Auteursrechterlijke overeenkomst

Opdat de Universiteit Hasselt uw eindverhandeling wereldwijd kan reproduceren, vertalen en distribueren is uw akkoord voor deze overeenkomst noodzakelijk. Gelieve de tijd te nemen om deze overeenkomst door te nemen, de gevraagde informatie in te vullen (en de overeenkomst te ondertekenen en af te geven).

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling met

Titel: Joint modeling of HCV and HIV from cross-sectional serological data
Richting: 2de masterjaar Biostatistics - icp Jaar: 2008

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Ik ga akkoord,

TESHOME AYELE, Birhanu

Datum: 5.11.2008

Joint modeling of HCV and HIV from cross-sectional serological data

Birhanu Teshome Ayele

promotor :
Drs. Harriet NAMATA,
Prof. dr. Ziv SHKEDY

Eindverhandeling voorgedragen tot het bekomen van de graad
Master in statistics Biostatistics

universiteit hasselt

Interuniversity Institute for Biostatistics and

Statistical Bioinformatics

Universiteit Hasselt

Joint modelling of HCV and HIV from cross-sectional serological data

By

Birhanu Teshome Ayele

Supervisors:

Internal supervisor: Mevrouw NAMATA Harriet

External supervisor: Prof. dr. SHKEDY Ziv

Thesis submitted in partial fulfilment of the requirements for the degree of Master of Science in Biostatistics

2007-2008

Certification

This is to certify that this report was written by *Birhanu Teshome Ayele* under our supervision.

…….…….. ……..... Date……

Mevrouw NAMATA Harriet **Internal Supervisor**

…….…….. ……..... Date……

Prof. dr. SHKEDY Ziv **External Supervisor**

Acknowledgments

Thanks to the Almighty God for his unlimited blessing.

I am deeply indebted to my supervisors Prof. dr. Shkedy Ziv and Drs. Namata Harriet for their guidance, suggestions and for sharing me their previous working papers. Ziv is a very kind person to work with. It was a real pleasure to work with him.

I would like to express my heartfelt appreciation for my professors at Hasselt University, my family members and my friends. Especially, I would like to thank the woman who shares my life, Selamawit. Her patient love enabled me to complete my studies.

The financial support from Vlaamse InterUniversitaire Raad (VLIR) is greatly acknowledged.

Birhanu Teshome

Diepenbeek

27, August 2008.

Abstract

In this study the co-infection of HCV and HIV is investigated by joint modeling of the two infections using serological data from Italy and Spain. The seroprevalence and force of infection of the diseases are estimated over exposure time using the alternating regression model (ALR) and shared random effect models. The marginal (ALR) and random effect models are fitted with the logit and complementary log-log (clog-log) links. On the basis of the AIC values, Weibull model was chosen as the best fitting model. Significant co-infection of HCV and HIV is observed. Known risk factors such as sharing of syringes and age at first injection were confirmed as risk factors.

Keywords: *co-infection: HCV: HIV: Alternating regression model (ALR): Shared random effect Models: Seroprevalence: Force of Infection:*

Table of Contents

List of tables

List of figures

1. Introduction

Hepatitis C (HCV) is a blood-borne viral infection that affects the liver. The World Health Organization (WHO) estimates that about 170 million people worldwide, and 8.9 million people in Europe, are infected with HCV (WHO, 2000) [12]. HCV is transmitted primarily by large or repeated exposures to contaminated blood (usually through the skin by a needle puncture). The strict screening procedures for blood products have succeeded in reducing the number of new infections in the developed world. But they remain groups in population that are at higher risk of contraction of HCV infection, most notably, IDUs (EMCDDA, 2004).

Injecting drug users (IDUs) are a hidden population for which entry to and exit from the population are hard to define and measure. According to the 2004 European Monitoring Centre for Drugs and Drug Addiction (EMCDDA) monograph, IDUs are the largest risk group for HCV infection in Europe. During their injecting career IDUs are exposed to infections like HCV and HIV due to their frequency of injection, sharing syringes or sharing other paraphernalia materials (Ziv *et al*, 2008). It is estimated that one in 10 new HIV infections worldwide is attributable to injecting drug use[14].

According to WHO, the number of people living with Human immune deficiency virus (HIV) worldwide in 2007 was estimated as 33.2 million. The EMCDDA estimated that, in the EU, there could be as many as 200,000 people living with HIV who are current or past drug injectors. The number of newly diagnosed cases of HIV among injecting drug users is estimated to be currently around 3,500 per annum in the EU, which still represents a considerable public health problem. A far more negative picture presents itself for rates of infection with HCV, which remain almost universally high among drug injectors: it is estimated that 1 million some-time injectors are infected with HCV, including a significant proportion that are no longer using drugs [11].

Literatures suggest that the natural history of HCV could be influence by a co-infection with HIV or hepatitis B (Ziv *et al,* 2008, Namata *et al*. 2008, EMCDDA, 2004). As those co-infections frequently occur in IDU population, there is a need to study the co-infection of HCV and HIV in IDUs population. The association between the risk factors, the disease status and transmission parameters can be studied using cross-sectional serological data.

This study aims at joint modeling of HCV and HIV from serological cross-sectional data. Identifying the potential risk factors for the transmission of HCV and HIV infections and

estimation of the seroprevalence and force of infection of the diseases over exposure time is also the target of the study.

The remainder of this thesis is organized as follows: Section 2 provides a description of the data set while the statistical methodologies used in actualizing the objectives of the study are explained in Section 3. Main results of the analyses are presented in Section 4 whereas the 5th section of the study is devoted for discussion.

2. Data Description

The data sets considered in this study are used in Harriet *et al*. 2008, Ziv *et al*.(2008) and others. The cross-sectional datasets consist of two seroprevalence samples of injecting drug users (IDUs) from Italy (N=1224) and Spain (N=629). The cross-sectional surveys were collected by taking serological tests and face-to-face interviews. All IDUs participated in the study were interviewed, information about demographic characteristic and injecting behavior were collected. Information about IDU status was collected by asking questions like "did you ever inject drugs?" and "did you inject drugs in the last 12 months?" The second question is a subset of the first one and only available for the Spain data. Hence this study focuses only on ever injectors.

Traditionally, for many infectious diseases, the force of infection is modeled as a function of the individual's age which is considered to be the exposure time. The exposure time (duration) is the length of the injecting career and it is considered to be the length of time (in years) in which the IDUs are in the risk group. It is defined as the difference between the age at test and the age at first injection. In this study we considered the continuous version of the exposure time unlike in the studies of Ziv *et al*, (2008) and Harriet *et al*, (2008) where exposure time was categorized.

A positive HCV status in IDU seems to be associated with syringes sharing , sharing of other injecting paraphernalia, number of injecting years, age at initial drug use, frequency of injecting, and older age [Harriet *et al.*, 2008 , Ziv *et al.*,2008, Stark *et al*.,1997, crofts *et al*.,1999a, Keppler and Stover, 1999]. Because of its high prevalence in IDUs and high infectivity, even short-term recreational injecting drug use may lead to HCV infection, (Novick, 2000). The possible influence of these risk factors on the prevalence of HCV and HIV is also of interest. Information about other risk factors like sharing syringes, sharing injecting paraphernalia and frequency of injections is only available for the Spain data.

3.0. Statistical Methodology

3.1. Exploratory Data Analyses

We started with an exploratory data analysis to gain insight into the dataset. Descriptive statistics were used to examine a possible association between the two diseases, and to investigate the common risk factors for both infections. Univariate associations between the two infections were assessed using Pearson chi-square test of independence and graphs.

3.2. Statistical Analyses

In the following sections statistical methods used to model the prevalence, $\pi(t)$ and the force of Infection, $\lambda(t)$ of seropositives IDUs are discussed.

3.2.1 Prevalence and force of infection

The prevalence of a disease in a statistical population is given as the ratio of seropositives at a given exposure time to the total number of individuals in the population. The force of infection is the risk per time unit for an uninfected (that is the seronegative) IDU to become infected. Under the assumption of lifelong immunity and that the disease is in a steady state, the seroprevalence and the force of infection can be estimated from seroprevalence data (Grenfell and Anderson, 1985).

Let $\pi(t)$ be the prevalence of a disease (HCV or HIV) at duration t. Then the force of infection is given by

$$\lambda(t)=\frac{\pi'(t)}{1-\pi(t)}$$

Where $\pi'(t)$ is the derivative of the prevalence with respect to duration (exposure time). $\pi(t)$ is the cumulative distribution function of exposure time at infection.

In this study Logistic regression model and a model fitted with the frame work of generalized linear models (GLM) with binomial error (McCullagh and Nelder, 1989) which was discussed by Beker (1989), Diamond and MCDonald (1992) and Keiding *et al* (1996) who used

complementary log-log link function in order to parameterize the prevalence and force of infection as a Weibull model are fitted.

The linear predictor of the logistic regression model is $\eta(t)=\text{logit}(\pi_t)=\beta_{oj}+\beta_j t$ where t represents the exposure time (duration). Here β_{oj} is the marginal parameter for the intercept of each disease and β_j represents the log (odds ratio) between the $t+1^{th}$ and the t^{th} exposure time effect of the j^{th} disease. $\frac{\exp\{\eta(t)\}}{1+\exp\{\eta(t)\}}$ and $\hat{\beta}\frac{\exp(\hat{\beta}_{0j}+\hat{\beta}_{1j}t)}{1+\exp(\hat{\beta}_{0j}+\hat{\beta}_{1j}t)}$ are the prevalence and force of infection of the logistic regression model respectively.

The linear predictor of a Weibull model is given by $\eta(t)=\beta_{oj}+\beta_j\log(t)$ with prevalence of $\pi(t)=1-\exp[-\exp\{\eta(t)\}]$ and force of infection of $\lambda(t)=\exp(\beta_{0j})\beta_j t^{\beta_j-1}$. In the case that other covariates are included in the model, the linear predictor becomes $\eta(t)=\beta_{oj}+\beta_j\log(t)+Z\gamma_j$ where Z is the design matrix and γ_j is the parameter vector to be estimated. In this case the prevalence and force of infection will be, $\pi(t)=1-\exp[-\exp\{\eta(t)\}*\exp(Z\gamma)]$ and $\lambda(t)=\exp(\hat{\beta}_{0j})\hat{\beta}_j t^{\hat{\beta}_j-1}*\exp(Z\hat{\gamma}_j)$ respectively.

Each IDU form a cluster for which the response is a vector of two repeated measurements (the serological status of HCV and HIV). The association between the two infections can be investigated by models that can handle the correlation between observations from the same IDU. This can be modeled using marginal models or by joint modeling of the binary responses.

3.2.2. Marginal model (Alternating Logistic regression model (ALR))

Marginal models are population-averaged models characterized with a marginal mean function. The models are used to study the association structure of the repeated measures, and the effect of the covariates accounting for the association between the two infections. The association structure is typically captured using a set of association parameters, such as correlation and odds ratio (Molenberghs and Verbeke, 2005).

The model has the form

$E(Y_{ij}) = \mu_i$ and $\eta_i(\mu_i) = X_i\beta_j$, where Y_{ij} is the vector of observed responses for the i^{th} IDU at infection j and μ_i is the vector of mean responses for the i^{th} IDU. The model for μ_i is specified via a vector of link function. $\eta_i(\mu_i) = \text{logit}(\mu_i)$, the marginal logit link can be used for binary outcomes. β_j is a vector of regression parameters and X_i is a known design covariate matrix.

Since the scientific interest of this study is to model the dependence structure, Alternating Logistic Regressions (ALR) algorithm of Carey, Zeger, and Diggle (1993) that models the association structure between pairs of responses with log odds ratios is preferable. Using ALR, inferences can be made not only about marginal parameters but also about pair-wise association. It has been stated that the odds ratio is a straightforward measure to capture association between binary outcomes (Molenberghs and Lesaffre, 1994). The ALR algorithm alternates between a GEE step to update the model for the mean and a logistic regression step to update the log odds ratio model. Upon convergence, the ALR algorithm provides estimates of the regression parameters for the mean, β, the regression parameters for the log odds ratios, α, their standard errors, and their covariance (Molenberghs and Verbeke 2005).

The odds ratio $\left(OR_{ij}\right)$ between responses Y_{i1} and Y_{i2} of the i^{th} IDU can be expressed as:

$$OR_{ij} = \frac{P(HCV=1,HIV=1)*P(HCV=0,HIV=0)}{P(HCV=1,HIV=0)*P(HCV=0,HIV=1)}$$

The odds ratio between k^{th} and j^{th} response greater than one indicates positive association and less than one indicates negative association. In the simplest case $\log(OR_{ij}) = \alpha$, is a constant pair wise log odds ratio.

3.2.3. Generalized Linear Mixed model (GLMM)

The generalized linear mixed model is the most frequently used random effects model for discrete outcomes. It is a rather straightforward extension of the generalized linear model for univariate data to the context of clustered measurements. In a random effects models it is assumed that there is natural heterogeneity across the subjects (IDUs) and that this heterogeneity can be modeled by a probability distribution which implies that the regression coefficients are varying from one subject(IDU) to the another (Molenberghs and Verbeke, 2005).

A joint random effects model that account for the correlation between the two infections can be fitted by a shared random effect frame work as given in the next section.

3.2.3.1. Joint Modeling of HCV and HIV

The joint modeling was chosen because it allows fitting a single model to both response variables (HCV and HIV) simultaneously while taking the correlation between the two into account. The joint modeling approach is warranted by the Pearson chi-square test of interdependence between the two infection (p-value <0.0001) and patterns observed in Figure 2. Possible consequences of analyzing correlated data as if it were independent are inconsistent inferences concerning regression parameters due to underestimated standard errors and inefficient estimators.

To correlate the two diseases, we can share a random effect between observations from the same IDU. The overall prevalence of HCV and HIV can be modeled taking in to account the association between the two diseases using the shared random effect model. The shared random effect assumes the same set of random effects for both outcomes.

The shared effect model can be formulated as

$$\text{logit}(\pi_{ij1})=\beta_{o1}+\beta_{11j}+b_i$$

$$\text{logit}(\pi_{ij2})=\beta_{o2}+\beta_{12j}+b_i$$

Where b_i is an IDU specific random effect $b_i \sim N(0,\sigma_b^2)$. The case with $\sigma_b^2=0$ implies that the diseases are independent at an IDU level. This model has stronger assumption about the association between the two outcomes. For our data this assumption is well supported by the chi-square test of independence and the pattern revealed in Figure 2.

A model with complementary log-log link function that implies Weibull model was also fitted. Weibull model is fitted by using log of the exposure time.

3.2.3.3 Marginalization of the GLMM

The regression coefficients of the GLMM need to be interpreted conditionally on the random effect b ,that is the parameters have IDU-specific interpretation. Incase population-averaged interpretations are of interest, additional computations are needed (Molenberghs and Verbeke, 2005).

The marginal expectation of the outcome Y_{ij} at time t for the HCV disease using the logit link is given by

$$\begin{aligned} E[Y_{ij}] &= E[E[Y_{ij}|b_i]] \\ &= E\left[\frac{\exp(\beta_o + b_i + \beta_1 t)}{1+\exp(\beta_o + b_i + \beta_1 t)}\right] \\ &\neq E\left[\frac{\exp(\beta_o + \beta_1 t)}{1+\exp(\beta_o + \beta_1 t)}\right] \end{aligned}$$

The marginal mean can be derived from the GLMM output by integrating out the random effect based on numerical integration techniques or based on numerical averaging. It is often much easier to use numerical averaging by sampling a large number M of the random effects vectors b_i from their fitted distribution $N(0,\sigma_b^2)$ (Molenberghs and Verbeke, 2005).

The estimate, $\hat{E}(Y_{ij})$ at specific exposure time, t is given by

$$\hat{E}(Y_{ij}) = \frac{1}{M}\sum_{i=1}^{M}\frac{\exp(\hat{\beta}_o + b_i + \hat{\beta}_1 t)}{1+\exp(\hat{\beta}_o + b_i + \hat{\beta}_1 t)}$$

We randomly generate M=920 (Italy) and M=470 (Spain) realized values of the random effect b taken from a normal distribution with mean zero and variance of σ_b^2 given in respective tables of the random effect models (Table 5 and 6) for the Italy and Spain data under both the logistic and Weibull regression models.

For example, considering the logistic regression model of the Italy data, an estimate for the unconditional mean at a given exposure time,t obtained from the 920 conditional mean is

$$\hat{E}[Y(t)] = \frac{1}{920}\sum_{i=1}^{1000}\frac{\exp[-0.4772 + b_i + 0.1633t]}{1+\exp[-0.4772 + b_i + 0.1633t]}, \text{ for the HCV disease.}$$

Graphical representations of these averages are given in Figure 5 and 6 of section four. The SAS code used for the generation of the random effect, b and used to plot the marginalized predicted seroprevalence plots of Figure 5 and 6 is given in the Appendix.

3.2.3 Model Selection

Model selection criteria provide a useful tool in selecting a suitable model from a candidate class to characterize the underlying data. Since models with different link function are not nested, the Akaike's information criterion (AIC) (Akaike, 1974) can used to select the best model. A model with the smallest AIC value is chosen to be the best (Ziv *et al*, 2006). The selected model was further used to examine the influence of other risk factors on the transmission of HCV and HIV.

3.3. Software used.

We mainly employed the GENMOD and NLMIXED procedures of SAS version 9.1 for the analysis. For this study, the type I error was controlled at a 5% level of significance.

4.0 Application to the Data

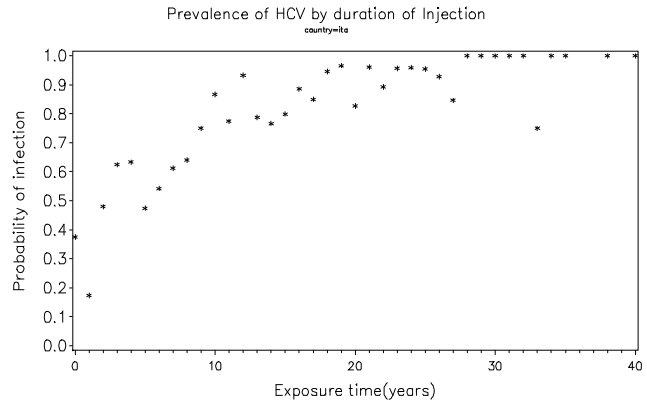
4.1 Descriptive Data Analyses

The overall seroprevalence and the demographic characteristic of IDUs from Italy and Spain data are presented in Table 1. The overall prevalence of HCV is 63% and 73.29% for Italy and Spain respectively. The overall prevalence of HIV is 7.68% and 25.59% for Italy and Spain respectively. For Italy, 62.78% of the males and 64.08% of the females are HCV seropositives. For Spain, 74.41% of the males and 70.12% of the females are HCV seropositives. Pearson chi-square test for independence shows that there is no difference in proportion of HCV seropositives between males and females. 6.95% of the males and 11.33% of the females from the Italy data and 23.82% of the males and 30.49% of the females from the Spain data are HIV seropositives. The average age at interview is 34.33(SD=7.67) and 26.11(SD=3.13) for Italy and Spain respectively. The average age at first injection is 21.44(SD=5.15) and 19.39(SD=3.81) for Italy and Spain respectively. The mean exposure time is 13.78(SD=7.94) years and 6.50 (SD=4.52) years for the Italy and Spain data respectively.

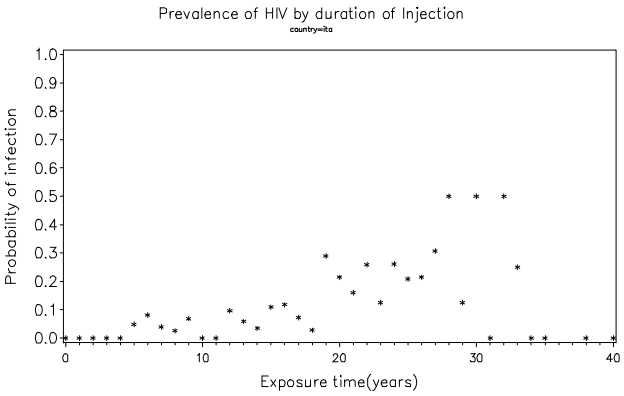
Table 1: *Descriptive statistics of HCV and HIV seroprevalence and demographic variables*

Variable	Italy	Spain
Total HCV+ (N, %)	772, 63.33%	461 , 73.29%
Total HIV+ (N, %)	93, 7.68%	161, 25.59%
Gender		
Male(N,%HCV+)	(1013, 62.78%)	(465,74.41)
Female(N,HCV+)	(206, 64.08%)	(164,70.12)
Chi-square(p-value)	0.1229(0.7259)	1.1381(0.2861)
Male(N,%HIV+)	(1007, 6.95%)	(466, 23.82)
Female(N,HIV+)	(203, 11.33%)	(164, 30.49)
Chi-square(p-value)	4.5653(0.0326)	2.8351(0.0922)
Age at Interview(mean, SD)	(34.33,7.67)	(26.11,3.13)
Age at first injection(Mean, SD)	(21.44,5.15)	(19.39,3.81)
OR(Estimate, (95%CL))	9.0785(3.9322,20.9602)	4.7276(2.6858,8.3216)

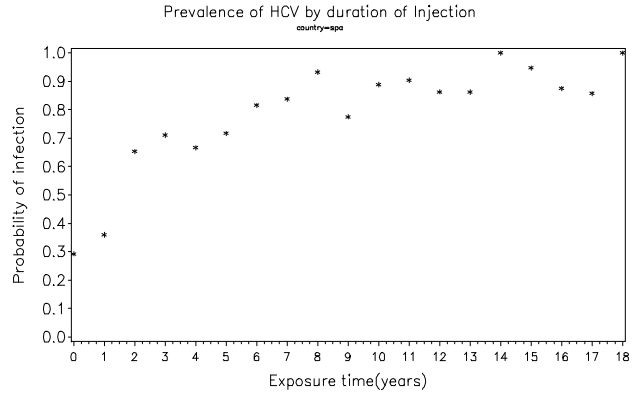
Figure 1 shows the empirical distribution for the seroprevalence of HCV and HIV by exposure time for Italy and Spain. Higher prevalence of HCV than HIV is observed in both countries. The plots indicate that the seroprevalence of HCV increases rapidly till the exposure year of around 30 and 10 for the Italy and Spain data respectively. Relatively lower prevalence of HIV is observed in the Italy data compared to that of Spain.



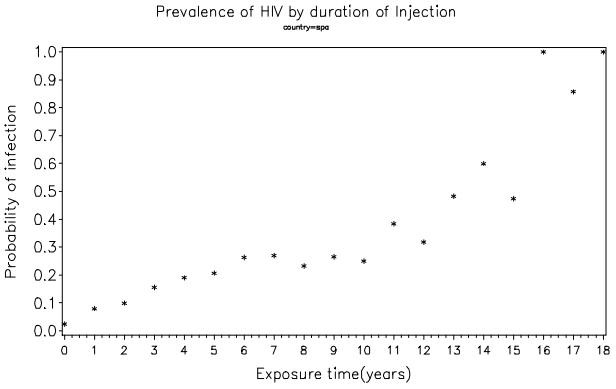
a. prevalence of HCV by duration, Italy



b. prevalence of HIV by duration, Italy



c. prevalence of HCV by duration, Spain



d. prevalence of HIV by duration, Spain

***Figure 1**: Seroprevalence of HCV and HIV by exposure time*

The association between the two diseases was examined by odds ratio, its corresponding confidence interval and the patterns revealed in Figure 2. The estimates of the odd ratio listed in Table 1 shows that an IDU who is infected by HCV is 9.08 and 4.73 times more likely to be infected by HIV than an IDU who is not infected by HIV for Italy and Spain studies respectively. 23.21% and 7.13% of the IDUs are co-infected by HCV and HIV for the Spain and Italy studies respectively. These proportions are much higher than what is expected under the assumption of independence. We expected 18.92% and 4.84% of co-infection for Spain and Italy respectively. The Pearson chi-square test of interdependence was significant (p-value<0.0001) and indicates that IDUs who are infected by one of the disease are more likely to be infected with the other disease too.

Figure 2 shows an upward relationship between HCV and HIV. As shown in Figure 2, HCV prevalence of below 40% for Italy and 60% for Spain have a respective HIV prevalence of near

zero, while for HCV prevalence of above these values a positive relationship between HCV and HIV prevalence is observed. This relationship is more observed in the Spain data.

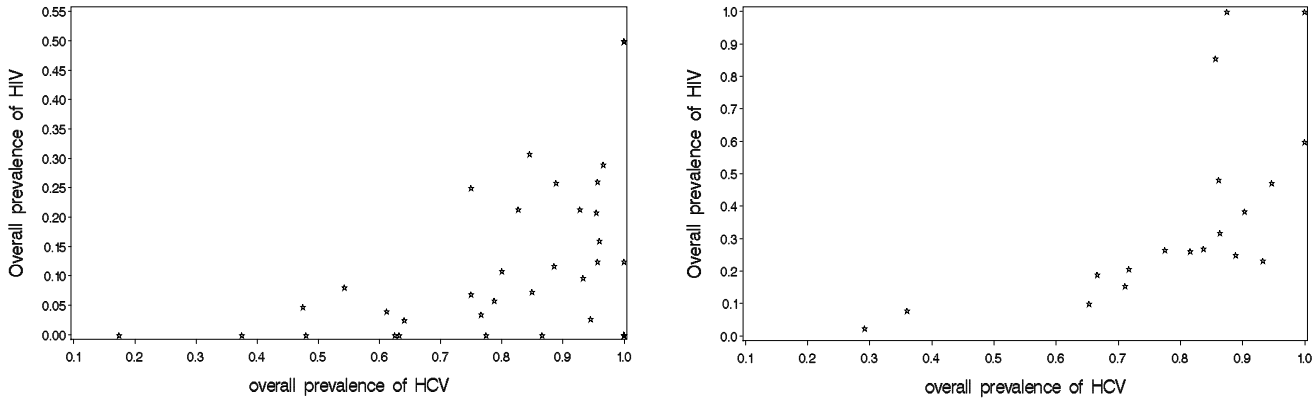


***Figure 2**: overall prevalence of HCV versus overall prevalence of HIV Italy (left panel) and Spain (right*

The influence of the risk behavior factors on the prevalence of HCV and HIV is explored by proportion and Pearson chi-square test of independence as shown in Table 2. The prevalence of HCV and HIV is higher among IDUs who share syringes. Pearson chi-square test of independence indicates that sharing syringes is a significant risk factor for the transmission of HCV and HIV. Although the proportion of IDUs who share other paraphernalia is higher among HCV and HIV seropositives, it is not a significant risk factor for the transmission of the two infections. Daily injectors are at a higher risk for HCV and HIV infection (higher proportion). Frequency of injections is found to be a significant risk factor for the HCV transmission.

***Table 2**: Distribution of HCV and HIV in relation to risk behavior factors-the Spain's study*

Variable	HCV^+ (%)	p-value	HIV^+ (%)	p-value
Sharing Syringes		**<0.0001**		**<0.0001**
No	65.28		16.07	
yes	87.35		40.24	
Sharing other paraphernalia		**0.2916**		**0.3988**
No	72.7		24.67	
yes	76.92		28.06	
Frequency of injection		**0.005**		**0.6295**
Daily	82.4		28.31	
1-6 days a week	72.3		24.76	
Less weekly	64		27.48	
Never	72		22.03	

4.2. Modeling the Prevalence and Force of Infection adjusting for the exposure time

In this section, statistical methods that were discussed in section three are applied to the Italy and Spain data. For each data set the GLM with the logit and clog-log links are fitted using marginal and random effect models. The best link function was selected by the AIC value.

4.2.1 Alternating Logistic Regression models (ALR)

Firstly, the ALR model was fitted by adjusting for the duration (exposure time). The pair wise log odds ratio was assumed to be constant. Conclusions were given using the empirical based estimates of the exchangeable working correlation structure. Parameter estimates (95% CL) are shown in Table 3 and 4 for the Italy and Spain data respectively.

***Table 3**: Parameter Estimates [95% CL] for ALR regression model of the Italy's study*

		Model 1(logit link)	**Model 2 (Clog-log link)**
Disease	Parameter	Estimate [95% CL]	Estimate [95% CL]
HCV	β_{01}	-0.4343[-0.7609, -0.1077]	-1.1990 [-1.5352, -0.8628]
	β_{11} (duration)	0.1399[0.1110,0.1688]	0.6806[0.5500, 0.8113]
HIV	β_{02}	-3.9486[-4.5169, -3.3803]	-6.3192[-7.7745, -4.8639]
	β_{12} (duration)	0.1071[0.0785, 0.1358]	1.5177[1.0201, 2.0153]
Log(OR)	α	0.9745[0.0436, 1.9055]	0.9459[0.0148, 1.8769]
AIC		659.9848	656.7683

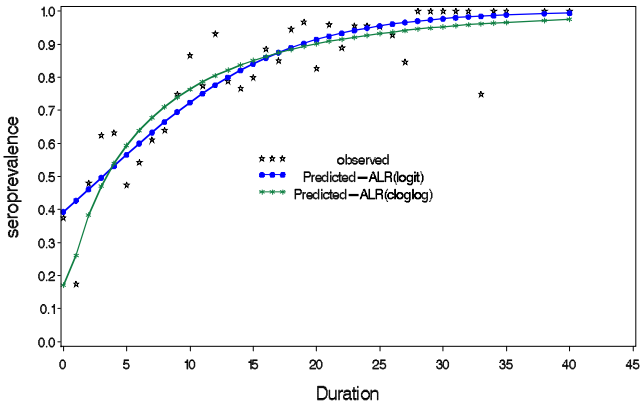
For both countries a significant effect of duration (time of exposure) is observed under the logit and clog-log link functions. Comparing the AIC values of the logistic and Weibull regression models, Weibull distribution with clog- log link has lower values for both datasets which makes it a better choice. The pair wise log odds ratio were found to be significant, 0.9459[95% CL: 0.0148, 1.8769] and 0.7377[95% CL: 0.0983, 1.3771] for Italy and Spain respectively implying strong common pair wise association between clusters.

Table 4**:** *Parameter Estimates [95% CL] for ALR regression model of the Spain's study*

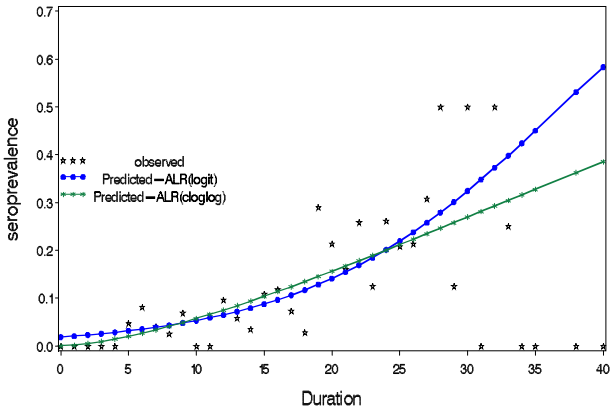
		Model 1(logit link)	**Model 2 (Clog log link)**
Disease	Parameter	Estimate [95% CL]	Estimate [95% CL]
HCV	β_{01}	-0.3765[-0.7245, -0.0285]	-0.6131[-0.8606, -0.3656]
	β_{11} (duration)	0.2638[0.1949, 0.3326]	0.6229[0.4872, 0.7587]
HIV	β_{02}	-2.6253[-3.0908, -2.1598]	-3.0720[-3.7404, -2.4036]
	β_{12} (duration)	0.2130[0.1640, 0.2619]	1.0395[0.7357, 1.3434]
Log(OR)	α	0.8061[0.1635, 1.4487]	0.7377[0.0983, 1.3771]
AIC		503.0867	499.4283

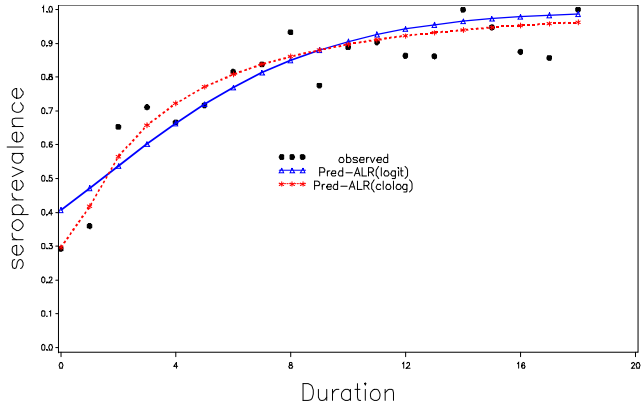
The predicted plots for the prevalence and force of infection are given in Figure 2 and 3 for the Italy and Spain data respectively. For HCV, the plot of the estimated prevalence of the two models indicate a rapid increase in prevalence from starting time of drug injection to the exposure time of around 30(15) years after which the prevalence stabilize for the Italy (Spain) studies. A maximum seroprevalence of HCV is shown in IDUs with longer duration of injections. An increase in prevalence of HCV is accompanied by increase in prevalence of HIV. This indicates the co-infection of HCV and HIV. The longer they inject drugs, the higher their chance to be infected by both infections.

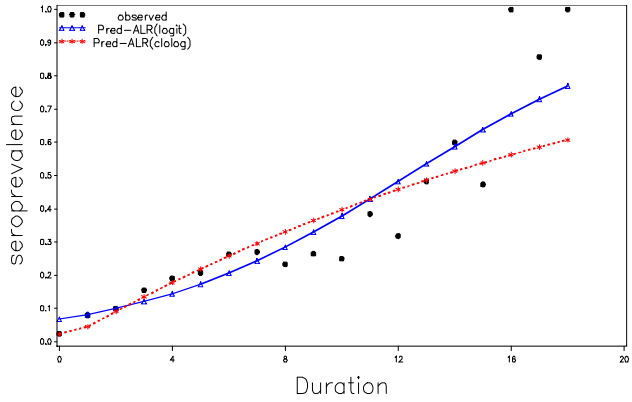
HCV-Italy



HIV-Italy







HCV-Spain

HIV-Spain

***Figure 3**: Estimated prevalence of HCV (left panels) and HIV (right panels) by exposure time obtained from the Weibull and logistic models-Italy & Spain data*

Force of infection of HCV is very high at the beginning of the injecting career and leveling off with the duration the injecting career, Figure 4. The predicted force of infection obtained from the Weibull model seems constant after duration of around 10(5) years while that of the logistic regression start increasing after this point for the Italy (Spain) data. Weibull model estimates a relatively constant force of infection of HIV for both countries. For HCV, the smaller the duration of exposure to injection, the higher the force of infection in both counties. From the predicted Force of infection by the Weibull model, HIV seems to have a constant force of infection.

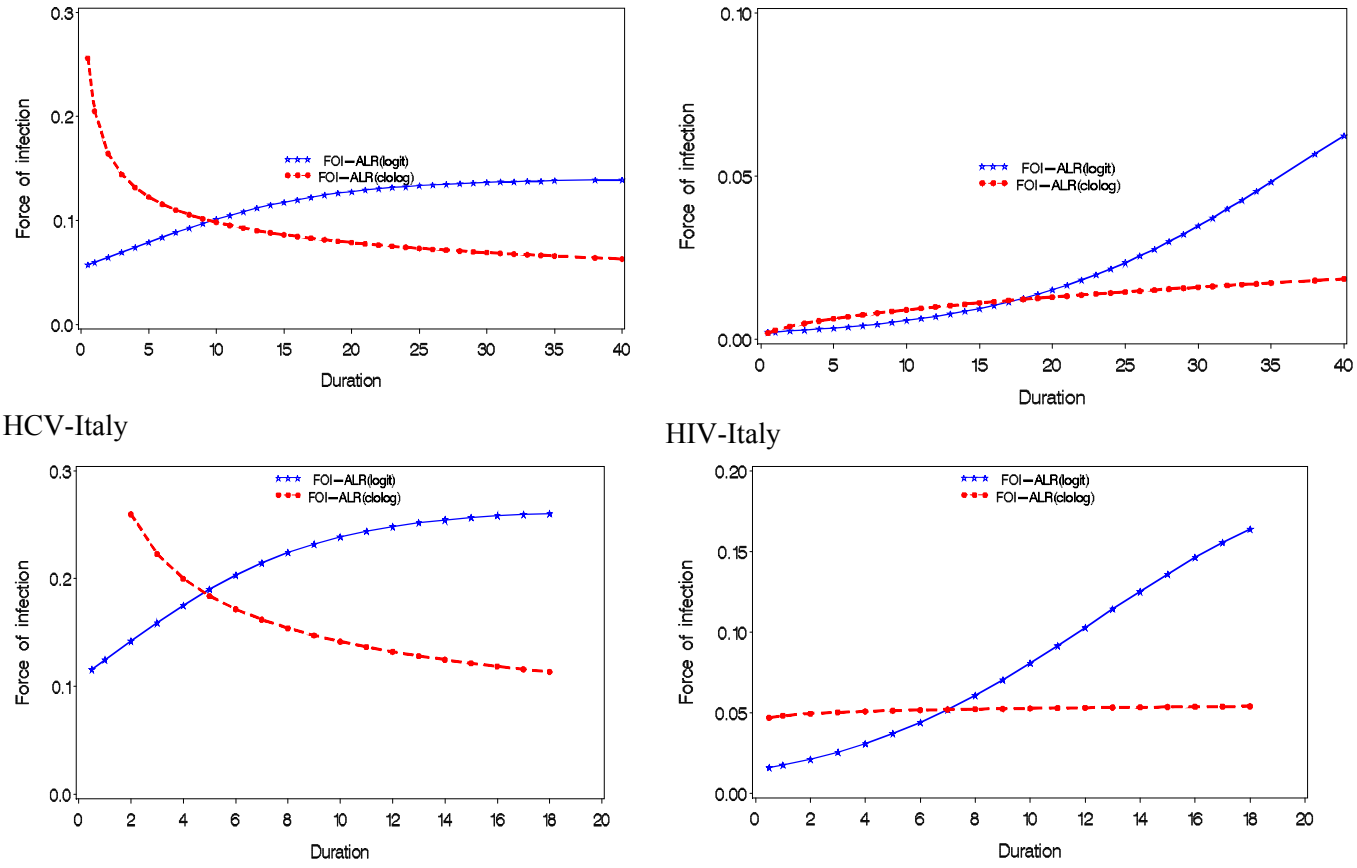


***Figure 4**: Force of Infection of HCV (left panels) and HIV (right panels) diseases by exposure time using the logistic and Weibull models-Italy data*

4.2.2 Random effect Model

A fixed effect and random effect models are fitted to test for the significance of σ_b^2. For the Italy data, the AIC values of the random effect models are equal to 1305.3 and 1287.4 for the logit and clog-log links while the AIC values of the fixed effect models with the respective links are 1308.0 and 1288.9 respectively. For the Spain data, the AIC value of the random effect models are 990.3 and 928.5 for the logit and clog-log links respectively. The AIC values of the fixed effect models for the Spain data are 993.2 and 930.3 for the logit and clog-log links respectively. The smaller AIC value of the random effects indicates that observations are correlated on the IDU level.

Parameter estimates 95% CL, standard error of the random effect and AIC values of the random effect model for both links are given in Table 5 and 6.

***Table 5**: Parameter Estimates [95% CL] for GLMM regression model of the Italy's study*

		Model 1(logit link)	Model 2 (Clog-log link)
Disease	Parameter	Estimate [95% CL]	Estimate [95% CL]
HCV	β_{01}	-0.4772[-0.8662, -0.0883]	-1.5442[-2.0717, -1.0166]
	β_{11} (duration)	0.1633[0.1231, 0.2035]	0.8852[0.6082, 1.1622]
HIV	β_{02}	-4.6095[-5.6245, -3.5945]	-6.6371[-8.1822, -5.0920]
	β_{12} (duration)	0.1231[0.0835, 0.1628]	1.5651[1.0537, 2.0765]
σ_b		1.0690[0.4272, 1.7108]	0.6736[0.1828, 1.1644]
AIC		1305.3	1287.4

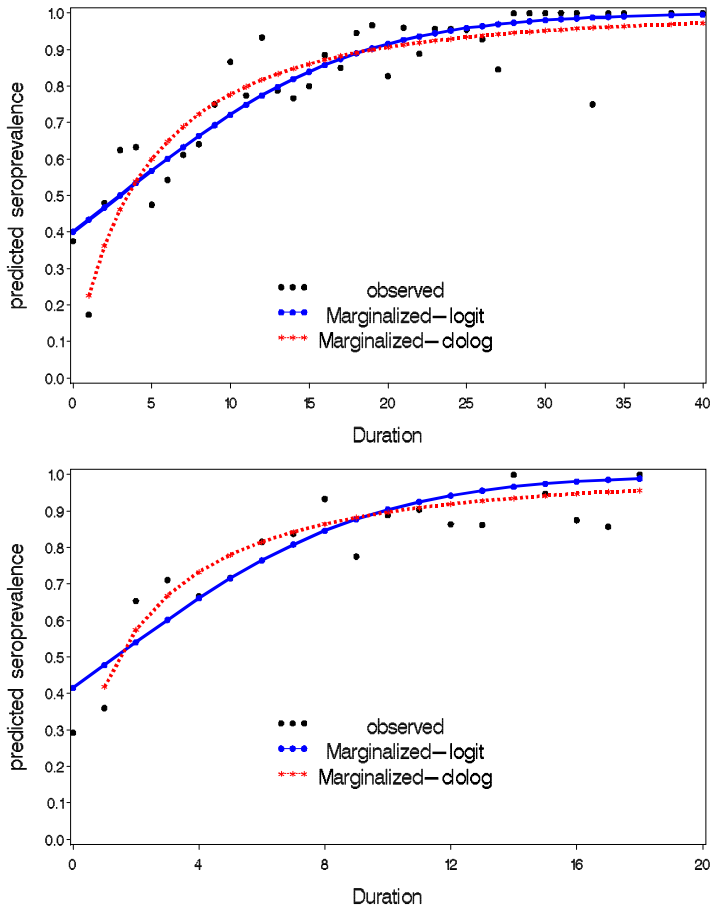
For both countries a significant relation is observed between the seroprevalence of the two diseases and exposure time.

***Table 6**: Parameter Estimates [95% CL] for GLMM regression model of the Spain's study*

		Model 1(logit link)	Model 2 (Clog-log link)
Disease	Parameter	Estimate [95% CL]	Estimate [95% CL]
HCV	β_{01}	-0.3907[-0.8009, 0.0195]	-0.6678[-1.0417, -0.2938]
	β_{11} (duration)	0.2963[0.2176, 0.3750]	0.7298[0.4823, 0.9772]
HIV	β_{02}	-3.0422[-3.7411, -2.3433]	-3.2561[-4.0109, -2.5013]
	β_{12} (duration)	0.2463[0.1765, 0.3161]	1.0839[0.7442, 1.4236]
σ_b		0.9511[0.4114, 1.4907]	0.5964[0.2017, 0.9910]
AIC		990.3	928.5

We generated a random sample of 920 random effects from a normal distribution with mean zero and variance of $(1.069)^2$, and $(0.6736)^2$ for the logit and clog-log of the Italy data respectively. For the Spain's study 470 random effects were simulated from a normal distribution with mean zero and variance of $(0.9511)^2$ and $(0.5964)^2$ for the logit and clog-log links respectively. The predicted seroprevalence plot of HCV and HIV are shown in Figure 5 for both countries.

HCV-Italy



HIV-Italy

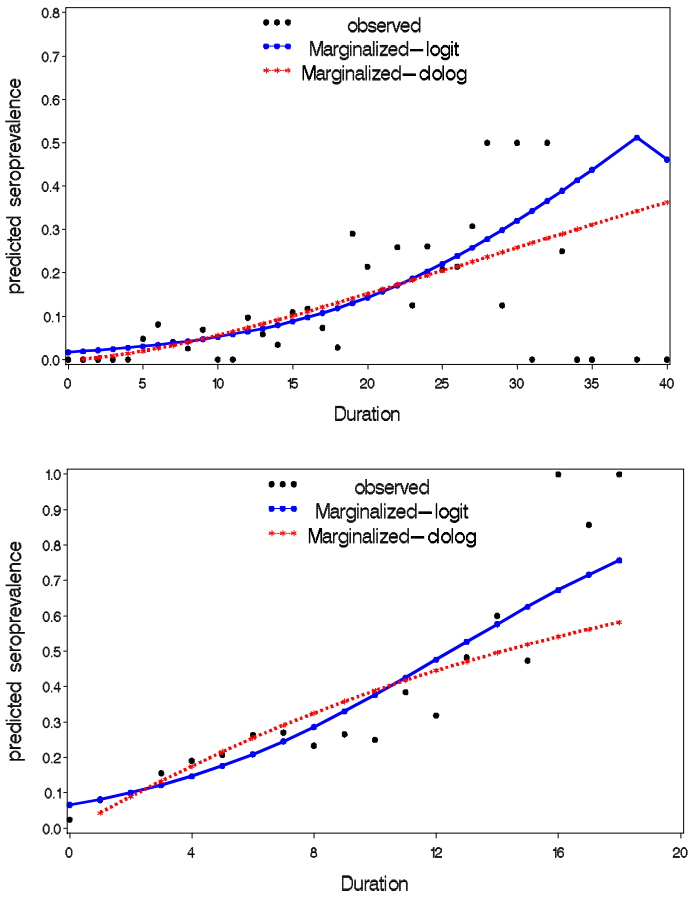


Figure 5: *Estimated prevalence of HCV (left panels) and HIV (right panels) by exposure time-the marginalized GLMM model*

The patterns observed in Figure 5 are similar to the ones observed in predicted prevalence plots of the ALR models. Plots from both link functions follow the data very well.

4.2.3. Influence of other risk behavior factors

The models discussed so far (the logistic and Weibull) only adjusted for the exposure time as a risk factor. Using these models we could identify the best link function and estimate the prevalence and force of infection. In this section we extend the Weibull model (with the smallest AIC) so that we can adjust for potential risk factors like age at first injection, sharing needles, and sharing other paraphernalia. Information about sharing needles and other paraphernalia is only available for the Spain data.

Table 7 shows the parameter estimates, the log odds ratio estimate of the ALR model and standard error of the random effect model of the Italy's study. Age at first injection is found to be a significant risk factor only for HIV infection. This indicates that the prevalence of HIV for IDUs who started injecting drugs in relatively older age is higher than those who started to inject in relatively younger ages. The log odds ratio is significant at 1% level of significance (P-value for alpha is 0.0568).

Table 7: *Parameter Estimates [95% CL] of the Italy data (Final models)*

		ALR(clog-log)	GLMM(clog-log)
Disease	Parameter	Estimate[95% CL]	Estimate[95% CL]
HCV	β_{01}	-1.6442[-2.2705, -1.0179]	-2.0981[-3.0031,-1.1931]
	β_{11} (duration)	0.7212[0.5808, 0.8616]	0.9296[0.6365, 1.2227]
	β_{21} (age at 1st injection)	0.0162[-0.0026, 0.0350]	0.02051[-0.0038, 0.0448]
HIV	β_{02}	-8.4402[-10.4542, -6.4262]	-8.8631[-11.2915, -6.4348]
	β_{12} (duration)	1.7922[1.2743, 2.3102]	1.8552[1.2716, 2.4387]
	β_{22} (age at 1st injection)	0.0647[0.0155, 0.1139]	0.0681[0.0175, 0.1187]
		α=0.8916[-0.0260, 1.8092]	σ_b=0.6582[0.1569, 1.1596]

For the Spain data, available risk factors were put in the previous ALR and GLMM models. Variables selection was done by removing a covariate with the highest p-value (most insignificant) sequentially. That is among the non significant p-values, the one with highest p-value was removed and the model was refitted with the rest covariates. Sharing syringes and exposure time are the only covariates left in the final model. Sharing syringes is found to a significant risk factor for the transmission of HCV and HIV. Using estimates of the ALR model, the risk of HCV for IDUs who share syringes is 1.47(exp (0.3)) times higher than the risk for those who don't share syringes. The risk of HIV for IDUs who share syringes is 2.16 (exp (0.77)) time the risk of those who don't share syringes. Hence for the Spain data sharing syringes can be claimed to the main route for the co-infection of HCV and HIV.

***Table 8**: Parameter Estimates [95% CL] of the Spain data (Final models)*

		ALR(clog-log)	GLMM(clog-log)
Disease	Parameter	Estimate[95% CL]	Estimate[95% CL]
HCV	β_{01}	0.4916[-0.7620, -0.2213]	-0.6746[-1.0449,-0.3042]
	β_{11} (duration)	0.1068[0.0681, 0.1454]	0.5984[0.3680, 0.8288]
	β_{21} (Sharing Syringes)	0.3900[0.1461, 0.6339]	0.4103[0.1139,0.7067]
HIV	β_{02}	-2.850[-3.2958, -2.4052]	-3.4512[-4.2174,-2.6850]
	β_{12} (duration)	0.1609[0.1221, 0.1997]	0.9783[0.6425,1.3142]
	β_{22} (Sharing Syringes)	0.7703[0.3879, 1.1526]	0.7991[0.4073,1.1910]
		α=0.6556[0.0252, 1.2859]	σ_b=0.4013[-0.0978,0.9005]

4.2.4. Marginal versus Shared random-effect Models

In this subsection we compare result from the ALR and the shared random effect models. Here the intention is not to compare models of different families, but to assess possible similarities or differences of the obtained estimates. Estimates of the covariates, standard errors and 95% confidence limits given in Tables 7 and 8 for the Italy and Spain studies respectively are used.

For a random-intercept logistic regression model, with normally distributed random intercepts like the one we fitted, it can be shown that the marginal model is well approximated by the shared random effect model, but with parameters satisfying

$\frac{\hat{\beta}^{GLMM}}{\hat{\beta}^{ALR}} \approx \sqrt{c^2\sigma_b^2+1}>1$, where σ_b^2 is the variance of the random effects and $c=\frac{16\sqrt{3}}{15\pi}$ (Diggle *et al*, 2002)[7]. This ratio implies that an estimate of the shared random effect model is not smaller than the ALR model. When the random-intercept variance is zero, the ratio will be one. For our study, these approximate factors are 1.120 and 1.0463 for the Italy and Spain studies respectively. These values provide good agreement between the two models. In absolute term estimates of the ALR models are lesser than their corresponding estimates of the GLMM.

The observed agreement between the two models is well revealed in Figure 6. Plots of the logit and clog-log links of the ALR and GLMM models overlaps in all panels with minor difference in higher durations for HIV exposure. The minor difference between corresponding plots might

show the effect of the random intercept. Both the ALR and GLMM models well predict the seroprevalence of HCV and HIV.

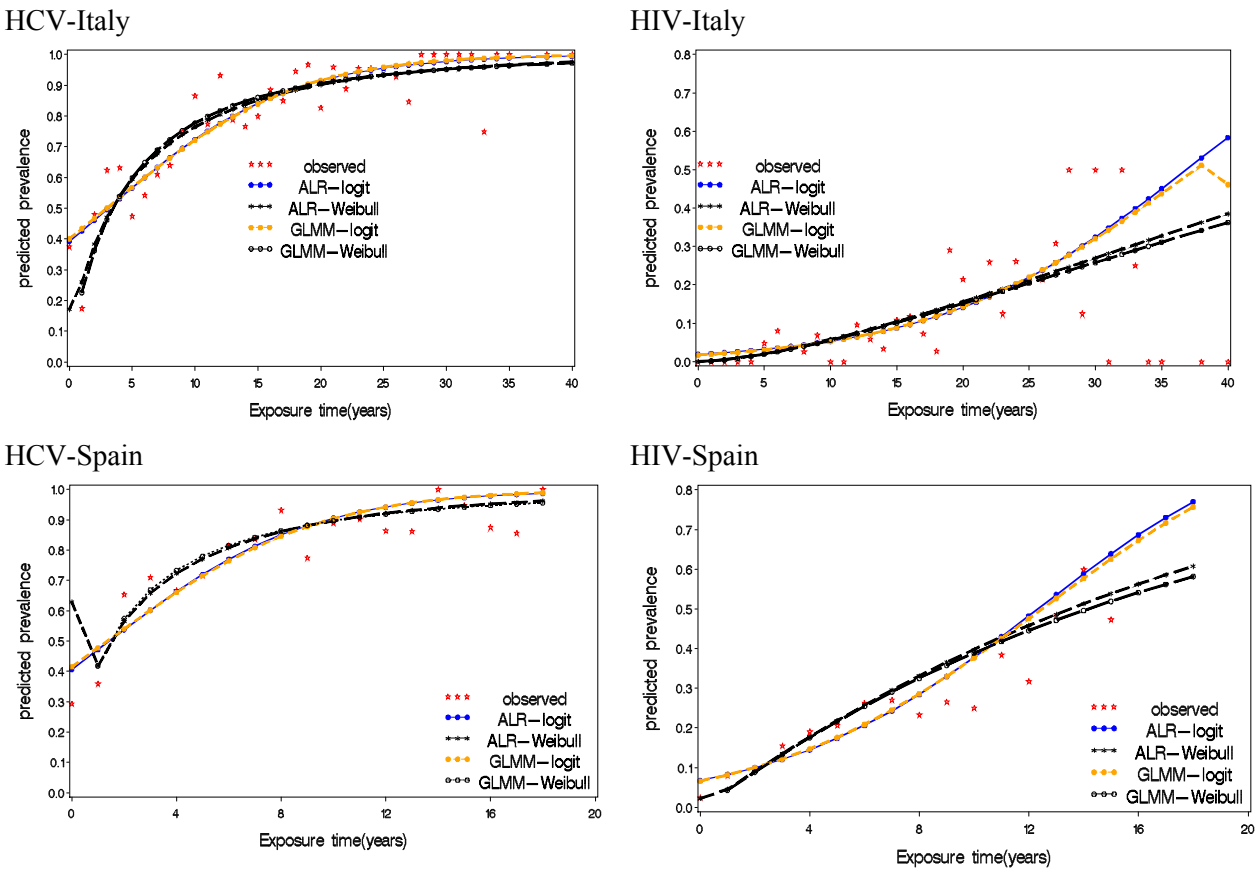


***Figure 6**: Estimated prevalence of HCV (left panels) and HIV (right panels) by exposure time marginalized for all models*

5. Conclusion

This study focused on joint modeling of HCV and HIV using serological data from Italy and Spain. Marked differences have been observed in the prevalence of HIV in IDUs across Italy and Spain, where as the prevalence of HCV infection is very high in both countries (Table 1).

Confection of HCV and HIV is investigated by fitting marginal and shared random effect model using the logit and clog-log links. The ALR model estimates the odds ratio to be infected by the two diseases. Shared random effect is one way of studying the correlation between two responses of the same IDU.

Firstly, we discussed the logistic regression and Weibull models under the marginal and shred random effect models by correcting for the exposure time (duration). The logistic regression model computed through GLM, appeared to be the poorest (highest AIC) one compared to the Weibull model for both countries. As shown in Figure 6 most of the graphs follow the data sets accurately. The same pattern is observed between the corresponding plots of ALR and the marginalized once. For both countries, HCV has the highest force of infection in the beginning of the injection career and decrease for IDUs with relatively long career. This might reflect high incidence in the years prior to recruitment of the sample.

To make decision about how to approach the problem of prevention and intervention in the IDUs group, one needs to have a deeper understanding on how different factors that lead to transmission act together, and which factors are the most influential for continued incidence and high prevalence (EMCDDA, 2004). Hence influence of other risk factors on the transmission of the two infections was the other target of the study. The age at first injection is shown to be a risk factor for HIV infection in Italy with an increasing force of infection for older ages at first injection. For the Italy's study, Prevalence has a positive association with the exposure time and age at first infection. This means that injecting drug users who start injecting at an older age have a higher risk of becoming infected soon after the beginning of their injecting career. The reason might be that they mix with older age groups of injecting drug users who have a higher prevalence already, or that they precede faster to high risk injecting behavior. Information about sharing syringes is only available for the Spain data. Sharing syringes is confirmed to be a

significant risk factor for the transmission of both HCV and HIV. The prevalence of HCV and HIV among IDUs who share syringes is higher than those who don't share as expected from the exploratory data analysis.

In conclusion, a significant co infection of HCV and HIV is observed in serological data of Italy and Spain. Known risk factors such as sharing of syringes and age at first injection were confirmed as risk factors.

References

1. Faes,C.,Hens,N.,Aerts,M.,Shkedy,Z.,Geys,H.,Mintiens,K.,Leavens,H.,and Boelaert,F.(2006) Estimating herd-specific force of infection by using random-effects models for clustered binary data and monotone fractional polynomials. *Appl.Statist*. **55**,Part5,pp.595-613.
2. Farrington, C.P. (1990), Modeling Forces of infection for measles, mumps and rubella. *Statistics in Medicine*, **9,**953-967.
3. Flom,L.P., McMahon,M.J.,and pouget,R.E.(2006) Using PROC NLMIXED and PROC GLMMIX to analyze dyadic data with binary outcomes. *NESUG Analysis.*
4. Jager,J., Limburg,W., Kretzschmar, M., Postma,M., and Weiessing,L. (2004)Hepatitis C and injecting drug use: impact, costs and policy options: *European Monitoring Centre for Drugs and Drug Addiction(EMCDDA) MONOGRAPHS*
5. James M. McMahon, Enrique R. Pouget, and Stephanie Tortu (2006). A guide for multilevel modeling of dyadic data with binary outcomes using SAS PROC NLMIXED. *Comput Stat Data Anal*. 2006 August; 50(12):3663-3680.
6. Keiding,N.,Begtrup, K.,Scheike,T.H., and Hasibeder (1996) Estimation from Current-Status Data in Continoous Time. *Lifetime Data Analysis*, 2,119-129(1996).
7. Molenberghs, G. and Verbeke, G (2005). *Models for Discrete Longitudinal Data*. Springer Series in Statistics, Verlag New York, Inc.
8. Molenberghs, G. and Verbeke, G. (2007). Project: Longitudinal Data Analysis, Diepenbeek, Hasselt University, unpublished course notes.
9. Namata, H., Shkedy, Z., Aerts, M., Faes, C., Mathei, C., Kretzschmar, M, Wiessing, L, Mravcik, V., Suligoi, B., Norden, L., and Vallejo, F.(2008) Estimation of the prevalence and force of infection of hepatitis C among injecting drug users in five European countries. *Working paper.*
10. Shekedy, Z., Aerts, M., Molenberghs, G., Beutels Ph., and Van Damme, p. (2003). Modelling force of infection by using monotone local polynomials. *Applied statistics*, volume 52, pages 469-485.
11. Shekedy, Z., Aerts, M., Molenberghs, G., Beutels Ph., and Van Damme, p. (2006). Modelling age-dependent force of infection from prevalence data using fractional polynomials. *Statistics in Medicine*, Volume 25, Issue 9, pages 1577-1591.
12. Shekedy, Z., Namata, H., Kasim, A., Maringwa, J.T., Aerts, M., Wiessing, L., Kretzschmar, M., and others. (2008). Cross Sectional and longitudinal Evidence for co infection of HCV and HIV. *Working paper*

World Wide Web

13. http://ec.europa.eu/health/ph_determinants/life_style/drug/drug_en.htm (retrieved date: July 15, 2008)
14. http://www.who.int/mediacentre/factsheets/fs164/en/ (retrieved date: August 7, 2008)
15. http://pjms.com.pk/issues/julsep07/article/article5.html (retrieved date: August 7, 2008)
16. http://www.who.int/hiv/mexico2008/idu/en/index.html (retrieved date: August 10, 2008)

Appendix

```
********Marginalization of NLMIXED*********;
/*Logit-Italy*/
data logit_Italy;
do subject=1 to 920 by 1;
do disease= 0 to 1 by 1;
do b=1.0690*rannor(-1);
do t=0 to 40 by 1;
if disease=0 then y=exp(-0.4772+b+0.1633*t)/(1+exp(-0.4772+b+0.1633*t));
else y=exp(-4.6094+b+0.1231*t)/(1+exp(-4.6094+b+0.1231*t));
                                        output;
                                end;
                        end;
                end;
end;

proc sort data=logit_Italy;
by t disease;run;

proc means data=logit_Italy;
var y;
by t disease;
output out=out;run;

proc print data=out;
where _stat_='MEAN';run;

proc gplot data=out;
plot y*t=disease/haxis=axis1 vaxis=axis2 legend=legend1;
axis1 label=(h=2 'Exposure time(years)') value=(h=1.5)
        order=(0 to 40 by 5)minor=none;
axis2 label=(h=2 A=90  'Prevalence') value=(h=1.5)
        order=(0 to 1 by 0.1)minor=none;
        legend1 label=(h=1.5 'Disease:')
                value=(h=1.5 'HCV'  'HIV');
                title h=2.5 'Marginalized NLMIXED for Italy-logit';
                symbol1 c=black i=join w=5 l=1 mode=include;
                symbol2 c=black i=join w=5 l=2 mode=include;
                where _stat_='MEAN';
                run;
```