for the variation within each year. Under the multiplicative approach, the random effects deal differently with the clustering and the overdispersion. While the random intercepts γ_{ik} induce the within region-association and are responsible for shifting the regional profiles, the overdispersion random effects θ_{ijk} act directly on the prevalence π_{ijk} and adjust the standard mean-variance relation of the binomial distribution by inflating the variance of the estimated prevalence according to the time-specific variations.

Several models have been proposed for each setting according to the dependence structure among the overdispersion random effects and therefore on their covariance matrix $\boldsymbol{D}_\theta$. In both settings, we parameterize both the correlation between HCV and HIV infection over time (via the joint distribution of the random intercepts γ_{ik}) and within a specific time point (via the joint distribution of the overdispersion parameters θ_{ijk}).

For model selection, the DIC (Spiegelhalter et al., 2002) cannot be considered a valid option when the condition $p_D << n$ does not hold (Plummer, 2008), because it tends to under-penalize more complex models. As concerns the other selection criteria here used, i.e., the PED and the difference in posterior deviances, they may lead to different conclusions, even though both methods claim that they penalize the more excessively parameterized models. To our experience, however, it seems that PED tends to favors less parameterized models, while the difference in posterior deviances selects more complex models. In our case, the posterior means estimated by the different best models look similar, as it appears from the graphical representation of prevalence.

Even though in this paper we presented several extensions to the basic model of Del Fava et al. (2011), still some enhancements are possible. A possible extension for the model with unstructured time-dependent overdispersion parameters could be a parametric

model for the correlation between HCV and HIV infections at the level of the overdispersion parameters over time. Furthermore, we could model the correlation between the overdispersion parameters for HCV and HIV using information collected at regional level, concerning risk factors related to injecting drug use (percentage of sharing syringes or other paraphernalia, etc.) or results of interventions (percentage of supplies of clean drug injection equipment, opioid substitution and other forms of drug dependence treatments, antiviral treatments for HIV, health promotion, etc.). An investigation of these type of models is a subject for an ongoing research and will be published in the near future.

References

Agresti, A. (2002) Categorical data analysis. 2nd Edition. New York, NY: John Wiley & Sons.

Aitkin, M. (2010) Statistical Inference. An Integrated Bayesian/Likelihood Approach. Boca Raton, FL: Chapman & Hall.

Breslow, N.E. (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.

Carey, V., Zeger, S.L., and Diggle. P. (1993). Modelling multivariate binary data with alternating logistic regression. Biometrika, 80, 517–526.

Celeux, G., Forbes, F., Robert, C.P., and Titterington, D.M. (2006) Deviance information criteria for missing data models. Bayesian Analysis, 1, 651–674.

Chen, J.J., and Ahn, H. (1997) Marginal models with multiplicative variance components for overdispersed binomial data. Journal of Agricultural, Biological, and Environmental Statistics, 2, 440–450.

Clayton, D. (1996) Generalized linear mixed models. In: Gilks, W.R., Richardson, S., Spiegelhalter, D.J. eds. Markov Chain Monte Carlo in Practice. London: Chapman & Hall.

Crofts, N., Dore, G., and Locarnini, S. (2001) Hepatitis C: an Australian perspective. Melbourne.

Del Fava, E., Kasim. A., Usman, M., Shkedy, Z., et al. (2011) Joint Modeling of HCV and HIV Infections among Injecting Drug Users in Italy Using Repeated Cross-Sectional

Prevalence Data. Statistical Communications in Infectious Diseases, 3: 1–24.

Gelman, A. and Rubin, D.B. (1992) Inference from iterative simulation using multiple sequences. Statistical Science, 7, 457–511.

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

Molenberghs, G. and Verbeke, G. (2005) Models for Discrete Longitudinal Data. Berlin: Springer-Verlag.

McLachlan, G.J. (1997) On the EM algorithm for overdispersed count data. Statistical Methods in Medical Research, 6, 76–98.

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

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

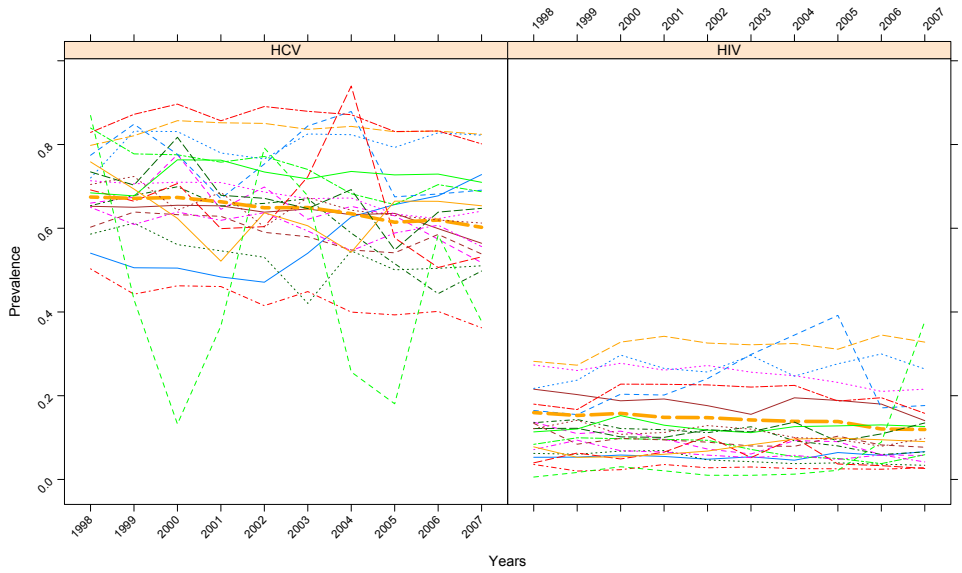
Plummer, M. (2003) JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling. Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), March 20–22, Vienna, Austria. ISSN 1609-395X.

Plummer, M. (2008) Penalized loss functions for Bayesian model comparison. Biostatistics, 9, 523–539.

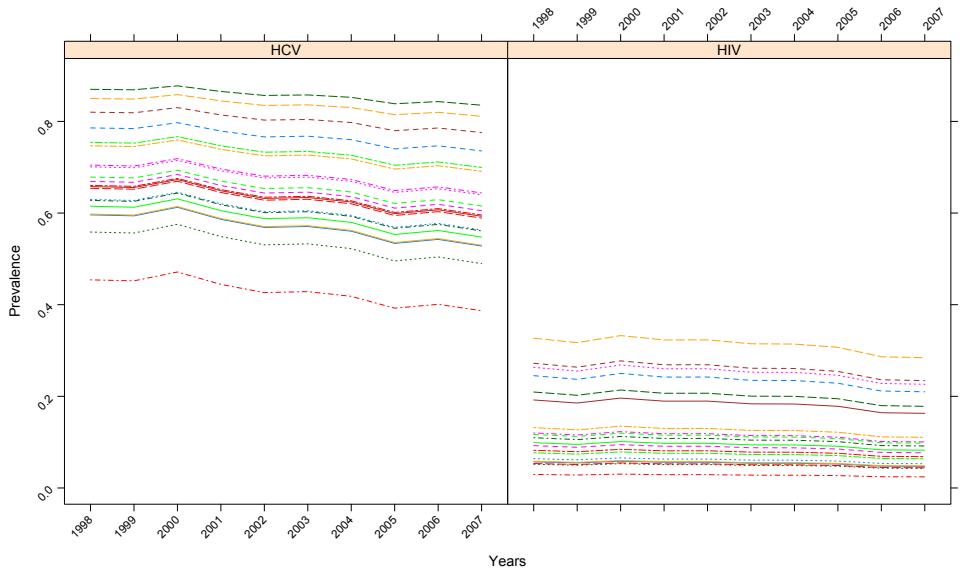
Qaqish, B.F., and Liang, K-Y. (1992) Marginal models for correlated binary responses with multiple classes and multiple levels of nesting. Biometrics, 48, 939–950.

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.

Spiegelhalter, D.J., Best, N.J., Carlin, B.P., and Van der Linde, A. (2002) Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society, series B, 64, 583–640.

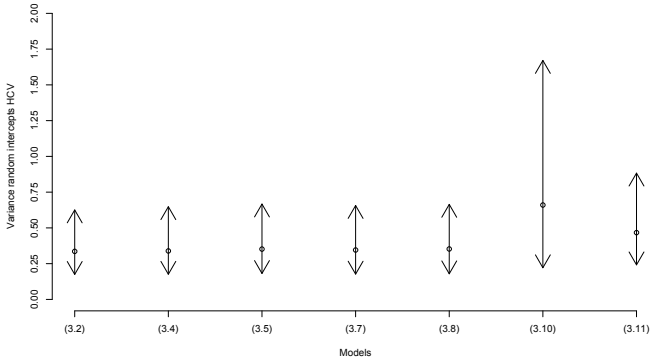


(a) Observed data

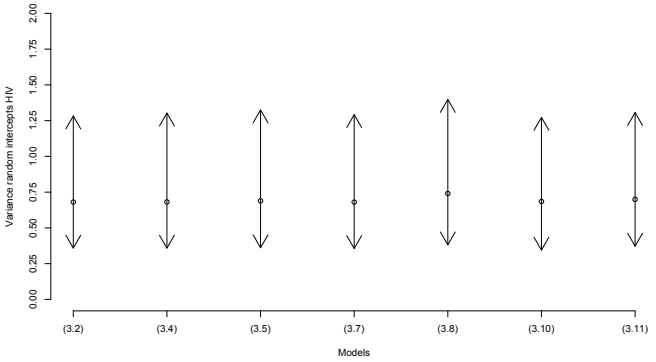


(b) Basic model

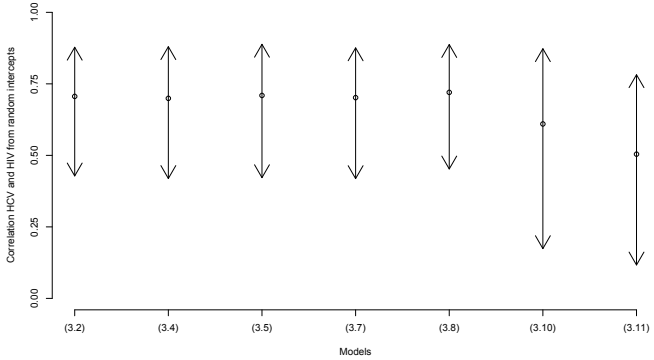
Figure 1: Observed (panel a) and estimated (basic model, panel b) individual prevalence profiles of HCV and HIV infections for the 20 Italian regions between 1998 and 2007. The bold line in the upper left panel stands for the overall prevalence, obtained by pooling together the regional results.



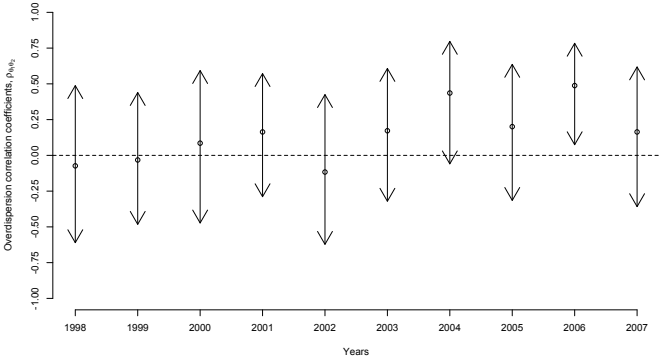
(a) Variance γ_{i1} per model



(b) Variance γ_{i2} per model

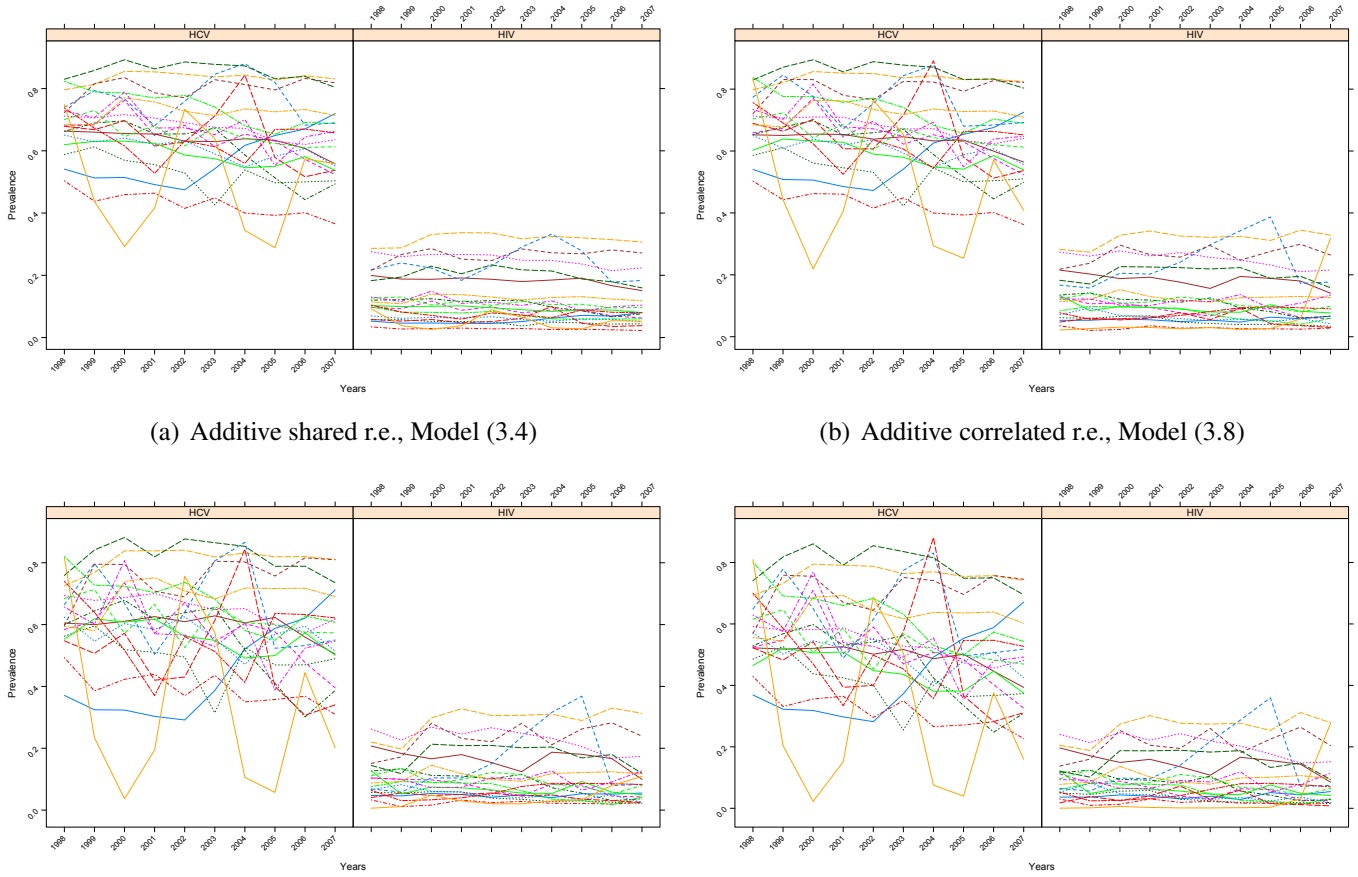


(c) Correlation $\rho_{\gamma_1\gamma_2}$ per model



(d) Correlation $\rho_{\theta_1\theta_2}$ from Model (3.8)

Figure 2: Posterior means and respective 95% CI of the infection-specific variances (panel a and b) and correlation (panel c) of random intercepts γ_{ik} per model, and of the time-dependent correlation (panel d) on the scale of overdispersion parameters θ_{ijk} from Model (3.8).



(a) Additive shared r.e., Model (3.4)

(b) Additive correlated r.e., Model (3.8)

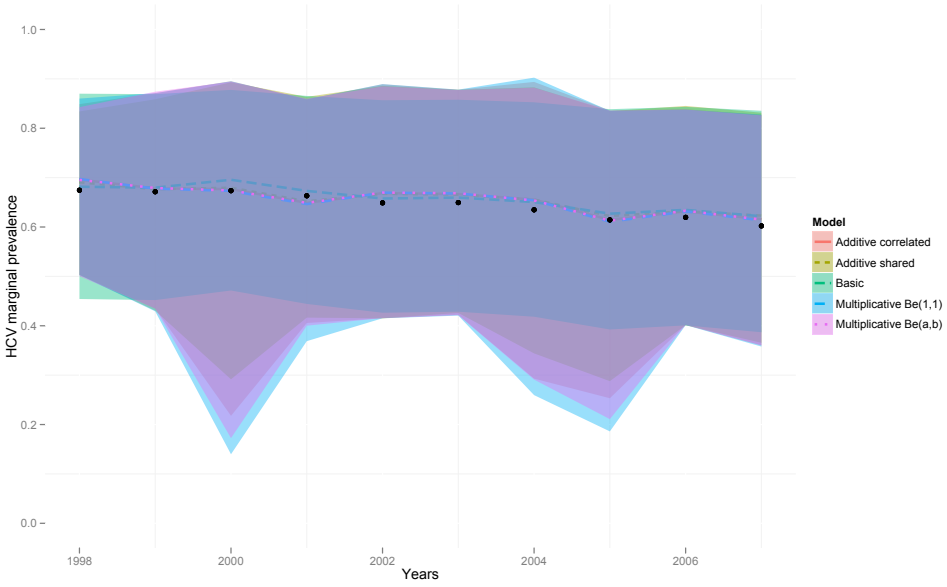
(c) Multiplicative Be(a,b), Model (3.11)

(d) Multiplicative Be(1,1), Model (3.10)

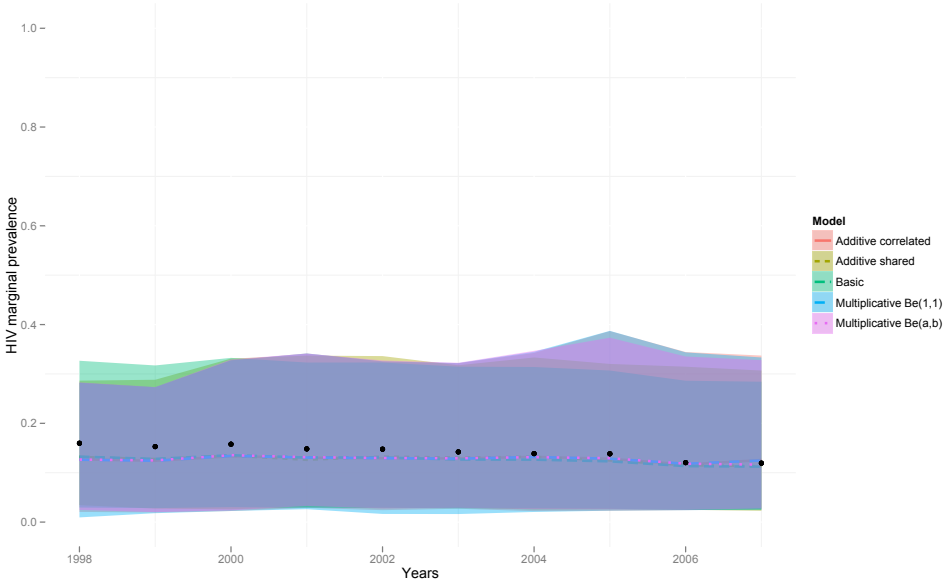
Figure 3: Individual fitted prevalence profiles of HCV and HIV infections for the 20 Italian regions between 1998 and 2007, resulting from the best models according to PED (panel a and c) and according to the difference in posterior deviances (panel b and d).

Table 1: Information criteria for model selection.

Type	Model	DIC	PED	Diff. $\overline{D(\theta)}$
Basic	(3.2)	10274	10351	-
Additive	(3.4)	4998	6176	$D_{3.4,3.2} = \overline{D_{3.4}} - \overline{D_{3.2}}$ $-5442(-5491,-5391)$ $P(D_{3.4,3.2} < 4.39) = 1$
	(3.5)	3835	7469	$D_{3.5,3.4} = \overline{D_{3.5}} - \overline{D_{3.4}}$ $-1304(-1378,-1227)$ $P(D_{3.5,3.4} < 4.39) = 1$
	(3.7)	3835	7515	$D_{3.7,3.5} = \overline{D_{3.7}} - \overline{D_{3.5}}$ $-1(-89,84)$ $P(D_{3.7,3.5} < 4.39) = 0.47$
	(3.8)	3775	8099	$D_{3.8,3.5} = \overline{D_{3.8}} - \overline{D_{3.7}}$ $-73(-157,11)$ $P(D_{3.8,3.5} < 4.39) = 0.94$
Multiplicative	(3.10)	3782	8270	$D_{3.10,3.8} = \overline{D_{3.10}} - \overline{D_{3.8}}$ $-7(-86,76)$ $P(D_{3.10,3.8} < 4.39) = 0.52$
	(3.11)	3869	7224	$D_{3.10,3.11} = \overline{D_{3.10}} - \overline{D_{3.11}}$ $-113(-200,-22)$ $P(D_{3.10,3.11} < 4.39) = 0.99$



(a) HCV



(b) HIV

Figure 4: Marginal prevalence π_{jk} for HCV (panel a) and HIV (panel b) with 95% CI, obtained by averaging out all over the regions the prevalence per year j and infection k, from the basic model (green line), the additive correlated model with time-dependent correlation coefficients (red line), the additive shared model (golden line), the multiplicative $Be(1,1)$ model (blue line), and the multiplicative $Be(a,b)$ model (pink line). The black dots stand for the national observed prevalence, obtained pooling together all the regional prevalences per each year.