Statistical Communications in Infectious Diseases

Volume 3, *Issue* 1 2011 *Article* 1

Joint Modeling of HCV and HIV Infections among Injecting Drug Users in Italy Using Repeated Cross-Sectional Prevalence Data

Emanuele Del Fava, *Hasselt University*
Adetayo Kasim, *Durham University*
Muhammad Usman, *University of Oxford*
Ziv Shkedy, *Hasselt University*
Niel Hens, *Hasselt University and University of Antwerp*
Marc Aerts, *Hasselt University*
Kaatje Bollaerts, *Scientific Institute of Public Health*
Gianpaolo Scalia Tomba, *University of Rome "Tor Vergata"*
Peter Vickerman, *London School of Hygiene & Tropical Medicine and University of Bristol*
Andrew J. Sutton, *University of Warwick*
Lucas Wiessing, *European Monitoring Centre for Drugs and Drug Addiction*
Mirjam Kretzschmar, *University Medical Centre Utrecht and the Center for Infectious Disease Control (RIVM)*

Recommended Citation:
Del Fava, Emanuele; Kasim, Adetayo; Usman, Muhammad; Shkedy, Ziv; Hens, Niel; Aerts, Marc; Bollaerts, Kaatje; Scalia Tomba, Gianpaolo; Vickerman, Peter; Sutton, Andrew J.; Wiessing, Lucas; and Kretzschmar, Mirjam (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*: Vol. 3: Iss. 1, Article 1.
DOI: 10.2202/1948-4690.1009
Available at: http://www.bepress.com/scid/vol3/iss1/art1

Joint Modeling of HCV and HIV Infections among Injecting Drug Users in Italy Using Repeated Cross-Sectional Prevalence Data

Emanuele Del Fava, Adetayo Kasim, Muhammad Usman, Ziv Shkedy, Niel Hens, Marc Aerts, Kaatje Bollaerts, Gianpaolo Scalia Tomba, Peter Vickerman, Andrew J. Sutton, Lucas Wiessing, and Mirjam Kretzschmar

Abstract

During their injecting career, injecting drug users (IDUs) are exposed to some infections, like hepatitis C virus (HCV) infection and human immunodeficiency virus (HIV) infection, due to their injecting behavioral risk factors, such as sharing syringes or other paraphernalia containing infected blood, or sexual behavior risk factors. If we consider that these IDUs might belong to a social network of people where these behavioral risk factors are spread, then HCV and HIV infections might be associated at both the individual and the population level. In this paper, we study the association between HCV and HIV infection at the population level using aggregate data. Our aim is to define a hierarchy of structured models with which the association between HCV and HIV infection at population level and the time trend of prevalence can be investigated. The data analyzed in the paper are "diagnostic testing data," which consist of repeated cross-sectional prevalence measurements from 1998 to 2006 for HCV and HIV infection, obtained from a sample of 515 drug treatment centers spread among the 20 regions in Italy, where subjects went for a serum diagnostic test. Since we do not have any individual data, it is not possible to relate these prevalence data to socio-demographic or behavioral risk data. Each region defines a cluster with repeated prevalence data for HCV and HIV infection over time. Several modeling approaches, such as generalized linear mixed models (GLMMs) and hierarchical Bayesian models are applied to the data. First, we test different covariance structures for the region-specific random effects in the GLMM context; second, a hierarchical Bayesian model is used to refit the best GLMM in order to obtain the posterior distribution for the parameters of primary interest. We found that the correlation at population level between HCV and HIV is approximately 0.68 and the prevalence of the two infections generally decreased over the years, compared to the situation in 1998.

KEYWORDS: association between HCV and HIV infection, IDUs, generalized linear mixed models, hierarchical Bayesian models

Author Notes: Emanuele Del Fava, Interuniversity Institute for Biostatistics and statistical Bioinformatics, Hasselt University, Belgium. Adetayo Kasim, Wolfson Research Institute, Durham University, UK. Muhammad Usman, Diabetes Trials Unit, OCDEM, University of

Oxford, UK. Ziv Shkedy, Interuniversity Institute for Biostatistics and statistical Bioinformatics, Hasselt University, Belgium. Niel Hens, Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Hasselt University and the Centre for Health Economics Research and Modeling Infectious Diseases, Centre for the Evaluation of Vaccination, Vaccine & Infectious Disease Institute, University of Antwerp, Belgium. Marc Aerts, Interuniversity Institute for Biostatistics and statistical Bioinformatics, Hasselt University, Belgium. Kaatje Bollaerts, Public Health & Surveillance, Scientific Institute of Public Health, Brussels, Belgium. Gianpaolo Scalia Tomba, Dept. Mathematics, University of Rome "Tor Vergata," Italy. Peter Vickerman, London School of Hygiene and Tropical Medicine, London and the University of Bristol, UK. Andrew J. Sutton, Dept. Biological Sciences, University of Warwick, UK. Lucas Wiessing, European Monitoring Centre for Drugs and Drug Addiction, Lisbon, Portugal. Mirjam Kretzschmar, Julius Centre for Health Sciences & Primary Care, University Medical Centre Utrecht and the Center for Infectious Disease Control (RIVM), Bilthoven, Netherlands.
We are grateful to the colleagues in the SER.T (public drug-treatment center) for collecting the data and to Istituto Superiore di Sanità - Centro Operativo AIDS for providing the data to EMCDDA. This study contributes to the work of the "European Study Group for Mathematical Modeling and Epidemiological Analysis of Drug-Related Infectious Diseases," coordinated by EMCDDA and RIVM with funding from WHO/ Europe and the government of the Netherlands. EDF thanks Professor Piero Manfredi for the precious and insightful scientific discussion on the matter that they had together in Pisa. NH gratefully acknowledges financial support from "SIMID", a strategic basic research project funded by the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), project number 060081. EDF, AK, ZS, NH, and MA gratefully acknowledge the IAP research network number P6/03 of the Belgian Government (Belgian Science Policy). EDF, ZS, NH, MA gratefully acknowledge support from the Belgian IUAP/PAI network "Statistical techniques and modeling for complex substantive questions with complex data." Finally, we thank the editor and the referees for the valuable comments that helped us to improve substantially the content and the presentation of the manuscript.

1 Introduction

During their injecting career, injecting drug users (IDUs) are exposed to several infections such as hepatitis C virus (HCV), human immunodeficiency virus (HIV), and hepatitis B virus (HBV). In the last years, a number of studies analyzed the prevalence of HCV infection within the IDU population and the statistical association of HCV infection with HIV infection in the same population (Hutchinson, 2004; Hope *et al.*, 2005; Mathëi *et al.*, 2006; Sutton *et al.*, 2006; Barrio *et al.*, 2007). It has been found that the risk of infection depends on several aspects of injecting, e.g., the length of injection career, the frequency of injection, sharing syringes and other paraphernalia (Hutchinson, 2004; Mathëi *et al.*, 2006). The association between risk factors and disease status can be studied using cross-sectional serological data, from which the prevalence and the force of infection can be estimated. Furthermore, in case that information about more than one infection is available, co-infection can be studied at individual level (Sutton *et al.*, 2008; Hens *et al.*, 2009; Del Fava *et al.*, 2011). In particular Del Fava *et al.* (2011) showed that there is a clear pattern of co-infection with HCV and HIV among IDUs in Italy and Spain based on cross-sectional serological survey data: they investigated how behavioral risk factors affect the association between the infections and concluded that IDUs who are infected by one virus are more likely to be infected by the other as well.

However, more often, only aggregate prevalence data are available, possibly collected over several years. The analysis of these yearly prevalence data is useful to investigate the time trend in prevalence, to establish intervention scenarios and evaluate the results of health policies aimed to reduce behavioral risks, and to study the evolution of the association between different infections. For instance, Vickerman *et al.* (2010) used prevalence data for IDUs from many geographical areas all over the world and estimated a strong positive correlation between the change in HIV infection prevalence and the change in HCV infection prevalence over time. Specifically, the time series suggest that, when the prevalence of HCV infection is low, any change in HIV prevalence over time is smaller than a change in HCV infection prevalence in the same period; however, this difference reduces at higher levels of HCV infection prevalence. Consequently, the authors postulate that HCV infection prevalence can be seen as a population-level marker of injection-related HIV risk, especially when the prevalence of HCV infection is high.

In this paper, we model the association between HCV and HIV infection at population level, using aggregate serological data from 20 regions in Italy, collected from 01/01/1998 to 31/12/2006. We focus the investigation on two points: (1) the change of HCV and HIV infection prevalence over time and (2) the correlation between HCV and HIV infection among the regions. In contrast with Vickerman *et* *al.* (2010), who modeled the dependency of HIV infection on HCV infection using a conditional model, we use a joint model with random effects for the binomial prevalence data of HCV and HIV infection. Two types of random-effects models are used: generalized linear mixed models (GLMM, McCulloch and Searle, 2001; Molenberghs and Verbeke, 2005) and hierarchical Bayesian models (Gilks *et al.*, 1996; Gelman *et al.*, 2004), the latter used to refit the best GLMM in order to obtain more information about the parameters of interest. Both modeling approaches can allow for overdispersion in binomial data, that is to say, they can deal with the variability in the data that is not adequately captured by the model's prescribed mean-variance link (Molenberghs *et al.*, 2010): in particular, we use them to capture the variability at the regional level. Note than we do not assume a priori that the two infections are associated, but we rather test different hypotheses for the covariance matrix of the random effects in order to find which one fits the data best.

The structure of this paper is as follows. In Section 2, we introduce and discuss the data. In Section 3, we focus on statistical methodology and formulate a sequence of GLMMs to model the association between HCV and HIV infection. The proposed models are fitted to the data and results are presented in Section 4. In Section 5, we formulate the hierarchical Bayesian model and we present the results of the analysis. Finally, we discuss and interpret all the results in Section 6.

2 The Prevalence Series from Italy

The data analyzed in the paper were reported to the European Monitoring Centre for Drugs and Drug Addiction (EMCDDA) and consist of diagnostic testing data providing information about the HCV and HIV infection status of IDUs in treatment from the 20 Italian regions in the period 01/01/1998–31/12/2006. Within the framework of a monitoring system established by the Italian government, these data were collected in 515 drug treatment centers (DTCs) spread all over Italy, from subjects who went there for a diagnostic test for HCV, HIV, and/or HBV. For each drug user, a serum specimen was taken and tested for antibodies against some of the three infections. Indeed, the fact that there is a difference among the sample sizes per year per infection implies that some subjects were not tested for all three infections. Note that individual data are not available for this study.

A first concern about these data is that we cannot distinguish between IDUs and non-IDUs. We have an unknown proportion of low-risk individuals included in the sample, thus we might underestimate the prevalence of both infections. This bias could in theory extend to underestimating the association between HCV and HIV infection among IDUs as well, because non-injectors are much less likely to get HCV, but still carry a sexual risk to be infected with HIV: what follows is that the infection probabilities for HCV and HIV infection might differ between the two behavioral risk groups. However, a recent study, based on individual serological data from a sample of 1330 drug users from the Italian DTCs in 2005 (Camoni *et al.*, 2010), obtained national HCV and HIV infection prevalence estimates comparable with those from the aggregate data used in this paper: as regards HCV infection, the estimated prevalence in 2005 from the individual data was 83.2% in IDUs and 22% in non-IDUs, whereas with aggregate data we estimated a HCV infection prevalence among drug users of 61.4%; as regards HIV infection, Camoni *et al.* estimated a prevalence in 2005 of 14.4% in IDUs and of 1.6% in non-IDUs, instead the aggregate data here used provided an estimated prevalence in drug users of 13.8%. This indicates that the dominant risk group in the aggregate data is that of IDUs. A second concern is that subjects usually self-selected (or were selected by physicians) and were not recruited in a follow up study. However, the surveillance system, based on a national protocol, is constant over time and across regions; as a consequence, it is very likely that the population over time is comparable, as the data mainly concern with drug users in longer term treatment, e.g., methadone substitution, who tend to stay in treatment for many years (up to life time), implying that the population has little turnover. A third problem regards the comparability of the population as concerns the uptake of the test and the retesting procedure, which may underestimate the prevalence. On the one hand, it is unknown, but likely, that someone with a positive test is not retested, and it is unknown to what extent known positive tests are re-reported to the national system in following years. On the other hand, since people asking for a test may be characterized by risky behaviors, the prevalence may be overestimated. In every case, we note that these potential biases may be more severe for prevalence and less severe for the correlation between HIV and HCV infection. Indeed, it seems reasonable to assume that these systematic biases work in the same direction on both infections, thus only the prevalence might be affected, not the correlation. In summary, although these diagnostic testing data are not the outcome of a designed study, they provide information about the prevalence of both HCV and HIV infection in Italy and can be used to model the change in the prevalence over time and to estimate the association between the two infections (see, for example, Vickerman *et al.*, 2010).

We begin with a preliminary exploratory data analysis. When taking into account the overall prevalence per region (see Figure 1), we notice a clear association between the prevalence of HCV and HIV infection: the Spearman's correlation coefficient between the overall prevalence of HCV and HIV infection at regional level is $\rho = 0.80$ and with the Spearman's rank test we can reject the null hypothesis $H_0 : r = 0$ with $p < 0.0001$. This finding is in agreement with Vickerman *et al.* (2010), who estimated a Pearson's correlation between HCV and HIV of 0.67 among many countries in the world. Figure 2 shows the prevalence of HCV and

HIV infection per region over time, where the regions are sorted by the average HIV infection prevalence over the years. Firstly, 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). Secondly, Figure 2 reveals a pattern of between-region and within-region variability. For instance, in 2000, HCV infection prevalence ranges from 13% (Valle d'Aosta) to 86% (Emilia Romagna), while, in 2006, HIV infection prevalence ranges from 0.3% (Campania) to 55% (Liguria). Instead the within-region variability is due to considerable differences in the sample size in successive time-points, e.g., a few regions are characterized by very large variability, like Valle d'Aosta and Molise, as concerns HCV, and Liguria, as concerns HIV.

The complete datasets, with the number of tested and infected individuals per region and year, are presented in Table 1 and 2 in the supplementary material of this paper.

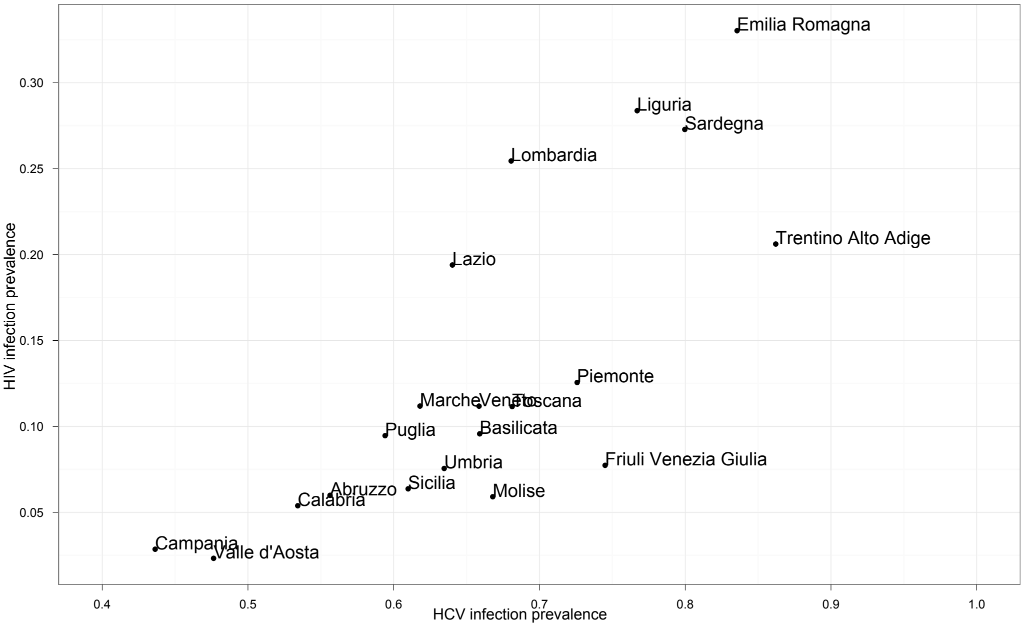

Figure 1: Overall regional observed prevalence (averaged over the years 1998-2006) for each of the 20 regions.

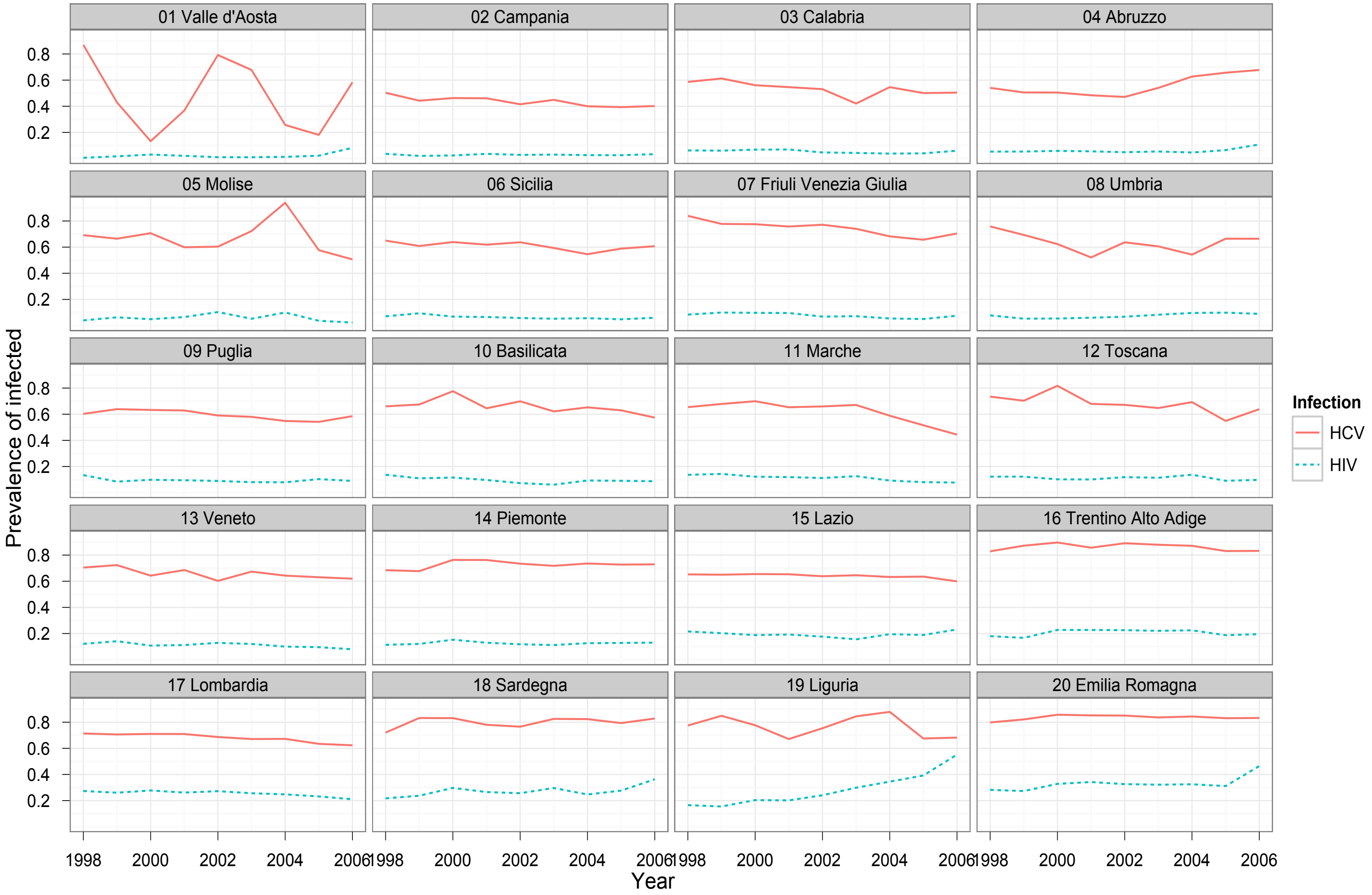

Figure 2: Observed prevalence profiles of HCV infection (continuous line) and HIV infection (dashed line) in Italy from 1998 to 2006 for the 20 regions, sorted by the average HIV infection prevalence over the years.

3 Statistical Methods: Joint Modeling of HCV and HIV Infection Prevalence with GLMMs

In this section, we formulate a sequence of nested random-effects models for binomial data which can allow for the correlation between HCV and HIV infection among the regions over the years, where the effect of time is included in the model as a set of unstructured time evolution slopes. The possible correlation among the observations from the same region and the deviation of the region-specific prevalence from the overall prevalence is captured by region-specific random intercepts.

3.1 The Independence Model

The data consist of aggregate repeated measurements over a period of 9 years, $j = 1,2,\ldots,9$. Let $\boldsymbol{y}_i = (y_{i1}, y_{i2})$ be the response vector for the i–th region, where y_{i1} denotes the number of reported cases of HCV infection and y_{i2} denotes the number of reported cases of HIV infection. Let $\boldsymbol{y}_{ik} = (y_{i1k}, \ldots, y_{iJk})$ be the response vector representing the number of infected individuals with infection k in the i–th region in year j. Let n_{ijk} be the sample size in the i–th region in year j for infection k. We assume that the distribution of y_{ijk} is binomial:

$$y_{ijk} \sim Bin(\pi_{ijk}, n_{ijk}) \qquad i = 1,\ldots,20, \qquad j = 1,\ldots,9, \qquad k = 1,2.$$

Here, $\pi_{ij1} = P(y_{ij1} = 1)$ and $\pi_{ij2} = P(y_{ij2} = 1)$ are the prevalence of HCV and HIV infection in the i–th region in year j, respectively. We further assume a set of unstructured means for the time effect, i.e., we fit infection-specific parameters for each year, but the first:

$$\begin{cases} g(\pi_{ij1}) = \beta_{01} + \beta_{11j}, \\ g(\pi_{ij2}) = \beta_{02} + \beta_{12j}. \end{cases}$$

Choosing the function $g(\cdot)$ to be the logit link, we can interpret the time evolution parameters β_{11j} and β_{12j}, with $j = 2,\ldots,9$, as the log odds ratios of being infected with HCV and HIV, respectively, in year j, compared to the reference year 1998. Note that this model (1) assumes that HCV and HIV infections are independent and there are no region-specific effects.

3.2 Generalized Linear Mixed Models (GLMMs)

3.2.1 The Independent Random-Effects Model

To capture the extra variability at the regional level, the first GLMM includes a region-specific effect in addition to the time effect, while keeping the independence between HCV and HIV infection. Hence, the independence model (1) is rewritten in the following way:

$$\begin{cases} g(\pi_{ij1}) = \beta_{01} + \beta_{11}j + a_i, \\ g(\pi_{ij2}) = \beta_{02} + \beta_{12}j + b_i. \end{cases}$$

Here, a_i and b_i are region and infection-specific random effects assumed to be independent from each other. More precisely, we assume a bivariate normal distribution with variance-covariance matrix for random effects given by

$$\begin{pmatrix} a_i \\ b_i \end{pmatrix} \sim MVN\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, D_{I_1} = \begin{pmatrix} \sigma_a^2 & 0 \\ 0 & \sigma_b^2 \end{pmatrix}\right].$$

The diagonal structure of the covariance matrix D_{I_1} of the random effects for model (2a) implies that the observations on a particular infection within the same region are associated over time, but HCV and HIV infection prevalences are independent. In addition, we test a model (2b) with a more restrictive structure for the independent covariance matrix, that is, we assume that HCV and HIV have independent infection-specific random effects, but with equal variances:

$$D_{I_2} = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}.$$

One can test the null hypothesis $H_0 : \sigma_a^2 = \sigma_b^2$ using the likelihood ratio test (Molenberghs and Verbeke, 2005).

3.2.2 The Shared Random-Effects Model

The random-effects models (2a) and (2b) assume that HCV and HIV infections are independent. Instead, the shared random-effects model takes into consideration the possible association between the two infections in the same region. In order to account for this association, we use the following set of random effects:

$$\begin{cases} g(\pi_{ij1}) = \beta_{01} + \beta_{11}j + b_i, \\ g(\pi_{ij2}) = \beta_{02} + \beta_{12}j + \gamma b_i. \end{cases}$$

Here, b_i is a region-specific random effect assumed to follow a normal distribution, $b_i \sim N(0, \sigma^2)$, and γ is a scale parameter. The underlying assumption behind the shared random-effects model (3) is that the correlation between the random effects is equal to 1. The parameter γ is used to relax the assumption of common variance between the random effects of HCV and HIV infection, since $\sigma^2_{HIV} = \gamma^2\sigma^2_{HCV}$. The case with $\sigma^2 = 0$ implies that the two infections are uniformly spread among the regions. Note that model (3) implies that regions with high levels of HCV have also high levels of HIV, if γ is positive.

3.2.3 The Correlated Random-Effects Model

As mentioned above, the shared random-effects model (3) assumes a perfect positive correlation between the infections at the level of the linear predictor. The next two GLMMs allow to estimate a more realistic value of the correlation. This can be incorporated in the model by specifying two possible covariance structures: an unstructured matrix (model 4a) and a Toeplitz matrix (model 4b). The former matrix allows for different variance parameters for the random effects; the latter assumes only two parameters, that is, equal variances for the random effects and the covariance between them. The two matrices are shown below:

$$\begin{aligned} D_U &= \begin{pmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{pmatrix} \quad \text{Unstructured matrix.} \\ D_T &= \begin{pmatrix} \sigma^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma^2 \end{pmatrix} \quad \text{Toeplitz matrix,} \end{aligned}$$

The two matrices imply that HCV and HIV infections are associated and this association can be modeled directly using the correlation coefficient between the region-specific random intercepts:

$$\rho_U = \frac{\sigma_{ab}}{\sigma_a \sigma_b} \qquad \rho_T = \frac{\sigma_{ab}}{\sigma^2}.$$

A positive correlation coefficient implies concordance between the prevalence of the two infections, that is to say, in a certain region, when the prevalence of HCV infection grows, also the prevalence of HIV infection grows, although with a different magnitude. Note that, for $\rho = 1$, the correlated random-effects models reduce to the shared random-effects model (3), while $\rho = 0$ implies that the models can be reduced to the independent random-effects models (2a) and (2b). We recall here that ρ measures the association between HCV and HIV infection at the level of the linear predictor.

4 Application to the Data: GLMMs

The GLMMs discussed above were fitted with SAS software, using the procedure NLMIXED with adaptive Gaussian quadrature method based on 10 quadrature points: this method is considered to give precise estimates at the price of being computationally intensive (Molenberghs and Verbeke, 2005). Table 1 presents the covariance parameter estimates, together with the Akaike's information criterion (AIC) for each model. Note that the model with the smallest AIC is the one with the best compromise between goodness-of-fit and model complexity. Model selection using the Bayesian information criterion (BIC) gives similar results, as shown in the supplementary material for the paper.

Table 1: Comparison of the fitted models with AIC and parameter estimates for the variance components with respective 95% asymptotic confidence intervals.

Model	Type	AIC	Covariance parameter estimates
1	Independence model	84753	-
3	GLMM shared RE	21032	$\hat{\gamma} = 1.96 \quad (1.92, 2.00)$ $\hat{\sigma}^2 = 0.13 \quad (0.04, 0.22)$
2b	GLMM independent RE Equal parameters	10616	$\hat{\sigma}^2 = 0.41 \quad (0.22, 0.61)$
2a	GLMM independent RE Different parameters	10615	$\hat{\sigma}_a^2 = 0.25 \quad (0.08, 0.42)$ $\hat{\sigma}_b^2 = 0.58 \quad (0.19, 0.98)$
4b	GLMM correlated RE Toeplitz covariance	10608	$\hat{\sigma}^2 = 0.41 \quad (0.18, 0.64)$ $\hat{\sigma}_{ab} = 0.26 \quad (0.03, 0.49)$ $\hat{\rho} = 0.64 \quad (0.35, 0.92)$
4a	GLMM correlated RE Unstructured covariance	10605	$\hat{\sigma}_a^2 = 0.25 \quad (0.08, 0.42)$ $\hat{\sigma}_b^2 = 0.57 \quad (0.19, 0.96)$ $\hat{\sigma}_{ab} = 0.26 \quad (0.04, 0.48)$ $\hat{\rho} = 0.69 \quad (0.43, 0.94)$

The abbreviation RE stands for ”random effects”.

According to the AIC, we discard the independence and the shared random-effects models. Considering the remaining GLMMs, the models which best fit in terms of model complexity are the two correlated random-effects. The likelihood ratio tests (LRT) indicate that we can reject the null hypothesis that the covariance $\hat{\sigma}_{ab} = 0$ (LRT=10 with 1 d.f., *p*=0.0016, for the models with equal variances; LRT=12 with 1 d.f., *p*=0.0005, for the models with different variances). Among the models with correlated random effects, the AIC for the model with unstructured covariance matrix is 10605 and is smaller than the AIC of the model with Toeplitz covariance matrix (10608). We formally test the null hypothesis $H_0 : \sigma_a^2 = \sigma_b^2$: the LRT statistic is 5.3 on 1 d.f. with $p = 0.025$, entailing that the null hypothesis should be rejected, therefore we conclude that the variability of the random effects for HCV and HIV infection is not the same. From the best model, we can estimate the correlation between the random effects of the two infections: $\hat{\rho} = 0.69$ with 95% CI (0.43, 0.94). This implies a strong positive correlation (but different from 1) exists between the infections among the regions at the level of the linear predictor, indicating a concordant association.

Figure 3 shows the odds ratios for HCV and HIV infection with their 95% asymptotic confidence intervals in function of time (the baseline is 1998), estimated by exponentiating the time evolution slopes β_{1kj} from the best model. All the odds ratios, apart from the one for HCV in 1999, are significantly different from 1. In general, the time evolution odds ratios decrease along the years with respect to 1998 for both infections. The exceptions are in 2000 and 2001 for HCV, when the odds ratios are bigger than 1, indicating a rise in the prevalence with respect to 1998.

Figure 4 shows the scatterplot of the random effects for HIV infection against those for HCV infection, obtained from the correlated model with unstructured covariance. The comparison between this graph and Figure 1 shows how the correlated model effectively translated the regional prevalence pattern to the regional-specific random effects pattern, while correcting for the time effect. Comparing the two figures, we see that we can divide the plot in 4 parts, with the regions mostly lying in the first and in the third quadrant: in the first quadrant we have the regions with higher levels of both infections, e.g., Emilia Romagna, Sardegna, and Trentino Alto Adige, whereas the third quadrant contains the regions with lower levels of both infections, e.g., Campania and Valle d'Aosta.

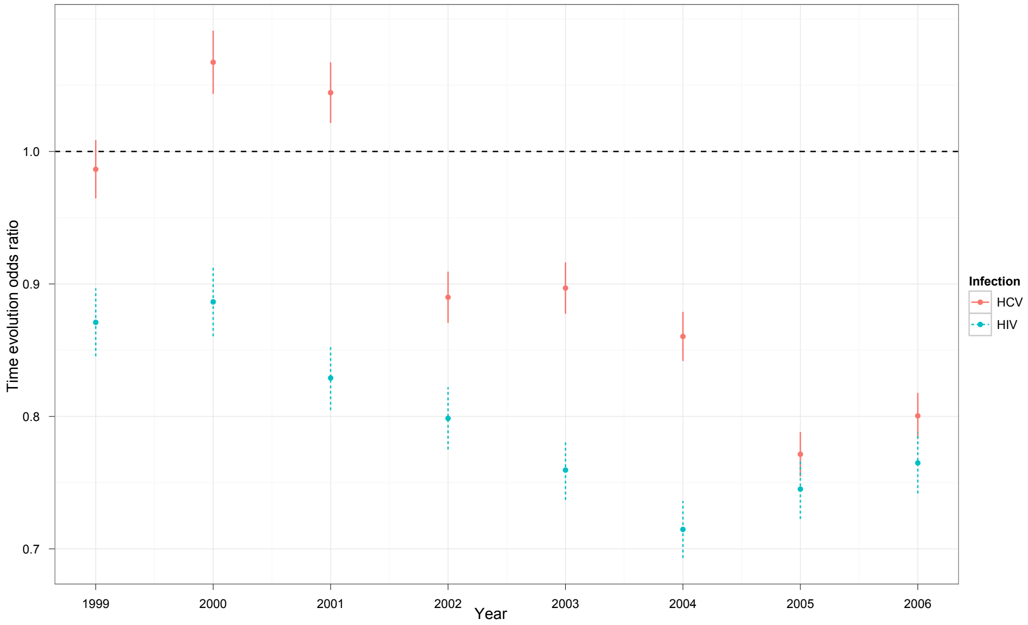

Figure 3: Time evolution odds ratios $\exp(\beta_{1kj})$ for HCV and HIV infection (baseline: year 1998) with 95% asymptotic CI, from the correlated model (4a).

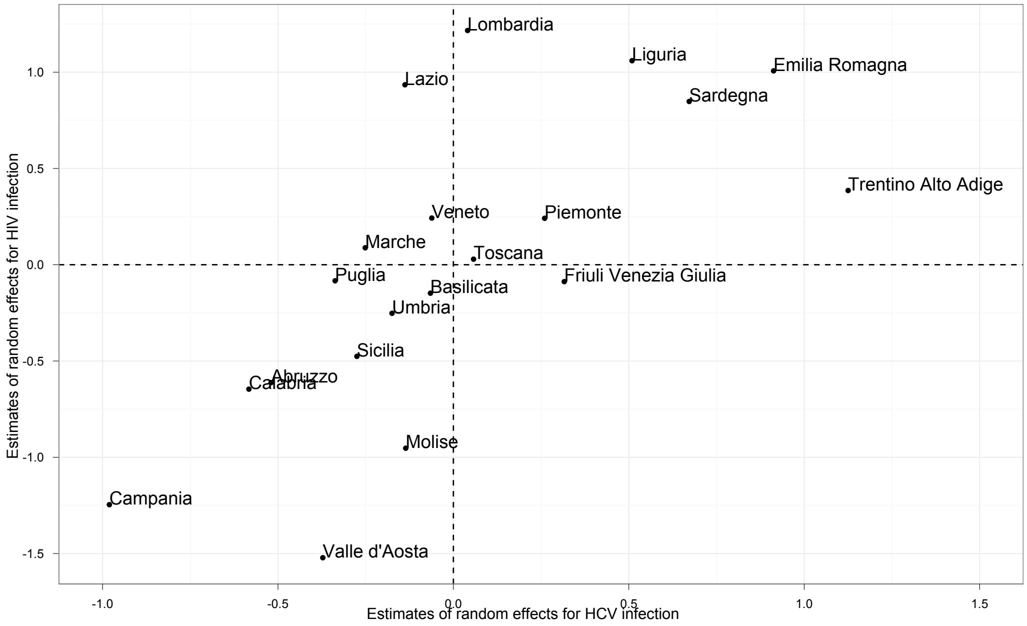

Figure 4: Estimates of the region-specific random effects for HCV and HIV infection from the correlated model (4a).

5 Hierarchical Bayesian Correlated Random-Effects Model for HCV and HIV Prevalence

In order to obtain more information about the correlation between the two infections, we refit the correlated random-effects model with unstructured covariance (model 4a) within the hierarchical Bayesian framework. The added value of this approach is that it provides not only with a point estimate of the parameter of interest and confidence interval, but it estimates also the posterior distribution of ρ. We assume that we do not have any prior knowledge about the true parameter values, thus we use flat prior distributions for the parameters such that their posterior distributions will be mostly determined by the likelihood of data. The model is parameterized in the following way. In the first stage of the model, a binomial likelihood is assumed for both HCV and HIV infection, with linear predictors given by

$$\begin{cases} g(\pi_{ij1}) = \alpha_{1i} + \beta_{11j}, \\ g(\pi_{ij2}) = \alpha_{2i} + \beta_{12j}. \end{cases}$$

For the joint prior distribution of the random intercepts α_{1i} and α_{2i}, we use the hierarchically centered parameterization (Gelfand *et al.*, 1996; Roberts and Sahu, 1997), that consists in specifying a distribution for the random effects which is not centered around zero but on other stochastic means β_{01} and β_{02},

$$\begin{pmatrix} \alpha_{1i} \\ \alpha_{2i} \end{pmatrix} \sim MVN\left[\begin{pmatrix} \beta_{01} \\ \beta_{02} \end{pmatrix}, D = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{pmatrix}\right].$$

This method is demonstrated to lead to a more efficient Gibbs sampling scheme, that is to say, the mixing of the stochastic chains is faster and the convergence requires a fewer iterations than the standard parameterization (Gelfand *et al.*, 1996). In order to complete the specification of the hierarchical model, we specify hyper prior distributions for the hyper parameters β_{01} and β_{02}:

$$\begin{cases} \beta_{01} \sim N(0, 1000), \\ \beta_{02} \sim N(0, 1000). \end{cases}$$

The hierarchical centering method is also used for the time evolution parameters of the unstructured means. We specify normal distributions for the parameters β_{11j} and β_{12j}, which are centered on the means $\mu_{\beta_{11}}$ and $\mu_{\beta_{12}}$ (uninformative prior distributions in the form of normal distributions with very large variances), with independent variances $\sigma^2_{\beta_{11}}$ and $\sigma^2_{\beta_{12}}$ (uninformative prior distributions in the form of inverse gamma distributions with small parameters). For example, the time evolution parameters for HCV infection are given by:

$$\begin{cases} \beta_{11j} \sim N(\mu_{\beta_{11}}, \sigma^2_{\beta_{11}}), & \text{prior for } \beta_{11j}, \\ \mu_{\beta_{11}} \sim N(0, 1000), & \text{hyperprior for } \mu_{\beta_{11}}, \\ \sigma^2_{\beta_{11}} \sim IG(0.01, 0.01), & \text{hyperprior for } \sigma^2_{\beta_{11}}. \end{cases}$$

Next, we specify a prior distribution for covariance matrix D. We used the Wishart distribution, which is usually employed in the estimation of the covariance matrix in case of multivariate normally distributed data (Gilks *et al.*; 1996, Congdon, 2003):

$$D \sim W_2 \left[R = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{pmatrix} \right];$$

the matrix R must be a positive definite matrix and thus we used the identity matrix for the starting values, in order to provide as less information as possible.

5.1 Application to the Data: Hierarchical Bayesian Model

The hierarchical Bayesian model was fitted in JAGS (Plummer, 2007) through the package `R2jags` (Yu-Sung and Masanao, 2011) in R. We ran the model with three chains of 20000 iterations and we discarded the first 10000 iterations for each chain as burn-in period. Employing multiple MCMC chains, we could use the ”potential scale reduction factor” (Gelman and Rubin, 1992) to check the convergence of each parameter. This diagnostic statistic compares the within-variability and the between-variability of the chains and it converges to 1 in case of MCMC convergence. For all the parameters of the hierarchical Bayesian model, we obtained values very close to 1. Figure 5 shows the posterior densities of the time evolution parameters for HCV infection, which coincide with the ML estimates from the correlated GLMM (4a) .

Table 2: Comparison of the results from the best models: the correlated random-effects model (4a) with unstructured covariance and the hierarchical Bayesian correlated model. We report the estimates for the variance parameters and for the correlation.

		GLMM		Bayesian model	
Effect	Parameter	Estimate	95% CI	Estimate	95% CI
Var RE HCV	$\hat{\sigma}_a^2$	0.25	(0.08, 0.42)	0.33	(0.18, 0.63)
Var RE HIV	$\hat{\sigma}_b^2$	0.57	(0.19, 0.96)	0.79	(0.41, 1.49)
Cov RE HCV & HIV	$\hat{\sigma}_{ab}$	0.26	(0.04, 0.48)	0.35	(0.14, 0.73)
Cor HCV & HIV	$\hat{\rho}$	0.69	(0.43, 0.94)	0.68	(0.38, 0.86)

The abbreviation RE stands for ”random effects”.

Table 2 shows the posterior means for the variance components obtained from the hierarchical Bayesian model, compared with the estimates from the GLMM. We notice that the posterior means of the variance components are larger than the ML estimates obtained from the GLMM, while the correlation coefficient remains the same ($\hat{\rho} = 0.68$ with 95% credible interval 0.38–0.86). Figure 6 shows the density estimate for the posterior distribution of the correlation coefficient. We notice that the distribution is left-skewed, with relatively few low values, implying that larger values of the correlation are more probable than smaller values.

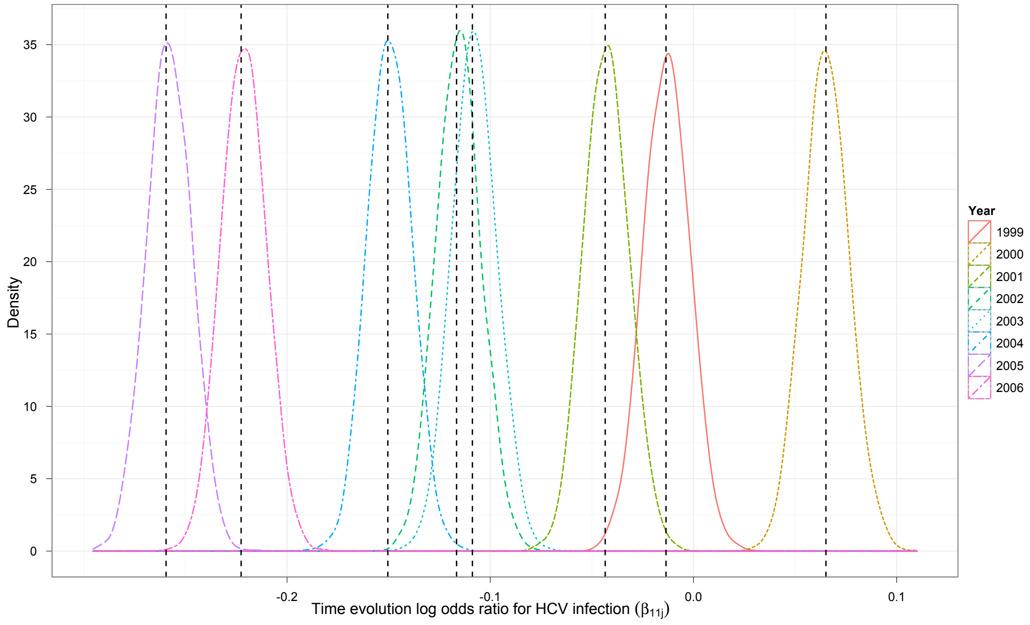

Figure 5: Density estimates for the posterior distribution of the time evolution log odds ratios of HCV infection β_{11j}, from 1999 to 2006 (baseline: year 1998), together with the maximum likelihood estimates from the correlated GLMM with unstructured covariance (dashed lines).

Figure 7 shows the posterior means for the random effects of HCV and HIV infection and reveals the same pattern observed in Figure 4. Note that the posterior means are given by the difference between the posterior means of the centered parameters $\bar{\alpha}_{ki}$ and the posterior means of their hyper parameters β_{0k}, i.e., $\bar{\alpha}_{1i} - \beta_{01}$ and $\bar{\alpha}_{2i} - \beta_{02}$, respectively. In this way, the random effects' estimates that we obtain are comparable with the ones from the correlated GLMM.

6 Discussion

Using repeated cross-sectional prevalence data for injection-related infections in IDUs in treatment in Italy from 1998 to 2006, we could define a hierarchy of structured models with which the association between HCV and HIV infection at population level can be investigated. We fitted several random-effects models for the prevalence of HCV and HIV infection, namely, five GLMMs with different covariance structures and a hierarchical Bayesian model. The models were conditioned on region-specific random intercepts, while correcting for the time effect. We tested different covariance matrices with increasing degree of association between the random effects in order to determine the structure that better fitted the data. The random effects served two purposes: firstly, their variance is a measure of the regional heterogeneity in the infection prevalence; secondly, the correlation between the region-specific random effects for each infection, a_i and b_i, is a measure of the association between the two infections.

We have shown in Table 2 that the estimated variance of the random effects for HIV infection is larger than the variance of the random effects for HCV infection, entailing a higher regional heterogeneity for HIV infection: this means that there are regions with prevalence levels of HIV infection much higher than the national level and others with much lower levels, while the prevalence of HCV infection is closer to the national levels. Looking at the time evolution odds ratios per year, with reference 1998, we observe that odds ratios of HCV and HIV infection are usually smaller than one and generally decrease over the years, except in 2000 and 2001, when there is an increase with respect to 1998. The decrease is more evident for HCV infection and less for HIV infection. The fact that the overall prevalence of HCV infection and, at a lesser extent, of HIV infection in Italy reduces suggests that strategies implemented at national and regional level and aimed at reducing risk behaviors among drug users in the last years have borne fruit. The pattern of HIV infection prevalence is confirmed by other studies as well: independent data in the form of case-reporting rates of newly diagnosed infections in drug users in Italy suggest that HIV infection diagnosis rates among drug users were declining until 2005 and remained relatively stable since (ECDC, 2009).

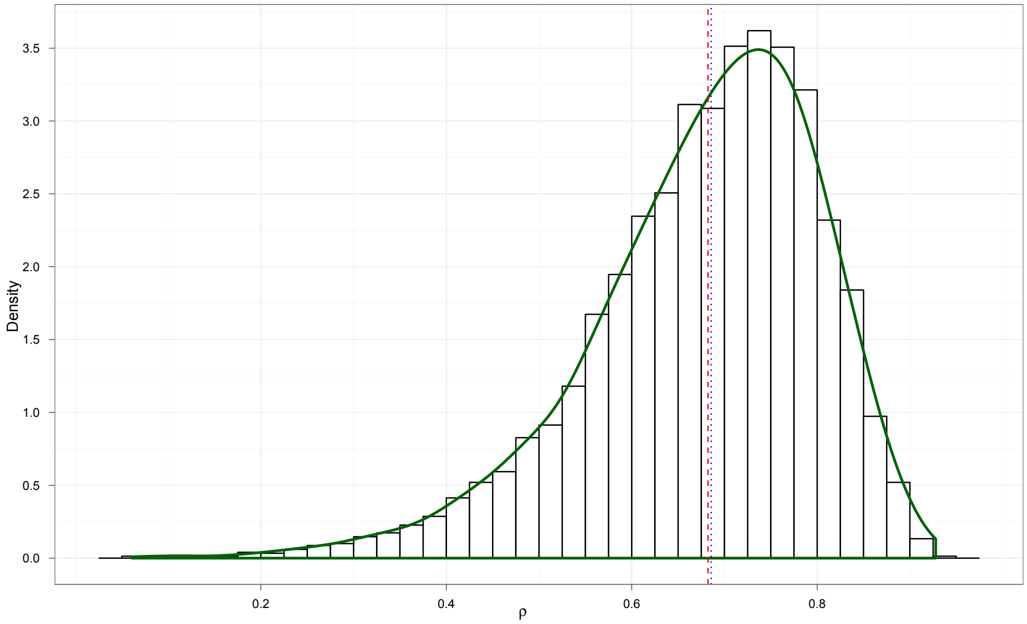

Figure 6: Density estimate for the posterior distribution of the correlation coefficient for the region-specific random effects of HCV and HIV infection, with over imposed the estimates from the Bayesian model (dashed line) and from the GLMM (dotted line).

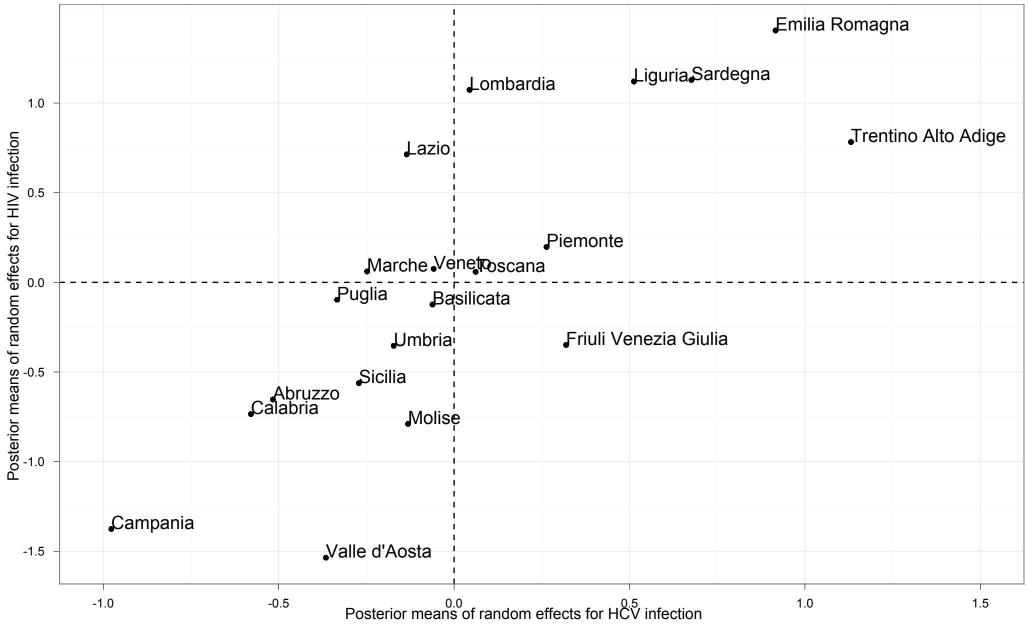

Figure 7: Posterior means of the random effects of HCV ($\bar{\alpha}_{1i} - \bar{\beta}_{01}$) and HIV ($\bar{\alpha}_{2i} - \bar{\beta}_{02}$) from the hierarchical Bayesian correlated model.

The point estimates of the correlation obtained from the GLMM and the Bayesian model are equal to 0.69 and 0.68, respectively. These results are very similar to the ones of Vickerman *et al.* (2010), who analyzed data from all over the world (Italy included) using a conditional model. All these results strongly suggest that there is a notable correlation between HCV and HIV infection at regional level, meaning that the infection prevalence tends to rise or decrease linearly and concordantly.

As we mentioned above, the variance of the random effects is significant and this provides evidence of regional heterogeneity, that is to say, there are regions characterized by high levels of prevalence for both infections (e.g., Emilia Romagna, Sardegna, and Trentino Alto Adige) and regions with lower profiles (e.g., Campania, Valle d'Aosta). Therefore, two important indications can be drawn from these findings for the health policy-makers at region levels. First, interventions to reduce behavioral risks among drug users ought to be carried out specifically for each region, because of the significant heterogeneity observed at such level. Second, it must be taken into account that the two infections usually move in the same direction, even though with different magnitude, since infectiveness of HIV is lower than that of HCV. This implies that the social network IDUs belong to and the associated risk factors (sharing syringes or other paraphernalia, higher or lower frequency of injection, presence of strangers in the network, unsafe sexual relations) ought to have a central importance in planning intervention policies (Vickerman *et al.*, 2009). Indeed, on the one hand, injections are the most likely way for IDUs to get infected with HCV, while the risk of sexual transmission is negligible (Neumayr *et al.*, 1999); on the other hand, an important transmission route for HIV infection is through unsafe sexual relationships, even though, among IDUs, the sharing of injecting equipment is a very likely way of transmission as well. Hence, given the strong correlation between HCV and HIV infection, it may be that IDUs normally belong to a network of subjects characterized by risky behaviors, more or less significant, either in terms of drug-related behavioral risks, e.g., sharing syringes or other paraphernalia, or in terms of sex-related behavioral risks, e.g., unprotected sex or prostitution. However, this hypothesis can only be tested using individual data, connected with information regarding drug and sex behavioral risks, not with the aggregate data used in this paper. Such an analysis is presented in Del Fava *et al.* (2011), where the effects of drug-related behavioral risks on the association between HCV and HIV infection are investigated.

The data analyzed in this paper have the limitation that they do not allow to link the prevalence with socio-demographic and behavioral risk information, due to the lack of the individual data. Therefore, we could only study the trend in prevalence over time and the association between the infections at population level. As we mentioned before, given the ”diagnostic testing” nature of these aggregate data, there may be a number of biases in the estimation of the prevalence. Although the proportion of IDUs among the tested individuals is unknown, the prevalence of HCV and HIV infection in the data is found to be similar to the prevalence among the IDUs reported by Camoni *et al.* (2010). In addition, not all subjects were tested for all infections, thus the sample sizes for HCV and HIV infection are generally different. It is also possible that people tested positive once are not retested again, thus it is not known to which extent positive tests are re-reported to the national system in the following years, and this might imply an underestimation of the prevalence. In addition, such data can provide a national picture of all drug users taking the tests. Finally, people who self-selected (or were selected by physicians) to be treated in a DTC are likely to present more risky behaviors and this can result in an overestimation of the prevalence. Nonetheless, such data really provide a national picture of all drug users taking the tests. Hence, we believe that they can be used for the estimation of the correlation between HCV and HIV infection, given that the biases likely affect more the prevalence of the infections rather than their correlation. For all these reasons, these types of data have already been used to model the association between HCV and HIV infection (see, for example, Vickerman *et al.*, 2010).

In this paper, we used models where the fixed effects for the time trends and the regional-specific random effects were kept separated. In the next stage, it might be interesting to analyze their combined effects, by including in the models region and time-specific random effects θ_{ijk}, for the region i, the year j, and the infection k, to fully take into account the overdispersion in these binomial data (Molenberghs *et al.*, 2010). Moreover, we are aware that generalized linear mixed models and hierarchical Bayesian models are not the only models available to study multivariate binary data. For future research, it would be interesting to explore the use of multivariate logit copula models (Nikoloupoulos and Karlis, 2008) to analyze these data: in such way, we could use other association measures, such as Kendall's tau, to jointly analyze HCV and HIV infection prevalence, and study the effect of important covariates, such as time and region.

References

Barrio G, De La Fuente L, Toro C, Brugal TM, Soriano V, Gonzalez F, Bravo MJ, Vallejo F, Silva TC, and the Project Itinere Group (2007). Prevalence of HIV infection among young adult injecting and non-injecting heroin users in Spain in the era of harm reduction programmes: gender differences and other related factors. *Epidemiology and Infection*, 135: 592–603.

Camoni L, Regine V, Salfa MC, Nicoletti G, Canuzzi P, Magliocchetti N, Rezza G, Suligoi B, and the SerT Study Group (2010). Continued high prevalence of HIV, HBV and HCV among injecting and noninjecting drug users in Italy. *Annali dell'Istituto Superiore di Sanitá*, 46: 59–65.

Congdon P (2003). *Applied Bayesian Modelling*. Chichester: Wiley.

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

Del Fava E, Shkedy Z, Hens N, Aerts M, Suligoi B, Camoni L, Vallejo F, Wiessing L, and Kretzschmar M (2011). Joint modeling of HCV and HIV co-infection among injecting drug users in Italy and Spain using individual cross-sectional data. *Statistical Communications in Infectious Diseases*. Accepted for publication.

European Centre for Disease Prevention and Control (2009). *Annual Epidemiological Report on Communicable Diseases in Europe*. Stockholm: European Centre for Disease Prevention and Control.

Gelfand AE, Sahu SK, and Carlin BP (1996). Efficient parametrizations for generalised linear mixed models (with discussion). In *Bayesian Statistics 5* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 165–180. Oxford University Press.

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

Gelman A, Carlin JB, Stern HS, and Rubin DB (2004). *Bayesian Data Analysis*. 2nd edn. London: Chapman & Hall/CRC.

Gilks WR, Richardson S, and Spiegelhalter DJ (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

Hens N, Wienke A, Aerts M, and Molenberghs G (2009). The correlated and shared gamma frailty model for bivariate current status data: An illustration for cross-sectional serological data. *Statistics in Medicine*, 28: 2785–2800.

Hope VD, Judd A, Hickman M, Sutton A, Stimson GV, Parry JV, and Gill ON (2005). HIV prevalence among injecting drug users in England and Wales 1990 to 2003: evidence for increased transmission in recent years. *AIDS*, 19: 1207–1214.

Hutchinson SJ (2004). *Modelling the hepatitis C virus disease burden among injecting drug users in Scotland*. PhD Thesis, University of Glasgow, UK.

Mathëi C, Shkedy Z, Denis B, Kabali C, Aerts M, Molenberghs G, Van Damme P, and Buntinx F (2006). Evidence for a substantial role of sharing of injecting paraphernalia other than syringes/needles to the spread of hepatitis C among injecting drug users. *Journal of Viral Hepatitis*, 13: 560–570.

McCulloch CE and Searle SR (2001). *Generalized, Linear and Mixed Models*. New York, NY: Wiley.

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

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

Neumayr G, Propst A, Schwaighofer H, Judmaier G, and Vogel W (1999). Lack of evidence for the heterosexual transmission of hepatitis C. *QJM*, 92: 505–508.

Nikoloulopoulos AK and Karlis D (2008). Multivariate logit copula model with an application to dental data. *Statistics in Medicine*, 27: 6393–6406.

Plummer M (2007). JAGS: A program for analysis of Bayesian. CiteSeerX - Scientific Literature Digital Library and Search Engine (United States). `URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=?doi=10.1.1.13.3406`

Roberts GO and Sahu SK (1997). Updating Schemes, Correlation Structure, Blocking and Parameterisation for the Gibbs Sampler. *Journal of the Royal Statistical Society Series B*, 59: 291–317.

Yu-Sung S and Masanao Y (2011). R2jags: A Package for Running jags from R. R package version 0.02-14. `http://CRAN.R-project.org/package=R2jags`

Sutton AJ, Gay NJ, Edmunds WJ, Hope VD, Gill ON, and Hickman M (2006). Modelling the force of infection for hepatitis B and hepatitis C in injecting drug users in England and Wales. *BMC Infectious Diseases*, 6: 93.

Sutton AJ, Hope VD, Mathëi C, Mravcik V, Sebakova H, Vallejo F, Suligoi B, Brugal MT, Ncube F, Wiessing L, and Kretzschmar M (2008). A comparison between the force of infection estimates for blood-borne viruses in injecting drug user populations across the European Union: a modelling study. *Journal of Viral Hepatitis*, 15: 809–816.

Vickerman P, Platt L, and Hawkes S (2009). Modelling the transmission of HIV and HCV among injecting drug users in Rawalpindi, a low HCV prevalence setting in Pakistan. *Sexually Transmitted Infections*, 85: ii23–ii30.

Vickerman P, Hickman M, May M, Kretzschmar M, and Wiessing L (2010). Can hepatitis C virus prevalence be used as a measure of injection-related human immunodeficiency virus risk in populations of injecting drug users? An ecological analysis. *Addiction*, 105: 311–318.