2012 | School voor Informatietechnologie
Kennistechnologie, Informatica, Wiskunde, ICT

DOCTORAATSPROEFSCHRIFT

Statistical methods for modeling of drug-related and close-contact infections

Proefschrift voorgelegd tot het behalen van de graad van doctor in wetenschappen, wiskunde, te verdedigen door:

Emanuele Del Fava

Promotor: prof. dr. Ziv Shkedy, Universiteit Hasselt
Copromotor: prof. dr. Piero Manfredi, Università di Pisa, Italy

D/2012/2451/32

Maastricht University

universiteit hasselt
KNOWLEDGE IN ACTION

To my parents, Umberto and Giuliana, and to my sister, Giulia.

”O frati,” dissi, ”che per cento milia
perigli siete giunti a l’occidente,
a questa tanto picciola vigilia

d’i nostri sensi ch’é del rimanente
non vogliate negar l’esperïenza,
di retro al sol, del mondo sanza gente.

Considerate la vostra semenza:
fatti non foste a viver come bruti,
ma per seguir virtute e canoscenza”.

(Dante, Inferno, vv. 112–120)

’O brothers, who amid a hundred thousand
Perils,’ I said, ’have come unto the West,
To this so inconsiderable vigil

Which is remaining of your senses still
Be ye unwilling to deny the knowledge,
Following the sun, of the unpeopled world.

Consider ye the seed from which ye sprang;
Ye were not made to live like unto brutes,
But for pursuit of virtue and of knowledge.’

(Translated by H. W. Longfellow, 1867)

Acknowledgements

I wish to thank a large number of people who contributed to this doctoral thesis, either directly or indirectly.

First and foremost, I really want to thank my promoter Ziv Shkedy. He gave me the opportunity to work with him in the last for years to explore the world of modeling infectious diseases, he helped me when I was uncertain about what to do and he told me: "Take it easy!" when I was feeling anxious for my work. I hope we will have other opportunities in the future to work together.

I am also eager to thank my co-promoter Piero Manfredi. He has been helping me since I was one of his students in Demography at the University of Pisa. He was my promoter for my Master thesis in Pisa. Afterwards, he strongly advised me to apply for the Master of Biostatistics at Hasselt University and then accepted to be my co-promoter for this PhD. Finally, he encouraged me to apply for the post-doc position in Milan at Bocconi University. Piero, I really owe you so much!

I am really grateful for the opportunity to prepare my PhD dissertation at two institutions, first and foremost at the Center for Statistics as the UHasselt side of the Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat), and second at the Department of Statistics and Mathematics for Economics at the University of Pisa.

I express my gratitude to all the co-authors with whom I worked together in these four years. In particular, I want to acknowledge the collaboration with the European Monitoring Center for Drugs and Drug Addictions (EMCDDA) in Lisbon (thanks to Lucas Wiessing), with the Department of Public Health at the University of Florence (thanks to Angela Bechini and Paolo Bonanni), with the European Center for Disease Control and Prevention (ECDC) in Stockholm, with the Bruno Kessler Foundation (FBK) in Trento (thanks to Laura Fumanelli, Piero Poletti, Giorgio Guzzetta, and Stefano Merler), and with Bocconi University in Milan (thanks to Alessia Melegaro).

I also would like to thank the colleagues here at CenStat for the useful discussions we had together on the matter of our research, most of all those in the Modeling Infectious Diseases group. In particular, I want to thank the colleagues with whom I shared for a while the office: Tina, Thomas, and Yves. Dank je, Yves, voor je hulp met mijn Nederlands, je hebt het geduld gehad om met mij in het Nederlands te praten. Bovendien heb je verdragen het lawaai in het kantoor wanneer mijn vrienden kwamen om me te bezoeken.

I am grateful for the feedback and suggestions I received from the jury members to an earlier version of this thesis.

Finally, I want to thank with all my heart a number of people who, indirectly, helped and encouraged

me a lot in order to get to this important achievement, which is the discussion of the PhD thesis: my friends and my family.

Many many thanks to the friends I met here in Belgium in different moments: Tanya, Carolina, Ambily, Hellen, Lulu, Nikolina, Shreosi, Martin, Ariel, Matuš, and others. You are very close to my heart, I spent so many great moments with all of you during these five years in Belgium. Meeting you and discovering your culture has really made a better person of me.

Querida Amparo, quiero agradecerte con todo mi corazón por la amistad que me has concedido durante estos cinco años en Bélgica, por el espíritu de colaboración que nos ha permitido de trabajar bien juntos en los proyectos de consultoría, por el tiempo disfrutado junto a tu familia durante los viajes y las vacaciones, por las sonrisas y la ayuda que siempre me das aunque estés cansada o ya tengas tus preocupaciones. Gracias de todo.

Un grazie di cuore agli amici italiani che ho conosciuto qua in Belgio, fra i migliori amici che abbiamo mai avuto: Fortunato, Filippo, Donato e Consuelo. Senza di voi l'esperienza di questi cinque anni passati in Belgio non sarebbe stata la stessa. Abbiamo condiviso molto, pranzi, cene, vacanze, e confidenze. Ci siamo dati una mano l'un con l'altro quando ce n'era bisogno. Grazie. Mi auguro veramente che le nostre strade continuino ad incrociarsi negli anni a venire.

Infine, il ringraziamento più sentito, quello verso la mia famiglia, per avermi dato la possibilità di trasferirmi in Belgio per cinque anni, per avermi supportato economicamente quando ne avevo bisogno e per avermi sempre sostenuto ed incoraggiato affinché facessi del mio meglio. Spero che siate fieri di questo traguardo che ho raggiunto, anche grazie a voi.

Emanuele Del Fava

Diepenbeek, September 23, 2012

Contents

List of Publications

The first part of the thesis, about the statistical modeling of HCV and HIV infections among injecting drug users, is partly based on the following publications:

Del Fava E., Kasim A., Usman M., Shkedy Z., Hens N., Aerts M., Bollaerts K., Scalia Tomba G., Vickerman P., Sutton A.J., Wiessing L., and Kretzschmar M. Joint modeling of HCV and HIV infections among injecting drug users in Italy using repeated cross-sectional prevalence data. *Statistical Communications in Infectious Diseases*, 3(1): 1, 2011. doi: 10.2202/1948-4690.1009.

Del Fava E., Shkedy Z., Hens N., Aerts M., Suligoi B., Camoni L., Vallejo F., Wiessing L., and Kretzschmar M. 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*, 3(1): 3, 2011. doi: 10.2202/1948-4690.1010.

Del Fava E., Shkedy Z., Aregay M.F. and Molenberghs, G. Modeling multivariate, overdispersed binomial data with additive and multiplicative random effects. *Technical report*.

The second part of the thesis, about Bayesian mixture models for antibody titers, is partly based on the following publications:

de La Fé Rodríguez P.Y., Coddens A., **Del Fava E.**, Cortiñas Abrahantes J., Shkedy Z., Maroto Martin L.O., Cruz Muñoz E., Duchateau L., Cox E., and Goddeeris B.M. High prevalence of F4(+) and F18 (+) Escherichia Coli in Cuban piggeries as determined by serological survey. *Tropical Animal Health and Production* (January 14), 2011. doi: 10.1007/s11250-011-9786-4.

Del Fava E., Shkedy Z., Bechini A., Bonanni P., and Manfredi P. Towards measles elimination in Italy: Monitoring herd immunity by Bayesian mixture modelling of serological data. *Epidemics*, 4(3): 124–131, 2012. doi: 10.1016/j.epidem.2012.05.001

Del Fava E., Shkedy Z. Estimating the prevalence and the force of infection of Parvovirus B19 in Belgium using hierarchical Bayesian mixture models. *Technical report*.

The third part of the thesis, about the estimation of transmission parameters for varicella in Europe, is partly based on the following publications:

Iozzi F., Trusiano F., Chinazzi M., Billari F.C., Zagheni E., Merler S., Ajelli M., **Del Fava E.**, and Manfredi P. Little Italy: an agent-based approach to the estimation of contact patterns-fitting predicted matrices to serological Data. *PLoS Computational Biology*, 6(12): e1001021, 2010. doi: 10.1371/journal.pcbi.1001021.

Guzzetta G., Poletti P., **Del Fava E.**, Ajelli M., Scalia Tomba G., Merler S. and Manfredi P. Hope-Simpson's progressive immunity hypothesis may explain Herpes Zoster incidence data. *American Journal of Epidemiology*, 0: 00–00, 2012.

Poletti P., Melegaro A., Ajelli M., **Del Fava E.**, Guzzetta G., Faustini L., Scalia Tomba G., Lopalco P., Rizzo C., Merler S. and Manfredi P. Perspectives on the impact of varicella immunization on herpes zoster. A model-based evaluation from three European countries. *BMC Medicine*, submitted for publication, 2012.

List of Abbreviations

AIC: Akaike's Information Criterion
ALR: Alternating Logistic Regression
B19: Parvovirus B19
BDM: Bivariate Dale Model
BPM: Bivariate Probit Model
BE: Belgium
BFGS: Broyden-Fletcher-Goldfarb-Shanno Method
BIC: Bayesian Information Criterion
BVS: Bayesian Variable Selection
cdf: cumulative density function
DE: Germany
DIC: Deviance Information Criterion
DTC: Drug Treatment Center
ECDC: European Center for Disease Control and Prevention
EMCDDA: European Monitoring Center for Drugs and Drug Addiction
ESEN2: European Sero-Epidemiology Network
ES: Spain
EU: European Union
FBK: Fondazione Bruno Kessler
FOI: Force of Infection
FP: Fractional Polynomial
GB: Great Britain
GCV: Generalized Cross-Validation
GLM: Generalized Linear Model
GLMM: Generalized Linear Mixed Model
HCV: Hepatitis C Virus
HIV: Human Immunodeficiency Virus
IDU: Injecting Drug User
IE: Ireland
IgG: Immunoglobulines G
IL: Israel

ILI: Influence-Like-Illness
IT: Italy
LP: Local Polynomial
LU: Luxembourg
MCMC: Monte Carlo Markov Chain
ML: Maximum Likelihood
NL: Netherlands
PAVA: Pool-Adjacent-Violators Algorithm
PC: Piecewise-Constant FOI
pdf: probability density function
PED: Penalized Expected Deviance
PL: Poland
SK: Slovakia
WAIFW: Whom Acquires Infection From Whom?

Chapter 1

Introduction

This thesis is formed by three parts. The first part is dedicated to statistical modeling of HCV and HIV infections among injecting drug users. The second part deals with hierarchical Bayesian mixture models applied to antibody titers in serological samples. In the third part, we discuss the serology of varicella in Europe and focus on the estimation of its transmission parameters.

1.1 Statistical Modeling of HCV and HIV Infections Among Injecting Drug Users

The first part of the thesis is devoted to statistical modeling of hepatitis C virus (HCV) and human immunodeficiency virus (HIV) infections among injecting drug users (IDUs) in Europe. This work is part of a joint collaboration with the European Monitoring Center for Drugs and Drug Addictions (EMCDDA), the European section of the World Health Organization (WHO) and the government of the Netherlands. The EMCDDA is an European institution with venue in Lisbon whose aim is to provide the European Union (EU) with an overview of the drug problems in the member states and a solid evidence base to support the drugs debate, by collecting data from all over the EU concerning drugs, drug addictions and drug-related infectious diseases. The purpose of this collaboration was to analyze data concerning the diffusion of drug-related infections in IDUs through statistical and mathematical modeling. In particular, we focussed on statistical methods developed in order to model jointly the HCV and the HIV infections in the specific risk group of the IDUs. The choice of these two infections is motivated by the high frequency of IDUs who are infected with both viruses. Indeed, the co-infection with HCV and HIV is a well known phenomenon which affects a large number of IDUs: Rockstroch and Spengler (2004) reports that 70% of HIV-positive IDUs will also be HCV-positive due to their shared transmission route. This implies that, at the general population level, among the HIV-infected persons, 4-5 million were also HCV-infected (Alter, 2006).

Different types of data are available regarding these infections. The most common data are surveillance data, coming from routine geographical repeated cross-sectional studies to assess the prevalence of HCV and HIV infection in a certain population, either a city, a region, or a country. These aggregated data are usually collected based on a random sample of IDUs from drug-treatment centers (DTCs). They provide us information on the burden of the infection in a specific population and allow us to monitor the effect of the health interventions aimed to reduce the spread of the infections over the years. However, this type of data can be used to assess the prevalence of the infection in the population, and must not be used in order to obtain information of the individuals who form that population. This is the "ecological fallacy", that is to say, if we found an association between two infections in certain area, that does not necessarily means that there exists an association between the two infections at the individual level (Piantadosi et al., 1988).

A second type of data, that better addresses the question of whether the two infections are associated at individual level, consist of cross-sectional individual data, arising from *ad hoc* cross-sectional studies designed to estimate the prevalence of the infections in the individuals and investigate the correlation among these infections.

The work presented in this part aims to show an extensive set of advances statistical models in order to model joint binomial data, accounting for possible dependences between the clusters present in the data. Even though none of the models is new in itself, what is new is the idea of combining these models together in order to extract more valuable information about the dependences between multiple infections, which are for this reason analyzed jointly.

Chapter 2 of this thesis gives an introduction to HCV and HIV infections among IDUs. In Chapters 3 and 4, we model the surveillance data using a joint random-effects model in order to estimate the prevalence of HCV and HIV infection in the twenty Italian regions from 1998 to 2006–2007 and, equally important, the correlation between the infections at the regional level. In Chapter 3, we used joint generalized linear mixed models (GLMMs) for the prevalence and we accounted for the regional clustering using the random effects (Del Fava et al., 2011a), estimating therefore the correlation between the infections at population level. In Chapter 4 we re-analyze the data adding a set of overdispersion parameters, in the form of random effects, to accommodate for possible extra variability in the binomial data in the regions over the years. The models are estimated using both maximum likelihood (ML) methods and Markov chain Monte Carlo (MCMC) methods.

Finally, in Chapter 5, we model cross-sectional individual data on the prevalence of HCV and HIV individuals from two studies in Italy (2005) and in Spain (2001–2003) in order to study the co-infection patterns between HCV and HIV infection among IDUs by assessing the effect of drug-use related risk factors on the correlation between HCV and HIV infection at individual level and by estimating the degree of individual heterogeneity in acquiring the infections (Del Fava et al., 2011b).

1.2 Bayesian Mixture Models for Antibody Titers

The second part of this thesis is devoted to statistical modeling of antibody titers of an infection assumed to provide lifelong immunity (even though, in reality, this lifelong immunity is prevented by waning immunity and/or possibility of reinfection). We use hierarchical Bayesian mixture models (Diebolt and Robert, 1994) to estimate age-specific prevalence and force of infection (FOI) of different childhood infections from the antibody titers. The source of the data are cross-sectional serological studies, whose aim is to study the prevalence of an infection in the individuals from data collected analyzing blood specimens.

By standard, data from serological studies, consisting in antibodies to a certain infection, are converted in binary data, used to classify individuals between susceptible and immune to that infection. These data are known as "current status data", because they inform on the current infection status of the individual. They can be seen as survival data with an extreme form of interval censoring (Jewell and Van Der Laan, 2002). Using mixture models for the data, we avoid the need of a conventional cut-off point to construct the current status data, but rather the mixture model classifies the subjects assigning them to the mixture component for which his/her posterior probability is maximized. The use of conventional cut-off points may lead to underestimate the prevalence of an infection (Greiner et al., 1994; Vyse et al., 2004), because they are diagnostic tools, chosen to have a lower sensitivity (ability to correctly identify the true positive cases, the immune) and thus have a higher specificity (ability to correctly identify the true negative cases, the susceptible). In contrast, mixture models, which are data-driven models, tend to have a higher sensitivity and therefore they are more adapted to estimate the prevalence of an infection from a serological study (Vyse et al., 2004).

The work presented in this part extends the previous work done in the application of mixture models to antibody data improving the methodology for the estimation the prevalence and the FOI directly from the antibody titers without using any given cut-off point. We develop a complex framework for these mixture models, taking into considerations different data setting, such as pre- and post-vaccination scenarios, but also several data assumptions, concerning their statistical distribution, the models for the prevalence and the FOI, and the age-specific profiles of the different mixture components.

Chapter 6 introduces general concepts about the estimation of the prevalence and the FOI from infectious diseases from both current status data and antibody data.

In Chapters 7 and 8 we focus on infections in pre-vaccination equilibrium. This implies that individuals can be simply classified either as susceptible or immune, following natural infection. These infections can be therefore possibly described by mixture models with two components, one for the susceptible and one for the naturally immune. In Chapter 7 we introduce hierarchical Bayesian mixture modeling for the estimation of two-components mixture models using several distributions for the data. We discuss several models, either parametric or nonparametric, for the prevalence and the FOI. For the data distributions, we consider both symmetrical and asymmetrical distributions, such as the normal and the skew-normal (Azzalini, 1985) distributions. We apply these models to antibody titers against par-

vovirus B19 infection in Belgium and in Italy.

In Chapter 8, we extend the hierarchical Bayesian mixture models presented in Chapter 7 by relaxing the assumption of constant mean of the immune component. Hence, in Chapter 8, the mean structure in the immune component changes with the age of individuals using a piecewise-constant model with five age groups. Considering each age group as a possible covariate for the model, we select the best subset of covariates using the Bayesian variable selection approach (BVS, George and McCulloch, 1993). In such a way, we can identify in which age groups the mean changes and in which groups it remains constant. In addition, since we can obtain the posterior probability of each model for the mean structure and in particular the model that assumes that the mean antibody level is constant in the immune component, we can test whether the mean structure in the immune component is constant or not. The methodology is applied to antibody titers to parvovirus B19 and VZV infections in five European countries, i.e., Belgium, Finland, Great Britain, Italy, and Poland.

Finally, in Chapter 9, we model an infection in post-vaccination equilibrium, e.g., measles in Tuscany (Italy). For this infection, immunity can follow either natural infection or vaccination. Thus, we do not restrain ourselves to a two-components mixture model, but rather the number of components become a further unknown parameter to be estimated. Given that we do not have any information in order to distinguish naturally immune from vaccinated, we do not aim to estimate the FOI. This multiple-component mixture model allows us to describe with enough flexibility the different degrees of immunity to measles and therefore to identify those groups of individuals who should be targeted more specifically with health interventions, such as aimed vaccination campaigns.

1.3 Estimation of Transmission Parameters for Varicella in Europe

The work presented in the third part of the thesis is part of a project supported by the European Center for Disease Control and Prevention (ECDC) and entitled: “Vaccine preventable disease modeling in the European Union and EEA/EFTA countries: Forecasting the effects of introducing a new vaccine in a national/regional program”.

Following VZV primary infection, the host gets a contagious illness called “chickenpox” or “varicella”. Varicella is a mild infection which in the developed countries targets mainly children and has the peak of incidence in the school ages, between 5 and 9 years old. However, after the primary infection, the virus lies dormant in the dorsal ganglia and may reactivate at older ages causing a disease called “herpes zoster” (HZ) or “shingles”. This disease, which is more severe than chickenpox, is also characterized by rash, but sufferers may experience nerve pain for months or years. HZ is currently the main concern in many European countries against the introduction of a vaccination against VZV in the routine national and regional programs. The fact is that the reasons behind the re-activation of the VZV are still largely unknown and the effects on HZ of reducing partly the circulation of the virus (unless a very high coverage is reached) are still unclear.

The aim of the project was therefore to develop a new mathematical model in order to investigate

the effects of introducing in Europe a vaccine immunization program against VZV, by simulating different vaccination programs. The model takes into account both the primary infection with VZV, causing varicella, and the subsequent boosting, that may be a cause of the re-activation of the dormant virus.

This part of the thesis is based on materials developed for two work packages of the project. According to the initial work plan, the first work package was devoted to data collection and a review of work on on social contact patterns. This research area experienced substantial advancements in recent years, due to the collection of survey data (Mossong et al., 2008), time-use data (Zagheni et al., 2008), or simulated data (Iozzi et al., 2010). Similarly, a substantial part of the third work package was devoted to the estimated, based on serological data, of "best" transmission rates to be used as main building blocks of mathematical models for varicella transmission in a number of European countries.

We start the third part of the thesis in Chapter 10 with an overview of the available serological data for VZV, coming from twelve countries (eleven EU countries, plus Israel) and we model them with flexible age-specific models for the prevalence and the FOI.

The most important development presented in this part of the thesis is shown in Chapter 11, where we focussed on the estimation of the basic reproduction number (R_0) and of the critical threshold (p_c) for varicella, using different contact structure to account for heterogeneous mixing among individuals: social contact data collected with a multicountry cross-sectional survey (POLYMOD matrices, Mossong et al., 2008), social contact data counted in a synthetic population based on socio-demographic census data (FBK matrices, Fumanelli et al., 2012), and newly proposed empirically based WAIFW matrices. We show indeed that several assumptions can be considered in order to account for the heterogenous mixing in the population. In this way the estimation of the prevalence and of the FOI is grounded directly on important concrete features of the population, i.e., the contact patterns, and is not a pure mathematical derivation from the data, as it happens in Chapter 10.

Part I

Joint Modeling of HCV and HIV Infections Among Injecting Drug Users

Chapter 2

Drug-Related Infectious Diseases: An Introduction

Hepatitis C is an infectious disease affecting the liver, caused by the hepatitis C virus (HCV). The infection is often asymptomatic, but once established, chronic infection can progress to scarring of the liver (fibrosis) and advanced scarring (cirrhosis). In some cases, those with cirrhosis will develop liver failure or other complications of cirrhosis, including liver cancer. No vaccine against HCV is available, due to the extensive genetic heterogeneity of the virus (Strickland et al., 2008). The main HCV transmission routes are blood transfusions from unscreened donors, injecting drug use, unsafe therapeutic injections, and other health-care-related procedures (Baker, 2002). It has been found that the sexual transmission of HCV is rare (Vandelli et al., 2004); however, in case of men having sex with men, it has been seen recently that HCV infection can occur in subjects already HIV-infected in absence of injecting drug use (Van de Laar et al., 2011), thus HCV infection emerges as a sexual transmitted infection in this population. The exposure to infected blood in the context of injecting drug use is the predominant path of transmission in the developed countries characterized by low prevalence (Alter, 2006). HCV infection seems to be acquired rapidly after the initiation of an injecting career and many people may have been infected as a result of occasional experimentation with illicit drugs (Mathëi et al., 2002; Shepard et al., 2005; Mathëi et al., 2006). The estimated prevalence of HCV infection worldwide is 2%, representing 123 million people (Perz et al., 2006).

Acquired immunodeficiency syndrome (AIDS) is a disease of the human immune system caused by the human immunodeficiency virus (HIV). This syndrome progressively reduces the effectiveness of the immune system and leaves individuals susceptible to opportunistic infections and tumors. HIV is transmitted through direct contact of a mucous membrane or the bloodstream with a bodily fluid containing HIV, such as blood, semen, vaginal fluid, preseminal fluid, and breast milk. This transmission can involve anal, vaginal or oral sex, blood transfusion, contaminated hypodermic needles, exchange between mother

and baby during pregnancy, childbirth, or breastfeeding, or other exposure to one of the above mentioned bodily fluids. According to the Joint United Nations Programme on HIV/AIDS (UNAIDS), the estimated number of people living with AIDS in 2008 is 33.4 million (0.8%), even though the number of newly infected is decreasing in recent years (UNAIDS, 2009).

Within the population of IDUs, HCV and HIV share the possibility of transmission by exposure with infected blood. HIV infection appears to accelerate HCV-related liver disease, while HCV infection does not appear to affect the rate of HIV disease progression. It is estimated that up to 70% of HIV-positive IDUs will also be HCV-positive due to their shared transmission route (Rockstroch and Spengler, 2004). Considering the general population, Alter (2006) reported that, among the HIV-infected persons, 4-5 million were also HCV-infected. If we compare the prevalence of HCV and HIV infections presented above, we notice that the former is much higher than the latter, both in the general population and in the group of IDUs. Moreover, it is estimated that the transmission risk from HIV-contaminated syringes is ten times less likely than if the syringes were contaminated with HCV, thus HIV requires more exposure to reach high prevalence (Hagan and Des Jarlais, 2000; Crofts et al., 2001; Alter, 2006).

Usually, the typical data available to investigate the association between HCV and HIV infection in IDUs consist of aggregate prevalence data, possibly collected over several years. The analysis of these repeated prevalence data is useful to investigate the time trend of 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.

Vickerman et al. (2010) modeled the association between HCV and HIV infection using prevalence data for IDUs from multiple geographical areas all over the world, by means of two different regression models (a fractional polynomial and a segmented linear model). Vickerman et al. (2010) reported a strong positive correlation between the average change in HIV infection prevalence and the average change in HCV infection prevalence over time. Specifically, the longitudinal data 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.

Following this approach, we conducted an analysis of the aggregated European prevalence data, provided by EMCDDA, collected between 1997 and 2007. The analysis of Vickerman (Vickerman et al., 2010) gave evidence of a threshold in HCV infection prevalence below which HCV infection is endemic, but HIV infection is not. Therefore our goal is to describe the relationship between the prevalence of HIV infection and of HCV infection and to estimate this threshold for European data. We fitted a change-point fractional polynomial (Royston and Altman, 1994) for HIV infection prevalence against HCV infection prevalence in order to estimate the HCV infection prevalence threshold ϕ below which HIV infection

prevalence is not associated with HCV infection prevalence:

$$\pi_{HIV} = \begin{cases} \alpha, & \pi_{HCV} < \phi, \\ \beta_0 + \sum_{j=1}^{2} \beta_j H_j(\pi_{HCV}) & \pi_{HCV} \geq \phi. \end{cases} \tag{2.0.1}$$

Here, π_{HIV} is the prevalence of HIV infection, π_{HCV} is the prevalence of HCV infection, and the function $H_j(\cdot)$ is given by the Box-Tidwell's transformation (Box and Tidwell, 1962). The results of this analysis are shown in Figure 2.1. Below the estimated HCV threshold $\hat{\phi}$, which is equal to 36%, HCV and HIV infection prevalences are not associated, as shown by the first derivative of the model with respect to HCV infection prevalence, which is equal to zero. Conversely, above the threshold, the first derivative of the model shows that any change in HCV infection prevalence is associated with increasing average changes in HIV infection prevalence up to a HCV infection prevalence of 70%; afterwards, the two prevalences are still associated, but now the changes in HCV infection prevalence are in relation with with decreasing changes in HIV infection prevalence.

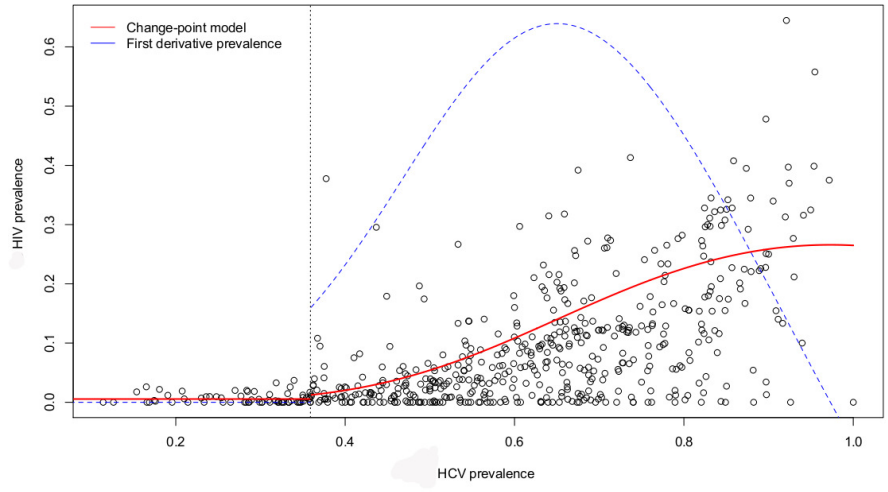


Figure 2.1: Scatter plot of HIV infection prevalence in European region between 1997 and 2007 against HCV infection prevalence with over imposed a change-point second order fractional polynomial. The red solid line represents the model for HIV infection prevalence against HCV infection prevalence, the blue dashed line represents the first derivative of the model with respect to HCV infection prevalence, the black dotted line is the estimated change point $\hat{\phi}$.

However, the conditional model (2.0.1) ignores the time trend in data and possible clustering effect of observations from the same location in Europe. In Chapter 3, we re-analyze the data (but reduced to the Italian data) using a joint generalized linear mixed model (GLMM) in order to describe the evolution of HCV and HIV infection prevalence in time and to estimate the correlation between the two infections infection among the regions. In Chapter 4 we extend the model further and include overdispersion

parameters to account for extra variability in the data.

The analysis presented in Chapters 3 – 4 is based on aggregated data. In case that individual data informing about the current infection status are available, it is the possible to model directly the prevalence of the infections within the IDU population and to estimate the statistical association of HCV infection with HIV infection in the same population (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 (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 for the same subject is available, the co-infection pattern can be studied at individual level (Sutton et al., 2008; Hens et al., 2009c; Del Fava et al., 2011b). In Chapter 5 we present an investigation of the pattern of co-infection with HCV and HIV among IDUs based on individual cross-sectional serological data from Italy and Spain, and we investigate which and how behavioral risk factors affect the association between the infections (Del Fava et al., 2011b).

Chapter 3

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

3.1 Introduction

In this chapter, we investigate the association between HCV and HIV infection from a population perspective by using aggregated repeated prevalence data. We aim to study how the the two infections are associated in Italy through the analysis of regional repeated measurements in time. Specifically, we want to model the correlation between HCV and HIV infection at population level, using aggregate serological data from 20 Italian regions, 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; Congdon, 2003; 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 the chapter is as follows. In Section 3.2, we introduce and discuss the data. In Section 3.3, 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 3.4. In Section 3.5, we formulate the hierarchical Bayesian model and we present the results of the analysis. Finally, we discuss and interpret all the results in Section 3.6.

3.2 The Prevalence Series from Italy

The data analyzed in this chapter 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 chapter: 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. (2010) 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 (Vickerman et al., 2010).

We begin with a preliminary exploratory data analysis. When taking into account the overall prevalence per region (see Figure 3.1), 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 3.2 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 3.2 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. We do not know the reason of the extremely large variability in Valle d'Aosta, but we expect that there could be problems in data collection.

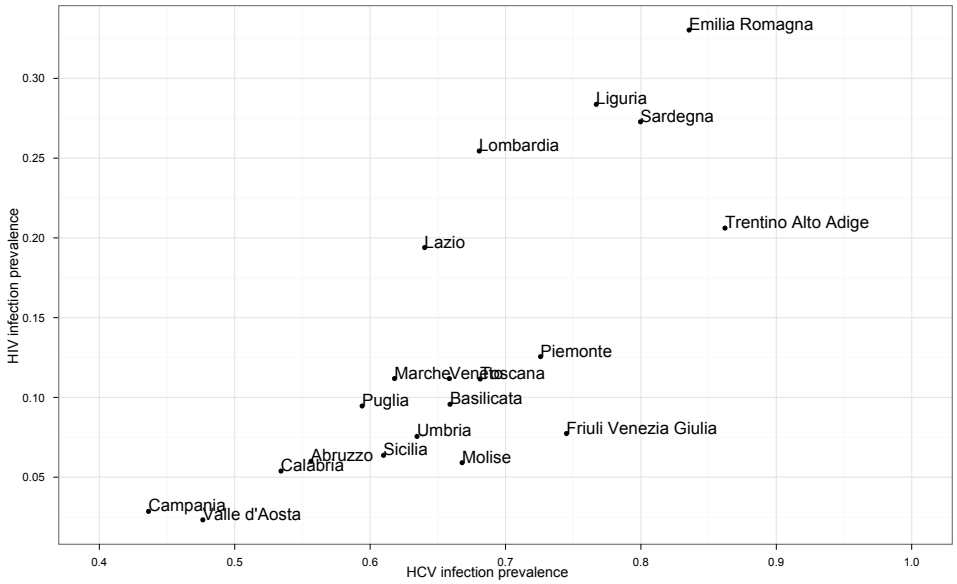


Figure 3.1: Overall regional observed prevalence (averaged over the years 1998-2006) for each of the 20 regions.

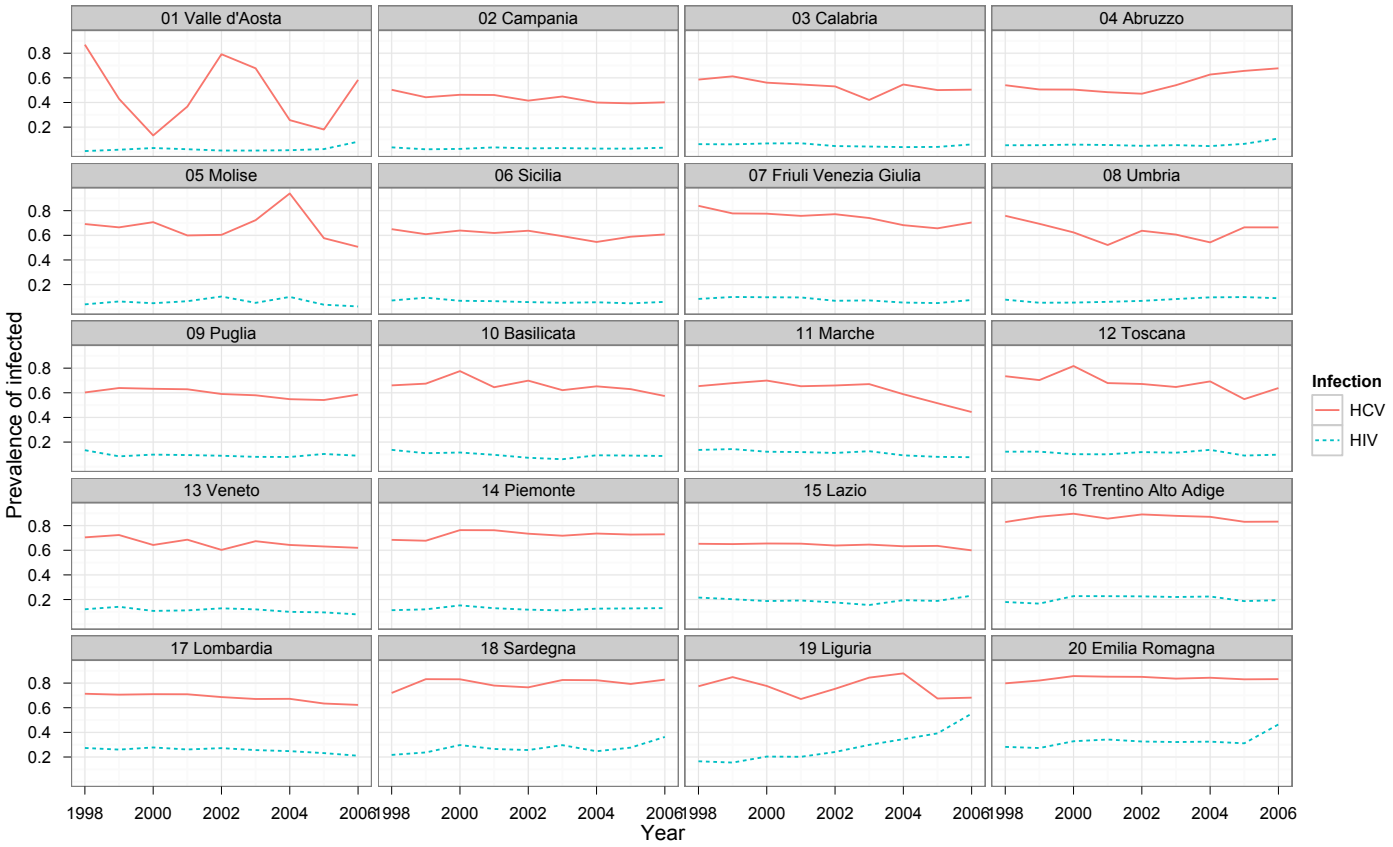


Figure 3.2: Observed prevalence profiles of HCV infection (continuous line) and HIV infection (dashed line) in Italy from 1998 to 2006 for the 20 regions, sorted by the average HIV infection prevalence over the years.

3.3 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.

3.3.1 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:

$$Y_{ijk} \sim Bin(\pi_{ijk}, n_{ijk}) \qquad i = 1,\ldots,20, \qquad j = 1,\ldots,9, \qquad k = 1,2. \tag{3.3.1}$$

Here, π_{ij1} and π_{ij2} 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:

$$\begin{cases} g(\pi_{ij1}) = \beta_{01} + \beta_{11j}, \\ g(\pi_{ij2}) = \beta_{02} + \beta_{12j}. \end{cases} \tag{3.3.2}$$

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.

3.3.2 Generalized Linear Mixed Models (GLMMs)

3.3.2.1 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:

$$\begin{cases} \text{logit}(\pi_{ij1}) = \beta_{01} + \beta_{11j} + a_i, \\ \text{logit}(\pi_{ij2}) = \beta_{02} + \beta_{12j} + b_i. \end{cases} \tag{3.3.3}$$

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

$$\begin{pmatrix} a_i \\ b_i \end{pmatrix} \sim MVN\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, D_{I_1} = \begin{pmatrix} \sigma_a^2 & 0 \\ 0 & \sigma_b^2 \end{pmatrix}\right]. \tag{3.3.4}$$

The diagonal structure of the covariance matrix D_{I_1} of the random effects for model (2a) implies that the measurements 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:

$$D_{I_2} = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}. \tag{3.3.5}$$

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

3.3.2.2 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:

$$\begin{cases} \text{logit}(\pi_{ij1}) = \beta_{01} + \beta_{11j} + b_i, \\ \text{logit}(\pi_{ij2}) = \beta_{02} + \beta_{12j} + \gamma b_i. \end{cases} \tag{3.3.6}$$

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 $\sigma_{HIV}^2 = \gamma^2 \sigma_{HCV}^2$. 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.

3.3.2.3 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. We firstly assume a model similar to the independent random-effects model,

namely,

$$\begin{cases} \text{logit}(\pi_{ij1}) = \beta_{01} + \beta_{11j} + a_i, \\ \text{logit}(\pi_{ij2}) = \beta_{02} + \beta_{12j} + b_i. \end{cases} \tag{3.3.7}$$

Secondly, we specify two possible covariance matrices for the random effects that assume a correlation between the infection which is neither equal to zero or to one: 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:

$$\begin{aligned} D_U &= \begin{pmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{pmatrix} \quad \text{Unstructured matrix.} \\ D_T &= \begin{pmatrix} \sigma^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma^2 \end{pmatrix} \quad \text{Toeplitz matrix,} \end{aligned} \tag{3.3.8}$$

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:

$$\rho_U = \frac{\sigma_{ab}}{\sigma_a \sigma_b} \qquad \rho_T = \frac{\sigma_{ab}}{\sigma^2}. \tag{3.3.9}$$

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.

3.4 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 3.1 presents the covariance parameter estimates, together with the deviance and the Akaike's information criterion (AIC) for each model. While the model with the lowest deviance is the one that fits the data the best without taking into account the number of parameters used, the model with the smallest AIC is the one with the best compromise between goodness-of-fit and model complexity. Moreover, we use the likelihood ratio test (LRT) to compare the covariance structures of the models.

For the fitted models, both the deviance and the AIC lead to the same conclusions. According to the AIC, we discard the independence and the shared random-effects models. Between the shared random-effects model and the independent and correlated random-effects model there is a large reduction in the

Table 3.1: Comparison of the fitted models with deviance and AIC and parameter estimates for the variance components with respective 95% asymptotic confidence intervals.

Model	Type	Deviance	AIC	Covariance parameter estimates
1	Independence model	84717	84753	-
3	GLMM shared RE	20992	21032	$\hat{\gamma} = 1.96$ $(1.92, 2.00)$ $\hat{\sigma}^2 = 0.13$ $(0.04, 0.22)$
2b	GLMM independent RE Equal parameters	10578	10616	$\hat{\sigma}^2 = 0.41$ $(0.22, 0.61)$
2a	GLMM independent RE Different parameters	10575	10615	$\hat{\sigma}_a^2 = 0.25$ $(0.08, 0.42)$ $\hat{\sigma}_b^2 = 0.58$ $(0.19, 0.98)$
4b	GLMM correlated RE Toeplitz covariance	10568	10608	$\hat{\sigma}^2 = 0.41$ $(0.18, 0.64)$ $\hat{\sigma}_{ab} = 0.26$ $(0.03, 0.49)$ $\hat{\rho} = 0.64$ $(0.35, 0.92)$
4a	GLMM correlated RE Unstructured covariance	10563	10605	$\hat{\sigma}_a^2 = 0.25$ $(0.08, 0.42)$ $\hat{\sigma}_b^2 = 0.57$ $(0.19, 0.96)$ $\hat{\sigma}_{ab} = 0.26$ $(0.04, 0.48)$ $\hat{\rho} = 0.69$ $(0.43, 0.94)$

AIC, likely due to the very restrictive constraint in the shared random-effects model.

Considering the remaining GLMMs, the models which best fit in terms of model complexity are the two correlated random-effects. The 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, and 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 3.3 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 3.4 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 3.1 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.

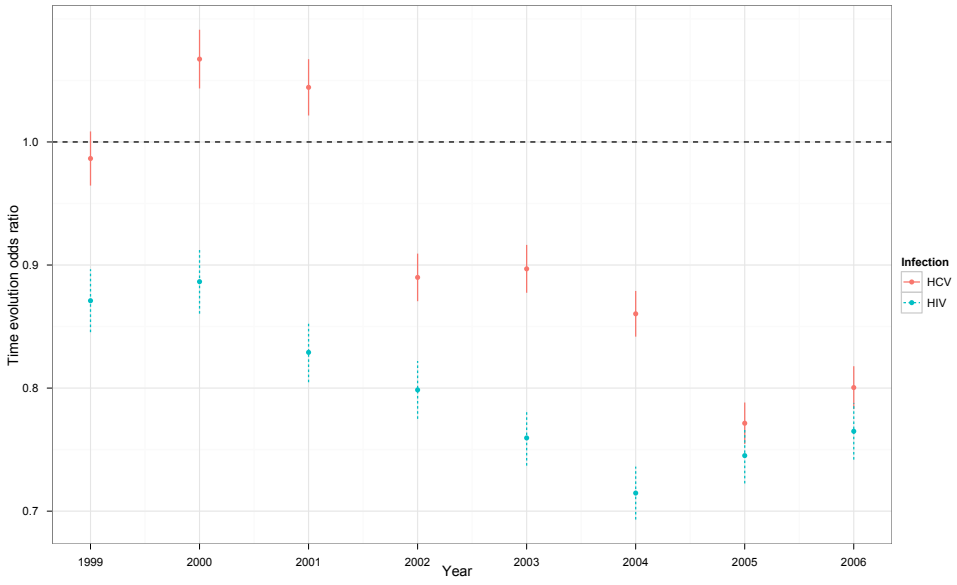


Figure 3.3: Time evolution odds ratios $\exp(\beta_{1kj})$ for HCV and HIV infection (baseline: year 1998) with 95% asymptotic CI, from the correlated model (4a).

3.5 Hierarchical Bayesian Random-Effects Model for HCV and HIV Infections

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) as a hierarchical Bayesian model (Gilks et al., 1996; Congdon, 2003; Gelman et al., 2004). 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

$$\begin{cases} g(\pi_{ij1}) = \alpha_{1i} + \beta_{11j}, \\ g(\pi_{ij2}) = \alpha_{2i} + \beta_{12j}. \end{cases} \tag{3.5.10}$$

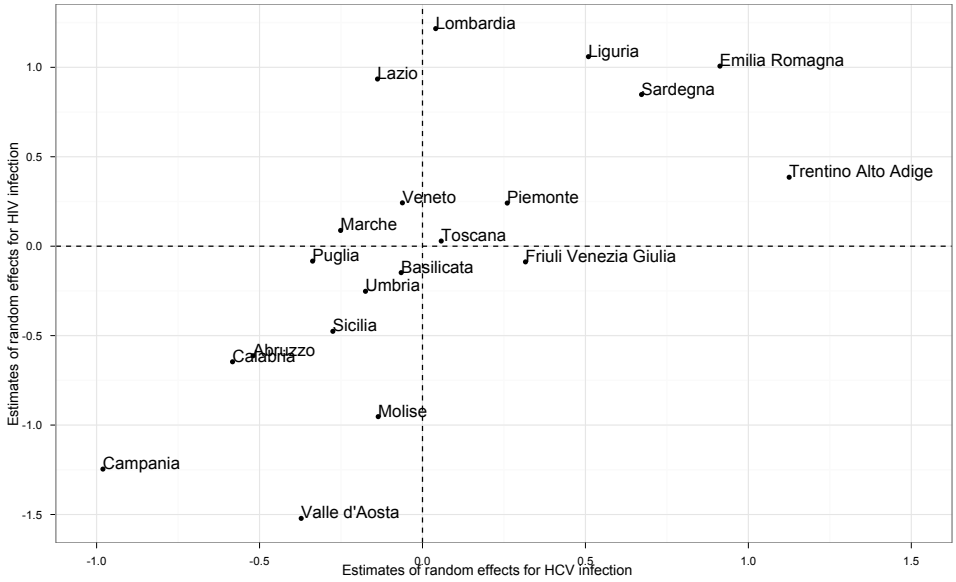


Figure 3.4: Estimates of the region-specific random effects for HCV and HIV infection from the correlated model (4a).

For the joint prior distribution of the random intercepts α_{1i} and α_{2i}, we use the hierarchically centered parameterization (Zhao et al., 2006; 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},

$$\begin{pmatrix} \alpha_{1i} \\ \alpha_{2i} \end{pmatrix} \sim MVN \left[\begin{pmatrix} \beta_{01} \\ \beta_{02} \end{pmatrix}, D = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{pmatrix} \right]. \tag{3.5.11}$$

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}:

$$\begin{cases} \beta_{01} \sim N(0, 1000), \\ \beta_{02} \sim N(0, 1000). \end{cases} \tag{3.5.12}$$

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 $\varepsilon = 0.01$ (Zhao et al., 2006). For example,

the time evolution parameters for HCV infection are given by:

$$\begin{cases} \beta_{11j} \sim N(\mu_{\beta_{11}}, \sigma^2_{\beta_{11}}), & \text{prior for } \beta_{11j}, \\ \mu_{\beta_{11}} \sim N(0, 1000), & \text{hyperprior for } \mu_{\beta_{11}}, \\ \sigma^2_{\beta_{11}} \sim IG(0.01, 0.01), & \text{hyperprior for } \sigma^2_{\beta_{11}}. \end{cases} \tag{3.5.13}$$

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):

$$D \sim W_2 \left[R = \begin{pmatrix} \sigma^2_a & \sigma_{ab} \\ \sigma_{ab} & \sigma^2_b \end{pmatrix} \right]; \tag{3.5.14}$$

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.

3.5.1 Application to the Data: Hierarchical Bayesian Model

The hierarchical Bayesian model was fitted in JAGS (Plummer, 2003) through the package `R2jags` (Su and Yajima, 2012) 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 3.5 shows the posterior densities of the time evolution parameters for HCV infection, which coincide with the ML estimates from the correlated GLMM (4a). Table 3.2 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 3.6 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 3.7 shows the posterior means for the random effects of HCV and HIV infection and reveals the same pattern observed in Figure 3.4. 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 $\bar{\beta}_{0k}$, i.e., $\bar{\alpha}_{1i} - \bar{\beta}_{01}$ and $\bar{\alpha}_{2i} - \bar{\beta}_{02}$, respectively. In this way, the random effects' estimates that we obtain are comparable with the ones from the correlated GLMM.

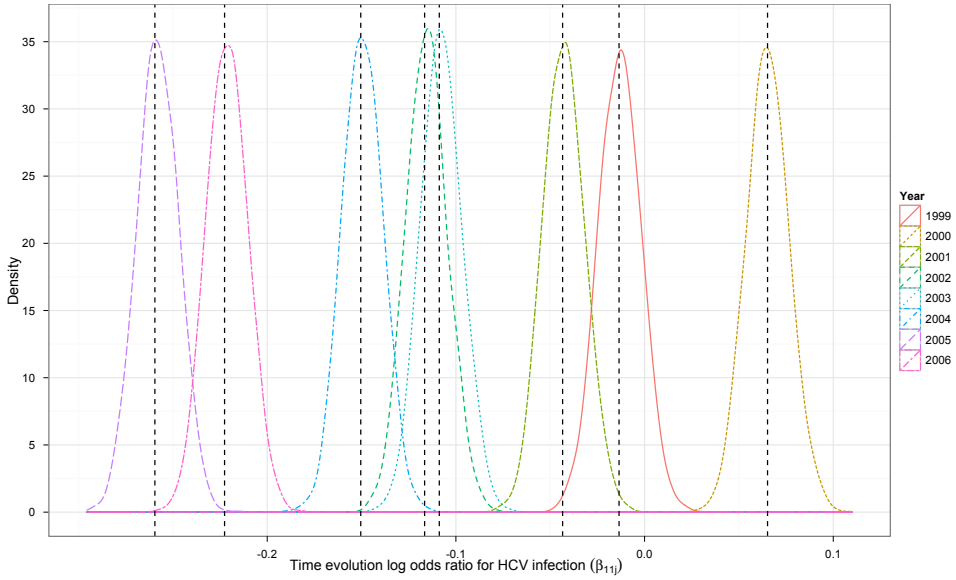


Figure 3.5: Density estimates for the posterior distribution of the time evolution log odds ratios of HCV infection β_{11j}, from 1999 to 2006 (baseline: year 1998), together with the maximum likelihood estimates from the correlated GLMM with unstructured covariance (dashed lines).

3.6 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

Table 3.2: 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”.

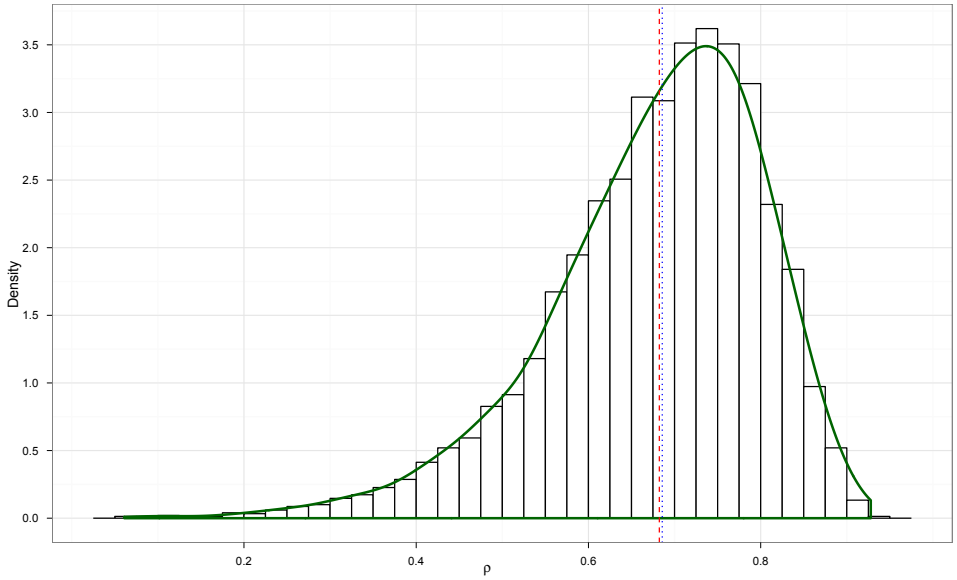


Figure 3.6: Density estimate for the posterior distribution of the correlation coefficient for the region-specific random effects of HCV and HIV infection, with over imposed the estimates from the Bayesian model (dashed line) and from the GLMM (dotted line).

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).

The point estimates of the correlation obtained from the GLMM and the Bayesian model are equal

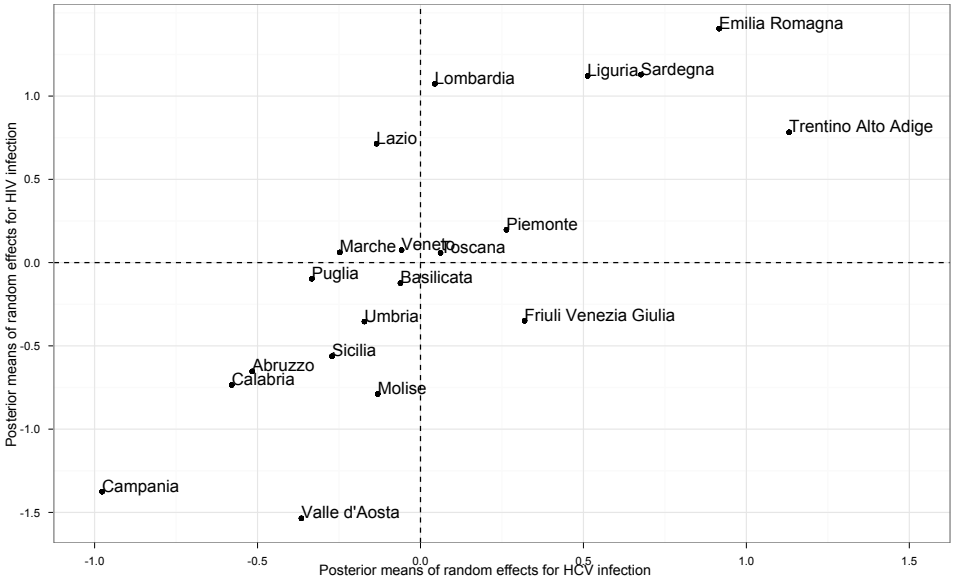


Figure 3.7: Posterior means of the random effects of HCV ($\bar{\alpha}_{1i} - \bar{\beta}_{01}$) and HIV ($\bar{\alpha}_{2i} - \bar{\beta}_{02}$) from the hierarchical Bayesian correlated model.

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 chapter. Such an analysis is presented in Del Fava et al. (2011b), where the effects of drug-related behavioral risks on the association between HCV and HIV infection are investigated.

The data analyzed in this chapter 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 (Vickerman et al., 2010).

In this chapter, 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 (Nikoloulopoulos 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.

Chapter 4

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

4.1 Introduction

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

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

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

binomial model (Skellam, 1948) for binomial data and the negative binomial model (Breslow, 1984) for count data, but also some versions of the GLMMs. A wide review of approaches able to deal with overdispersion can be found in Hinde and Demétrio (1998).

In particular cases, depending on the data at hand, interest may lie in simultaneously combining these two phenomena, clustering and overdispersion. Both marginal and random-effects models can be used to address these concerns. If the interest lies in the estimation of the fixed effects rather than in the correlation structure, marginal models can be used to adjust the variance-covariance structure in order to accommodate for clustering and overdispersion. For instance, the GEE2 approach (Qaqish and Liang, 1992) and, better suited for binomial data, the alternating logistic regression (Carey et al., 1993) can both model the marginal means of the outcomes as well as the correlation between pairs of within-cluster measurements. Chen and Ahn (1997) go further in this direction and develop a marginal model for multivariate overdispersed binomial data where the mean structure depends on two multiplicative nested random effects. On the other hand, conditional models depending on random effects are a better choice if we are more interested in modeling the individual profiles rather than the population mean and in estimating the correlation among the measurements. Molenberghs et al. (2007, 2010) proposed a class of GLMMs that accommodate for clustering and overdispersion, making use of two separate sets of random effects, which then are estimated with maximum likelihood (ML) methods. These GLMMs are meant to be used for modeling normal, binomial, Poisson and time-to-event data.

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

We apply the proposed methodology to prevalence data of hepatitis C virus (HCV) and human immunodeficiency virus (HIV) infection for injecting drug users (IDUs) in treatment from the 20 Italian regions from 1998 to 2007. The chapter is organized as follows. In Section 4.2 we present the data. In Section 4.3 we introduce the proposed methodology, giving details about the GLMMs with additive and multiplicative overdispersion parameters and about model selection. Section 4.4 is dedicated to the presentation of the main results, while in Section 4.5 we wrap up with a discussion of the proposed methodology.

4.2 Data

The longitudinal data for HCV and HIV prevalence discussed in Chapter 3 will be used for the analysis presented in this chapter, even though they are now updated to 2007. Figure 4.3 (panel a) shows the

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

4.3 Methodology

4.3.1 Joint Model with Additive Overdispersion Parameters

Our starting point for the analysis is the joint model discussed in Chapter 3, which is parameterized in the following way:

$$\begin{cases} \text{logit}(\pi_{ij1}) = \alpha_1 + \beta_{j1} + \gamma_{i1}, \\ \text{logit}(\pi_{ij2}) = \alpha_2 + \beta_{j2} + \gamma_{i2}. \end{cases} \tag{4.3.1}$$

The random effects γ_{ik}, accounting for regional clustering, are multivariate normally distributed with mean $(0,0)$ and with the following covariance matrix:

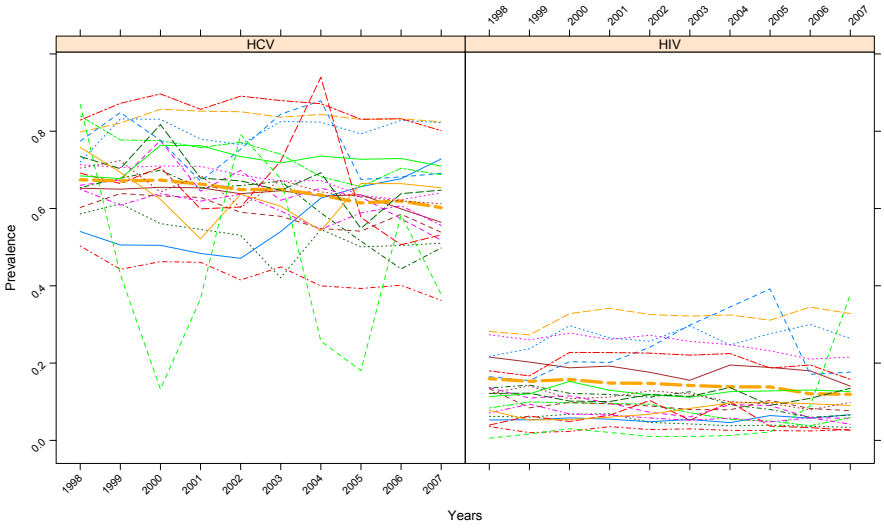
$$\boldsymbol{D}_\gamma = \begin{pmatrix} \sigma^2_{\gamma_1} & \rho_{\gamma_1\gamma_2}\sigma_{\gamma_1}\sigma_{\gamma_2} \\ \rho_{\gamma_1\gamma_2}\sigma_{\gamma_1}\sigma_{\gamma_2} & \sigma^2_{\gamma_2} \end{pmatrix}. \tag{4.3.2}$$

We now propose extensions to the basic joint model (4.3.1) to deal simultaneously with clustering and overdispersion. This is achieved by using separate sets of random effects for overdispersion and for clustering. We opt for a set of random slopes θ_{ijk}, which, for convenience, are assumed to be independent of the random intercepts γ_{ik}. In this section, we focus on additive overdispersion parameters θ_{ijk} (McLachlan, 1997), which are introduced on the same scale of the linear predictor. We consider four possible situations of interest for the overdispersion random effects: (1) they are shared by the two infections; (2) they are differentiated by infection and independent; (3) they are differentiated by infection and allowed to be dependent; (4) they are differentiated by infection and by year, leading thus to time-specific covariance matrices, $\boldsymbol{D}_{\theta j}$.

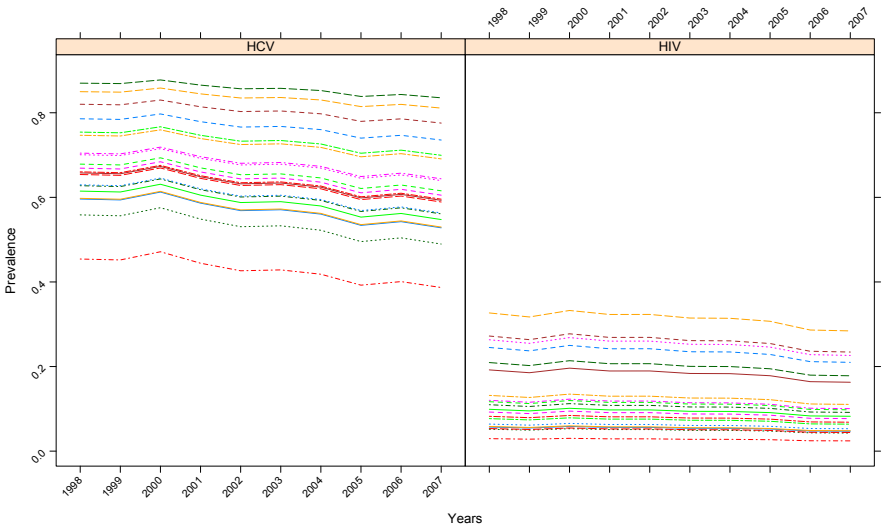
4.3.1.1 Shared Overdispersion Parameters

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

$$\begin{cases} \text{logit}(\pi_{ij1}) = \alpha_1 + \beta_{j1} + \gamma_{i1} + \theta_{ij}, \\ \text{logit}(\pi_{ij2}) = \alpha_2 + \beta_{j2} + \gamma_{i2} + \delta\theta_{ij}. \end{cases} \tag{4.3.3}$$



(a) Observed data



(b) Basic model (4.3.1)

Figure 4.1: Observed (panel a) and estimated (from the basic model, panel b) individual prevalence profiles of HCV and HIV infections for the 20 Italian regions between 1998 and 2007. The bold line in panel (a) stands for the overall prevalence, obtained by pooling together the regional results.

We assume that θ_{ij} has a normal prior distribution, $\theta_{ij} \sim N(0, \sigma^2_\theta)$, where σ^2_θ is an hyperparameter with a flat inverse Gamma (IG) prior distribution, $\tau_\theta = 1/\sigma^2_\theta \sim \Gamma(\varepsilon, \varepsilon)$. For ε we choose to use the value 0.01, because it has been found to provide more stable results with respect to smaller values (Zhao et al., 2006).

The underlying assumption behind the shared overdispersion model (4.3.3) is that the correlation between the infections described by the overdispersion parameters is equal to one. We use the parameter δ to relax the assumption of common variance between the random slopes of HCV and HIV infections, since $\sigma^2_{\theta_{HIV}} = \delta^2 \sigma^2_{\theta_{HCV}}$. Note that the case with $\sigma^2_\theta = 0$ implies the absence of regional-specific evolution patterns during the years.

4.3.1.2 Independent Overdispersion Parameters

Another possible model is the independent overdispersion model, which, differently from Model (4.3.3) includes infection-specific random slopes, θ_{ijk}:

$$\begin{cases} \text{logit}(\pi_{ij1}) = \alpha_1 + \beta_{j1} + \gamma_{i1} + \theta_{ij1}, \\ \text{logit}(\pi_{ij2}) = \alpha_2 + \beta_{j2} + \gamma_{i2} + \theta_{ij2}. \end{cases} \tag{4.3.4}$$

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

$$\begin{pmatrix} \theta_{ij1} \\ \theta_{ij2} \end{pmatrix} \sim MVN \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{D}_\theta = \begin{pmatrix} \sigma^2_{\theta_1} & 0 \\ 0 & \sigma^2_{\theta_2} \end{pmatrix} \right]. \tag{4.3.5}$$

Regarding the variances of the overdispersion parameters θ_{ijk}, we assume that $\sigma^2_{\theta_1}$ and $\sigma^2_{\theta_2}$ are independently distributed according to a flat IG prior distribution, that is to say, $\tau_{\theta_1} = 1/\sigma^2_{\theta_1} \sim \Gamma(0.01, 0.01)$ and $\tau_{\theta_2} = 1/\sigma^2_{\theta_2} \sim \Gamma(0.01, 0.01)$ (Zhao et al., 2006).

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

4.3.1.3 Correlated Overdispersion Parameters

As a further extension, Model (4.3.4) can be expressed as a correlated overdispersion model. The new model is similar to the independent overdispersion model (4.3.4), except for the overdispersion parameters that are assumed to be correlated between the infections:

$$\begin{pmatrix} \theta_{ij1} \\ \theta_{ij2} \end{pmatrix} \sim MVN \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{D}_\theta = \begin{pmatrix} \sigma^2_{\theta_1} & \rho_{\theta_1\theta_2}\sigma_{\theta_1}\sigma_{\theta_2} \\ \rho_{\theta_1\theta_2}\sigma_{\theta_1}\sigma_{\theta_2} & \sigma^2_{\theta_2} \end{pmatrix} \right]. \tag{4.3.6}$$

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

Assuming an unstructured covariance matrix for $\boldsymbol{D}_\theta$, it is possible to estimate an additional corre-

lation $\rho_{\theta_1\theta_2}$ between the infections. This implies that, after having accounted for the correlation between the infections using the region-specific random intercepts γ_{ik}, we find that there still is some constant correlation between HCV and HIV infection in the time evolution, which is captured by the overdispersion parameters.

4.3.1.4 Correlated Overdispersion Parameters with Time-Dependent Correlation

The last additive model that we consider extends Model (4.3.6) relaxing the hypothesis of a constant correlation between HCV and HIV infection captured by the overdispersion parameters. Rather, we now let the covariance matrix $\boldsymbol{D}_\theta$ free to change year after year:

$$\begin{pmatrix} \theta_{ij1} \\ \theta_{ij2} \end{pmatrix} \sim MVN\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{D}_{\theta j} = \begin{pmatrix} \sigma^2_{\theta_1|j} & \rho_{\theta_1\theta_2|j}\sigma_{\theta_1|j}\sigma_{\theta_2|j} \\ \rho_{\theta_1\theta_2|j}\sigma_{\theta_1|j}\sigma_{\theta_2|j} & \sigma^2_{\theta_2|j} \end{pmatrix}\right]. \tag{4.3.7}$$

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

4.3.2 Joint GLMM with Multiplicative Overdispersion Parameters

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

$$\begin{cases} Y_{ijk} \sim Bin(\pi_{ijk} = \theta_{ijk} \cdot \kappa_{ijk}, n_{ijk}), \\ \text{logit}(\kappa_{ij1}) = \alpha_1 + \beta_{j1} + \gamma_{i1}, \\ \text{logit}(\kappa_{ij2}) = \alpha_2 + \beta_{j2} + \gamma_{i2}. \end{cases} \tag{4.3.8}$$

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

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

$$\begin{aligned} \theta_{ij1} &\sim Be(1,1), \\ \theta_{ij2} &\sim Be(1,1). \end{aligned} \tag{4.3.9}$$

However, in general, we can assume that the parameters of the beta prior distributions are hyperparameters to be estimated:

$$\begin{aligned} \theta_{ij1} &\sim Be(a,b), \\ \theta_{ij2} &\sim Be(a,b). \end{aligned} \tag{4.3.10}$$

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

$$\sigma^2 = \frac{ab}{(a+b)^2(a+b+1)}. \tag{4.3.11}$$

As concerns the prior distribution of the hyperparameters a and b in (4.3.11), we choose independent diffuse Uniform distributions within the range [0,100]:

$$\begin{aligned} a &\sim U(0,100), \\ b &\sim U(0,100). \end{aligned} \tag{4.3.12}$$

4.3.3 Model Checking and Model Selection

Under the Bayesian approach, the assessment of the discrepancy between the observed data and the replicated data from the model is usually performed by checking the predictive posterior distribution of the data. In this way, similarly to the frequentist approach, it is possible to compute a Bayesian $P-$value (Gelman et al., 1996). However, when we are dealing with hierarchical models containing latent parameters, such as random-effects models, it may be quite hard to use this approach (Gelfand, 2003). Several modifications for hierarchical models have been therefore constructed (Bayarri and Castellanos, 2007; Steinbakk and Storvik, 2009), but they have been not used in this work. Instead, we rather focused on model selection.

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

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

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

$$PED = \overline{D(\theta)} + p_{opt}. \tag{4.3.13}$$

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

with random-effects models. Similarly to the DIC, the smaller the PED, the better.

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

$$\left\{D_{1,2}^{(m)} = D_1^{(m)} - D_2^{(m)} : m = 1, \ldots, M\right\}, \tag{4.3.14}$$

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

$$P(D_{1,2}^{(m)} < -2\log 9) = \frac{1}{M}\sum_{m=1}^{M} I(D_{1,2}^{(m)} < -2\log 9 = -4.39), \tag{4.3.15}$$

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

4.4 Results

Table 4.1: Information criteria for model selection.

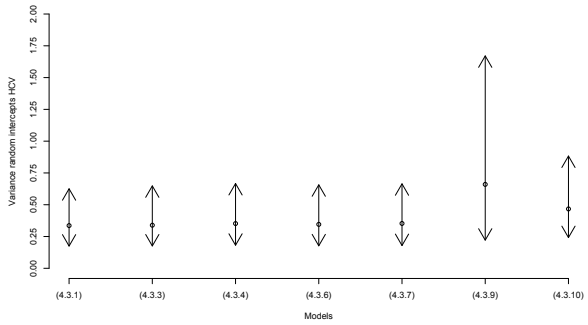
Type	Model	DIC	PED	Diff. $\overline{D(\theta)}$
Basic	(4.3.1)	10274	10351	-
Additive	(4.3.3)	4998	6176	$D_{4.3.3,4.3.1} = \overline{D_{4.3.3}} - \overline{D_{4.3.1}}$ $-5442(-5491,-5391)$ $P(D_{4.3.3,4.3.1} < -4.39) = 1$
	(4.3.4)	3835	7469	$D_{4.3.4,4.3.3} = \overline{D_{4.3.4}} - \overline{D_{4.3.3}}$ $-1304(-1378,-1227)$ $P(D_{4.3.4,4.3.3} < -4.39) = 1$
	(4.3.6)	3835	7515	$D_{4.3.6,4.3.4} = \overline{D_{4.3.6}} - \overline{D_{4.3.4}}$ $-1(-89,84)$ $P(D_{4.3.6,4.3.4} < -4.39) = 0.47$
	(4.3.7)	3775	8099	$D_{4.3.7,4.3.4} = \overline{D_{4.3.7}} - \overline{D_{4.3.6}}$ $-73(-157,11)$ $P(D_{4.3.7,4.3.4} < -4.39) = 0.94$
Multiplicative	(4.3.9)	3782	8270	$D_{4.3.9,4.3.7} = \overline{D_{4.3.9}} - \overline{D_{4.3.7}}$ $-7(-86,76)$ $P(D_{4.3.9,4.3.7} < -4.39) = 0.52$
	(4.3.10)	3869	7224	$D_{4.3.9,4.3.10} = \overline{D_{4.3.9}} - \overline{D_{4.3.10}}$ $-113(-200,-22)$ $P(D_{4.3.9,4.3.10} < -4.39) = 0.99$

The hierarchical Bayesian models presented in the previous section are fitted to data using MCMC methods, specifically Gibbs sampling implemented through JAGS software (Plummer, 2003). For each model,

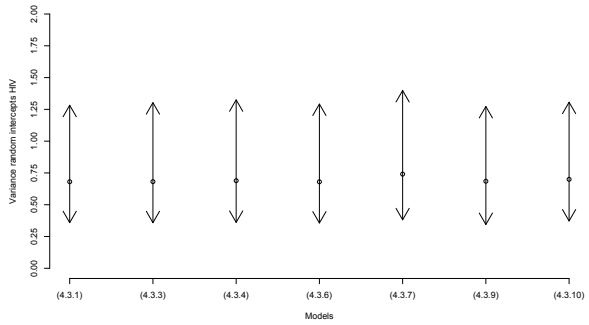
we used three chains of 250000 iterations each, burn-in of 125000 and thinning of 125. Convergence for all parameters was assessed with the potential scale reduction factor (Gelman and Rubin, 1992), for which approximate convergence is diagnosed when the factor approaches one. For each model the DIC and the PED are computed, based on further 20000 iterations; furthermore, we compute the difference in posterior deviances for each pair of models. We refer to Table 4.1 for a summary of our main results. For each model, we give the values of each selection criterion. We notice that the PED and the difference in posterior deviances lead us to favor different models. As expected, the worst model is the basic joint GLMM (Model 4.3.1). Among the models with additive overdispersion parameters, the PED indicates that the shared random-effects model is the best model (Model 4.3.3), whereas the difference in posterior deviances favors the model with correlated overdispersion parameters and time-dependent correlation coefficients $\rho_{\theta_1\theta_2}$ between HCV and HIV infection (Model (4.3.7)). We discard the model with correlated overdispersion parameters (Model (4.3.6)), implying that the assumption of a constant correlation between HCV and HIV infection over the years on the scale of the overdispersion parameters is not tenable. The same results given by the difference in posterior deviances are obtained when the DIC is used for model selection. Indeed, from Figure 4.2 (panel d), showing the posterior means of the time-dependent correlation coefficients with their respective 95% CI, we observe that the correlation is never significantly different from zero, except for 2006, when it becomes significantly positive. For the multiplicative models, according to the difference in posterior deviances and DIC, Model (4.3.9) outperforms Model (4.3.10), while the PED ranks the models in the other way around.

Finally, when comparing the best additive model and the best multiplicative model, the PED favors the additive model with shared random effects, while, according to the difference in posterior deviance we do not have enough confidence to choose between Model (4.3.7) and Model (4.3.9). The DIC criterion favors Model 4.3.7. Figure 4.2, (panels a–c), for each fitted model, presents the posterior means of the variance components (with 95% CI), for HCV and HIV infection, respectively, and of the correlation for the clustering random effects γ_{ik}. For the models with additive random effects only, (Models 4.3.1 – 4.3.7), we notice that misspecifying the overdispersion random effects or even ignoring them affects neither the estimates of the variances components for the clustering random effects, nor the length of their credible intervals (panels a and b). This may be in relation to the fact the γ_{ik} and θ_{ik} are assumed to be independent and are both introduced on the same scale of the linear predictor, therefore the former accommodate the within-region association all over the years, while the latter capture the unexplained additional variability. However, the same argument does not hold for the models with multiplicative random effects. For instance, for Model (4.3.9) and, to a lesser extent, for Model (4.3.10), we notice that $\hat{\sigma}^2_{\gamma_1}$ (but not $\hat{\sigma}^2_{\gamma_2}$) is larger with a wider CI. This is reflected in smaller values of the correlation $\rho_{\gamma_1\gamma_2}$ and longer CI, as can be observed in Figure 4.2 (panel c). We refer to Figures 4.1 – 4.4 for a graphical representation of the results. Figure 4.1 displays the observed regional profiles and the fitted profiles from the basic model. We notice that Model (4.3.1) in Figure 4.1 (panel b) shows parallel regional profiles, because only the cluster-specific random effects are specified, thus failing to describe all the excess variability within the data. This is instead accomplished by the best additive and multiplicative models (according to PED and

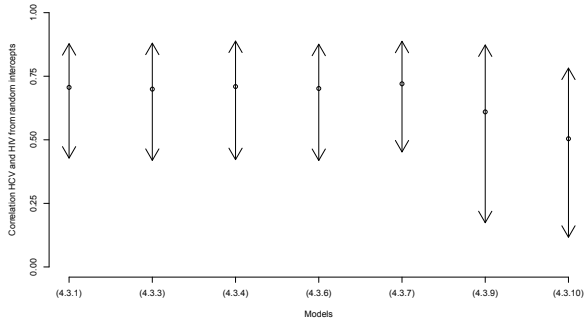
the difference in posterior deviances), plotted in Figure 4.3. What is most striking from the graphical representation is that the fitted regional profiles from the pair of additive models look very similar, as well as it happens with the pair of multiplicative models. Finally, Figure 4.4 displays the marginal prevalence of HCV and HIV infection with the respective 95% CI for the basic model as well as the overdispersion models with additive and multiplicative random effects. For comparison, we plotted also the national prevalence per year, obtained by pooling together all the regional results. We notice that the five estimated prevalence profiles are fairly close to the national observed prevalence. What really distinguishes the marginal prevalence estimates from each other is their credible intervals, which account for all the variability shown by the regional profiles.



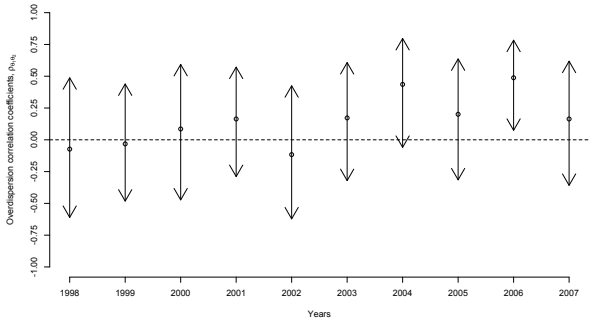
(a) Variance γ_{i1} per model



(b) Variance γ_{i2} per model

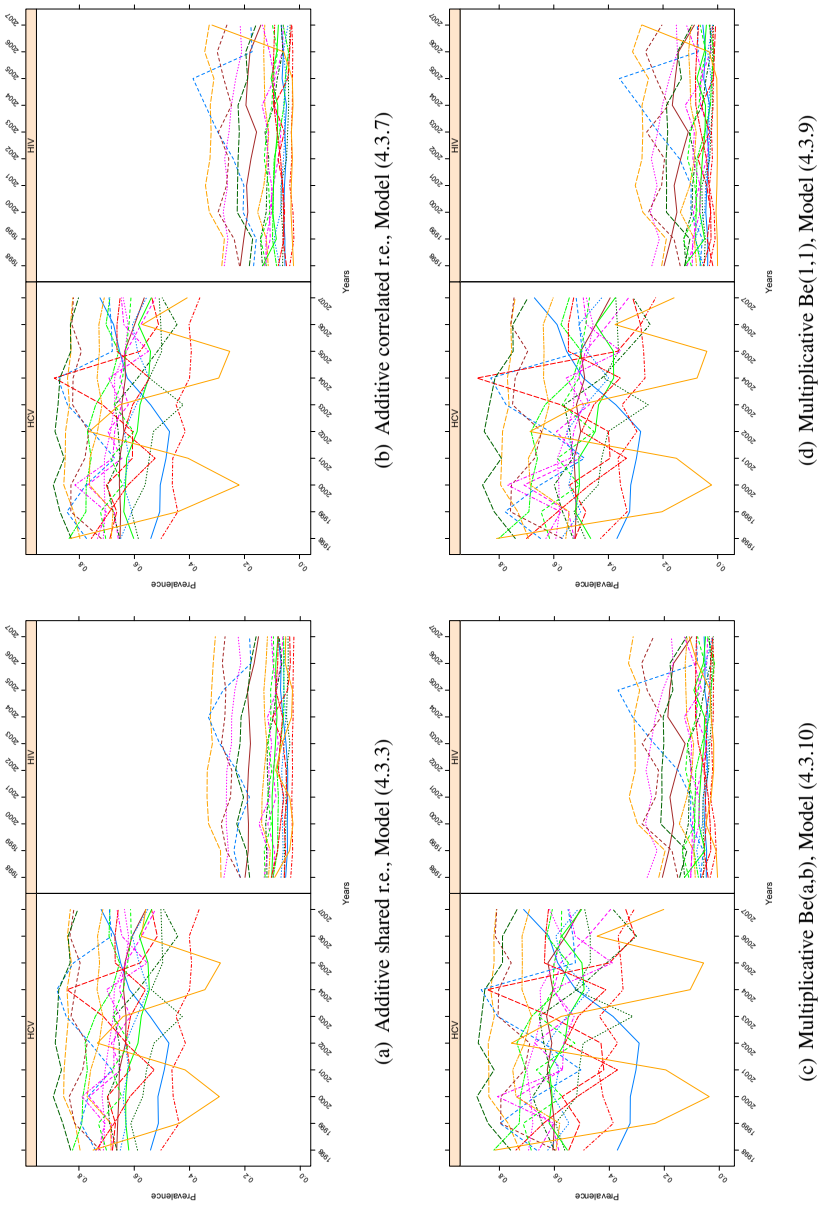


(c) Correlation $\rho_{\gamma_1\gamma_2}$ per model



(d) Correlation $\rho_{\theta_1\theta_2}$ from Model (4.3.7)

Figure 4.2: Posterior means and respective 95% CI of the infection-specific variances (panel a and b) and correlation (panel c) of random intercepts γ_{ik} per model, and of the time-dependent correlation (panel d) on the scale of overdispersion parameters θ_{ijk} from Model (4.3.7).



(a) Additive shared r.e., Model (4.3.3)

(b) Additive correlated r.e., Model (4.3.7)

(c) Multiplicative Be(a,b), Model (4.3.10)

(d) Multiplicative Be(1,1), Model (4.3.9)

Figure 4.3: Individual fitted prevalence profiles of HCV and HIV infections for the 20 Italian regions between 1998 and 2007, resulting from the best models according to PED (panel a and c) and according to the difference in posterior deviances (panel b and d).

4.5 Discussion

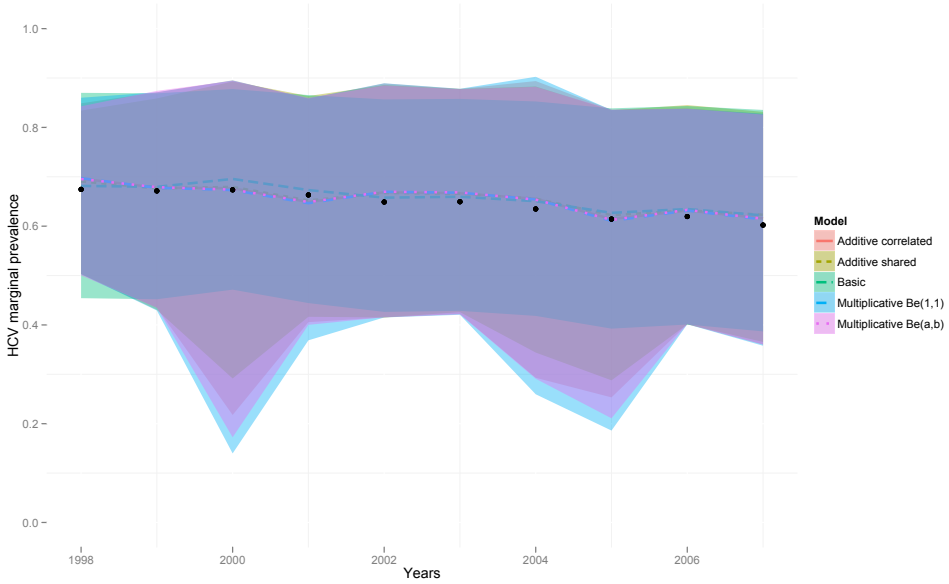
In this chapter, we addressed the issue of dealing simultaneously with clustering and overdispersion for binomial data, using a series of joint GLMMs to model HCV and HIV infection prevalence data, repeatedly measured from 1998 to 2007. This objective has been achieved using two separate sets of random effects, one to induce association among the within-region measurements and between the infections, the other to account for the extra binomial variability in the data. For the latter set of random effects, two different settings have been considered: (1) the overdispersion parameters are included additively into the model, on the same scale of the linear predictor; (2) the overdispersion parameters are included multiplicatively in the model, correcting the prevalence in order to encompass the additional variability. All fitted combined models outperform the so-called basic joint model (Del Fava et al., 2011a), which accommodates only the clustered nature of the data. This result was expected since the random-intercept model cannot capture the complete association structure in the data.

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

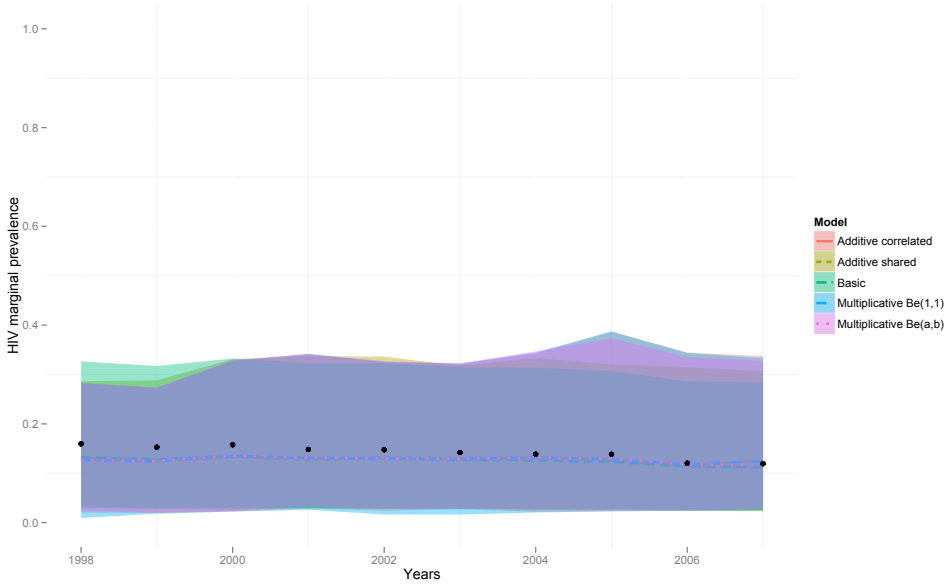
Several models have been proposed for each setting according to the dependence structure among the overdispersion random effects and therefore on their covariance matrix $\boldsymbol{D}_\theta$. In both settings, we parameterize both the correlation between HCV and HIV infection over time (via the joint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

Notwithstanding the high degree of complexity of the GLMM model-specification, the Bayesian paradigm successfully manages to handle all of it. Differently from the frequentist approach, we do not have to worry about simplifying assumptions in order to specify the distribution of the random effects. For instance, different prior distributions for the random effects can be tried and tested within a sensitivity analysis. Moreover, the MCMC method provides us with the full posterior distribution of each parameter of interest and summary measures as the posterior mean and the credible intervals are easily calculated.

On the other hand, what is more challenging within the Bayesian framework is model selection. The DIC (Spiegelhalter et al., 2002) cannot be considered a valid option when the condition $p_D << n$ does not hold (Plummer, 2008), because it tends to under-penalize more complex models. As concerns the other selection criteria here used, i.e., the PED (Plummer, 2008) and the difference in posterior deviances (Aitkin et al., 2009; Aitkin, 2010), they may lead to different conclusions, even though they both claim to



(a) HCV



(b) HIV

Figure 4.4: Marginal prevalence π_{jk} for HCV (panel a) and HIV (panel b) with 95% CI, obtained by averaging out all over the regions the prevalence per year j and infection k, from the basic model (green line), the additive correlated model with time-dependent correlation coefficients (red line), the additive shared model (golden line), the multiplicative $Be(1,1)$ model (blue line), and the multiplicative $Be(a,b)$ model (pink line). The black dots stand for the national observed prevalence, obtained pooling together all the regional prevalences per each year.

penalize the more excessively parameterized models. To our experience, however, it seems that PED tends to favors less parameterized models, while the difference in posterior deviances selects more complex models. Fortunately, in our case, the results provided by the different best models look quite similar, as it appears from the graphical representation of the estimated prevalence.

Chapter 5

Joint Modeling of HCV and HIV Co-Infection among Injecting Drug Users in Italy and Spain Using Individual Cross-Sectional Data

5.1 Introduction

In Chapters 3 and 4, we analyzed aggregated prevalence data from Italy to investigate the association at population level between HCV and HIV infection prevalence. In this chapter we focus on joint modeling of HCV and HIV co-infection using individual data. The models that we will present allow us to estimate the influence of risk factors on the prevalence (and on the their association), taking into account possible individual heterogeneity.

Italy and Spain are among the Western European countries with the highest prevalence for HCV and HIV infection, both in the general population and among the IDUs. In the general population, HCV infection prevalence is respectively equal to 0.5% and 2% (WHO, 1999), while among injecting drug users (IDUs) it is equal to 59.2% and 59.1%-73.3% (EMCDDA, 2010), respectively. Mathëi et al. (2002) report similar figures for the prevalence of HCV infection in IDUs. The prevalence of HIV infection in the general population in Italy is between 0.1% and 0.5%, while the prevalence in Spain ranges between 0.5% and 1% (UNAIDS, 2008); in contrast, available data suggest a prevalence of HIV infection among IDUs of 11.7% and 34.5% (EMCDDA, 2010) in Italy and Spain, respectively.

In recent years, many epidemiological studies focussed on the co-infection with HCV and HIV in IDUs (Hagan and Des Jarlais, 2000; Alter, 2006). Kretzschmar and Wiessing (2008) remarked on the

importance of using mathematical and statistical modeling in order to get a deeper insight in the epidemiological processes of this co-infection. A number of studies investigated the topic by analyzing data from cross-sectional surveys either in the general population (Sherman et al., 2002; Roca et al., 2003) or among IDUs from countries with different prevalence settings, usually concentrating on the effects of sexual and drug-related risk factors on the prevalence of both infections (Miller et al., 2004; Rhodes et al., 2005; Zocratto et al., 2006; Bollepalli et al., 2007; Dumchev et al., 2009; Vickerman et al., 2009; Camoni et al., 2010; Rahimi-Movaghar et al., 2010; Vickerman et al., 2010). In this chapter we focus on the association between HCV and HIV infections in the IDU population, using individual bivariate binomial data (the HCV and HIV infection statuses). Hence, we use statistical models that can take into account the clustered nature of these binomial data, and model the association either as a constant or as a function of known drug-related behavioral risk factors, such as the length of the injecting career, the age at first injection, the frequency of current injecting and the sharing of syringes (Mathëi et al., 2006). Our focus on the association between HCV and HIV infections is motivated by two reasons. First, we are interested in the association itself in order to understand which risk factors can enhance or diminish it. Second, we want to estimate the degree of individual heterogeneity in the acquisition of the infections. The idea is that each IDU carries a different risk of being infected, due, for instance, to a stronger or weaker genetical resistance to infections or due to residual confounding of unobserved behavioral risk factors that might work similarly for HIV and HCV infections or differentially between both infections. A way to obtain evidence of this individual heterogeneity is by analyzing data on co-infection with multiple viruses. If these infections are strongly associated within subjects, this provides evidence of heterogeneity in the way different subjects become infected, even though we do not identify the source of this heterogeneity (Coutinho et al., 1999; Farrington et al., 2001). To achieve this purpose, we consider two families of models. First, we use marginal models for binary data, namely, the alternating logistic regression (Carey et al., 1993), the bivariate Dale model (Dale, 1986), and the bivariate probit model (Ashford and Sowden, 1970; Morimune, 1979; Molenberghs and Verbeke, 2005), which allow for direct estimation of association measures between HCV and HIV infections (Hens et al., 2008). Second, we consider the family of mixed-effects models, i.e., the generalized linear mixed models (GLMM, McCulloch and Searle, 2001; Agresti, 2002; Molenberghs and Verbeke, 2005) and the Gamma frailty models (Farrington et al., 2001; Sutton et al., 2006, 2008; Hens et al., 2009c), which are conditional on random effects that capture the individual behavior in the acquisition of infections. With the latter models, we do not estimate directly the association, but we study the effect of the individual heterogeneity in becoming infected by testing the significance of the variance of the individual-specific random effects. All these statistical methods are applied to two serological cross-sectional samples of IDUs with exposure to several behavioral risk factors from Italy and Spain.

The chapter is organized in the following way. Section 5.2 presents the data used in the analysis and descriptive statistics are used to assess the degree of the co-infection. In Section 5.3 we present a joint model for HCV and HIV infections using the length of the injecting career as exposure time. The different modeling approaches mentioned above are used to model the association between the infections

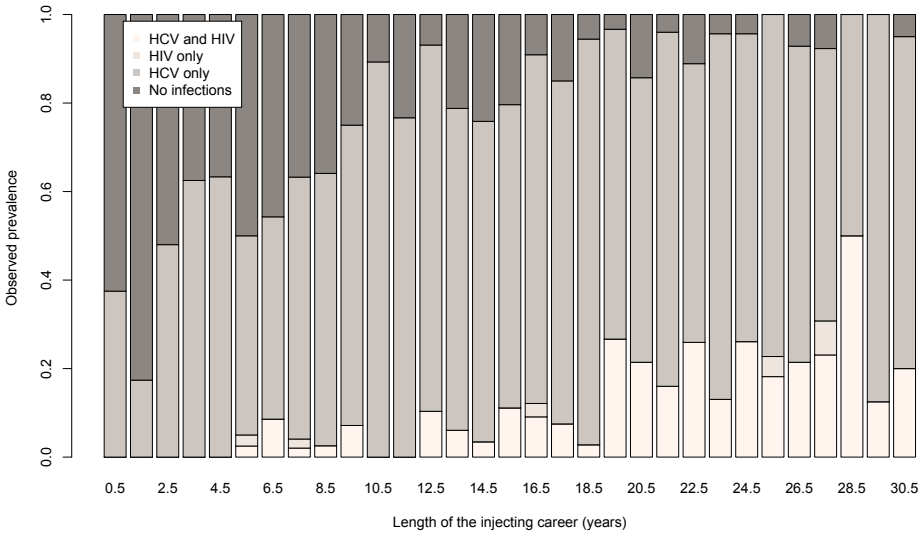
and are discussed in Section 5.3.1 and Section 5.3.2. In Section 5.4, the proposed methods are applied to two cross-sectional samples from Italy and Spain. Finally, in Section 5.5 we discuss the results.

5.2 Data

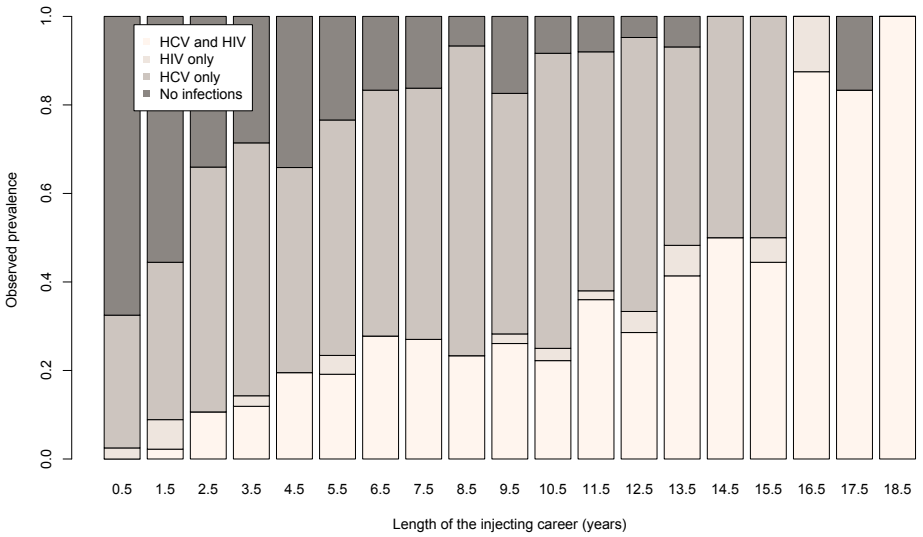
The cross-sectional data used in this analysis consist of two seroprevalence samples of IDUs from Italy ($N = 856$) and Spain ($N = 589$).

The data from Italy come from a cross-sectional survey carried out in 2005 among IDUs and non-IDUs at public drug-treatment centers (DTCs) in 14 Italian regions (Camoni et al., 2010). The original survey included 1330 persons, of whom 1009 injected at least once in the last year. As concerns the serological status for HCV and HIV infections, the persons were not tested directly during the survey, but rather the information was taken from their clinical records in the DTC, where subjects' blood specimens were tested for antibodies against HCV and HIV. Since in this study we focus on the association between HCV and HIV infection, the following inclusion criteria were applied: (1) all subjects have serological results for both infections, (2) all subjects have information about their length of injecting career, (i.e., the age at first injection should be reported). Using these inclusion criteria, we obtained a final dataset of 856 IDUs. Note that in this dataset information about behavioral risk factors in Italy is available for the length of the injecting career and for the age at first injection.

The data from Spain were collected in a cohort study on 961 subjects, the "Itinere Project", which was carried out between April 2001 and December 2003 in the metropolitan areas of Madrid, Barcelona, and Seville, among both street-recruited current injecting and non-injecting heroin users. Subjects from DTCs were excluded (Barrio et al., 2007). All the subjects in the sample were administered a questionnaire for the collection of socio-demographic, sexual behavior and drug use data; moreover, a blood specimen was obtained and tested for antibodies against HBV, HCV, and HIV. The application of the inclusion criteria mentioned above for Italy lead to a total sample of 589 IDUs with complete information about their drug-related behavioral risk factors: the length of the injecting career, the age at first injection, the sharing of syringes, and the frequency of current injecting. The latter risk factor consists of four categories, namely, "every day", "1–6 days per week", "less than weekly", and "never"; the last category refers to those drug users who have not injected in the last month. For the sharing of syringes we present the analysis on a "ever/never" basis. This implies that we assume time-homogeneity for this behavioral risk factor, which might be not the case. For this reason, in the supplementary material file (Section 5.2), we also present the results for sharing syringes based on the "current status" information (last month before the interview) for a comparison of the results. The bar plots in Figure 5.1 present the four observed joint prevalences of HCV and HIV infections depending on the length of the injecting career, revealing a clear pattern of association between the prevalence and the duration of injecting, as already reported in the literature (Pallás et al., 1999; Hagan and Des Jarlais, 2000; Mathëi et al., 2006): the probability of either having a single infection or being co-infected with both viruses increases quickly with the length of injecting career, most of all in the first 5 years, while the probability of still being susceptible decreases.



(a) Italy



(b) Spain

Figure 5.1: Bar plots with the observed joint probabilities for HCV and HIV infections among IDUs in Italy and Spain depending on the length of the injecting career.

We also notice that the prevalence of HCV infection is far higher than the prevalence of HIV infection and almost all of those who are HIV-infected are HCV-infected as well (Rhodes et al., 2005).

Table 5.1 shows descriptive statistics for the prevalence of the two infections, for their co-infection and for the risk factors as well as association measures between HCV and HIV infections. The proportion of males is 0.73 in Italy and 0.74 in Spain. As regards the prevalence estimates, they are fairly comparable to the figures in the literature, considering the high variability of the estimates (Aceijas and Rhodes, 2007; Mathers et al., 2008; EMCDDA, 2010): we notice that, although the median length of the injecting career in Italy is higher than in Spain (14 versus 6 years, respectively), which may be due to the exclusion in the Spanish study of IDUs in drug treatment, the overall prevalence of HIV infection in Spain is higher (26%) than in Italy (10%), indicating higher incidence in Spain and/or higher mortality in Italy. The estimated odds ratios for co-infection are equal to 5.5 with 95% confidence interval (CI) 2.2–17.5 for Italy and to 4.2 (95% CI: 2.4–8) for Spain. Both tetrachoric correlations (Molenberghs and Verbeke, 2005) are equal to 0.4 and highly significant. Similarly, the χ^2-test statistics are found to be significant for both countries, indicating that there is significant association between HCV and HIV infections.

Table 5.1: The prevalence of HCV infection, HIV infection, and their co-infection in the sample (with 95% CI). Descriptive statistics for the behavioral risk factors and association measures for the co-infection with HCV and HIV. For more information about the tetrachoric correlation, see Section 3 in the supplementary material file.

		Italy	Spain
Prev. HCV		77% (CI: 74%–80%)	74% (CI: 70%–77%)
Prev. HIV		10% (CI: 8%–12%)	26% (CI: 22%–30%)
Prev. co-inf. HCV–HIV		10% (CI: 8%–12%)	23% (CI: 20%–27%)
Median age first inj. ($Q_{0.25}, Q_{0.75}$)		20 (18,24)	19 (17,22)
Median age at interview ($Q_{0.25}, Q_{0.75}$)		35 (30,41)	27 (24,29)
Median leng. inj. car. ($Q_{0.25}, Q_{0.75}$)		14 (7,19)	6 (3,10)
Ever Shared Syringes	*Yes*	-	41%
	No	-	59%
Frequency of current injecting	*Every day*	-	27%
	1-6 days/week	-	35%
	Less than weekly	-	21%
	Never	-	17%
Pearson's χ^2 for co-infection		16.3 ($P < 0.0001$)	27.6 ($P < 0.0001$)
Odds ratio for co-infection, ψ		5.5 (CI: 2.2–17.5)	4.2 (CI: 2.4–8)
Tetrachoric corr. for co-infection, ρ		0.4 ($P < 0.0001$)	0.4 ($P < 0.0001$)

5.3 Statistical Methods

The data collected in the two surveys are bivariate. Each IDU is a cluster with two measurements corresponding to the serological status of HCV and HIV infection. Let $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2})$, $i = 1, \ldots, n$, be a vector of indicator variables representing the serological status for HCV and HIV infections, respectively, for the ith subject. Then, let d_i be the length of the injecting career for the ith subject, calculated as the difference between the age at interview and the age at first injection.

In what follows we present different statistical models for clustered data, focussing on the estimation of the association between the two infections. For this purpose, we fitted both marginal and mixed effects models. In particular, for the marginal models we considered the alternating logistic regression model (Carey et al., 1993), the bivariate Dale model (Dale, 1986), and the bivariate probit model (Ashford and Sowden, 1970; Morimune, 1979; Molenberghs and Verbeke, 2005). The mixed effects models used are the GLMM for binary data (McCulloch and Searle, 2001; Agresti, 2002; Molenberghs and Verbeke, 2005) and the shared Gamma frailty model (Farrington et al., 2001; Sutton et al., 2006, 2008; Hens et al., 2009c).

5.3.1 Modeling the Association Between HCV and HIV Infection Using Marginal Models

5.3.1.1 Alternating Logistic Regression (ALR) and Bivariate Dale Model (BDM)

Carey et al. (1993) argued that the objectives of a multivariate analysis of binary data should include (1) the description of the dependency of each binary response on some covariates and (2) the characterization of the degree of association between pairs of responses and the dependence of this association on covariates. In this sense, the first two marginal models we consider, the ALR (Carey et al., 1993) and the BDM (Dale, 1986), model the association between HCV and HIV infections using the odds ratio, the former using a quasi-likelihood approach, the latter using full likelihood. Let ψ_i be the pairwise odds ratio between responses Y_{i1} and Y_{i2} and let the log odds ratio be defined as

$$\log(\psi_i) = \log\left[\frac{P(Y_{i1}=1,Y_{i2}=1)P(Y_{i1}=0,Y_{i2}=0)}{P(Y_{i1}=1,Y_{i2}=0)P(Y_{i1}=0,Y_{i2}=1)}\right] = \log\left[\frac{\pi_{11}\pi_{00}}{\pi_{10}\pi_{01}}\right]. \tag{5.3.1}$$

Here, $P(Y_{i1}=1,Y_{i2}=1)=\pi_{11}$ is the joint probability to be infected with both viruses. In the simplest case, $\log(\psi_i)=c$, i.e., the pairwise log odds ratio is constant. However, the marginal models considered allow to model the dependence of the association upon other behavioral risk factors for HCV and HIV infections. For example, let X_i be a binary behavioral risk factor, e.g., ever sharing syringes. The log odds ratio can be modeled as $\log(\psi_i)=\alpha_0+\alpha_1 X_i$. The case in which the null hypothesis $H_0:\alpha_1=0$ cannot be rejected is consistent with the idea that the association between HCV and HIV infection is the same across the levels of X_i. Furthermore, if X_i is continuous, e.g., the age at first injection, a more flexible model such as a smooth differentiable function $h(\cdot)$ can be used to model the log odds ratio, $\log(\psi_i)=h(X_i)$.

Since the prevalence of HCV and HIV infections depends on the length of the injecting career d_i as well as on other risk factors X_i, we follow the modeling approach of Mathëi et al. (2006) and fit a parametric model for the prevalence:

$$\begin{cases} g(P(Y_{i1}=1)) &= \beta_{01}+\beta_{11}\log(d_i)+\gamma X_i, \\ g(P(Y_{i2}=1)) &= \beta_{02}+\beta_{12}\log(d_i)+\gamma X_i, \\ \log(\psi_i) &= \alpha X_i. \end{cases} \tag{5.3.2}$$

For the marginal prevalence models, $g(\cdot)$ is a link function, the parameters β_{0j} and β_{1j} ($j = 1,2$) are infection-specific intercepts and slopes, respectively, X_i is a covariate representing a behavioral risk factor and γ is its coefficient. Concerning the odds ratio model, the parameter α is the coefficient of X_i, which can be one the following behavioral risk factors: length of the injecting career and age at first injection (continuous), ever sharing syringes (binary), and frequency of current injecting (categorical), the latter two available only for Spain.

Three standard link functions for binary data were considered: the logit link, $\log(\pi/(1-\pi))$, which implies a log-logistic distribution for the time spent in the susceptible class, when the time covariate is expressed on a log scale; the probit link, $\Phi^{-1}(\pi)$, which implies a log-normal distribution for the time spent in the susceptible class; the complementary log-log link, $\log(-\log(1-\pi))$, which implies a Weibull distribution for the time spent in the susceptible class.

5.3.1.2 Bivariate Probit Model (BPM)

Another possibility to model the association between responses within the marginal models family is to use a BPM (Ashford and Sowden, 1970; Morimune, 1979; Molenberghs and Verbeke, 2005). We assume that the current status of the infections $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2})$ is related to two latent variables $\boldsymbol{W}_i = (W_{i1}, W_{i2})$ that represent the unknown individual antibody levels of HCV and HIV infections. It is further assumed that the antibody levels of HCV and HIV infections follow a bivariate normal distribution given by

$$\begin{pmatrix} W_{i1} \\ W_{i2} \end{pmatrix} \sim MVN \left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma \right], \tag{5.3.3}$$

where Σ is a 2x2 covariance matrix given by

$$\Sigma = \begin{pmatrix} \sigma^2_{W_1} & \rho\,\sigma_{W_1}\sigma_{W_2} \\ \rho\,\sigma_{W_1}\sigma_{W_2} & \sigma^2_{W_2} \end{pmatrix}. \tag{5.3.4}$$

The serological status of each individual for the jth disease, $j = 1,2$, is defined by

$$y_{ij} = \begin{cases} 1 & W_{ij} > \tau_j \quad \text{seropositive,} \\ 0 & W_{ij} \leq \tau_j \quad \text{seronegative.} \end{cases} \tag{5.3.5}$$

Here, τ_j is a infection-specific cut-off point used to classify individuals either as seropositive or as seronegative. The BPM allows to model the correlation ρ between the latent normal variables $\boldsymbol{W}_i$ underlying the binary data $\boldsymbol{Y}_i$. The model can be formulated using a probit link function for each of the two marginal probabilities, while the correlation is fitted with the "rhobit" link (Yee, 2010), which is

commonly used for parameters that lie between -1 and 1,

$$\begin{cases} \Phi^{-1}(P(Y_{i1}=1)) &= \beta_{01}+\beta_{11}\log(d_i)+\gamma X_i, \\ \Phi^{-1}(P(Y_{i2}=1)) &= \beta_{02}+\beta_{12}\log(d_i)+\gamma X_i, \\ \log\frac{1+\rho_i}{1-\rho_i} &= \alpha X_i. \end{cases} \tag{5.3.6}$$

Similarly to the odds ratio ψ in the ALR and BDM, ρ_i can be either constant or depending upon some behavioral risk factors X_i.

5.3.2 Modeling the Association Between HCV and HIV Infections Using Mixed-Effects Models

The marginal models discussed in the previous sections were focussed on the estimation of odds ratio and correlation as association measures for co-infection with HCV and HIV. In this section, we discuss a second modeling approach whereby we concentrate on individual heterogeneity in the acquisition of infections. Two types of mixed-effects models are used: (1) a GLMM (McCulloch and Searle, 2001; Agresti, 2002; Molenberghs and Verbeke, 2005) and (2) a shared Gamma frailty model (Farrington et al., 2001; Sutton et al., 2006, 2008; Hens et al., 2009c).

5.3.2.1 Generalized Linear Mixed Models (GLMM)

A GLMM is an alternative modeling approach for multivariate data that captures individual heterogeneity by conditioning on subject-specific random effects in the model. For this reason, we refer to it as a conditional model. We assume that infections may be associated at individual level and that this association could be explained by individual random effects. The variance of these random effects accounts for the variability not explained by the covariates in the model, and indicates that there are differences among individuals, giving consequently evidence of heterogeneity in the transmission. For the analysis presented here, we used the so-called shared random intercept model. Let b_i be a subject-specific random intercept assumed to be normally distributed, $b_i \sim \mathrm{N}(0,\sigma_b^2)$. Conditional on b_i, we model the prevalence of HCV and HIV infections in the following way:

$$\begin{cases} g(P(Y_{i1}=1|b_i)) &= \beta_{01}+\beta_{11}\log(d_i)+b_i, \\ g(P(Y_{i2}=1|b_i)) &= \beta_{02}+\beta_{12}\log(d_i)+b_i. \end{cases} \tag{5.3.7}$$

The parameter of primary interest in the GLMM specified in (5.3.9) is the variance σ_b^2 of the shared random intercepts b_i. Whenever the null hypothesis $H_0 : \sigma_b^2 = 0$ is rejected, it means that there is significant heterogeneity among the subject-specific intercepts, and this means that infections are associated at individual level. If the null hypothesis cannot be rejected, then the data are consistent with the infections being acquired independently: IDUs could still vary in their risks for HCV and HIV infections, but not jointly. Note that this model implies that, for a given value of a covariate, there is no difference among

individuals in their risk for HCV or HIV infections. In this case, the population-averaged prevalence models are more appropriate.

In order to compare the estimated population-averaged prevalence obtained from the GLMM with the prevalence obtained from the marginal models presented in Section 3.1, we need to marginalize the mixed-effects model. The marginalization consists in the integration of the random effects based on numerical integration techniques or numerical averaging. However, a more practical approach is to generate randomly a large number M of realized values for the random effects b_i from $N(0,\sigma_b^2)$, where for σ_b^2 we plug in the estimated variance of the random effects $\hat{\sigma}_b^2$ (Verbeke and Molenberghs, 2000). Then, the marginalized prevalences of HCV and HIV infections, given an injecting career of length d_i, the estimated coefficients $\hat{\boldsymbol{\beta}} = (\beta_{0j},\beta_{1j})$, $j = 1,2$, and the realized values of $\hat{b}_i$, are provided by the following formulae:

$$\begin{cases} P(Y_{i1}=1) &= (1/M)\sum_{i=1}^{M} P(Y_{i1}=1|d_i;\hat{\boldsymbol{\beta}},\hat{\boldsymbol{b}}_i), \\ P(Y_{i2}=1) &= (1/M)\sum_{i=1}^{M} P(Y_{i2}=1|d_i;\hat{\boldsymbol{\beta}},\hat{\boldsymbol{b}}_i). \end{cases} \tag{5.3.8}$$

5.3.2.2 Shared Gamma Frailty Models

A second conditional model, proposed by Farrington et al. (2001), assumes that every IDU is infected differently from the others, that is, the force of infection (FOI), $\lambda(d_i,b_i)$, depends on subject-specific random effects b_i, or "frailties", which represent to what extent IDUs carry a higher or lower risk of infection (Coutinho et al., 1999), and on the length of the injecting career d_i. Given the assumption of proportional hazards, the FOI can be written as $b_i\lambda_0(d_i)$, where λ_0 is the baseline hazard. Therefore, the susceptible proportion for the infection j is given by

$$S_j(d_i|b_i) = \exp\left(-b_i\int_0^{d_i}\lambda_{0j}(t)dt\right), \qquad j = 1,2. \tag{5.3.9}$$

Similar to the GLMM, the use of multiple sera allows the estimation of the distribution of these frailties, providing us with a measure of the individual heterogeneity in acquiring infections. We assume b_i to be shared between the two infections and to have Gamma distribution, $\Gamma(\theta,1/\theta)$, where the heterogeneity parameter θ measures the association between the two infections. Using shared frailties, we assume perfect correlation among them at the level of the linear predictors and common variance. The frailties b_i have expected value $E(b_i) = 1$ and variance $Var(b_i) = 1/\theta$: the smaller θ, the larger the heterogeneity, and the more different the way individuals acquire the infections. Moreover, the variance of the frailty, $1/\theta$, is related to the cross-ratio function – a measure of local dependence, evaluated at a single time point (Clayton, 1978) –, and hence to Kendall's τ (Farrington et al., 2012).

For the current version of the model, we follow Sutton et al. (2006, 2008) and Hens et al. (2009c). Let $\pi_{00}(d_i)$ be the probability that individual i with a length of injecting career of d_i years has not been infected with either virus; let $\pi_{10}(d_i)$ be the probability that individual i with a length of injecting career of d_i years has been infected with HCV, but not with HIV; similarly we define the joint probabilities $\pi_{01}(d_i)$ and $\pi_{11}(d_i)$. The model for the probability of being still susceptible to both diseases, derived

from (5.3.11) by applying the Laplace transform (Sutton et al., 2006), is given by

$$\pi_{00}(d_i) = \left[S_1^{-1/\theta} + S_2^{-1/\theta} - 1\right]^{-\theta}. \tag{5.3.10}$$

Reparameterizing the joint probability in terms of the cumulative FOI, $\Lambda_j(d_i) = \int_0^{d_i} \lambda_{0j}(t)dt$, we obtain the following set of equations for the four joint probabilities:

$$\begin{cases} \pi_{00}(d_i) &= \left[\exp\left(\frac{\Lambda_1(d_i)}{\theta}\right) + \exp\left(\frac{\Lambda_2(d_i)}{\theta}\right) - 1\right]^{-\theta}, \\ \pi_{10}(d_i) &= \exp\left(\frac{\Lambda_2(d_i)}{\theta}\right) - \pi_{00}(d_i), \\ \pi_{01}(d_i) &= \exp\left(\frac{\Lambda_1(d_i)}{\theta}\right) - \pi_{00}(d_i), \\ \pi_{11}(d_i) &= 1 - \pi_{10}(d_i) - \pi_{01}(d_i) - \pi_{00}(d_i). \end{cases} \tag{5.3.11}$$

To solve these equations, it is necessary to assume a model for the FOI, for instance, a log-logistic, a log-normal, or a Weibull model. The unknown parameters can be estimated by maximizing the log-likelihood of the observations,

$$L = \sum_d \{n_{00d}\log[\pi_{00}(d)] + n_{10d}\log[\pi_{10}(d)] + n_{01d}\log[\pi_{01}(d)] + n_{11d}\log[\pi_{11}(d)]\}. \tag{5.3.12}$$

The likelihood ratio test (LRT) can be used to check whether the frailty b_i can be included in the model or not. Similarly to GLMMs, in order to compare the prevalence from the shared Gamma frailty model with the prevalences from the previous marginal models, we can compute the marginal prevalence of the infections using the estimated joint probabilities: for instance, the marginal prevalence of HCV infection, $P(Y_{i1} = 1|d_i)$, is given by the sum of the probability of being only HCV-infected and the probability of being co-infected, $\pi_{10}(d_i) + \pi_{11}(d_i)$.

5.4 Results

5.4.1 Constant Measures for Association

In this section, we present the results obtained from the models wherein the association measures were assumed to be constant. The initial mean structure of the models included infection-specific intercepts and coefficients for the logarithm of the length of injecting career, $\log(d_i)$, without any other risk factors. Model selection was performed with the Akaike's information criterion (AIC) and the Bayesian information criterion (BIC), except for the ALR, for which the Quasi-likelihood information criterion (QIC, Pan, 2001) was used. Based on the information criteria, the best mean structure and the best link function were selected. Furthermore, we used the LRT for the random-effects models to assess whether the random intercept is needed or not. Regarding models' computation, the ALR was fitted in SAS with the procedure GENMOD, while the BDM and the BPM were fitted in R with package `VGAM` (Yee, 2010); the GLMM

was fitted in SAS with procedure GLIMMIX, whereas the shared Gamma frailty model was fitted in R.

Table 5.2 presents the estimates of the association measures for the models with the smallest information criteria. For all the models but the GLMM, the best mean structure remained the starting one, while for the GLMM the mean structure was reduced by constraining the coefficient of $\log(d_i)$ to be shared between the infections. Besides, since the AIC and the BIC for the GLMMs with logit and probit link functions were very close to each other, we showed the estimates of the variance of the random effects for both models. For each prevalence model, results indicate that the association between the infections is significant, even though the confidence interval for the variance parameter of the shared Gamma frailty model in Italy is quite close to zero. We notice that odds ratios obtained from the ALR and BDM are similar in both countries. For GLMM, the model with smallest AIC and BIC for Italy is the model with probit link, while for Spain it is the model with logit link. The variance of the random effects was found to be significant for both countries (P-values of LRT for the inclusion of random effects are equal to 0.014 and 0.0014 for Italy and Spain, respectively). In the same way, the heterogeneity parameters for the shared Gamma frailty model were found to be significant ($P-$values for the LRT for the inclusion of random effects are equal to 0.023 and 0.0023 for Italy and Spain, respectively). We notice that, due to scarcity of HIV infection cases in Italy, the 95% profile-likelihood CI for θ is very wide (1.07–176.43).

Table 5.2: Estimates of the association measures and 95% CI. For each model and country, we reported the results for the model with the smallest AIC and BIC (except for the ALR, for which the QIC was used). The 95% CI for BDM and BPM are studentized bootstrap CI, based on $B = 999$ replicates; for ALR, we used asymptotic CI; for GLMM and shared Gamma frailty model, we used a profile-likelihood CI. For GLMM and shared Gamma frailty model, we report the result of LRT for the inclusion of a random intercept, as well. Information criteria are presented in Section 4 of the supplementary material file.

Model	Association	Italy	Spain
ALR	ψ	2.58 (1.02, 6.53)	2.43 (1.35, 4.36)
	link function	cloglog	cloglog
BDM	ψ	2.56 (1.43, 6.68)	2.42 (1.41, 4.30)
	link function	cloglog	cloglog
BPM	ρ	0.23 (0.02, 0.44)	0.26 (0.12, 0.64)
	link function	probit	probit
GLMM	σ_b^2	1.34 (0.23, 3.27)	1.15 (0.32, 2.44)
	LRT	6.53 ($P = 0.0053$)	8.97 ($P = 0.0014$)
	link function	logit	logit
	σ_b^2	0.31 (0.026, 0.78)	0.36 (0.099, 0.74)
	LRT	4.81 ($P = 0.014$)	8.72 ($P = 0.016$)
	link function	probit	probit
Shared Gamma Frailty Model	σ_b^2	0.39 (0.01, 0.94)	0.37 (0.09, 0.73)
	θ	2.59 (1.07, 176.43)	2.70 (1.38, 10.71)
	LRT	3.99 ($P = 0.023$)	8.01 ($P = 0.0023$)
	link function	cloglog	cloglog

5.4.2 Testing for Common Variance for Subject-Specific Random Effects Between Countries

In this section we discuss an extension of the GLMM and the shared Gamma frailty model to the case of two populations. The parameter estimates for σ_b^2 reported in Table 5.2 suggest that the variance of random intercepts in the GLMM and in the shared Gamma frailty model may be equal for Italy (IT) and Spain (ES). Therefore, the two random-effects models were re-fitted for the two countries jointly and the null hypothesis of common variance of the random effects was tested.

For the GLMM, let I_i be an indicator variable which takes value 1 if the IDU is from Italy and 0 otherwise. In order to test the null hypothesis, $H_0 : \sigma_{b,IT}^2 = \sigma_{b,ES}^2$, for $j = 1,2$, we formulated a joint linear predictor given by

$$h_j = \beta_{0j_{IT}} I_i + \beta_{0j_{ES}}(1-I_i) + \beta_{1_{IT}} I_i \log(d_i) + \beta_{1_{ES}}(1-I_i)\log(d_i) + b_{iIT} I_i + b_{iES}(1-I_i). \tag{5.4.13}$$

The parameters $\beta_{0j_{IT}}, \beta_{0j_{ES}}, \beta_{1_{IT}}$ and $\beta_{1_{ES}}$ are infection and country-specific intercepts and slopes, respectively. Under the null hypothesis, the random effects b_{iIT} and b_{iES} are assumed to follow a bivariate normal distribution with independent covariance matrix and common variance σ_b^2. The alternative hypothesis states that the variances of random effects are country-specific. Hence, the joint distribution of the random effects is given by

$$\begin{pmatrix} b_{iIT} \\ b_{iES} \end{pmatrix} \sim MVN\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, D\right]. \tag{5.4.14}$$

where

$$D = \begin{pmatrix} \sigma_b^2 & 0 \\ 0 & \sigma_b^2 \end{pmatrix} \qquad \text{and} \qquad D = \begin{pmatrix} \sigma_{b,IT}^2 & 0 \\ 0 & \sigma_{b,ES}^2 \end{pmatrix}, \tag{5.4.15}$$

under H_0 and H_1, respectively. The joint shared Gamma frailty model is formulated in the following way. Under the null hypothesis, the heterogeneity parameter θ is shared by the countries, while, under the alternative, the parameter is country-specific. Note that, under the two alternatives, the cumulative FOI remains country/disease-specific as before. Hence the probability of being susceptible for both infections under the alternative can be rewritten as

$$\pi_{00}(d_i) = \begin{cases} \left[\exp\left(\frac{\Lambda_{1,IT}(d_i)}{\theta_{IT}}\right) + \exp\left(\frac{\Lambda_{2,IT}(d_i)}{\theta_{IT}}\right) - 1\right]^{-\theta_{IT}} & I_i = 1, \\ \left[\exp\left(\frac{\Lambda_{1,ES}(d_i)}{\theta_{ES}}\right) + \exp\left(\frac{\Lambda_{2,ES}(d_i)}{\theta_{ES}}\right) - 1\right]^{-\theta_{ES}} & I_i = 0. \end{cases} \tag{5.4.16}$$

Note that under the null hypothesis $\theta_{IT} = \theta_{ES} = \theta$. The remaining three probabilities are defined in a similar way.

Table 3 and 4 present the parameter estimates for the variance of random effects, goodness-of-fit measures (AIC and BIC), and the result of LRT for the three link functions, for GLMM and shared Gamma frailty model, respectively. For both random-effects models, regardless of the link function, the

null hypothesis cannot be rejected, therefore we conclude that the level of individual heterogeneity in the countries is the same.

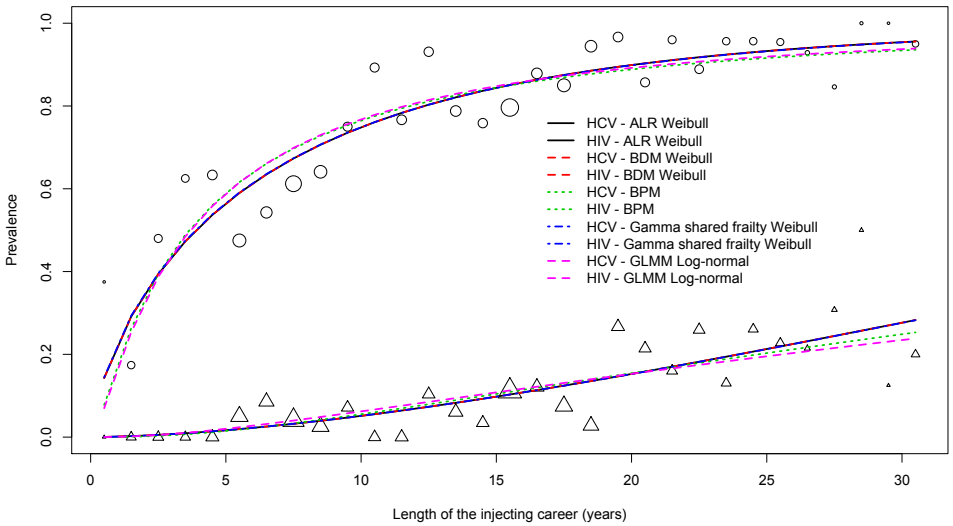
Table 5.3: Parameter estimates for the σ_b^2 and 95% profile-likelihood CI for Italy, for Spain, and for the joint model; LRT for testing for common variance; information criteria for the joint model.

Model	Country	σ_b^2 (95% C.I.)	LRT c.v.	AIC/BIC
	Italy	0.78 (0.17, 1.93)		
Weibull	Spain	0.60 (0.20, 1.24)	0.14	AIC=2490
	IT & ES	0.65 (0.30, 1.18)	($P = 0.71$)	BIC=2527
	Italy	1.39 (0.27, 3.33)		
Log-logistic	Spain	1.15 (0.31, 2.43)	0.07	AIC=2486
	IT & ES	1.24 (0.52, 2.24)	($P = 0.79$)	BIC=2523
	Italy	0.32 (0.03, 0.78)		
Log-normal	Spain	0.36 (0.10, 0.74)	0.03	AIC=2486
	IT & ES	0.34 (0.14, 0.62)	($P = 0.86$)	BIC=2523

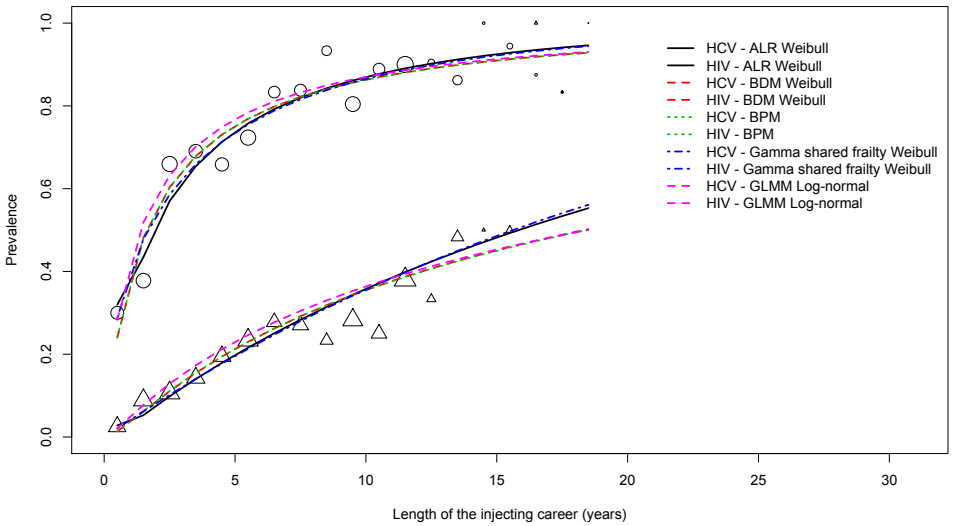
Table 5.4: Parameter estimates of the heterogeneity parameters θ and 95% profile-likelihood CI for the shared Gamma frailty models for Italy, for Spain, and for the joint model; LRT for testing for common variance; information criteria for the joint model.

Model	Country	θ (95% C.I.)	σ_b^2 (95% C.I.)	LRT c.v.	AIC/BIC
	Italy	2.59 (1.07, 176.43)	0.39 (0.01, 0.94)		
Weibull	Spain	2.70 (1.38, 10.71)	0.37 (0.09, 0.73)	0.006	AIC= 2483
	IT & ES	2.68 (1.51, 7.03)	0.37 (0.14, 0.66)	(P = 0.94)	BIC= 2529
	Italy	2.39 (1.01, 52.11)	0.42 (0.02, 0.99)		
Log-logistic	Spain	2.52 (1.31, 8.93)	0.40 (0.11, 0.76)	0.006	AIC=2490
	IT & ES	2.48 (1.43, 6.15)	0.40 (0.16, 0.70)	(P = 0.94)	BIC=2538
	Italy	2.76 (1.15, 102.81)	0.36 (0.01, 0.87)		
Log-normal	Spain	2.99 (1.57, 10.12)	0.33 (0.1, 0.64)	0.013	AIC=2491
	IT & ES	2.92 (1.69, 7.18)	0.34 (0.14, 0.59)	(P = 0.91)	BIC=2539

Figure 5.2 shows the estimated prevalences obtained from the best models for each family. Note that for the random-effects models the marginalized prevalence (discussed in Section 5.3.2) obtained from the models with common variance is presented. As expected, we notice that the proportions of IDUs infected with HCV and HIV increase according to the length of injecting career. In both countries, the prevalence of HCV infection increases very steeply until about 5 years of injecting drug use, while afterwards it levels out. In contrast, the prevalence of HIV infection presents a more linear profile over the years of injecting drug use, with different slopes for Italy and in Spain. Finally, we notice that the prevalence in Spain is generally higher than in Italy for both infections: averaging the prevalences obtained from the five models, we found that, after 10 years of injecting, in Italy the estimated HCV infection prevalence equals 77% and the estimated HIV infection prevalence equals 6%, whereas in Spain the respective values are 87% and 36%.



(a) Italy



(b) Spain

Figure 5.2: Observed (points) and estimated (lines) prevalence for HCV infection (upper curve) and for HIV infection (lower curve) for Italy and Spain. The fitted models are the ALR (black solid lines), the BDM (red dashed lines), the BPM (green dotted lines), the shared Gamma frailty model with common variance (blue dash-and-dot lines), and the GLMM with common variance (purple dashed lines). The area of the points is proportional to the sample size per length of the injecting career group.

5.4.3 Modeling the Association between HCV and HIV Infections as Function of Behavioral Risk Factors Using Marginal Models

We discuss now the results of the marginal models, when we let the association depend on the risk factors. The initial mean structure of the prevalence models include infection-specific intercepts and slopes for the logarithm of the length of injecting career, $\log(d_i)$, as well as infection-specific coefficients for the risk factors used in turn for the association model. Then, the mean structure is simplified based on the Wald's test for coefficients.

We first examine the results for the continuous behavioral risk factors, namely, the length of injecting career and the age at first injection. Note that this information is available for both Italy and Spain. Figure 5.3 shows the odds ratios ψ from ALR and BDM, while Figure 5.4 shows the correlation coefficients ρ obtained from BPM. The 95% pointwise CI for the odds ratios from ALR are asymptotic, while the intervals for the odds ratios from BDM and for the correlations from BPM are 95% studentized bootstrap CI (Davison and Hinkley, 1997), based on $B = 999$ bootstrap replicates.

Table 5.5: Parameter estimates for the association measures (OR for ALR and BDM, ρ for BPM) across the levels of the behavioral risk factors. For ALR we report 95% asymptotic CI, whereas for BDM and BPM we report 95% bootstrap studentized CI. The results presented in the table have been obtained from a complete case analysis.

Risk factor		ALR (Log-logistic)	BDM (Weibull)	BPM (Log-normal)
Ever Shared Syringes	*Yes*	0.92 (0.38, 2.22)	0.93 (0.37, 2.10)	-0.031 (-0.36, 0.26)
	No	3.69 (1.67, 8.16)	3.76 (1.16, 7.53)	0.37 (0.13, 0.55)
Frequency of current injecting	*Every day*	2.21 (0.72, 6.76)	2.13 (1.06, 5.21)	0.24 (-0.14, 0.55)
	1-6 days/week	1.64 (0.68, 3.95)	1.63 (0.53, 2.60)	0.15 (-0.39, 0.29)
	Less than weekly	2.70 (0.69, 10.52)	2.64 (1.09, 8.41)	0.29 (-0.12, 0.64)
	Never	5.32 (0.99, 28.41)	5.37 (3.48, 43.03)	0.46 (0.23, 0.93)

Figures 5.3 – 5.4 (panels *a* and *b*) show the estimated functions for the odds ratio and the correlation for the length of injecting career. The final mean structure of the prevalence contains infection-specific intercepts and slopes for the length of injecting career. The two types of association generally rise with the length of injecting career in both countries. Based on the estimates of the association measures with their 95% CI, when injecting less than two years, odds ratios are lower than 1 and the estimated correlation is negative, implying that it is difficult to be already co-infected by HCV and HIV in the first years of injecting. However, after already two years of injecting, the estimated correlation becomes positive and the odds ratio larger than one. Hence, with increasing years of injecting, being infected with only one virus becomes less common and individuals tend to be increasingly infected with both viruses or not to be infected at all.

The results obtained for the age at first injection are shown in panel *c* and *d* of Figures 5.3 – 5.4. The final mean structure of the prevalence models includes the infection-specific intercepts and slopes, but not the risk factor, because it is not significant: indeed, the *P*-values for this risk factor from the ALR are 0.13 and 0.38 in Italy and Spain, respectively. As regards the association measures, the results differ between the two countries. In Spain both association measures increase with IDU's age at first injection

and the 95% pointwise CI never include the value of independence (one for the odds ratio and zero for the correlation); instead, in Italy we found that the odds ratio computed with ALR and the correlation are not significant, while the odds ratio computed with BDM is significant.

We turn now to examine the results for the categorical behavioral risk factors, i.e., sharing syringes on a ever/never basis and frequency of current injecting, information only available for Spain. The parameter estimates for the association measures are shown in Table 5.5. Note that the results presented in this table have been obtained from a complete case analysis.

For sharing of syringes, the final mean structure for the prevalences includes infection-specific intercepts and slopes for $\log(d_i)$ and sharing syringes as additional covariate. As expected from the literature (Mathëi et al., 2006), sharing syringes is a significant behavioral risk factor for both infections. This indicates that, adjusting for the length of the injecting career, the prevalence of HCV and HIV infections among IDUs never sharing syringes is significantly lower than among IDUs ever sharing syringes. However, the odds ratios are significantly greater than 1 and the correlation coefficients are significantly positive only for IDUs who have never shared syringes (see Table 5.5).

Finally, the effect of frequency of current injecting is not significant in the prevalence models; in the association model, the association parameters are significant only for current non-IDUs (persons who have not injected in the last month), as we can see from BDM and BPM, but not from ALR.

5.5 Discussion

In this chapter, we presented a general statistical framework to study the association between multiple infections, and we applied it to the case of co-infection with HCV and HIV in IDUs. The marginal models (ALR, BDM, and BPM) allowed us to estimate association measures between HCV and HIV infection, i.e., odds ratios and correlation coefficients, while the random-effects models gave us a measure of the individual heterogeneity in the acquisition of the infections.

We have shown that the length of injecting career, the age at first injection, ever sharing of syringes, and, at a lesser extent, the frequency of current injecting are behavioral risk factors which may influence either the prevalence of HCV and HIV infection, their association or both. The length of injecting career is known to be a main determinant of prevalence: indeed, we found that the prevalence of both infections increases with exposure time. In addition, we have shown that this risk factor affects the association between both infections as well. In particular, the odds ratio and the correlation slightly increase with the length of injecting use, indicating that people with a longer history of injecting have higher odds of experiencing both infections (or none), as opposed to having just one infection, or, in other words, the two infections become more correlated at the individual level.

The age at first injection is not a significant behavioral risk factor for the prevalence of the infections, but has a significant positive effect on the association measures, at least in Spain: older beginners have lower odds of experiencing just one infection and show a stronger correlation between both infections than younger beginners.

We have shown that ever sharing syringes is a significant behavioral risk factor for the prevalence of both infections; in particular, IDUs who admit ever sharing syringes are characterized by higher prevalence for HCV and HIV infections. However, the association between these infections (as measured by the odds ratio and the correlation coefficient) is significant only in the group of IDUs who report having never shared syringes. One possible explanation of this finding can be found in the concept of individual heterogeneity (Coutinho et al., 1999; Farrington et al., 2001). Sharing syringes implies more homogeneous mixing of IDUs and thus more homogeneous transmission of viruses among individuals. The lack of association in those who have ever shared syringes is due to a balance between concordant statuses (mostly co-infection) and discordant ones (one infection, mostly with HCV). For this group of IDUs, the joint distribution of the infection statuses of HCV and HIV is not statistically different from the joint distribution of the infection statuses under independence. On the contrary, in the group of IDUs who report having never shared syringes, there is a preponderance of subjects with concordant statuses (mostly neither of the infections) over those with discordant ones (one infection, again mostly with HCV). A possible reason for the observed individual heterogeneity in the association of HIV and HCV infections in those IDUs who report having never shared syringes may be that this group consists of subgroups differing strongly in their actual behavioral risk, one of them being the group who truthfully reports having never shared, while the other one is made of those IDUs who do not admit their sharing behavior. In addition, these differences could be due to some unobserved effects, such as paraphernalia sharing and unhygienic injecting in general, which have been seen as possible transmission routes for HCV infection, but not for HIV infection (Lucidarme et al., 2004; Mathëi et al., 2006).

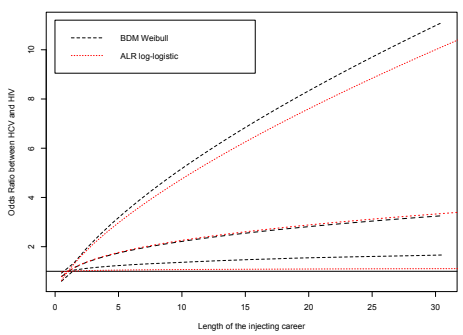
Finally, we found the effect of the frequency of current injecting on the association to be not very strong, except for the group of non-current injectors, who show higher odds ratios and correlations between HCV and HIV infections. As well for sharing syringes, this effect may be explained in terms of individual heterogeneity. A higher frequency of injecting, possibly related to the sharing of syringes, may imply more homogeneous mixing of IDUS and thus more homogeneous transmission of infections among IDUs, meaning that the probabilities of concordant and discordant statuses are similar. In contrast, the group of non-current injectors shows more heterogeneous mixing. This could be due to the merging of IDUs who gave up injecting with different timing, thus the probability of concordant statuses is higher than the probability of discordant ones. For this outcome, we found a discordance in the conclusions between the three marginal models. While this disagreement can be in part due to the different characteristics of the models, we nonetheless think that this may imply that the data do not strongly support the dependence of the odds ratio and the correlation on the frequency of current injecting, as instead they do with the ever sharing syringes.

The second group of models discussed in the paper deals more directly with the issue of individual heterogeneity by including subject-specific random effects in the model. For both GLMM and shared Gamma frailty model, a test for the variance of random effects is used to assess the degree of heterogeneity in the way IDUs acquire the infections. We have shown that, for both models, subject-specific random effects are needed, indicating that, in Italy and in Spain, HCV and HIV infection statuses within IDUs

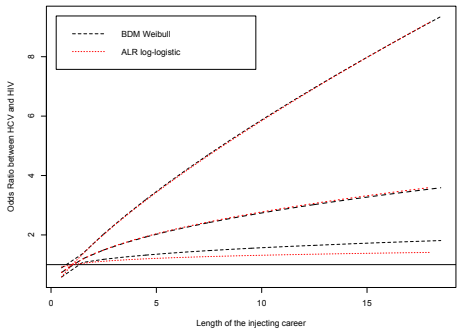
are correlated. Furthermore, we have shown that the null hypothesis of common variance in Italy and in Spain cannot be rejected: this means that the same level of individual heterogeneity can be assumed for both countries.

The study presents some limitations related to the data. It may be that, to a certain extent, national differences are confounded by differences in study methods. First, we consider that in Italy the sampled subjects were IDUs who self-selected for a treatment in a DTC, while in Spain the sampled subjects were street-recruited IDUs. Second, in Spain only drug users younger than 30 years were included, while in Italy there was no age limit. Notwithstanding these differences, the authors of both surveys recognized that the prevalence results of the surveys were in line with other national prevalence results and that similar prevalence rates had been reported in Spain, Italy, Portugal and France (Camoni et al., 2010; Barrio et al., 2007).

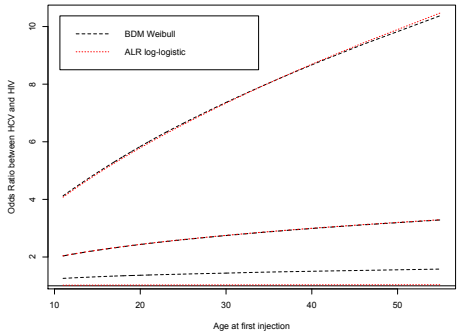
In this chapter we have used marginal models and random-effects models to estimate both prevalence and association measures. The marginal models focus on population-averaged quantities, such as the prevalence in the population or the association measures (odds ratios, correlation coefficients). When subject-specific questions are of interest, such as the individual heterogeneity in the population, GLMMs and shared Gamma frailty models ought to be considered, because of their use of individual-specific random effects. As shown in Section 5.3.2, since such conditional models allow derivation of population-averaged prevalence, these models can also be used for the same purpose of the marginal models. However, an advantage of the marginal models in such a setting is that the effects of direct interest (population prevalence and association measures) are captured directly by model parameters or simple functions thereof, unlike GLMMs.



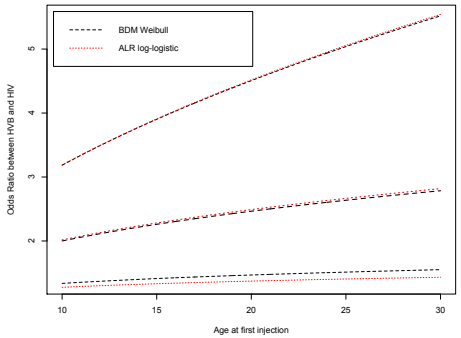
(a) Italy



(b) Spain

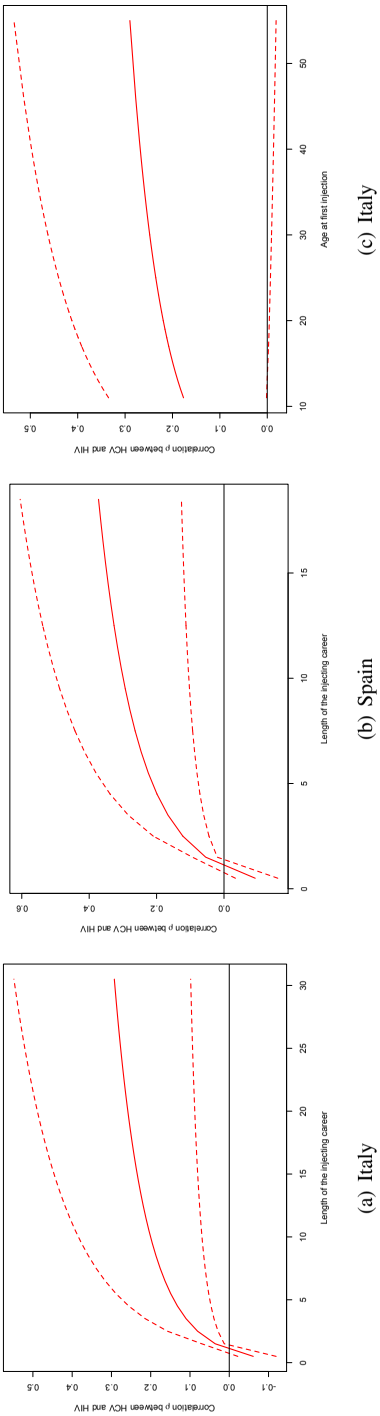


(c) Italy



(d) Spain

Figure 5.3: Odds ratio ψ between HCV and HIV infection depending on the length of the injecting career (upper part) and on the age at first injection (lower part), obtained from BDM Weibull with 95% pointwise bootstrap studentized CI ($B = 999$ replicates, black dashed lines), and from ALR log-logistic with 95% pointwise asymptotic CI (red dotted lines). The solid horizontal line is equal to $\psi = 1$, where there is no association.



(a) Italy (b) Spain (c) Italy (d) Spain

Figure 5.4: Correlation ρ between HCV and HIV infection depending on the length of injecting career (upper part) and on the age at first injection (lower part), obtained from BPM with 95% pointwise bootstrap studentized CI ($B = 999$ replicates, dashed lines). The solid horizontal line is equal to $\rho = 0$, where there is no association.

Part II

Bayesian Mixture Models for Antibody Titers

Chapter 6

Estimating Epidemiological Parameters of Infectious Diseases from Antibody Titers

The prevalence of immune and the force of infection (FOI) are key factors in the epidemiology of an infection. The prevalence, π, is the proportion of immune individuals at a specific time in a certain population, and the FOI, λ, is the instantaneous per capita rate at which susceptible individuals acquire infection (Farrington, 1990). For many airborne infections, both the prevalence and the FOI are assumed to be dependent on the age of the host a and on the calendar time t (Anderson and May, 1991). For the analyses presented in this part of the thesis we assume that the disease is in a steady state, i.e., the FOI and the transmission parameters are time-independent. We further assume that the disease is irreversible, meaning that immunity is assumed to be lifelong, and that the mortality caused by the infection is negligible and can be ignored. However, we are aware that these assumptions may be unrealistic for some infections. For instance, waning immunity and/or reinfection are likely to occur for parvovirus B19 (Goeyvaerts et al., 2011), as well as there is waning for measles immunity arising from vaccination (Kremer et al., 2006). These issues are ignored in Chapter 7, but are taken into account in Chapter 8 and Chapter 9.

The prevalence and the FOI are commonly estimated using current status data (Keiding et al., 1996; Jewell and van der Laan, 2004) from cross-sectional serological surveys. It is assumed that the current infection status Z_i of a subject aged $a_i, i = 1, \ldots, n$, is known and has Bernoulli distribution,

$$Z_i = \begin{cases} 0 & 1-\pi(a_i) \quad \text{susceptible,} \\ 1 & \pi(a_i) \quad \text{immune,} \end{cases} \tag{6.0.1}$$

where $\pi(a_i)$ is the population prevalence at age group a. The variable Z_i is obtained by dichotomizing the individual antibody counts Y_i using the cut-off point ϕ. This means that the prevalence $\pi(a_i)$ is defined by $\pi(a_i) = P(Z_i = 1|a_i) = P(Y_i > \phi|a_i)$. In order to estimate the prevalence $\pi(a_i)$, we can maximize the log-likelihood function of the current status data Z_i in its parameters, denoted by the vector $\boldsymbol{\pi}$:

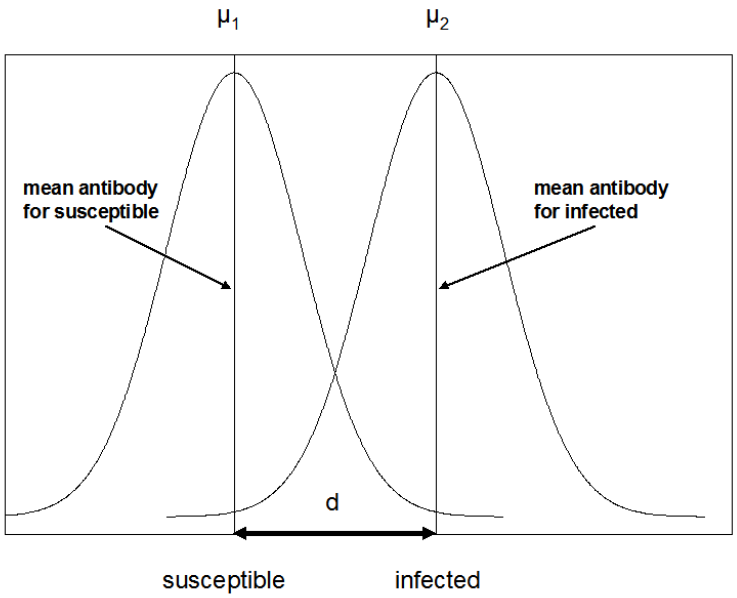
$$\ell(\boldsymbol{\pi}) = \sum_{i=1}^{N} \{Z_i \log[\pi(a_i)] + (1 - Z_i) \log[1 - \pi(a_i)]\}. \tag{6.0.2}$$

In literature, several approaches have been discussed to model both $\pi(a)$ and $\lambda(a)$ using current status data. For example, constrained nonlinear models (Farrington, 1990; Farrington et al., 2001), generalized linear models (GLM) for binary data (Becker, 1989; Keiding et al., 1996), nonparametric models (Keiding, 1991; Shkedy et al., 2003), semiparametric models (Hens et al., 2012), and constrained fractional polynomials (Shkedy et al., 2006). A general review of statistical methods for the estimation of the force of infection from current status data can be found in Hens et al. (2012). Since the prevalence and the FOI are related by the following formulas (Keiding, 1991),

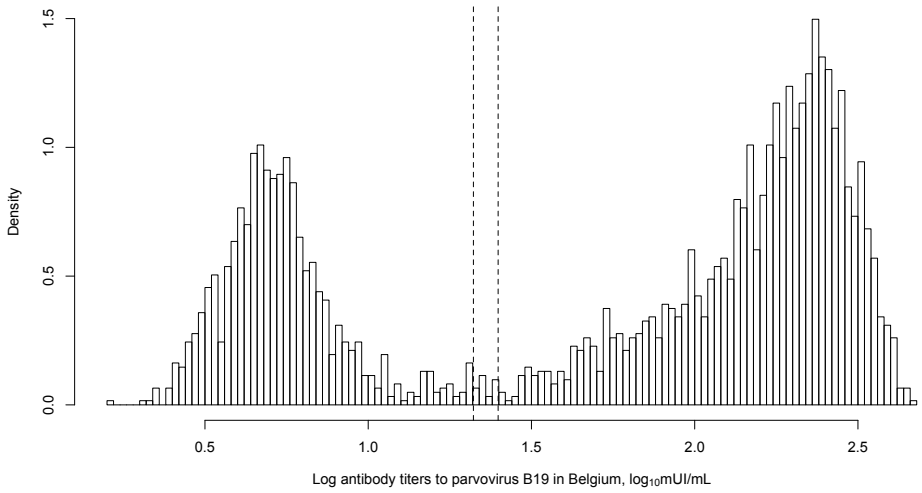
$$\pi(a) = 1 - \exp\left[-\int_0^a \lambda(u)du\right] \qquad \text{and} \qquad \lambda(a) = \frac{d\pi(a)}{da}\frac{1}{1-\pi(a)}. \tag{6.0.3}$$

both epidemiological functions can be estimated by maximizing the log-likelihood in Eq. 6.0.2.

All the methods mentioned above assume that the current status of the disease Z of a subject is known, given a fixed cut-off point ϕ. The application of a fixed cut-off to the distribution of antibody titers implies that what lies at its right is considered as signal (evidence of past infection), while what lies at its left is just noise (residual of maternal protection, giving no evidence of past infection). A more correct point of view is to consider the entire distribution of the antibody titers as signal, assuming thus that individuals with different infection statuses (susceptible or immune) have different distributions of antibody titers, described by different parameters. In this case, the membership of an individual to a particular distribution can be determined by other methods, rather than a cut-off point. In Figure 6.1 we compared the distribution of an hypothetical infection in pre-vaccination status with the actual antibody data for parvovirus B19 in Belgium. In particular, Figure 6.1a shows an illustrative example of the distribution of the antibodies: the distribution on the left side, with mean μ_1, account for the antibody levels of the susceptible, while the distribution on the right side, with mean μ_2, account for the antibody levels of the immune; the distance between the two means is given by d and is positive, implying that mean antibody level of immune is larger than the mean antibody level of susceptible, i.e., $\mu_1 \leq \mu_2$. The histogram of antibodies to parvovirus B19 in Belgium, shown in Figure 6.1b, shows indeed two groups separated by two cut-off points: the antibodies lying at the left side of the lower cut-off arise from the susceptible, while the those lying at the right side of the upper cut-off arise from the immune. The antibodies lying between the two cut-off points are generally considered as equivocal cases. In most cases, these dubious cases are re-tested. After re-testing, if they are still found to be equivocal, depending on the purpose, they are either excluded from the analysis or they are classified as seropositive.



(a) Illustrative example of the distribution of antibodies of hypothetical infection in pre-vaccination status, reproduced from Figure 11.2 in Hens et al. (2012)



(b) Histogram of of the log antibody titers to parvovirus B19 in Belgium

Figure 6.1: Example of the distribution of antibody levels for susceptible and immune individuals.

In addition, the cut-off point ϕ is chosen and given as a fixed threshold by the assay's manufacturer, based on some conventions, e.g., mean plus three-fold standard deviation (Greiner et al., 1994, 1997). Considering that the sensitivity is the ability to correctly identify the seropositive cases and the specificity is the ability to correctly identify the seronegative cases (Lalkhen and McCluskey, 2008), it has been argued (Vyse et al., 2004, 2006) that conventional cut-offs imply that test specificity is larger than test sensitivity. The reason is that the assays are mainly addressed to diagnostic purposes, where the focus is on the detection of true negative cases (susceptible), in order to better exclude that a person is infected. As argued by Vyse et al. (2006), such an approach allows to have a high positive predictive value (which is the proportion of cases, classified as positive, that are actually true positive), which is appropriate in a diagnostic setting.

However, the focus of population serological surveys is more on detecting true positive cases (immune), because it is of interest not the individual patient management (as in diagnoses), but rather to estimate the seroprevalence in the population. For this purpose, a larger sensitivity, rather than a larger specificity, is requested. Indeed, the use of a conventional cut-off would lead to underestimate the prevalence, because many seropositive cases with low antibody response would be classified as seronegative. In order to avoid the need to base the classification of the infectious statuses on a pre-specified cut-off point, a modeling approach based on mixture modeling was firstly proposed by Greiner et al. (1994) and refined by Gay (1996). Considering that a certain population can be divided in K subpopulations, mixture modeling is a data-driven approach that aims to assign the observations to these K subpopulations (components) and allows, differently from cluster analysis methods, to characterize their distribution within the obtained clusters. Being a data-driven approach, the method tends to have a higher sensitivity, in the sense that it is able to assign the observations to their correct component.

The basic approach, discussed by Gay (1996) and Vyse et al. (2004), sees a two-components mixture model stratified by age class, one component for the susceptible and one for the immune, where data are distributed according to a normal distribution or some skewer or more robust distributions (Truncated Normal, Gamma, Student's t) and the estimation is done using maximum likelihood methods. Developing models for vaccine-preventable infections, other authors do not constrain the mixture to two components only, but the number of groups is rather an additional parameter of the model that has to be estimated (Vyse et al., 2006; Gay et al., 2003; Baughman et al., 2006; Rota et al., 2008). In all the mixture models mentioned above, the component with the lower mean antibody level represents the log antibody level for the susceptible individuals, while the remaining components account for increasing levels of antibodies, arising either from vaccination or from natural infection. The prevalence can be estimated by adding the mixture probabilities (i.e., the probability to belong to a specific component) of all the components labelled as immune. Furthermore, Grün and Leisch (2008) discussed finite mixture of regression models, where the means of the mixture components and/or the mixture probabilities depend on a set of fixed and random effects and which are estimated using a Bayesian approach (Evans and Erlandson, 2004; Ødegård Et al., 2005; Nielsen et al., 2007; Hardelid et al., 2008). Finally, Bollaerts et al. (2012) followed a different approach: after having shown that the choice of the fixed threshold can be troublesome, since the optimal

cut-off point should be different for the estimation of the true prevalence and of the FOI, they resorted to estimate the true prevalence and the FOI combining the use of a mixture model (for the estimation of the means of the components) and of a flexible regression model (for the estimation of the age-dependent mean of the antibody titers).

This part of the thesis is structured in the following way. In Chapter 7 we present a two-component hierarchical Bayesian mixture model, with varying data density distributions, to fit antibody titers to parvovirus B19, an infection that is currently in pre-vaccination status. The mixture model's components have age-independent mean and variance and, differently from previous works on mixture models, age-dependent mixture weights, which are constrained to follow an increasing monotonic model specified by the user. In this way, we do not account anymore for the dependence on the age through stratification (Gay, 1996; Vyse et al., 2004, 2006), but we rather fit a flexible model for the prevalence. Moreover, whereas the previous works aimed to estimate only the prevalence of the immune, we propose a model from which both prevalence and FOI can be estimated. Further, we extend the parametric models for the prevalence discussed in Hens et al. (2012) and propose a set of flexible models for the mixture probability, from which the FOI is estimated. In Chapter 8 we extend the mixture model of Chapter 7 by allowing the mean structure of the immune component to vary according to age, in order to account for possible decays or boosting of the immunity response. The model for the mean of the immune is chosen using the Bayesian Variable Selection approach, which, in this case, permits the user to select the best model and also to formally test the null hypothesis of age independence. The models here presented are applied to antibody titers against parvovirus B19 infection and varicella. Finally, in Chapter 9 we present an application for measles antibody titers. The mixture here described differs from the previous ones for the fact that we do not have necessarily only two components and the number of components becomes a further unknown to estimate. Being measles an infection in a post-vaccination phase, with immunity proceeding from both natural infection and vaccination, it is not possible anymore to estimate the FOI, therefore the focus is on the characterization of the different immunity patterns and on the estimation of the age-specific prevalence of the immune.

Chapter 7

Estimating the Prevalence and the Force of Infection of Parvovirus B19 in Belgium and Italy Using Hierarchical Bayesian Mixture Models

7.1 Introduction

In this chapter, we focus on a finite mixture model with two components using a hierarchical Bayesian model (Diebolt and Robert, 1994). The mixture probability (the probability to belong to the infected component) is equivalent to the population prevalence, and is assumed to be age-dependent. Several models, either parametric or nonparametric, will be used in order to model the dependence of the mixture probability on age. In contrast with earlier work, that was focussed on the estimation of the prevalence, the methods discussed here in the chapter allow to model both the prevalence and the force of infection (FOI). Concerning the data distribution, we select the best fitting model to the data among two symmetric distributions (normal, Student's t) and their skewed versions (skew-normal, skew-t). The hierarchical Bayesian mixture model is fitted through Markov Chain Monte Carlo (MCMC) methods with Gibbs sampling (Frühwirth-Schnatter, 2007).

The chapter is structured as follows. In Section 7.2 we introduce the data used in the analysis. We discuss the hierarchical Bayesian mixture model and a criterion for model selection in Section 7.3. In Section 7.4 we present the results of the analysis of real data, two serological samples of antibodies to parvovirus B19 in Belgium and Italy. In Section 7.5 we present the results of a simulation study. Finally, the findings are discussed in Section 7.6.

7.2 Data

Parvovirus B19 is a human virus that causes a childhood rash called "fifth disease" or "slapped cheek syndrome". The virus is primarily spread by infected respiratory droplets. Symptoms usually begin around six days after exposure and last around a week. Infected individuals with normal immune systems are infective before they become symptomatic, but probably not after then (Corcoran and Doyle, 2004). Individuals with B19 IgG antibodies are generally considered immune to recurrent infection, although re-infection is possible in a minority of cases (Lehmann et al., 2003). Infection with parvovirus B19 may be dangerous during pregnancy: intrauterine parvovirus B19 infection can lead to both miscarriage and intrauterine fetal death (Tolfvenstam et al., 2001), even though these problems may occur in less than 5% of all pregnant women infected with parvovirus B19 (CDC, 2005). There is currently no vaccine available against this virus (Servey et al., 2007).

Serological samples were collected and tested for antibodies to B19 as part of the European Sero-Epidemiology Network (ESEN2) (Nardone et al., 2007) and POLYMOD (Mossong et al., 2008) projects. Data from Belgium and Italy are used for illustration of the methodology discussed in this chapter. For each subject, the antibody count (expressed as $\log_{10}$ mUI/mL) and the age are available. Sample sizes and age ranges of the samples analyzed are as follows: Belgium (1–64 years, $n = 3072$), and Italy (1–65 years, $n = 2365$). Children below one year, who may still be protected from infection by the inherited maternal antibodies, are discarded from the analysis. Children under 10 years old were oversampled in each country. Figure 7.1 shows a scatter plot of the antibody titers for parvovirus B19 in the two countries by age and the conventional set of cut-off points given by the assay's manufacturer.

Figure 7.1: Histogram and scatter plot of the log antibody titers to parvovirus B19 in Belgium and Italy by age with over imposed the two cut-off points given by the assay's manufacturer.

7.3 Methods

7.3.1 Hierarchical Bayesian Mixture Models with Age-Independent Prevalence

In order to model the heterogeneity in the antibody levels, we assume that the sample is drawn from a population consisting of an unknown number K of subpopulations (components), resulting in a density function g that can be represented as a mixture-distribution density of K unobserved densities. Thus, given a sample of antibody counts for subject i, Y_i, $i = 1,\ldots,n$, we wish to undertake inference for the parameters of a finite K-component mixture distribution of the form

$$g(Y_i) = \sum_{k=1}^{K} \pi_k f(Y_i|\boldsymbol{\xi}_k), \tag{7.3.1}$$

where $f(Y_i|\boldsymbol{\xi}_k)$ are the mixture components, π_k are the mixture weights (or mixture probabilities) with $\sum_{k=1}^{K} \pi_k = 1$, and $\boldsymbol{\xi}_k$ are the vectors of density parameters to be estimated (McLachlan and Peel, 2000). In particular, we focus on a hierarchical Bayesian mixture model (Congdon, 2003; Gelman et al., 2004; Gilks et al., 1996) of two components with possibly different mean and variance parameters. The first component is the distribution of the antibody counts of the susceptible and the second component is the distribution of the antibody counts of the immune. Hence, the mixture model in (7.3.1) can be rewritten as

$$g(Y_i) = (1-\pi) f(Y_i|\boldsymbol{\xi}_1) + \pi f(Y_i|\boldsymbol{\xi}_2). \tag{7.3.2}$$

We introduce a latent indicator variable Z_i, which represents the individual unknown infection status (Diebolt and Robert, 1994), assumed to be Bernoulli distributed,

$$Z_i = \begin{cases} 1 & \pi \quad \text{individual } i \text{ is immune,} \\ 0 & 1-\pi \quad \text{individual } i \text{ is susceptible.} \end{cases} \tag{7.3.3}$$

Thus, we can reformulate the mixture model in (7.3.2) in terms of Z_i, in the following way:

$$g(Y_i) = (1-Z_i) f(Y_i|\boldsymbol{\xi}_1) + Z_i f(Y_i|\boldsymbol{\xi}_2). \tag{7.3.4}$$

The probabilities π and $1-\pi$ are the mixing weights of the infected and the susceptible components, respectively. In particular, π is the probability that an individual in the population belongs to the "immune" component and can be interpreted as the prevalence of the infection in the population (Evans and Erlandson, 2004).

Following Diebolt and Robert (1994), we assume a Dirichlet prior for π. Note that the Dirichlet distribution is a multivariate version of the Beta distribution, with parameter $\alpha = 1$, to give a uniform prior over the probabilities:

$$(1-\pi, \pi) \sim \text{Dir}(\alpha = 1, \alpha = 1). \tag{7.3.5}$$

7.3.2 Hierarchical Bayesian Mixture Models with Age-Dependent Prevalence

As we mentioned previously, it is widely documented that the prevalence of childhood infections may depend on the age of individuals (Anderson and May, 1991). Therefore, we express the mixture probabilities as a function of the age and we rewrite consequently the mixture model of (7.3.2) as

$$g(Y_i) = (1-\pi(a_i))f(Y_i|\boldsymbol{\xi}_1) + \pi(a_i)f(Y_i|\boldsymbol{\xi}_2). \tag{7.3.6}$$

As we mentioned before, $\pi(a_i)$, which is the age-dependent probability governing the Bernoulli distribution of $Z(a_i)$, is interpreted as the prevalence of immune in the population:

$$Z(a_i) = \begin{cases} 1 & \pi(a_i) \quad \text{individual } i \text{ of age } a \text{ is immune,} \\ 0 & 1-\pi(a_i) \quad \text{individual } i \text{ of age } a \text{ is susceptible.} \end{cases} \tag{7.3.7}$$

In what follows, we define a model either for the prevalence or for the FOI, and we derive the other function according to (6.0.3). The two functions are estimated simultaneously with the mixture parameters and with the latent classification variables Z_i or $Z_i(a_i)$. The chosen models ought to guarantee a nonnegative FOI, thus we have to constrain the prevalence function to be monotonically increasing. We consider three possible age-dependent models for the prevalence, the first two being parametric and the third nonparametric.

7.3.2.1 Log-Logistic Model

The log-logistic model for the mixture probability (Hens et al., 2012) can be defined as a generalized linear model (GLM) in the following way:

$$\ln\left[\frac{\pi(a_i)}{1-\pi(a_i)}\right] = \beta_0 + \beta_1 \log(a_i), \qquad \beta_1 > 0. \tag{7.3.8}$$

This GLM implies the following mixture probability (prevalence) function $\pi(a_i)$,

$$\pi(a_i) = \frac{e^{\beta_0} a_i^{\beta_1}}{1+e^{\beta_0} a_i^{\beta_1}}, \tag{7.3.9}$$

and the FOI function $\lambda(a_i)$ is given by

$$\lambda(a_i) = \frac{e^{\beta_0}\beta_1 a_i^{\beta_1-1}}{1+e^{\beta_0} a_i^{\beta_1}}. \tag{7.3.10}$$

For the prior distributions of the coefficients of the models, we use diffuse "noninformative" distributions, namely, zero-centered normal distributions with large variance, $\beta_k \sim N(0,1000)$. Then we use an exponential transformation to constraint the coefficient β_1 to be positive. Note that the log-logistic model

is presented here only as an example, other GLMs can be formulated for the mixture probability.

7.3.2.2 Piecewise-Constant FOI Model

Since Anderson and May (1985), this basic, but flexible model, has been widely used for the estimation of the prevalence and the FOI of childhood infections. It is based on the idea that the different age groups for individuals in school age are characterized by different FOI levels. Assuming that the FOI is constant in each age class $[a_{j-1}, a_j)$, we formulate the model for the prevalence at exact age a in the following way:

$$\pi(a) = 1 - \exp\left\{ -\sum_{j=1}^{J-1} \lambda_j (a_{j+1} - a_j) - \lambda_J (a - a_J) \right\}, \tag{7.3.11}$$

where λ_j, with $j = 1, \ldots, J$, is the constant FOI in the jth age group. Note that, whereas the FOI is constant within each age class, the susceptibility profile $S(a) = 1 - \pi(a)$ is piecewise exponential. To analyze the seroprevalence of parvovirus B19, the following eight 5-years age groups (but the last one) were used: 0–4, 5–9, 10–14, 15–19, 20–29, 30–34, 35–40, over 40. For the last age group, over 40 years old, we aggregate all the remaining age groups, because we expect the FOI not to change anymore, unless for variability issues related to the small sample sizes. Note that we split up the age group 30–39 in two groups, because the age 35 was found as best threshold for B19 waning immunity by Goeyvaerts et al. (2011). In their work, (Goeyvaerts et al., 2011) hypothesized a MSIRW mathematical model for parvovirus B19 infection, assuming that individuals are initially protected by the inherited maternal antibodies (M), than enter the susceptible (S) class, they get infected (I) and then they recover (R); at this point, they may lose antibody protection due to waning (W), that is assumed to be age-dependent. The authors assumed a piecewise-constant model for waning, with best threshold (waning starting point) at age 35 years.

We complete the specification of the hierarchical Bayesian model by imposing a noninformative Uniform prior distribution on the piecewise-constant FOI in each age group, $\lambda_j \sim U(0,5)$.

7.3.2.3 Constrained Nonparametric Model with Product-Beta Prior

Considering $j = 1, \ldots, J$ one-year age groups, we assume that $\boldsymbol{\pi}$ is a right-continuous nondecreasing function defined on $[0, \delta]$, with $\pi_J \leq \delta \leq 1$. In the previous sections, we formulated a GLM for the mixture probabilities. In contrast, in this section we do not assume any deterministic relationship between π_j and $a_i = j$, but we rather specify a probabilistic model for π_j at each distinct level of $a_i = j$. Gelfand and Kuo (1991) proposed the product-beta prior for $\boldsymbol{\pi}$, given by

$$P_B(\boldsymbol{\pi}|\boldsymbol{\alpha}, \boldsymbol{\beta}) \propto \prod_{j=1}^{J} (\pi_j)^{\alpha_j - 1} (1 - \pi_j)^{\beta_j - 1}, \quad (\alpha_j > 0, \beta_j > 0), \tag{7.3.12}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_J)$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_J)$. Note that the product-beta is a conjugate prior for the binomial likelihood. Thus, the posterior distribution of $\boldsymbol{\pi}|\boldsymbol{Y}$ is also beta.

Since we assume that the prevalence is monotonically increasing, in order to get a nonnegative estimate of the FOI, we require that $\pi_1 \leq \pi_2 \leq \cdots \leq \pi_J$. Thus, the J-dimensional parameter vector is a subset S^J of R^J. As pointed out by Gelfand and Kuo (1991), the constraint can be defined by using a restricted prior distribution. Gelfand et al. (1992) show that the posterior distribution of $\boldsymbol{\pi}$ given the order constraints is the unconstrained posterior distribution normalized such that

$$P(\boldsymbol{\pi}|\boldsymbol{Y}) \propto \frac{P(\boldsymbol{Y}|\boldsymbol{\pi})P(\boldsymbol{\pi}|\boldsymbol{\alpha},\boldsymbol{\beta})}{\int_{S^J} P(\boldsymbol{Y}|\boldsymbol{\pi})P(\boldsymbol{\pi}|\boldsymbol{\alpha},\boldsymbol{\beta})d\boldsymbol{\pi}}, \qquad \boldsymbol{\pi} \in S^J. \tag{7.3.13}$$

Hence, in our setting,

$$\begin{cases} P(\pi_j|\boldsymbol{Y},\alpha_j,\beta_j,\boldsymbol{\pi}_{-j}) \propto P(\boldsymbol{Y}|\boldsymbol{\pi})P(\boldsymbol{\pi}|\boldsymbol{\alpha},\boldsymbol{\beta}) & \pi_j \in S_j^J, \\ P(\pi_j|\boldsymbol{Y},\alpha_j,\beta_j,\boldsymbol{\pi}_{-j}) = 0 & \pi_j \notin S_j^J. \end{cases} \tag{7.3.14}$$

Here, $\boldsymbol{\pi}_{-j} = (\pi_1,\ldots,\pi_{j-1},\pi_{j+1},\ldots,\pi_J)$. Therefore, the conditional posterior distribution of $\pi_j|\boldsymbol{Y},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\pi}_{-j}$ is the standard posterior distribution, that is to say, Beta$(y_j+\alpha_j, n_j-y_j+\beta_j)$ restricted to the interval $\left[\pi_{j-1},\pi_{j+1}\right]$ (Gelfand and Kuo, 1991). This means that during the MCMC simulation, the sampling from the full conditional distribution can be reduced to interval-restricted sampling from the standard posterior distribution (Gelfand et al., 1992).

We consider the following hierarchical model:

$$\begin{cases} Y(a_i = j) \sim \text{Bin}(n_j,\pi_j) & \text{likelihood} \\ \pi_j \sim \text{Beta}(\alpha_j,\beta_j)I(\pi_{j-1},\pi_{j+1}) & \text{prior,} \end{cases} \tag{7.3.15}$$

where $I(\pi_{j-1},\pi_{j+1})$ is an indicator variable which takes the value of one if $\pi_{j-1} \leq \pi_j \leq \pi_{j+1}$, and zero elsewhere. In order to complete the specification of the hierarchical model in (7.3.15), we need to specify a hyperprior distribution for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. In the special case that $\alpha_j = \beta_j = 1$, it follows that the prior distribution of the prevalence in the jth age group is a Uniform distribution over the interval $[\pi_{j-1},\pi_{j+1}]$. However, if there is no reason to fix $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to be equal to one, there is no clear way how to choose the hyperprior distribution for the components in $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ either. For the analysis presented below, we specify noninformative distributions for the hyperparameters by specifying a left truncated (at zero) normal distribution with variance equal to 1000 for each one of the components in $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ at the third stage of the hierarchical model.

7.3.3 Selecting the Optimal Density for Data

Different densities for the distribution of the data can be taken into account to model the data. An obvious choice is the normal distribution. Other distributions can be considered as well, which can allow either for skewness (skew-normal) or for thicker tails (Student's t), thus being more robust, or for both (skew-t).

7.3.3.1 Normal Distribution

A two-component normal mixture model is given by (Diebolt and Robert, 1994)

$$g(Y_i|\mu_k,\sigma_k^2) = (1-Z_i)N(Y_i|\mu_1,\sigma_1^2) + Z_i N(Y_i|\mu_2,\sigma_2^2). \tag{7.3.16}$$

For the unknown mixture parameters, we specify flat prior distributions, so that they contain as less information as possible (Gilks et al., 1996). For the location parameter μ_k, we use a Uniform distribution within the range given by a lower bound, Y_l, and an upper bound, Y_u. In practice, we take these two bounds to be equal to the minimum and the maximum of the data, respectively:

$$\mu_k \sim U(Y_l, Y_u) = U(Y_{\min}, Y_{\max}), \qquad \mu_1 < \mu_2. \tag{7.3.17}$$

The order restriction on the prior for μ_j reflects the idea that the mean antibody levels of immune individuals are higher than the mean antibody levels of susceptible and is required in order to avoid label switching during MCMC sampling, ensuring thus the identifiability of the parameters (McLachlan and Peel, 2000). As it is implemented in the software JAGS (Plummer, 2011), the order restriction is obtained in the MCMC procedure by sorting, at each iteration, the two values of the mean, (μ_1, μ_2), drawn from the prior distribution in (7.3.17). For the precision parameter τ_j, which is the reciprocal of the variance, we assume a Gamma distribution with small parameters $\varepsilon = 0.01$ (Zhao et al., 2006):

$$\tau_k = 1/\sigma_k^2 \sim \Gamma(0.01, 0.01). \tag{7.3.18}$$

7.3.3.2 Student's t Distribution

Another possible symmetric distribution for the data is the Student's t distribution, which, differently from the normal distribution, allows for thicker tails. The degrees of freedom, ν_k, account for this feature, adjusting for the excess kurtosis in the data: the smaller ν_k, the thicker tails. Therefore, a large estimate for the degrees of freedom implies that the Student's t density reduces to the normal density. Using Student's t distribution, the mixture model is given by

$$g(Y_i|\mu_k,\sigma_k^2,\nu_k) = (1-Z_i)t_{\nu_1}(Y_i|\mu_1,\sigma_1^2) + Z_i t_{\nu_2}(Y_i|\mu_2,\sigma_2^2). \tag{7.3.19}$$

The same prior distributions, as in (7.3.17) and in (7.3.18), were assumed for the parameters μ_k and σ_k^2, respectively. For the prior distribution of the degrees of freedom ν_k, we assume an Exponential distribution with hyperparameter κ, where κ is assumed to uniformly distributed (Evans and Erlandson, 2004):

$$\nu_j \sim Exp(\kappa) \text{ and } \kappa \sim U(0.01, 0.5). \tag{7.3.20}$$

7.3.3.3 Skew-normal Distribution

The skew-normal distribution (Azzalini, 1985, 1986) is an extension of the normal distribution which allows for skewness in the data. The probability density function (pdf) of this distribution is given by

$$f(X|\mu,\sigma^2,\alpha) = \frac{2}{\sigma}\phi\left(\frac{X-\mu}{\sigma}\right)\Phi\left(\alpha\frac{X-\mu}{\sigma}\right), \tag{7.3.21}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are, respectively, the pdf and the cumulative density function (cdf) of the standard normal distribution. The parameters μ and σ are the location and the scale parameter, respectively, and α is the skewness parameter, which can lead to a skewness coefficient in [-0.9953, 0.9953].

In order to fit a Bayesian mixture model with skew-normally distributed components, we follow the approach of Frühwirth-Schnatter and Pyne (2010), using a stochastic representation of the distribution, namely, a random-effect model with truncated normal random effects S (Azzalini, 1986). We define the variable Y as

$$Y = \mu + \sigma\delta S + \sigma\sqrt{1-\delta^2}\varepsilon, \tag{7.3.22}$$

where S is a random effect with truncated normal distribution, $S \sim TN_{[0,\infty]}(0,1)$, ε is the measurement error with normal distribution, $\varepsilon \sim N(0,1)$, independent from S, and $\delta = \alpha/\sqrt{1+\alpha^2}$.

Following Frühwirth-Schnatter and Pyne (2010), in order to implement the Bayesian approach, the parameter vector $\boldsymbol{\theta}_k = (\mu_k, \psi_k, \omega_k)$ is parameterized as $\boldsymbol{\theta}_k = (\mu_k, \sigma_k\delta_k, \sigma_k^2(1-\delta_k^2))$. The mixture model for Y_i can be rewritten as

$$g(Y_i|\mu_k,\psi_k,\omega_k^2) = (1-Z_i)N(Y_i|\mu_1+\psi_1 S_i, \omega_1^2) + Z_i N(\mu_2+\psi_2 S_i, \omega_2^2). \tag{7.3.23}$$

Then, the parameters σ_k^2 and α_k can be recovered through (Frühwirth-Schnatter and Pyne, 2010)

$$\alpha_k = \frac{\psi_k}{\omega_k}, \qquad \sigma_k^2 = \omega_k^2 + \psi_k^2. \tag{7.3.24}$$

For the prior distributions of the parameters μ_k, ω_k, and ψ_k, we chose the following diffuse distributions (Frühwirth-Schnatter and Pyne, 2010):

$$\mu_k \sim U(Y_{\min}, Y_{\max}), \text{ with } \mu_1 \leq \mu_2; \tag{7.3.25}$$

$$\tau_k = 1/\omega_k^2 \sim \Gamma(2.5, C_0), \text{ with } C_0 \sim \Gamma(1, 2/Var(Y)); \tag{7.3.26}$$

$$\psi_k \sim N(0, 100). \tag{7.3.27}$$

7.3.3.4 Skew-t Distribution

The skew-t distribution is an extension of the skew-normal and it accounts for the excess in kurtosis present in the data (Azzalini and Capitanio, 2003). The pdf of this distribution is given by

$$f(X|\mu,\sigma^2,\alpha,\nu) = \frac{2}{\sigma} t_\nu(Y_X) T_{\nu+1}\left(aY_X\sqrt{\frac{\nu+1}{\nu+Y_X^2}}\right), \tag{7.3.28}$$

where $Y_X = (X-\mu)/\sigma$ and t_ν and T_ν denote, respectively, the pdf and the cdf of a standard Student's t distribution with ν degrees of freedom. Similar to the skew-normal, the skew-t can be formulated using a stochastic representation. We define the variable Y as

$$Y = \mu + \sigma\delta S + \frac{\sigma\sqrt{1-\delta^2}}{W}\varepsilon, \tag{7.3.29}$$

where $W \sim \Gamma(\nu/2,\nu/2)$, $S \sim TN_{[0,\infty]}(0,1/W)$, and W_i, S_i, and $\varepsilon_i|W_i$ are jointly independent. Following the parameterization of Frühwirth-Schnatter and Pyne (2010), the mixture model is given by

$$g(Y_i|\mu_k,\psi_k,\omega_k^2,\nu_k) = (1-Z_i)t_{\nu_1}(Y_i|\mu_1+\psi_1 S_i,\omega_1^2/W_i) + Z_i t_{\nu_2}(\mu_2+\psi_2 S_i,\omega_2^2/W_i). \tag{7.3.30}$$

For the prior distributions of the parameters μ_k, ω_k, ψ_k and ν_k, we assume the following diffuse prior distributions:

$$\begin{aligned}
\mu_k &\sim U(Y_{\min},Y_{\max}), \text{ with } \mu_1 \le \mu_2; && (7.3.31)\\
\tau_k = 1/\omega_k^2 &\sim \Gamma(2.5,C_0), \text{ with } C_0 \sim \Gamma(1,2/Var(Y)); && (7.3.32)\\
\psi_k &\sim N(0,100); && (7.3.33)\\
\nu_k &\sim Exp(\kappa); && (7.3.34)\\
\kappa &\sim U(0.02,0.5). && (7.3.35)
\end{aligned}$$

7.3.4 Model Selection

A common choice for model selection in Bayesian models is the Deviance information criterion (DIC, Spiegelhalter et al., 2002), which can be computed with the following formula:

$$DIC = \overline{D(\theta)} + p_D = 2\overline{D(\theta)} - D(\bar{\theta}), \tag{7.3.36}$$

where $\overline{D(\theta)}$ is posterior mean of the deviance, easily calculated using MCMC methods, and $D(\bar{\theta})$ is the so-called "plug-in deviance", which is the deviance evaluated at the posterior mean.

However, according to Celeux et al. (2006), it is not possible to base model selection on the DIC with mixture models as well as with other missing data problems (we had the same issue with the hierarchical random-effects models in Chapter 4). With mixture models, the missing information is the current status

of infection of the individuals, Z_i or $Z(a_i)$. For these models it has been shown that the penalty p_D, representing the effective dimension of the model, can behave badly, turning even negative. The main problem is a poor identifiability of the model parameters, leading to the posterior mean being a bad estimator for the plug-in deviance. Unfortunately, it is unclear which parameter estimators to plug into the deviance $D(\bar{\theta})$ More in general, Celeux et al. (2006) show that the DIC lacks a natural generalization outside exponential families. For this reason, Celeux et al. (2006) come up with several alternatives for DIC, according to the way the missing information Z_i is treated, and propose different estimators for the parameters.

Other authors, aware of this issue with DIC, proposed further alternative selection criteria to the DIC. In this chapter, we use two of them, namely, the penalized expected deviance (PED, Plummer, 2008) and the difference in posterior deviance (Aitkin, 2010). We refer to Section 4.3.3 for an elaborate presentation of these two selection criteria.

7.3.5 Determination of the Current Status of Infection

After having obtained the posterior means $\bar{\theta}_k = (\bar{\pi}_k(a), \bar{\mu}_k, \bar{\sigma}_k)$ of the mixture parameters, we need to classify the components either as susceptible or as immune, and then assign the individuals to one of the two components. First, according to the estimated locations $\bar{\mu}_k$, we label each component. Second, we assign each subject in the sample to one of the two components through the posterior mean of the mixture probability $\bar{\pi}_k(a_i)$, i.e., the component for which $\bar{\pi}_k(a_i)$ is maximal and larger than 0.5. The current status of the infection, $\bar{Z}(a_i)$, is estimated by

$$\bar{Z}(a_i) = \begin{cases} 1, & \bar{\pi}_1(a_i) < \bar{\pi}_2(a_i) \quad \text{individual } i \text{ of age } a \text{ is immune,} \\ 0, & \bar{\pi}_1(a_i) > \bar{\pi}_2(a_i) \quad \text{individual } i \text{ of age } a \text{ is susceptible.} \end{cases} \tag{7.3.37}$$

Finally, when each individual has been classified either as susceptible or immune, one can estimate the proportion of immune individuals per age group by averaging $\bar{Z}(a_i)$ in the jth age group,

$$\hat{\pi}^I(a = j) = \frac{\sum_{i=1}^{n_j} \bar{Z}(a_i)}{n_j}, \text{ with } j = 1, \ldots, n_j. \tag{7.3.38}$$

We remark that $\hat{\pi}^I(a)$ are the estimates of the age-specific proportions seropositive in the serological sample and can be compared with the proportions seropositive in the sample obtained using the conventional cut-off points. In contrast, the mixture probabilities $\bar{\pi}^I(a)$ depend on the chosen model for the prevalence and they are an estimate of the age-specific prevalence in the population.

7.4 Application to Parvovirus B19

The hierarchical Bayesian mixture models were fitted to parvovirus B19 antibodies with age-dependent mixture probabilities. We compared two symmetric distributions (normal and Student's t) and their

respective skewed versions (Skew-normal (Azzalini, 1985, 1986) and Skew-t (Azzalini and Capitanio, 2003; Frühwirth-Schnatter and Pyne, 2010)). The mixture models were fitted with MCMC through Gibbs sampling, using JAGS (Plummer, 2003) in R, through the package `R2jags` (Su and Yajima, 2012). We used 3 chains of 10000 iterations each, with burn-in 5000 and thinning 10. Since we used multiple chains, we could check the convergence of the parameters using three different tests: Gelman and Rubin (1992), Geweke (1992), and Heidelberger and Welch (1983) convergence diagnostics, all available in the R package `coda` (Plummer et al., 2006). For each parameter of interest, both the posterior mean and the 95% credible interval (CI) are shown. The choice of the best data distribution and the best age-dependent prevalence model was done using the PED (Plummer, 2008) and the difference in posterior deviances (Aitkin, 2010).

7.4.1 Mixture Models with Age-Dependent Mixture Probabilities

Table 7.1 reports, for each country, model, and density, the posterior mean (with 95% CI) of all the estimated mixture parameters. We notice that the estimated means are similar under normal and Student's t distributions, while the standard deviations are smaller for the Student's t distribution. The same occurs with the skew-normal and the skew-t distributions. In both countries, as expected, the estimated means for the immune component are higher under the skewed densities, as the model accounts for the negative skewness. This negative skewness is consistent with the left-tailed shape of the immune component, due to the mass of the antibody distribution being concentrated on the right side. This implies that higher antibody counts for the immune are more probable than lower counts. Also from Figure 7.2, which shows the estimated mean antibody levels of the immune component under the assumption of normal and skew-normal distributions, we notice that the mean under the skew-normal is higher than under the normal.

In order to select the best model for Belgium and Italy, in Tables 7.2 and 7.3 we report the values of the two selection criteria, which are the PED and the difference in posterior deviances.

Table 7.2 shows the PED per each age-dependent mixture model, according to country, prevalence model and data distribution. For the PED, the normal distribution outperforms the other three distributions in both countries, followed by the skew-normal distribution. Both Student's t and skew-t have very high PED values, implying that there is no necessity to account for thick distribution tails. For the prevalence models, the PED indicates the nonparametric model with product-beta prior as the best model for parvovirus B19 in Italy and Belgium.

Table 7.3 shows, for several models, the difference in posterior deviance (with 95% CI) and the respective posterior probability that the difference between two models in smaller than 4.39. This is the value to which corresponds a likelihood ratio test favoring the first model with a posterior probability of 0.9 (Aitkin, 2010). We notice that this selection criterion leads to different conclusion with respect to PED. For the data distribution, the skew-normal distribution outperforms the normal and the Student's t distribution in both countries, while it is only slightly better than the Skew-t distribution. For the prevalence model, we cannot confidently choose one model, but rather all the models fit in a similar

way.

Despite the fact that the two selection criteria lead to different conclusions, we notice from Figure 7.3 that the estimates of the proportions seropositive, of the population prevalence, and of the FOI from the skew-normal and the normal distribution are highly correlated. This means that using different data distributions for these data does not affect the way individuals are classified between susceptible and infected. A simulation study, presented in Section 7.5 was conducted in order to investigate the effect of misspecifying the distribution on the estimation of the current status infection and of the prevalence.

Finally, in Figures 7.4 and 7.5, we display the estimated prevalence and FOI of parvovirus B19 in Belgium and Italy, respectively, obtained from the three age-dependent models with 95% pointwise CIs, estimated under the assumption of skew-normal (solid curves) and normal (dashed curves) distributions. In the same figures, we show the proportions seropositive estimated through classification with the mixture models, using Eq. (7.3.38). As previously mentioned, the best prevalence model for the PED is the nonparametric with product-beta prior. Since this model does not assume any deterministic function, but rather estimates the prevalence in each age group separately, it results being more flexible than the other models, also as regards the FOI. As concerns the parameter δ for this model, it should generally be a value larger than the maximal proportion seropositive in the sample. However, since we do not know in advance these proportions seropositive (which are estimated from the mixture model), we decided to set its value equal to 0.9. However, looking at the graphs in Figures 7.4 and 7.5, we notice that the value was probably too low, most of all for Belgium. The piecewise-constant FOI model fits similarly to the nonparametric model, while the log-logistic model shows a bad fitting to the data in almost the whole age range. The nonparametric and the piecewise-constant FOI models fit quite adequately the data in the first 25 years, then they miss the drop in the prevalence, which however is very difficult to catch, considering the increasing monotonic constraints on our curves. In the last part of the age range, both models provide an upward trend, more marked under the nonparametric model. We notice that this upward pattern is different from the one predicted by the MSIRW model of Goeyvaerts et al. (2011), which predicts a decrease in prevalence after age 35 due to waning of antibody protection.

As concerns the FOI, the nonparametric model shows clearly the different peaks in the force of infection during the first 10 years of age, differently from the the piecewise-constant FOI model, where these peaks are lower, because averaged over the ages within the age groups. Conversely, the log-logistic model, because of its shape, results the less flexible model and assumes a peak at age zero, followed by an exponential decline.

Table 7.1: Posterior mean with 95% CI for the parameters $\boldsymbol{\xi}_k$ of the age-dependent mixture models, according to the different distributions, for Belgium and Italy.

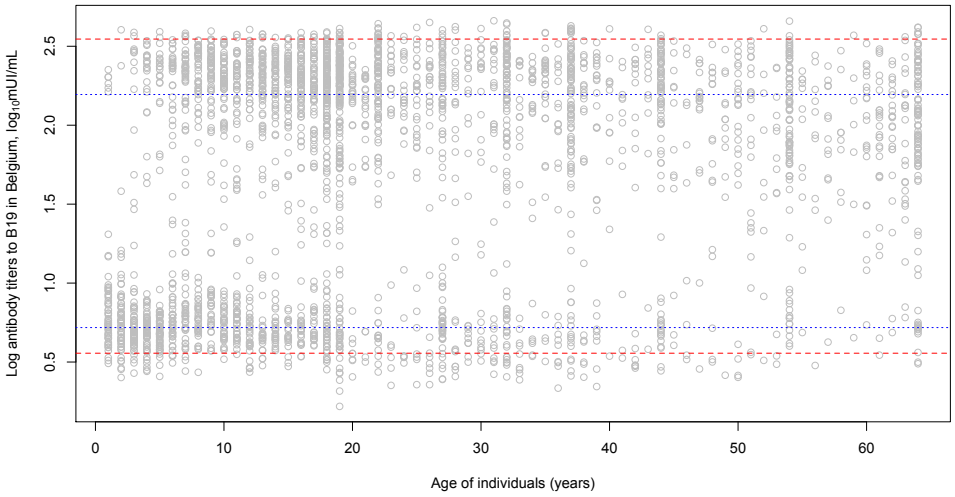
Density	$\boldsymbol{\xi}$	Log-logistic		Piecewise-constant FOI		Product-beta	
		Belgium	Italy	Belgium	Italy	Belgium	Italy
		Post. Mean 95% CI	Post. Mean 95% CI	Post. Mean 95% CI	Post. Mean 95% CI	Post. Mean 95% CI	Post. Mean 95% CI
normal	μ_1	0.72 (0.71,0.73)	0.86 (0.85,0.88)	0.72 (0.71,0.73)	0.86 (0.85,0.88)	0.72 (0.71,0.73)	0.86 (0.85,0.88)
	μ_2	2.19 (2.18,2.21)	2.05 (2.04,2.07)	2.19 (2.18,2.21)	2.06 (2.04,2.07)	2.19 (2.18, 2.21)	2.05 (2.04,2.07)
	σ_1	0.17 (0.16,0.18)	0.23 (0.22,0.24)	0.17 (0.16,0.18)	0.23 (0.22,0.24)	0.17 (0.16, 0.18)	0.23 (0.22,0.24)
	σ_2	0.29 (0.28,0.30)	0.23 (0.22,0.25)	0.29 (0.28,0.30)	0.23 (0.22,0.24)	0.29 (0.28, 0.30)	0.23 (0.22,0.25)
Student's t	μ_1	0.71 (0.70,0.72)	0.87 (0.86,0.88)	0.71 (0.70,0.72)	0.87 (0.86,0.88)	0.71 (0.70,0.72)	0.87 (0.86,0.88)
	μ_2	2.24 (2.22,2.25)	2.07 (2.05,2.08)	2.24 (2.22,2.25)	2.07 (2.05,2.09)	2.24 (2.22, 2.25)	2.07 (2.05,2.09)
	σ_1	0.13 (0.12,0.14)	0.20 (0.19,0.22)	0.13 (0.12,0.14)	0.20 (0.19,0.22)	0.13 (0.12, 0.14)	0.20 (0.19,0.21)
	σ_2	0.23 (0.22,0.25)	0.22 (0.20,0.23)	0.23 (0.22,0.25)	0.22 (0.20,0.23)	0.23 (0.22, 0.25)	0.22 (0.20,0.23)
	ν_1	4.55 (3.05,7.05)	6.28 (4.14,10.17)	4.27 (2.93,6.68)	6.13 (4.03,9.70)	4.31 (2.94, 6.72)	6.09 (4.07,9.67)
	ν_2	5.17 (3.75,7.44)	31.03 (8.38,100.86)	5.27 (3.82,7.62)	32.60 (8.54,109.59)	5.25 (3.80,7.48)	30.13 (7.95,105.74)
Skew-normal	μ_1	0.56 (0.54,0.58)	1.05 (1.02,1.08)	0.56 (0.53,0.58)	1.05 (1.03,1.08)	0.56 (0.53,0.58)	1.05 (1.02,1.08)
	μ_2	2.55 (2.54,2.55)	2.34 (2.33,2.36)	2.55 (2.54,2.55)	2.34 (2.33,2.36)	2.55 (2.54,2.55)	2.34 (2.33,2.36)
	ω_1	0.22 (0.20,0.25)	0.30 (0.27,0.32)	0.23 (0.20,0.25)	0.30 (0.28,0.32)	0.22 (0.20,0.25)	0.29 (0.27,0.32)
	ω_2	0.47 (0.45,0.48)	0.41 (0.39,0.43)	0.46 (0.45,0.48)	0.41 (0.39,0.43)	0.47 (0.45,0.49)	0.41 (0.39,0.44)
	α_1	1.87 (1.41,2.35)	-1.81 $(-2.29, -1.36)$	1.90 (1.42,2.40)	$-1.85(-2.28, -1.44)$	1.88 (1.35,2.43)	-1.79 $(-2.27, -1.34)$
	α_2	-7.49 $(-8.66, -6.39)$	-4.74 $(-5.62, -3.97)$	-7.44 $(-8.57, -6.39)$	$-4.78(-5.59, -4.05)$	-7.49 $(-8.65, -6.44)$	-4.79 $(-5.67, -4.04)$
Skew-t	μ_1	0.59 (0.56,0.64)	1.03 (0.99,1.06)	0.59 (0.56,0.63)	1.03 (0.98,1.06)	0.59 (0.56,0.63)	1.02 (0.96,1.06)
	μ_2	2.54 (2.52,2.55)	2.33 (2.31,2.35)	2.54 (2.52,2.55)	2.33 (2.31,2.35)	2.54 (2.52,2.55)	2.33 (2.31,2.35)
	ω_1	0.17 (0.15,0.21)	0.27 (0.23,0.30)	0.18 (0.15,0.21)	0.26 (0.22,0.30)	0.18 (0.15,0.21)	0.26 (0.21,0.30)
	ω_2	0.40 (0.38,0.43)	0.37 (0.33,0.41)	0.40 (0.38,0.43)	0.36 (0.32,0.40)	0.40 (0.38,0.43)	0.36 (0.33,0.40)
	ν_1	11.94 (5.16,33.06)	31.74 (9.41,88.15)	11.58 (5.03,30.23)	29.98 (7.51,97.24)	10.65 (24.97,31.26)	20.48 (6.97,53.48)
	ν_2	6.50 (4.58,9.70)	6.24 (3.80,11.43)	6.66 (4.70,9.98)	5.20 (3.39,8.50)	7.02 (4.80,11.18)	5.65 (3.69,9.30)
	α_1	1.22 (0.65,1.81)	-1.77 $(-2.36, -1.10)$	1.26 (0.69,1.79)	-1.67 $(-2.33, -1.01)$	1.26 (0.69,1.82)	-1.54 $(-2.20, -0.78)$
	α_2	-6.43 $(-7.73, -5.29)$	-4.30 $(-5.35, -3.43)$	-6.43 $(-7.85, -5.23)$	-4.24 $(-5.22, -3.40)$	-6.42 $(-7.70, -5.27)$	-4.17 $(-5.06, -3.37)$

Table 7.2: PED for model selection. The focus of the analysis is the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\xi}, \boldsymbol{\pi}, \boldsymbol{Z})$.

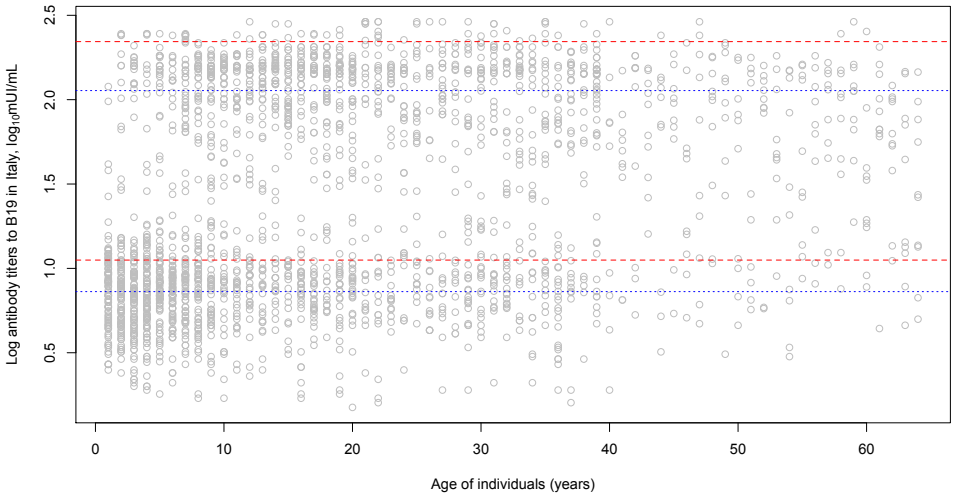
Country	pdf	Log-logistic	Piecewise-constant FOI	Product-beta
Belgium	Normal	4614	4519	4228
	Student's t	15137	15369	15185
	Skew-normal	12636	12630	12433
	Skew-t	17859	17884	18145
Italy	Normal	6057	6347	5567
	Student's t	12789	12486	13046
	Skew-normal	7077	7216	7195
	Skew-t	11072	12126	10378

Table 7.3: Difference in posterior deviance for model selection.

Country	Prev. Model	Distribution	Diff. $\overline{D(\boldsymbol{\theta})}$	P(diff<-4.39)
Belgium	Log-logist.	Student's t vs. Normal	$10(-95, 86)$	0.36
		Skew-normal vs. Normal	$-7241(-7841, -6631)$	1
		Skew-normal vs. Skew-t	$-559(-1501, 386)$	0.87
	P.-const. FOI	Student's t vs. Normal	$9(-53, 82)$	0.37
		Skew-normal vs. Normal	$-7236(-7828, -6652)$	1
		Skew-normal vs. Skew-t	$-542(-1528, 410)$	0.86
	Prod.-beta prior	Student's t vs. Normal	$12(-53, 90)$	0.34
		Skew-normal vs. Normal	$-7240(-7839, -6628)$	1
		Skew-normal vs. Skew-t	$-545(-1442, 358)$	0.87
Italy	Log-logist.	Student's t vs. Normal	$52(-12, 129)$	0.044
		Skew-normal vs. Normal	$-3450(-3897, -2983)$	1
		Skew-normal vs. Skew-t	$-64(-849, 687)$	0.56
	P.-const. FOI	Student's t vs. Normal	$53(-12, 130)$	0.047
		Skew-normal vs. Normal	$-3496(-3930, -3091)$	1
		Skew-normal vs. Skew-t	$-184(-845, 479)$	0.70
	Prod.-beta prior	Student's t vs. Normal	$58(-12, 136)$	0.040
		Skew-normal vs. Normal	$-3442(-3905, -2969)$	1
		Skew-normal vs. Skew-t	$-240(-999, 499)$	0.73
Country	Distribution	Prev. Model	Diff. $\overline{D(\boldsymbol{\theta})}$	P(diff<-4.39)
Belgium	Skew-normal	Piecewise-constant FOI vs. Log-logistic	$4(-817, 820)$	0.49
		Prod.-beta prior vs. P.-const. FOI	$-2(-786, 829)$	0.51
		Prod.-beta prior vs. Log-logist.	$1(-865, 868)$	0.50
Italy	Skew-normal	Piecewise-constant FOI vs. Log-logistic	$-44(-721, 573)$	0.54
		Prod.-beta prior vs. P.-const. FOI	$57(-557, 678)$	0.42
		Prod.-beta prior vs. Log-logist.	$13(-679, 619)$	0.47

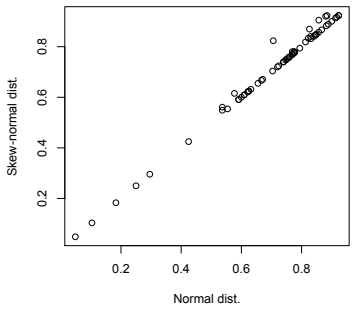


(a) Belgium

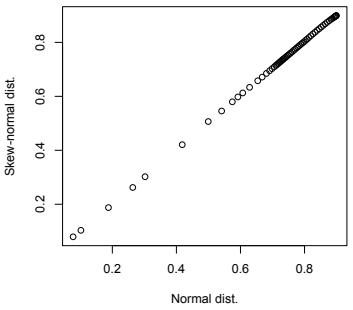


(b) Italy

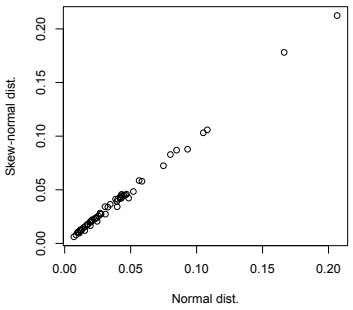
Figure 7.2: Scatter plot of the log antibody titers to parvovirus B19 in Belgium and Italy by age with over imposed the posterior means of the location parameters μ_k, $k = 1,2$, identifying the susceptible component (lower component), and the infected component (upper component), according to age-dependent mixtures with normal (blue dashed line) and skew-normal distribution (red dashed line).



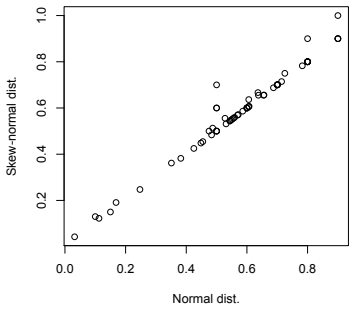
(a) Proportions seropositive in Belgium



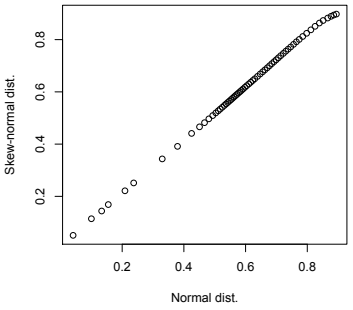
(b) Population prevalence in Belgium



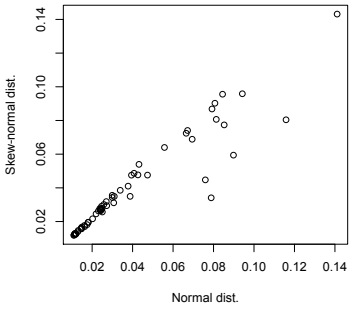
(c) FOI in Belgium



(d) Proportions seropositive in Italy



(e) Population prevalence in Italy



(f) FOI in Italy

Figure 7.3: Comparison between the skew-normal and the normal distribution for proportions seropositive (panels a and d), the estimates of the population prevalence (panels b and e), and FOI (panels c and f), in Belgium and in Italy, using the nonparametric model with product-beta prior.

7.5 Simulation Study

7.5.1 Simulation Setting

A simulation study was conducted in order to gain more insight in the performance of the hierarchical Bayesian two-components mixture models. The aim of the simulation study was to investigate the impact of misspecification of the density on the estimates of the prevalence and FOI and the determination of the current status of the infection for each individual.

Firstly, we simulated data from a two-component mixture of normal densities and then we fitted to these data a two-component hierarchical Bayesian mixture model for the estimation of prevalence and FOI, with the following densities: normal, Student's t, skew-normal and skew-t. The true prevalence and FOI were obtained from a piecewise-constant FOI model estimated from current status data. We performed the simulation for six settings given by different combinations of location and scale parameters: we fixed the mean of the susceptible component ($\mu_1 = 1.5$) and then we considered three different locations for the immune component, ($\mu_2 = (2.5, 3.5, 4.5)$). For the scale parameter, we considered two settings, namely, $\sigma = (0.3, 0.5)$ and $\sigma = (0.5, 0.75)$.

A second setting was considered in which data were generated from a two-component skew-normal mixture model. In this case we use $\alpha = (1.5, -8)$ as skewness parameters. Hence, we assume low positive skewness in the susceptible component and large negative skewness in the immune component.

For each setting, a hundred data sets were generated and analyzed. Since our main interest is to investigate the performance of the mixture model in terms of correct classification of the infection status, the following performance measures were monitored: the accuracy, the sensitivity and the specificity (Lalkhen and McCluskey, 2008). The sensitivity, which is related to the ability to correctly identifying seropositive cases, is given by the ratio between the number of true positive cases (TP) and the total number of positive cases (given by the sum of true positive and false negative, $TP+FN$):

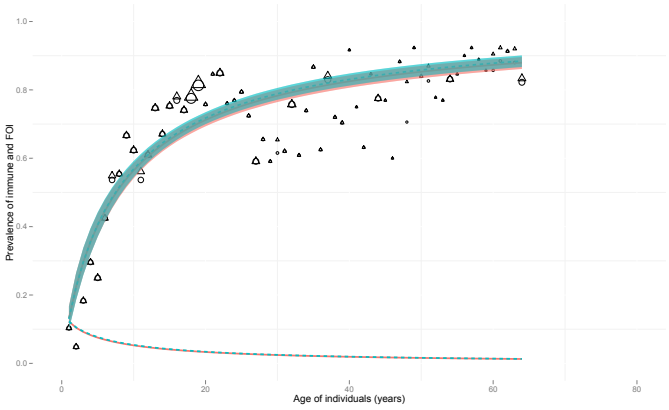
$$\text{Sensitivity} = \frac{TP}{TP+FN}. \tag{7.5.39}$$

The specificity, which is related to the ability of correctly identifying seronegative cases, is given by the ratio between the number of true negative cases (TN) and the total number of negative cases (given by the sum of true negative and false positive, $TN+FP$):

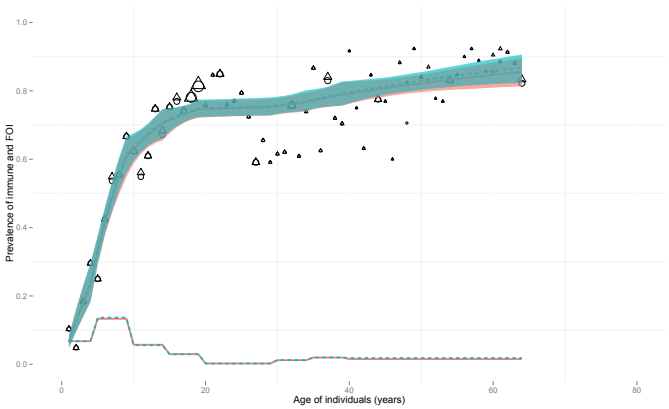
$$\text{Specificity} = \frac{TN}{TN+FP}. \tag{7.5.40}$$

Finally, the accuracy, which is related to the ability of correctly classifying the cases, is given by the ratio between the total number of cases correctly classified by the mixture model ($TP+TN$) and the total number of cases ($TP+TN+FP+FN$):

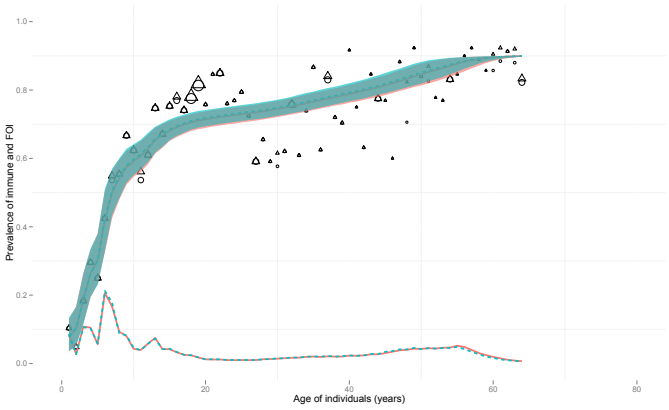
$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}. \tag{7.5.41}$$



(a) Log-logistic model

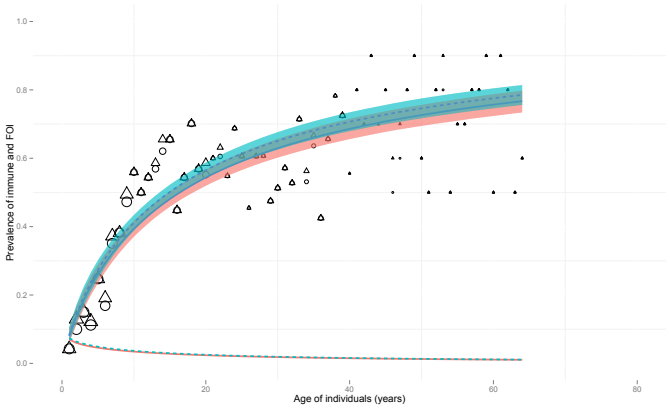


(b) Piecewise-constant FOI model

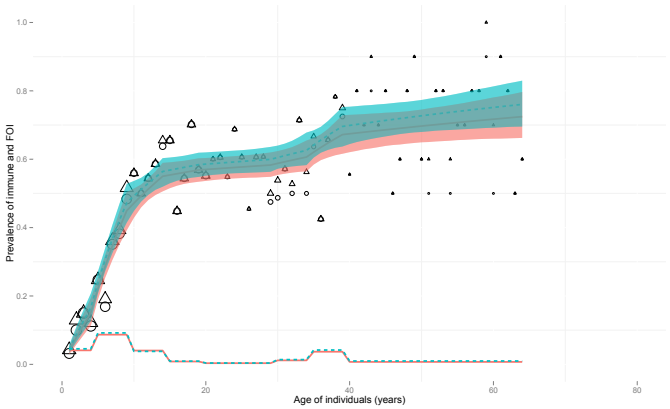


(c) Nonparametric beta-product model

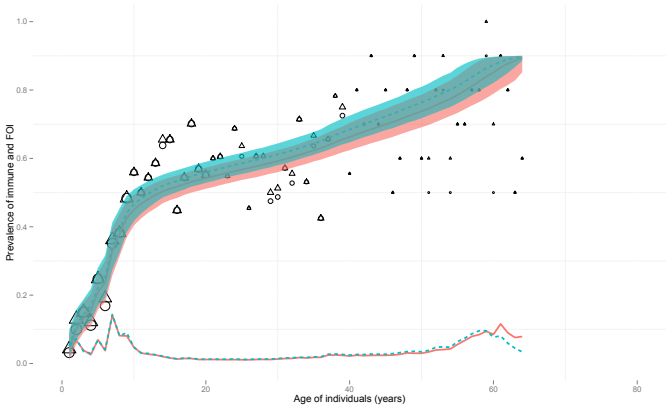
Figure 7.4: Posterior means of prevalence and FOI (with 95% pointwise CIs) for parvovirus B19 in Belgium from the normal mixture (red solid curve) and from the skew-normal mixture (blue dashed curve), with over imposed the proportions seropositive estimated through classification with the normal mixture (circles) and the skew-normal mixture (triangles).



(a) Log-logistic model



(b) Piecewise-constant FOI model



(c) Nonparametric beta-product model

Figure 7.5: Posterior means of prevalence and FOI (with 95% pointwise CIs) for parvovirus B19 in Italy from the normal mixture (red solid curve) and from the skew-normal mixture (blue dashed curve), with over imposed the proportions seropositive estimated through classification with the normal mixture (circles) and the skew-normal mixture (triangles).

7.5.2 Results of the Simulation Study

Tables 7.4 –7.7 summarize the results of the simulation study, presenting the average over 100 datasets of the accuracy, specificity and sensitivity of the classification procedure using the mixture model. In each table, the data, which are generated by a two-component mixture with normal and skew-normal distribution, are fitted with two-components mixture with a different distribution.

In general, also under misspecification of the data distribution, the mixture model reports very high accuracy, specificity and sensitivity, usually close to 100%, implying that the misspecification of the data distribution does not affect the estimation of the current infection status and the prevalence.

We notice that when the means of the components are relatively closed, $\mu = (1.5, 2.5)$, and for relatively high variability, $\sigma = (0.5, 0.75)$, there is a reduction in both specificity and efficiency. This result was expected since, in the case that the two components are not well separated, i.e., the distribution of the susceptible and the immune components are overlapping, the mixture model will not be able to identify correctly the true disease status of individuals in the overlapping area.

Table 7.4: Measures of test performance of a two-component normal mixture model under different settings for μ and σ, with data generated from a normal and a skew-normal mixture, respectively. Results were averaged out over 100 simulated data sets.

Distribution	σ	μ	Accuracy	Specificity	Sensitivity
Normal	(0.3,0.5)	(1.5,2.5)	0.930 (0.902–0.939)	1 (0.999–1)	1 (1–1)
		(1.5,3.5)	0.996 (0.994–0.997)	1 (1–1)	1 (1–1)
		(1.5,4.5)	0.999 (0.999–1)	1 (1–1)	1 (1–1)
	(0.5,0.75)	(1.5,2.5)	0.807 (0.777–0.873)	0.803 (0.585–1)	1 (1–1)
		(1.5,3.5)	0.964 (0.959–0.969)	1 (1–1)	1 (1–1)
		(1.5,4.5)	0.994 (0.992–0.997)	1 (1–1)	1 (1–1)
Skew-normal	(0.3,0.5)	(1.5,2.5)	0.821 (0.764–0.871)	0.848 (0.563–1)	0.994 (0.912–1)
		(1.5,3.5)	0.988 (0.983–0.992)	1 (0.999–1)	0.999 (0.996–1)
		(1.5,4.5)	1 (0.999–1)	1 (1–1)	1 (0.999–1)
	(0.5,0.75)	(1.5,3.5)	0.886 (0.859–0.907)	1 (1–1)	0.979 (0.960–0.992)
		(1.5,4.5)	0.985 (0.980–0.989)	1 (0.999–1)	0.998 (0.996–1)

Table 7.5: Measures of test performance of a two-component Student's t mixture model under different settings for μ and σ, with data generated from a normal and a skew-normal mixture, respectively. Results were averaged out over 100 simulated data sets.

Distribution	σ	μ	Accuracy	Specificity	Sensitivity
Normal	(0.3,0.5)	(1.5,2.5)	0.932 (0.926–0.939)	1 (1–1)	1 (1–1)
		(1.5,3.5)	0.996 (0.994–0.997)	1 (1–1)	1 (1–1)
		(1.5,4.5)	1 (1–1)	1 (1–1)	1 (1–1)
	(0.5,0.75)	(1.5,2.5)	0.816 (0.800–0.880)	0.849 (0.613–1)	1 (1–1)
		(1.5,3.5)	0.964 (0.959–0.969)	1 (1–1)	1 (1–1)
		(1.5,4.5)	0.994 (0.992–0.996)	1 (1–1)	1 (1–1)
Skew-normal	(0.3,0.5)	(1.5,3.5)	0.988 (0.985–0.993)	1 (1–1)	0.999 (0.997–1)
	(0.5,0.75)	(1.5,4.5)	0.985 (0.981–0.989)	1 (1–1)	0.999 (0.996–1)

Table 7.6: Measures of test performance of a two-component skew-normal mixture model under different settings for μ and σ, with data generated from a normal and a skew-normal mixture, respectively. Results were averaged out over 100 simulated data sets.

Distribution	σ	μ	Accuracy	Specificity	Sensitivity
Normal	(0.3,0.5)	(1.5,2.5)	0.931 (0.923–0.938)	1 (1–1)	1 (1–1)
		(1.5,3.5)	0.996 (0.994–0.997)	1 (1–1)	1 (1–1)
		(1.5,4.5)	1 (1–1)	1 (1–1)	1 (1–1)
	(0.5,0.75)	(1.5,2.5)	0.798 (0.776–0.831)	0.764 (0.566–1)	1 (1–1)
		(1.5,3.5)	0.963 (0.956–0.968)	1 (1–1)	1 (1–1)
		(1.5,4.5)	0.994 (0.992–0.996)	1 (1–1)	1 (1–1)
Skew-normal	(0.3,0.5)	(1.5,2.5)	1 (1–1)	1 (1–1)	1 (1–1)
		(1.5,3.5)	1 (1–1)	1 (1–1)	1 (1–1)
		(1.5,4.5)	1 (0.999–1)	1 (1–1)	1 (1–1)
	(0.5,0.75)	(1.5,2.5)	0.998 (0.996–0.999)	1 (1–1)	1 (0.999–1)
		(1.5,3.5)	1 (1–1)	1 (1–1)	1 (1–1)
		(1.5,4.5)	1 (0.999–1)	1 (1–1)	1 (1–1)

Table 7.7: Measures of test performance of a two-component skew-t mixture model under different settings for μ and σ, with data generated from a normal and a skew-normal mixture, respectively. Results were averaged out over 100 simulated data sets.

Distribution	σ	μ	Accuracy	Specificity	Sensitivity
Normal	(0.3,0.5)	(1.5,2.5)	0.930 (0.922–0.937)	1 (1–1)	1 (1–1)
		(1.5,3.5)	0.996 (0.994–0.997)	1 (1–1)	1 (1–1)
		(1.5,4.5)	1 (0.998–1)	1 (1–1)	1 (1–1)
	(0.5,0.75)	(1.5,2.5)	0.806 (0.774–0.872)	0.792 (0.568–1)	1 (1–1)
		(1.5,3.5)	0.962 (0.957–0.967)	1 (1–1)	1 (1–1)
		(1.5,4.5)	0.994 (0.992–0.996)	1 (1–1)	1 (1–1)
Skew-normal	(0.3,0.5)	(1.5,2.5)	1 (0.999–1)	1 (1–1)	1 (1–1)
		(1.5,3.5)	0.999 (0.988–1)	1 (1–1)	1 (1–1)
		(1.5,4.5)	0.995 (0.957–1)	1 (1–1)	1 (1–1)
	(0.5,0.75)	(1.5,2.5)	0.998 (0.996–0.999)	1 (1–1)	1 (1–1)

7.6 Discussion

In this chapter, we introduced a new methodology, based on hierarchical Bayesian mixture models, in order to estimate the prevalence and the FOI of an infection directly from the antibody titers, rather then using the binary current status data obtained from the antibodies using a fixed cut-off point. Compared to standard methods to estimate prevalence and FOI based on cut-off points, mixture model present many advantages.

The main advantage of mixture models is that they use all available information present in the antibody titers for the estimation of the prevalence and the FOI. We have shown that different types of age-specific models for the prevalence and the FOI can be incorporated in the hierarchical Bayesian mixture model: the log-logistic model is a GLM model with a completely deterministic shape; the model with piecewise-constant FOI is another GLM that, however, has a higher degree of flexibility than the log-logistic model, because it assumes independent constant regimes of FOI for each age groups. Finally,

the model with product-beta prior distributions does not assume any deterministic relation for the age, but rather yields order-constrained prevalence estimates for each age groups, separately. Note that not only the age, but also other information available with the data can be included in the model as covariates for the mixture parameters and/or the mixture probabilities. See, for example, Evans and Erlandson (2004); Ødegård Et al. (2005); Nielsen et al. (2007); Hardelid et al. (2008).

We have also shown that several distributions can be used to model the data. In this chapter we compared the performances of the normal, the Student's t, the skew-normal and the skew-t distributions. However, other distributions might be assumed for the data, e.g., the Gamma distribution. For parvovirus B19 data we found that the estimates of the current status infection and of the prevalence are not very different from each other under the four data distributions. In general, we expect this phenomenon to occur (1) when the infection is in a pre-vaccination status, thus only two components are necessary, one for susceptible and one for immune, (2) the mean of the two components are distant enough, and (3) the variances of the components are not too large compared to their respective means. The importance of the two latter conditions becomes evident after the simulation study we conducted, which reveals that the misspecification of the distribution has no effect on the classification of individuals and on the estimation of the prevalence, when these conditions are respected.

Another advantage of mixture models is that they have a higher sensitivity than conventional cut-off methods, which instead tend to be built with a higher specificity. This property is shown also by the simulation study, from which we see that the sensitivity is always equal to 100%. This higher ability of correctly classifying the seropositive cases is more appropriate for a serological study, whose aim is essentially to estimate the prevalence in the population. In this way, using a mixture model, we do not have issues of underestimation of the prevalence any longer, as it occurred with methods based on the fixed cut-off point (Greiner et al., 1994; Vyse et al., 2004, 2006). For a future research, it would be interesting to investigate, through a simulation study, how effectively the age-dependent mixture model performs better than the conventional cut-off point in classifying the individuals between susceptible and immune, similarly to what has been done by Bollaerts et al. (2012).

Furthermore, the mixture approach has the advantage of using the whole serological datum: since the method is able to classify any individual in the components, we do not have issues with equivocal cases any longer, as it usually happens using the conventional couple of cut-off points (measles, varicella, parvovirus B19, etc.). However, note that, if we chose to assign individuals to components when their posterior probability exceeds a certain threshold bigger than 0.5, say 0.6, that would imply that we are not able anymore to classify all the individuals, but rather we should categorize some of them as equivocal.

Using a hierarchical Bayesian approach with MCMC methods, it is possible to fit quite complex mixture models. However, an open problem in the application of Bayesian methods to mixture models is the assessment of the goodness-of-fit of the models and the model selection. Considering that the DIC cannot be used for model selection with mixture models (Celeux et al., 2006), we used two alternative selection criteria, namely, the PED and the difference in posterior deviance. However, as it may happen also with other selection criteria under the frequentist approach, these two criteria lead, in our case, to

different conclusions about the best data distribution and the best prevalence model. The PED used in Table 7.2 for the selection of the best distribution and the mean structure of the prevalence is in contrast with the original use of Plummer (2008), who developed the criterion for the selection of the number of components. This may be the reason of the misleading results. An investigation of the performance of different criteria for model selection is outside the scope of this thesis, but it surely is an important issue that necessitates further research.

In conclusion, if interest lies in describing the serology and not in the transmission process of the infection, we believe that mixture models are the best way to estimate the prevalence and the FOI, because they are able to characterize the entire distribution of the antibody titers, taking into account the heterogeneity in the individual antibody responses.

Chapter 8

Modeling Antibody Titers Using Bayesian Mixture Model with Age-Dependent Mean of the Immune Component

8.1 Introduction

In Chapter 7 we described the estimation method for the prevalence and the FOI of an infection by modeling directly the antibody titers, using a hierarchical Bayesian mixture model. For such model, we assumed that the location and the scale parameters of the mixture components are constant, whereas the mixture weights vary according to the age of individuals.

However, looking at Figures 7.4 and 7.5 in Chapter 7, we notice that the estimated prevalence models for parvovirus B19 are not able to capture the drop in prevalence observed in the young adult group, between age 20 and 40 (Goeyvaerts et al., 2011). Therefore, in this chapter, we relax the assumption of a constant location parameter for the immune component. Our purpose is to investigate whether the waning of natural immunity for parvovirus B19 and for VZV, even of relatively low magnitude, can be responsible for the drop observed in prevalence between age 20 and 40. The idea is that a decrease in the mean antibody level of the immune component, which indicates an average diminution in the antibody titers, can affect the mean structure of the prevalence. This chapter is structured in the following way. In Section 8.2 we introduce the methodology used to estimate and test for an age-dependent profile of the mean of the immune, i.e., the Bayesian Variable Selection (BVS) approach (George and McCulloch, 1993; O'Hara and Sillanpää, 2009). In Section 8.3 we apply the methodology to antibody titers for B19

and VZV from five European countries, namely, Belgium (BE), Finland (FI), Italy (IT), Poland (PL), and Great Britain (GB) (Nardone et al., 2007; Mossong et al., 2008). Finally, in Section 8.4, a short discussion of the methodology and of the results is presented.

8.2 Methodology

Given a sample of individuals aged a_i, whose antibody titers are denoted with Y_i, $i = 1, \ldots, n$, we consider a two-component hierarchical Bayesian mixture model for the antibody data Y_i, whose distribution g with age-dependent probabilities $\pi(a_i)$ is given by

$$g(Y_i) = \sum_{k=1}^{2} \pi_k(a_i) f(\boldsymbol{\xi}_k), \tag{8.2.1}$$

where $f(\cdot)$ is a specific pdf and $\boldsymbol{\xi}_j$ is the mixture parameter vector. The component with the parameter vector $\boldsymbol{\xi}_1$ accounts for the antibody data of the susceptible individuals, while the component with $\boldsymbol{\xi}_2$ accounts for the immune individuals. For the mean structure of the prevalence, we specify a piecewise-constant FOI model, as discussed in Section 7.3.2.2, but other mean structures are possible as well.

As shown in Section 7.3.2, a two-component Bayesian normal mixture model is given by

$$g(Y_i) = (1 - \pi(a_i))N(\mu_1, \sigma_1^2) + \pi(a_i)N(\mu_2(a_i), \sigma_2^2). \tag{8.2.2}$$

Note that the mixture model (8.2.2) assumes that the mean antibody level of the susceptible component, μ_1, is constant, while the mean antibody level of the immune component, $\mu_2(a_i)$ (from now on, $\mu^I(a_i)$) can change according to the age of individuals. We specify a piecewise-constant model for the mean antibody level of the immune, $\mu^I(a)$ and we assume that the mean antibody level of the immune is constant within each of the T age groups (a_{t-1}, a_t) considered:

$$\mu^I(a_i) = \mu_1^I + \sum_{t=2}^{T} \mu_t^I, \text{ for } a_i \in (a_{t-1}, a_t). \tag{8.2.3}$$

Using a matrix notation, we can express the mean antibody level in the immune component as $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$, where $\boldsymbol{X}$ is a known design matrix of size $n \times T$ (given below), and $\boldsymbol{\beta}$ is a parameter vector,

$\boldsymbol{\beta} = (\mu_1^I, \ldots, \mu_T^I)$ to be estimated (Kasim et al., 2012).

$$\boldsymbol{X} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ \ldots & & & & \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ \ldots & & & & \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ \ldots & & & & \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ \ldots & & & & \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ \ldots & & & & . \end{pmatrix} \tag{8.2.4}$$

Since we do not have any *a priori* knowledge about the direction of the changes over age, we use an unrestricted model, that is to say, the age-dependent parameters μ_t^I may either increase or decrease with respect to the preceding value μ_{t-1}^I. Within the Bayesian approach, we need to specify a prior distribution for the parameters μ_t^I, namely,

$$\mu_t^I \sim N(\mu_{\mu_t}, \sigma_{\mu_t t}^2). \tag{8.2.5}$$

To complete the specification, we assume flat hyperprior distributions for the hyperparameters of (8.2.5), i.e., $\mu_{\mu_t} \sim N(0, 1000)$ and $\sigma_{\mu_t}^{-2} \sim \Gamma(0.01, 0.01)$.

8.2.1 The Bayesian Variable Selection Approach

The previous model formulation for $\mu(a)^I$ implies that the mean antibody level in the immune component changes across the age groups. Therefore, to find the best model, a selection criterion such as the PED (Plummer, 2008) can be used. However, the number of possible models is known to increase exponentially with the number of columns in $\boldsymbol{X}$. In particular, for $T = 5$, we would have $2^{T-1} = 16$ different possible models. Thus, a model selection procedure would be computer-intensive and not possible in practice. The Bayesian Variable Selection framework (BVS, George and McCulloch, 1993; O'Hara and

Sillanpää, 2009; Kasim et al., 2012) allows to estimate the posterior probability of each model, and in particular to test the hypothesis that the mean antibody level in the immune component is constant. Even though the models are fitted within the Bayesian context and not within the frequentist one, we are able to compute a posterior probability for each model and, in particular, to use the posterior probability of the null model (constant mean antibody level of the immune) to test the hypothesis that the mean is constant.

We rewrite the model for the mean antibody level of the immune component in (8.2.5) as

$$\mu^I(a_i) = \mu_1^I + \sum_{t=2}^{T} \mu_t^I = \mu_1^I + \sum_{t=2}^{T} z_{t-1}\delta_{t-1}, \text{ for } a_i \in (a_{t-1}, a_t). \tag{8.2.6}$$

where z_t is a latent indicator variable such that

$$z_t = \begin{cases} 1, & \delta_t \text{ is included in the model,} \\ 0 & \delta_t \text{ is not included in the model.} \end{cases} \tag{8.2.7}$$

To complete the specification of the model, we define a prior distributions for z_t and δ_t, i.e., for $t = 1, \ldots, T-1$,

$$\begin{aligned} \delta_t &\sim N(\mu_\delta, \sigma_\delta^2), \\ z_t &\sim \text{Bernoulli}(\pi_t), \\ \pi_t &\sim U(0,1). \end{aligned} \tag{8.2.8}$$

Again, for the hyperparameters of δ_t, we use the following flat priors: $\mu_\delta \sim N(0, 1000)$ and $\sigma_\delta^{-2} \sim \Gamma(0.01, 0.01)$.

The posterior mean of the indicator z_t corresponds to the posterior inclusion probability, which is the probability that a specific δ_t is included in the mean structure of the immune component (O'Hara and Sillanpää, 2009), i.e.,

$$P(z_t = 0) = P(\delta_t \text{ is not included in the model}). \tag{8.2.9}$$

Each quadruplet $\boldsymbol{z} = (z_1, z_2, z_3, z_4)$ defines uniquely each one of the sixteen plausible models. For instance, with $\boldsymbol{z} = (z_1 = 0, z_2 = 0, z_3 = 0, z_4 = 0)$ we obtain a constant mean structure in the immune component $(\mu_1^I, \mu_1^I, \mu_1^I, \mu_1^I, \mu_1^I)$. Instead, with the quadruplet $\boldsymbol{z} = (z_1 = 1, z_2 = 0, z_3 = 0, z_4 = 1)$ we obtain the corresponding mean structure $(\mu_1^I, \mu_1^I + \delta_1, \mu_1^I + \delta_1, \mu_1^I + \delta_1, \mu_1^I + \delta_1 + \delta_4)$, for which the mean of the immune changes in the second age group and then in the last age group as well. In general, using the indicator variable z_t, we can specify a transformation that permits to identify uniquely each model (Ntzoufras,

2002). Let $\boldsymbol{c} = (1,2,4,8)$ and let $\boldsymbol{Z}$ be a $2^{T-1} \times (K-1)$ matrix given by

$$\boldsymbol{Z} = \left(\begin{array}{cccc} 0 & 0 & 0 & 0 \\ \hline 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \hline 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ \hline 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ \hline 1 & 1 & 1 & 1 \end{array} \right). \tag{8.2.10}$$

Then we define the transformation $M_r = 1 + \boldsymbol{Z}\boldsymbol{C}^T$. Each element of the vector M_r corresponds to one of all the possible 16 models, namely,

$$M_r = \begin{cases} 1, & \text{for } \boldsymbol{z} = (z_1 = 0, z_2 = 0, z_3 = 0, z_4 = 0), & \text{null model,} \\ 2, & \text{for } \boldsymbol{z} = (z_1 = 1, z_2 = 0, z_3 = 0, z_4 = 0), & \text{model 1,} \\ 3, & \text{for } \boldsymbol{z} = (z_1 = 0, z_2 = 1, z_3 = 0, z_4 = 0), & \text{model 2,} \\ 5, & \text{for } \boldsymbol{z} = (z_1 = 0, z_2 = 0, z_3 = 1, z_4 = 0), & \text{model 3,} \\ 9, & \text{for } \boldsymbol{z} = (z_1 = 0, z_2 = 0, z_3 = 0, z_4 = 1), & \text{model 4,} \\ 4, & \text{for } \boldsymbol{z} = (z_1 = 1, z_2 = 1, z_3 = 0, z_4 = 0), & \text{model 5,} \\ 6, & \text{for } \boldsymbol{z} = (z_1 = 1, z_2 = 0, z_3 = 1, z_4 = 0), & \text{model 6,} \\ 10, & \text{for } \boldsymbol{z} = (z_1 = 1, z_2 = 0, z_3 = 0, z_4 = 1), & \text{model 7,} \\ 7, & \text{for } \boldsymbol{z} = (z_1 = 0, z_2 = 1, z_3 = 1, z_4 = 0), & \text{model 8,} \\ 11, & \text{for } \boldsymbol{z} = (z_1 = 0, z_2 = 1, z_3 = 0, z_4 = 1), & \text{model 9,} \\ 13, & \text{for } \boldsymbol{z} = (z_1 = 0, z_2 = 0, z_3 = 1, z_4 = 1), & \text{model 10,} \\ 8, & \text{for } \boldsymbol{z} = (z_1 = 1, z_2 = 1, z_3 = 1, z_4 = 0), & \text{model 11,} \\ 12, & \text{for } \boldsymbol{z} = (z_1 = 1, z_2 = 1, z_3 = 0, z_4 = 1), & \text{model 12,} \\ 14, & \text{for } \boldsymbol{z} = (z_1 = 1, z_2 = 0, z_3 = 1, z_4 = 1), & \text{model 13,} \\ 15, & \text{for } \boldsymbol{z} = (z_1 = 0, z_2 = 1, z_3 = 1, z_4 = 1), & \text{model 14,} \\ 16, & \text{for } \boldsymbol{z} = (z_1 = 1, z_2 = 1, z_3 = 1, z_4 = 1), & \text{model 15,} \end{cases} \tag{8.2.11}$$

As pointed out by Ntzoufras (2002), the posterior probability of $\tilde{M}_r = 1 + \sum_{t=1}^{T-1} z_t 2^{t-1}$ defines uniquely each one of the possible models. In particular,

$$P(\tilde{M}_r = 1|\boldsymbol{z}, Y) = P(\boldsymbol{z} = (0,0,0,0)|Y) = P(\text{null model}|\boldsymbol{z}, Y), \tag{8.2.12}$$

where the null model is characterized by $\mu = (\mu_1^I, \mu_1^I, \mu_1^I, \mu_1^I)$, i.e., a constant mean antibody level of the immune component. Hence, the BVS method allows us to select the best model for the mean structure in the immune component and, at the same time, to test the hypothesis that the mean antibody level in the immune component is constant. We notice that $P(\text{null model}|\boldsymbol{z}, Y)$ is the posterior probability that the mean structure in the immune component is constant. Therefore, one can reject the null hypothesis of constant mean structure in the immune component if $P(\text{null model}|\boldsymbol{z}, Y) < \tau$, where τ is a pre-specified threshold level, e.g., $\tau = 0.05$, similar to a significance level $\alpha = 5\%$ in the frequentist approach.

8.3 Application to the Data

The hierarchical Bayesian mixture model discussed above is implemented in JAGS through R. We fitted both normal and skew-normal mixture models and we based model selection on the PED (Plummer, 2008) and the difference in posterior deviances (Aitkin, 2010). Model convergence is assessed using Gelman and Rubin's diagnostic (Gelman and Rubin, 1992). We used the following age groups t to model the mean structure of the immune component: 1–9, 10–19, 20–29, 30–34, 35–39, 40+. The grouping was chosen based on the work of Goeyvaerts et al. (2011), who assumed a MSIRW model with age-dependent antibody waning after age H, whose best value was found to be at 35 years. This model assumes that after recovering from the infection, the antibody protection may wane with age, more likely after age 35 years. Therefore, we tried to use the same value for our model in order to compare our results with those of (Goeyvaerts et al., 2011).

Table 8.1 shows the results for the selection criteria, the PED and the difference in posterior deviance. As it happened in Chapter 7, the two criteria lead to different conclusions about the best data distribution, except for a few cases. According to PED, the normal mixture model outperforms the skew-normal mixture model, except in Great Britain for parvovirus B19, and in Finland and Poland, for VZV. On the contrary, the difference in posterior deviances always favors the skew-normal mixture model. Table 8.1 also shows the posterior mean (with 95% credible interval (CI)) of the skewness parameter α_2 for the mean of the immune: the parameter is always negative, larger for parvovirus B19, indicating that the immune component under the skew-normal mixture model is left-skewed, with a longer right tail and a higher mean than the normal mixture model, with high antibody titers more probable than low titers.

Inference for the age-dependent mean structure of the immune component with the BVS model is shown in Tables 8.2 and 8.3, where the posterior probabilities of the null model and of the most probable model are shown. Under the normal mixture model, we reject the null hypothesis of constant mean structure of the immune for all the five countries, for parvovirus B19, and for Belgium and Italy, for VZV.

Under the skew-normal mixture model, we only reject the null hypothesis of constant mean structure of the immune in Belgium and Italy, for both infections. An explanation to this difference can be found in the negative skewness estimated by the skew-normal mixture model: with a higher mean and a longer right tail than the normal, there is less need for an age-dependent mean μ^I. Figures 8.1 and 8.3 show the scatter plot of the antibody data, with over imposed the mean antibody levels for the susceptible and for the immune from both the normal and the skew-normal mixture models, for parvovirus B19 and VZV, respectively. Tables 8.4 and 8.5 show both the posterior means of the mean antibody level (with 95% CI) obtained from the BVS models and the posterior mean and median for the latent indicator z_t, under normal and skew-normal distribution, respectively. For parvovirus B19, we generally notice, with the normal mixture model, a decrease in the mean antibody level after age 40, except for Poland, that also presents a rise in the antibodies after age 10. For VZV, we observe again a decrease in the mean antibody level, in Belgium after age 10 and in Italy after age 40. Comparing Table 8.2 with Table 8.3 that often the actual model is dissimilar from the most probable model. This happens because the actual is a model averaging of the most probable models: this is visible, for instance, in Finland for VZV infection under the skew-normal, where the final model is averaged between Model 1 and Model 7.

However, these variations in the mean structure of the immune component are relatively small compared to the large variability of the antibody levels, and have a little impact on the estimates of the prevalence and the FOI, which are displayed in Figures 8.2 – 8.4 for parvovirus B19 and VZV, respectively. When looking at the impact of the differences between distributions on the estimates of the prevalence and the FOI, we note that the main differences are found only in Italy, for the prevalence of parvovirus B19 infection, and in Belgium, Italy and Poland, for the FOI of varicella. In the latter case, the age groups affected by the differences are the oldest ones, where the models are more sensitive to the variability in data caused by the small sample sizes.

The prevalence of B19 infection usually grows slower or even plateau in the age groups of teenagers and young adults, before it rises again between age 30 and 35. The FOI has a main peak in the age group [5,10) in Belgium and Italy, and in the age group [10,15) in the other three countries. Afterward, following the plateau between adolescents and young adults, there is usually a second peak in the FOI between age 30 and 40, i.e., in the age group of parents with children attending school. These results are very close to those obtained of Goeyvaerts et al. (2011) under the models assuming waning of immunity to B19 and possible reinfection. For varicella, we observe a different profile. The prevalence of VZV, which is characterized by a higher transmission than parvovirus B19, goes up very fast with a near-linear pattern until age 10, reaching 90%, and afterwards it plateaus, until age 30, when it slightly increases again. The main peak of the FOI is in the age group [5,10), but a second peak can be observed after age 30 years, likely in the age group of parents leaving with their children.

8.4 Discussion

We applied hierarchical Bayesian mixture models, with normal and skew-normal distribution, to antibody data for parvovirus B19 and VZV infection in five European countries, in order to estimate the prevalence and the force of infection and, at the same time, testing the hypothesis of age-dependence of the mean antibody level of the immune component. Under the BVS approach, the age groups chosen for the analysis formed the set of possible covariates. In this way, the selected covariates allowed to describe the age-specific profile of the mean structure of the immune component. The BVS models allow us to calculate the inclusion probability for every covariate in the model and the posterior probability of all the piecewise-constant models that can be fitted to the data. We have shown that, using the posterior probability of the null model, we can test the hypothesis that the mean antibody level in the immune component is constant.

For parvovirus B19, under the two distributions we obtained different models for the mean antibody levels of the immune component. The skew-normal mixture model, which estimated a large negative skewness parameter for the immune component, gave high values for the mean structure of the immune, which compensated for null or relatively small change between the age groups. On the other hand, the normal mixture model yielded a lower mean structure for the mean of the immune with larger variations between the age groups. For the case of BVS model with normal mixture for parvovirus B19, we always rejected the null hypothesis of constant mean antibody level of the immune, observing a slight waning of antibody titers after age 40. Nonetheless, despite these dissimilarities between the skew-normal and the normal distribution, we did not find major differences in the estimates of the prevalence and of the FOI, except for Italy, where the normal mixture model gave a lower prevalence and FOI. Our results for parvovirus B19 are similar to those obtained by Goeyvaerts et al. (2011), using the same data. However, note that, while Goeyvaerts et al. (2011) used a mathematical transmission model for parvovirus B19 infection, assuming age-dependent waning, loss of infection-acquired immunity and possible reinfection, we obtained this prevalence and FOI using a statistical mixture modeling to describe the age-specific immunity patterns given by the antibody data.

For VZV, the situation was somehow different, since we found that the two distributions gave similar results, with the null hypothesis of constant mean structure of the immune component rejected only in Belgium and Italy. In these two countries, we observe a slight decline of the mean antibody levels for the immune, but again with small impact on the prevalence and on the FOI.

Table 8.1: Inference using BVS models: PED for model selection and skewness parameter α_2 of the immune component under skew-normal distribution.

Infection	Country	Distribution	PED	Diff. $\overline{D(\theta)}$	α_2
Parvovirus B19	Belgium	Normal	3875	$D_{SN,N} = -7024(-7748, -6365)$	-
		Skew-normal	12423	P(diff<4.39)=1	$-7.17(-8.63, -5.99)$
	Finland	Normal	-1409	$D_{SN,N} = -5171(-5589, -4755)$	-
		Skew-normal	6950	P(diff<4.39)=1	$-5.85(-6.76, -5.01)$
	Great Britain	Normal	30127	$D_{SN,N} = -4972(-5410, -4522)$	-
		Skew-normal	11784	P(diff<4.39)=1	$-6.02(-6.85, -5.27)$
	Italy	Normal	6881	$D_{SN,N} = -4032(-4994, -3317)$	-
		Skew-normal	9621	P(diff<4.39)=1	$-5.35(-7.48, -4.05)$
	Poland	Normal	3752	$D_{SN,N} = -5268(-5776, -4759)$	-
		Skew-normal	219469	P(diff<4.39)=1	$-3.27(-3.91, -2.71)$
VZV	Belgium	Normal	6132	$D_{SN,N} = -1973(-2416, -1522)$	-
		Skew-normal	9600	P(diff<4.39)=1	$-1.82(-2.13, -1.52)$
	Finland	Normal	724536	$D_{SN,N} = -15106(-16317, -13802)$	-
		Skew-normal	-2032	P(diff<4.39)=1	$-3.96(-4.79, -3.27)$
	Great Britain	Normal	9416	$D_{SN,N} = -2951(-3420, -2483)$	-
		Skew-normal	10756	P(diff<4.39)=1	$-2.86(-3.56, -2.25)$
	Italy	Normal	6924	$D_{SN,N} = -2375(-2834, -1921)$	-
		Skew-normal	9218	P(diff<4.39)=1	$-2.35(-2.83, -1.92)$
	Poland	Normal	5193	$D_{SN,N} = -1057(-1396, -736)$	-
		Skew-normal	5009	P(diff<4.39)=1	$-2.64(-3.43, -1.92)$

Table 8.2: Inference using BVS models. Posterior median for z and posterior probability of the null model.

Infection	Distribution	Country	z	Posterior probability
Parvovirus B19	Normal	Belgium	(0 0 0 0)	0.00
		Finland	(0 0 0 0)	0.00
		Great Britain	(0 0 0 0)	0.00
		Italy	(0 0 0 0)	0.00
		Poland	(0 0 0 0)	0.00
	Skew-Normal	Belgium	(0 0 0 0)	0.00
		Finland	(0 0 0 0)	1.00
		Great Britain	(0 0 0 0)	0.93
		Italy	(0 0 0 0)	0.04
		Poland	(0 0 0 0)	0.06
VZV	Normal	Belgium	(0 0 0 0)	0.00
		Finland	(0 0 0 0)	0.77
		Great Britain	(0 0 0 0)	1.00
		Italy	(0 0 0 0)	0.00
		Poland	(0 0 0 0)	0.99
	Skew-Normal	Belgium	(0 0 0 0)	0.00
		Finland	(0 0 0 0)	0.13
		Great Britain	(0 0 0 0)	1.00
		Italy	(0 0 0 0)	0.00
		Poland	(0 0 0 0)	0.99

Table 8.3: Inference using BVS models. We report the number of the most probable model, the posterior median for $\boldsymbol{z}$, and the posterior probability.

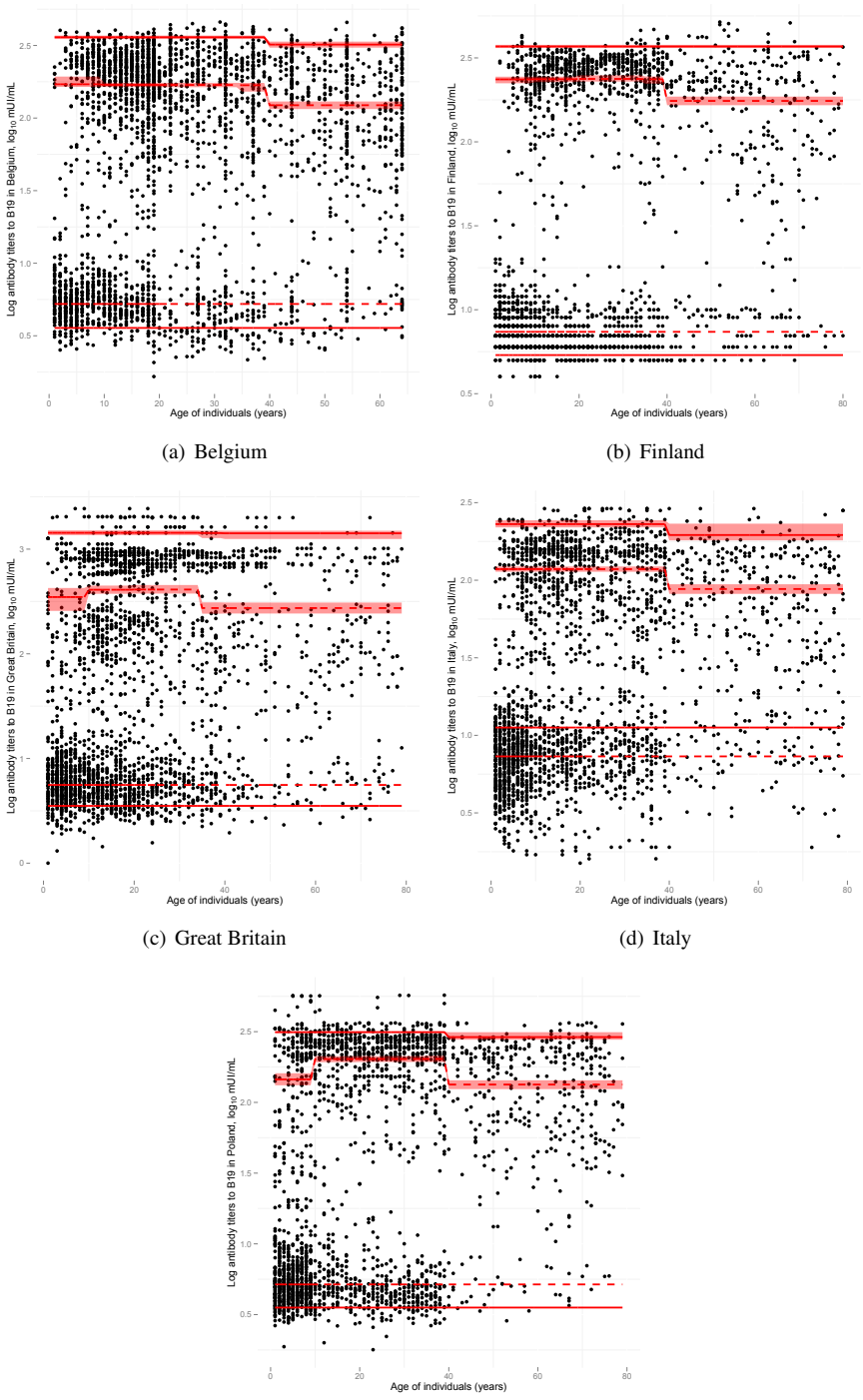
Infection	Distribution	Country	Model	$\boldsymbol{z}$	Posterior probability
Parvovirus B19	Normal	Belgium	4	(0 0 0 1)	0.90
		Finland	4	(0 0 0 1)	0.93
		Great Britain	3	(0 0 1 0)	0.53
		Italy	4	(0 0 0 1)	0.99
		Poland	7	(1 0 0 1)	0.96
	Skew-Normal	Belgium	4	(0 0 0 1)	1.00
		Finland	null	(0 0 0 0)	1.00
		Great Britain	null	(0 0 0 0)	0.93
		Italy	4	(0 0 0 1)	0.95
		Poland	4	(0 0 0 1)	0.93
VZV	Normal	Belgium	2	(0 1 0 0)	0.96
		Finland	null	(0 0 0 0)	0.77
		Great Britain	null	(0 0 0 0)	1.00
		Italy	7	(1 0 0 1)	0.55
		Poland	null	(0 0 0 0)	0.99
	Skew-Normal	Belgium	2	(0 1 0 0)	0.95
		Finland	1	(1 0 0 0)	0.33
		Finland	7	(1 0 0 1)	0.32
		Great Britain	null	(0 0 0 0)	1.00
		Italy	4	(0 0 0 1)	0.86
		Poland	null	(0 0 0 0)	0.99

Table 8.4: Posterior means and 95% CI for $\mu^I(a)$, obtained from the BVS model, using a normal mixture model.

Country	Age group	Parvovirus B19			VZV		
		$\bar{\mu}^I$ (95% CI)	Post. mean z	Post. median z	$\bar{\mu}^I$ (95% CI)	Post. mean z	Post. median z
Belgium	1–9	2.23 (2.21,2.29)	-	-	2.86 (2.82,2.90)	-	-
	10–19	2.23 (2.21,2.24)	0.08	0	2.71 (2.69,2.74)	0.02	0
	20–34	2.23 (2.21,2.24)	0.01	0	2.71 (2.69,2.74)	1.00	1
	35–39	2.23 (2.18,2.24)	0.04	0	2.71 (2.69,2.73)	0.01	0
	40+	2.09 (2.06,2.11)	1.00	1	-	-	-
Finland	1–9	2.37 (2.35,2.39)	-	-	3.07 (3.05,3.14)	-	0
	10–19	2.37 (2.36,2.39)	0.02	0	3.06 (3.04,3.08)	0.18	0
	20–34	2.37 (2.36,2.40)	0.04	0	3.06 (3.04,3.08)	0.01	0
	35–39	2.37 (2.35,2.39)	0.03	0	3.06 (3.04,3.08)	0.03	0
	40+	2.24 (2.22,2.27)	0.00	1	3.06 (3.02,3.07)	0.02	0
Great Britain	1–9	2.54 (2.41,2.63)	-	-	1.83 (1.81,1.84)	-	-
	10–19	2.61 (2.57,2.65)	0.46	0	1.83 (1.81,1.84)	0.00	0
	20–34	2.61 (2.58,2.66)	0.04	0	-	-	-
	35–39	2.44 (2.38,2.49)	1.00	1	-	-	-
	40+	2.44 (2.38,2.49)	0.03	0	-	-	-
Italy	1–9	2.07 (2.05,2.09)	-	-	3.09 (3.03,3.14)	-	-
	10–19	2.07 (2.05,2.09)	0.00	0	3.03 (3.00,3.07)	0.73	1
	20–34	2.07 (2.05,2.09)	0.00	0	3.02 (2.98,3.06)	0.15	0
	35–39	2.07 (2.05,2.09)	0.00	0	3.03 (2.99,3.09)	0.12	0
	40+	1.94 (1.91,1.97)	1.00	1	2.89 (2.85,2.93)	1.00	1
Poland	1–9	2.16 (2.12,2.21)	-	-	2.97 (2.94,3.00)	-	-
	10–19	2.31 (2.28,2.33)	1.00	1	2.97 (2.94,3.00)	0.00	0
	20–34	2.31 (2.29,2.33)	0.02	0	-	-	-
	35–39	2.31 (2.29,2.33)	0.02	0	-	-	-
	40+	2.13 (2.09,2.16)	1.00	1	-	-	-

Table 8.5: Posterior means and 95% CI for $\mu^I(a)$, obtained from the BVS model, using a skew-normal mixture model.

Country	Age group	Parvovirus B19			VZV		
		$\bar{\mu}^I$ (95% CI)	Post. mean z	Post. median z	$\bar{\mu}^I$ (95% CI)	Post. mean z	Post. median z
Belgium	1–9	2.56 (2.54,2.57)	-	-	3.30 (3.25,3.36)	-	-
	10–19	2.56 (2.54,2.57)	0.00	0	3.30 (3.25,3.36)	0.02	0
	20–34	2.56 (2.54,2.57)	0.00	0	3.16 (3.11,3.20)	1.00	1
	35–39	2.56 (2.54,2.57)	0.00	0	3.16 (3.11,3.20)	0.03	0
	40+	2.51 (2.49,2.53)	1.00	1	3.16 (3.11,3.20)	0.01	0
Finland	1–9	2.57 (2.56,2.58)	-	-	3.53 (3.48,3.58)	-	0
	10–19	2.57 (2.56,2.58)	0.00	0	3.49 (3.45,3.53)	0.68	1
	20–34	2.57 (2.56,2.58)	0.00	0	3.49 (3.45,3.53)	0.09	0
	35–39	2.57 (2.56,2.58)	0.00	0	3.49 (3.45,3.53)	0.02	0
	40+	2.57 (2.56,2.58)	0.00	0	3.48 (3.45,3.51)	0.49	0
Great Britain	1–9	3.16 (3.13,3.18)	-	-	2.18 (2.15,2.21)	-	-
	10–19	3.16 (3.13,3.18)	0.00	0	2.18 (2.15,2.21)	0.00	0
	20–34	3.16 (3.13,3.18)	0.00	0	-	-	-
	35–39	3.16 (3.13,3.18)	0.04	0	-	-	-
	40+	3.15 (3.10,3.18)	0.03	0	-	-	-
Italy	1–9	2.36 (2.34,2.39)	-	-	3.49 (3.45,3.54)	-	-
	10–19	2.36 (2.34,2.39)	0.00	0	3.49 (3.44,3.53)	0.06	0
	20–34	2.36 (2.34,2.39)	0.00	0	3.48 (3.43,3.53)	0.10	0
	35–39	2.36 (2.34,2.39)	0.00	0	3.48 (3.43,3.53)	0.02	0
	40+	2.29 (2.26,2.36)	0.96	1	3.33 (3.27,3.38)	1.00	1
Poland	1–9	2.50 (2.49,2.50)	-	-	3.42 (3.46,3.48)	-	-
	10–19	2.50 (2.49,2.50)	0.00	0	3.42 (3.46,3.48)	0.00	0
	20–34	2.50 (2.49,2.50)	0.00	0	-	-	-
	35–39	2.50 (2.49,2.50)	0.00	0	-	-	-
	40+	2.46 (2.44,2.50)	0.94	1	-	-	-

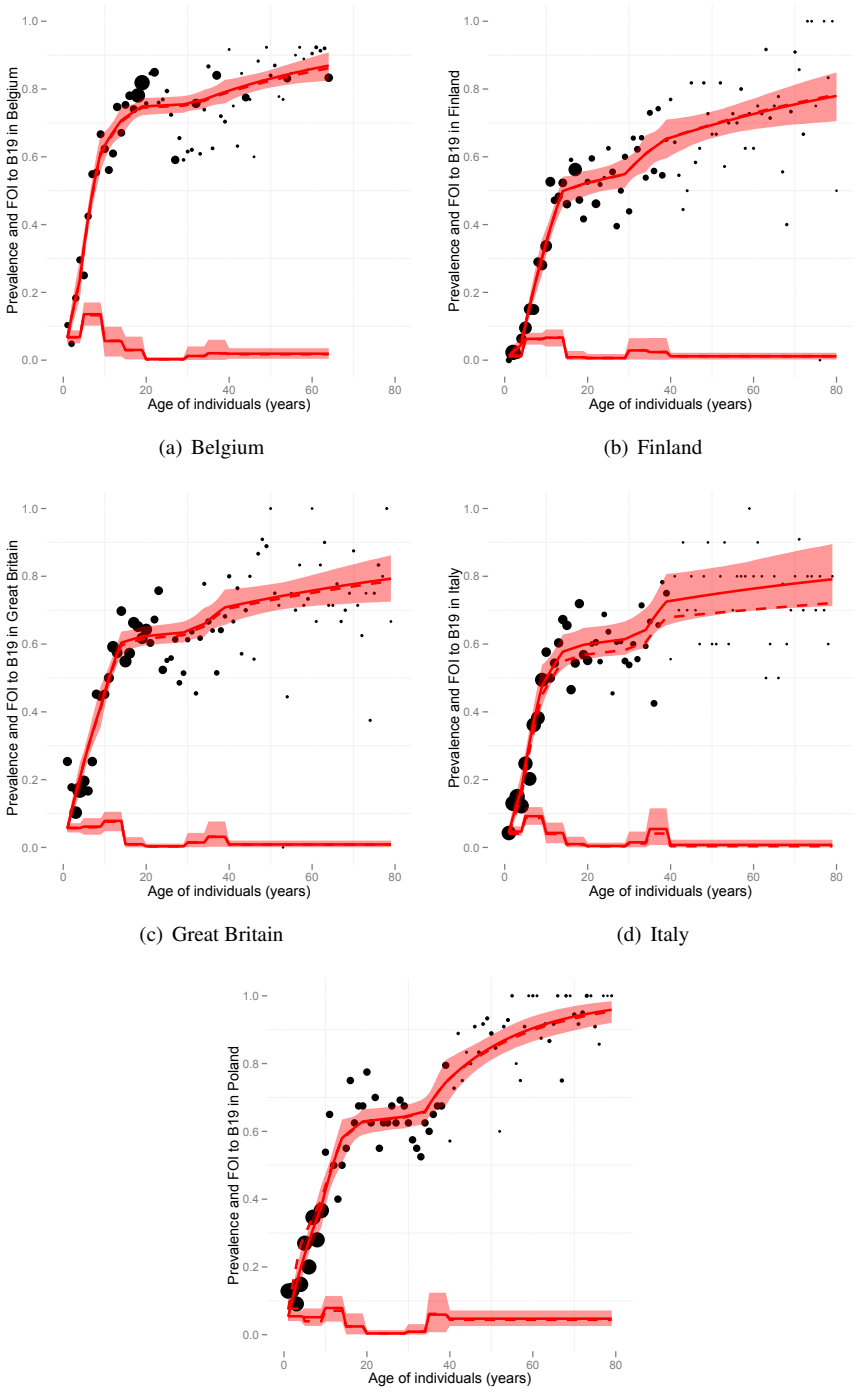


(a) Belgium (b) Finland

(c) Great Britain (d) Italy

(e) Poland

Figure 8.1: Parvovirus B19 in five European countries: scatter plot of the log antibody titers with over imposed the posterior means of the mean of susceptible and of the age-dependent mean of immune (with 95% CI), from a normal mixture (dashed curves) and a skew-normal mixture (solid curves).

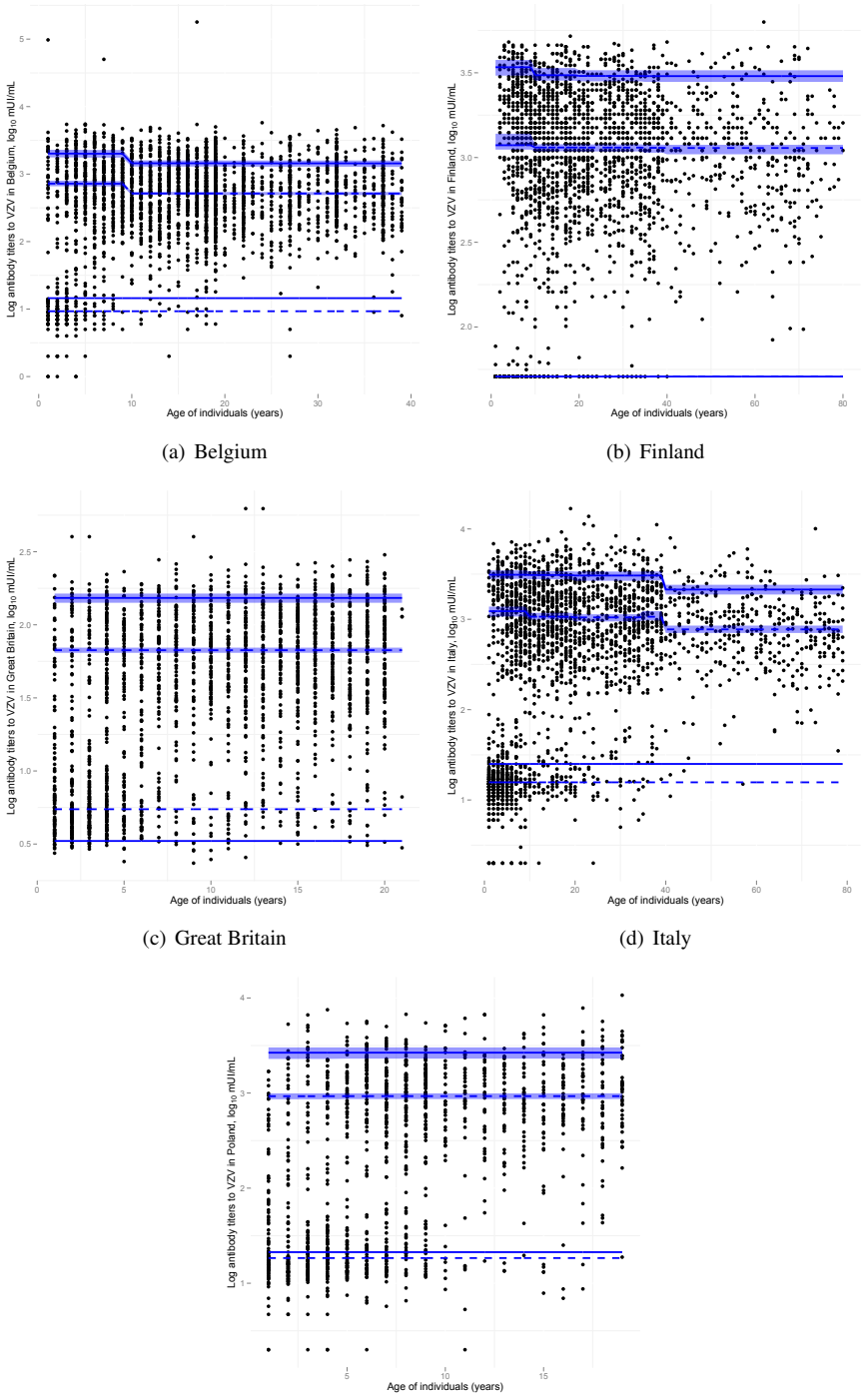


(a) Belgium (b) Finland

(c) Great Britain (d) Italy

(e) Poland

Figure 8.2: Parvovirus B19 in five European countries: posterior mean of the prevalence and the FOI, with 95% CI, and proportions seropositive by age estimated from a normal mixture (dashed curves) and from a skew-normal mixture (solid curves).

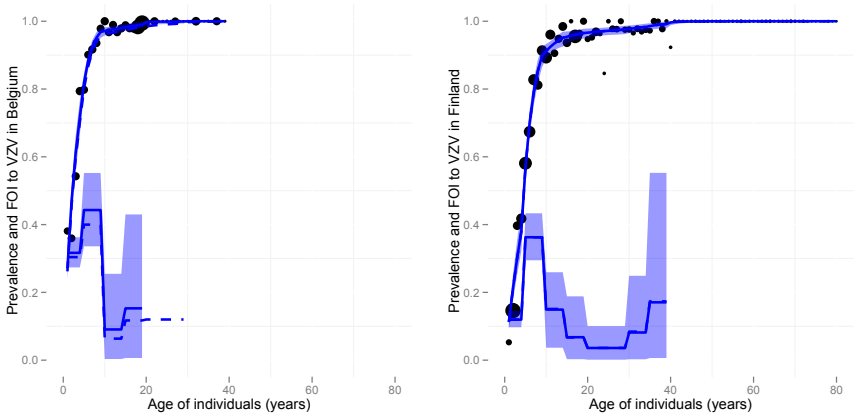


(a) Belgium (b) Finland

(c) Great Britain (d) Italy

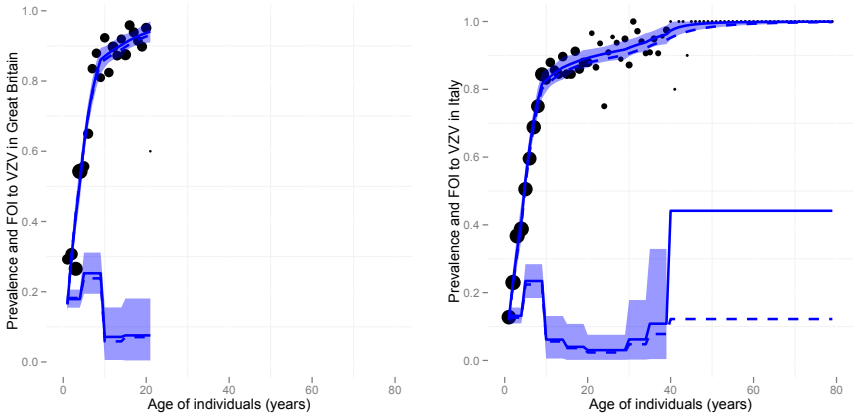
(e) Poland

Figure 8.3: VZV in five European countries: scatter plot of the log antibody titers with over imposed the posterior means of the mean of susceptible and of the age-dependent mean of immune (with 95% CI), from a normal mixture (dashed curves) and a skew-normal mixture (solid curves).



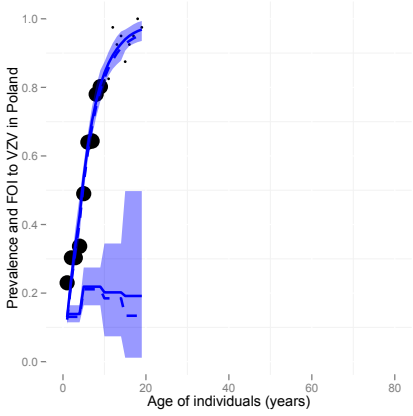
(a) Belgium

(b) Finland



(c) Great Britain

(d) Italy



(e) Poland

Figure 8.4: VZV in five European countries: posterior mean of the prevalence and the FOI, with 95% CI, and proportions seropositive by age estimated from a normal mixture (dashed curves) and from a skew-normal mixture (solid curves).

Chapter 9

Towards Measles Elimination in Italy: Monitoring Immunity by Bayesian Mixture Modeling of Serological Data

9.1 Introduction

Though traditionally considered a major public health problem (Anderson and May, 1991), measles has recently been brought on the road to eradication thanks to the availability of a highly effective vaccine (WHO, 2009). The path to elimination however is a complex one, mainly due to the difficulty to persistently keep vaccine uptake at the high levels required by measles (Anderson and May, 1991), even if incidence should eventually decline to negligible level (which is not currently the case, for some measles resurgence has been recently observed in Europe (CDC, 2011; WHO, 2011)). In these circumstances it is critical to continuously monitor the degree of herd immunity in the population by, e.g., repeated serological surveys. Therefore it follows that accurate modeling of serological data becomes a critical supporting tool of all local elimination plans.

In Italy, where measles immunization had traditionally been low and spatially inhomogeneous (ICONA, 2003, 2008), a National Measles Elimination Plan was introduced in 2003, aimed to rapidly approach WHO targets by (a) increasing first dose coverage, (b) initiating the second dose program, and (c) launching a national campaign in 2004 aimed to all children less than 15 years old. In order to assess the impact of the campaign two serological surveys were conducted in Tuscany (Italy), in the pre- (2003) and post-campaign (2005–2006) period (Bechini et al., 2010). Interpretation of the serological data based on the classical cut-off methods, though indicating in this case a substantial amount of susceptibility in age groups 10–20, is certainly weakened by the incapability of the conventional threshold to describe the

different degrees of immunity of those classified as seropositive. Moreover, conventional cut-off methods face the fact that their test specificity is larger than test sensitivity (Gay et al., 2003; Vyse et al., 2004, 2006), possibly leading to overestimated seronegativity rates. Instead, a higher value of the test sensitivity is necessary to accurately estimate the prevalence of immune people. Therefore, we believe that methods consisting in directly modeling the antibodies can gain us a deeper insight in the age-specific immunity profiles, in order to identify those groups of individuals who should be targeted by specific catch-up vaccination campaigns. Mixture models have been already successfully applied to the analysis of antibody count data, both in context of diagnostics and serological surveys (Vyse et al., 2004, 2006; Greiner et al., 1994, 1997).

In order to use in a fully appropriate manner the available serological information given by antibody titers, without the loss of information implicit in current status data analyses, we resorted to a Bayesian mixture model (Diebolt and Robert, 1994) with mixing weights depending on the age of individuals (Gay et al., 2003; Vyse et al., 2004, 2006; Gay, 1996; Hardelid et al., 2008). The model is estimated by a Bayesian approach using Markov chain Monte Carlo (MCMC) methods. The estimated model is subsequently used to provide an interpretation of the differences in the prevalence observed in the 2003 and 2005–2006 Tuscany surveys. The chapter is structured in following way. In Section 9.2 we give information about the current situation of measles and the vaccination in Tuscany, and we introduce the data that will be analyzed. In Section 9.3, we present the methodology used to fit K-component hierarchical Bayesian normal mixture models to antibody titers for measles. In Section 9.4 we apply several mixture models to antibody levels in the two periods with varying K and we summarize the results. Finally, in Section 9.5, we discuss the results and their consequences for monitoring the herd immunity of measles in Tuscany.

9.2 Data

9.2.1 Measles and Vaccination in Tuscany

In Tuscany (Italy), prior to mass immunization, outbreaks of measles occurred every three or four years, as we can see from Table 9.1. According to the regional notification system, during the Nineties there were two epidemics in the region and measles incidence reached values of 203.1 and 75.8 cases per 100,000 inhabitants, respectively in 1992 and 1995. Instead, the lowest historical levels were reached during 2004–2005, although a recrudescence of the illness occurred in 2006 (1.5 per 100,000), but with a low total number of cases (55). Incidence of measles in Tuscany in 2002–2006 by age group is shown in Figure 9.1. Nonetheless, it is important to remark that notifications of measles cases in Tuscany, as well in the whole Italy, in the pre-vaccination period suffered from severe problems of under-reporting (Williams et al., 2003). For this reason, we decided not to use these data in relation to the serological data.

Measles vaccine was made available in Italy since 1976 and it was recommended for all newborns at

Table 9.1: Incidence of measles cases ($\times 100,000$) in Tuscany (2002–2006).

Years	1992	1993	1994	1995	1996	1997	1998	1999
Incidence	203.1	17.8	14.9	75.8	40.8	31.8	1.7	1.6
Years	2000	2001	2002	2003	2004	2005	2006	2007
Incidence	1.5	1.4	9.4	6.2	0.4	0.3	1.5	0.1

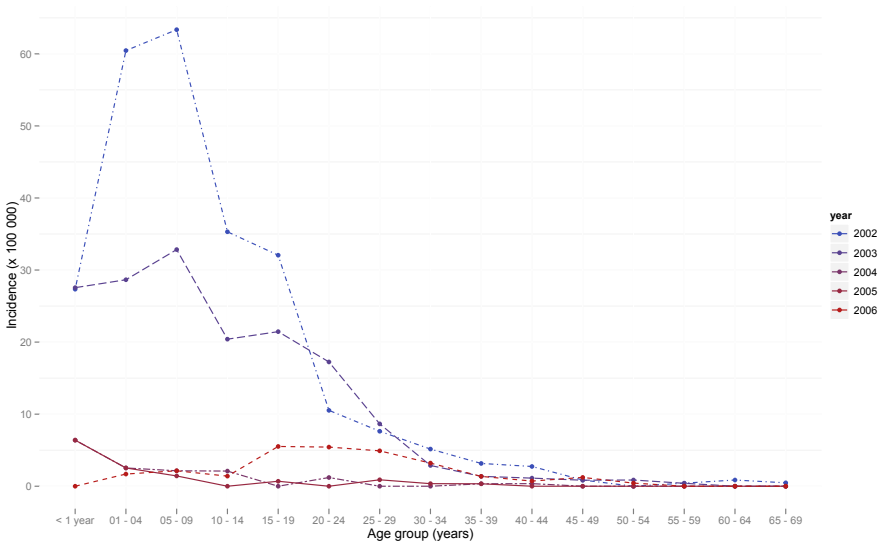


Figure 9.1: Incidence of measles cases ($\times 100,000$) in Tuscany, by age groups (2002–2006).

15 months of age, but it was underused for many years in Italy and also in Tuscany, so that large epidemics continued to occur, at least until 1997, when the last peak was observed (Bonanni et al., 2005). Data on vaccination coverage are not available for those years because measles vaccine was usually administered by private pediatricians and records were not collected at regional level (Bechini et al., 2010). By the end of the Nineties and with the implementation in 2003 of the National Plan for Elimination of Measles and Congenital Rubella, vaccination coverage in Tuscany has reached high values in newborns (24 months of age) and in school-aged children for the first dose (Table 9.2). School-aged children were also targeted by a catch-up immunization campaign performed in 2004–2005, which involved children in the primary and secondary schools (7–14 years). The aim of this vaccination campaign was to catch up those subjects who had never been vaccinated before and to administer a second dose of MMR (Measles, Mumps and Rubella) vaccine as well (Table 9.2). However, these high vaccination coverages were not sufficient to interrupt the indigenous transmission of the virus. Indeed, measles outbreaks have continued to occur in Tuscany since 2006, even though more than 95% of the cases are notified in people over 15 years old (Boncompagni et al., 2006). Moreover, analyzing incidence data from the recent years and matching them with data on vaccination coverage, Bechini et al. (2010) predicted that the accumulation of susceptible

Table 9.2: Routine vaccination coverage (VC) at 24 months of age (2003–2006) and vaccination coverage for first and second dose in school-aged children (7–14 years, born in 1991–1997) targeted by the immunisation campaign with MMR vaccine in Tuscany (2004–2005).

VC	2003	2004	2005	2006
24 months of age	87.02%	87.58%	89.2%	91.2%
First dose	-	81.6%	88.2%	-
Second dose	-	42.4%	64.7%	-

people mostly among adolescents (>15 years) may produce with high chances measles outbreaks in the future.

9.2.2 Seroprevalence Data in Tuscany

Two serum samples were collected in Tuscany (Italy) in 2003 and in 2005–2006, respectively, before and after the school-immunization campaign, from individuals aged 1–49 years (Bechini et al., 2010). For each subject, the antibody count (evaluated quantitatively as an antibody concentration and expressed as $\log_{10}$ mUI/mL) and the age were collected. The original 2003 sample contains 1030 subjects from the provinces of Florence, Siena and Lucca, whereas the original 2005–2006 sample contains 927 subjects from the province of Florence only.

All sera were stored at -20 °C and tested in different periods at the Department of Public Health of the University of Florence, Italy. Sera collected in 2003 were tested during 2004, while sera collected in 2005-2006 were tested in December 2006 and during the first months of 2007. The commercial enzyme linked immunosorbent assay (ELISA) Enzygnost Anti-Masern-Virus/IgG (Dade Behring, Germany) was used for detection of measles IgG antibodies. The same assays were used for testing all samples to ensure the comparability of the results in different years. Since the sera analyzed were anonymous, it was not known whether an individual had been vaccinated or not before.

9.3 Methods

9.3.1 Estimating Prevalence Using Normal Mixture Models

For measles in Tuscany, it cannot be assumed that the infection is in a steady state, because of the perturbation due to the National Elimination Plan, therefore the assumption of time homogeneity is untenable. Moreover, given that we have no information about the vaccination status of subjects, we cannot estimate the FOI, since it is not possible to distinguish between naturally infected and vaccinated. Therefore our purpose is to estimate the age-specific population prevalence on the assumptions that immunity is lifelong and that mortality caused by the infection is negligible and can thus be disregarded. A possible pitfall in the assumption of lifelong immunity is the ascertained waning of measles immunity following vacci-

nation. However, we believe that the mixture model can account for this decrease in antibody titers by classifying the involved subjects in mixture components that accommodate different degrees of immunity.

In order to model the heterogeneity in the antibody levels, we assume that the sample is drawn from a population consisting of an unknown number K of normal subpopulations, resulting in a density function g that can be represented as a mixture-distribution density of K unobserved Normal densities. These densities account for increasing levels of antibody counts, thus separating susceptible individuals, with low antibody levels, from the immune ones, with high antibody levels.

For the estimation of the parameters of the Normal mixture model, we follow the methods already introduced in Chapter 7, Section 7.3. However, the current methodology differ for what concerns the model specified for the prevalence and for the fact that more than two mixture components are taken into account. For the probabilities $\pi_k(a_i)$, we do not assume any deterministic relationship between the mixture probabilities and the age (Hardelid et al., 2008), but we rather specify a probabilistic model for π_k at each distinct level of the age, consisting in a Uniform distribution in the range (0,1) for each k. In our multivariate setting, these K Uniform priors imply a Dirichlet prior distribution (Diebolt and Robert, 1994) for each age group:

$$(\pi_1(a),\ldots,\pi_K(a)) \sim \mathrm{Dir}\,(\alpha_1 = 1,\ldots,\alpha_K = 1)\,, \tag{9.3.1}$$

under the constraint $\sum_{k=1}^{K} \pi_k(a) = 1$.

As selection criterion for the model with the best goodness-of-fit, we use the penalized expected deviance (PED, Plummer, 2008). A small value of PED indicates better goodness-of-fit.

9.3.2 Determination of the Current Status of Infection

After having established the number K of components and obtained the posterior means $\bar{\theta}_k = (\bar{\pi}_k(a), \bar{\mu}_k, \bar{\sigma}_k)$, we classify the components either as susceptible or as immune, and then assign the sampled individuals to one of the components. The procedure here presented is slightly different from that of Section 7.3.5, since it has to account for the fact that more than two components are used. Thus, firstly, according to the estimated locations $\bar{\mu}_k$, we label each component, e.g., based on previous biological knowledge. Secondly, t out of K components, i.e., those which have been labelled as components with low antibody counts, are classified as susceptible, with probability of being susceptible (S) $\bar{\pi}^S(a) = \sum_{k\in S} \bar{\pi}_k(a)$; the remaining $K-t$ components, i.e., those labelled as components with high antibody counts, are then classified as immune, with probability of being immune (I) $\bar{\pi}^I(a) = \sum_{k\in I} \bar{\pi}_k(a)$. Thirdly, we assign each subject in the sample to one of the components through the posterior mean of the mixture probability $\bar{\pi}_k(a_i)$, i.e., the component for which $\bar{\pi}_k(a_i)$ is maximal. Fourthly, we estimate the current status of the infection for each subject using the estimator $\bar{Z}(a_i)$, given by

$$\bar{Z}(a_i) = \begin{cases} 0, & \text{if } k \in S, \quad \text{individual } i \text{ is susceptible,} \\ 1, & \text{if } k \in I, \quad \text{individual } i \text{ is immune.} \end{cases} \tag{9.3.2}$$

Finally, when each individual has been classified either as susceptible or immune, one can estimate the proportion of immune per age group by averaging the membership variable $\bar{Z}(a_i)$ in each age group, that is to say,

$$\hat{\pi}^I(a=j) = \frac{\sum_{i=1}^{n_j} \bar{Z}(a_i)}{n_j}, \text{ with } j = 1,\ldots,n_j. \tag{9.3.3}$$

Note that this individual classification procedure may be of interest in case the mixture model is used as a diagnostic test for patient management (Greiner et al., 1994, 1997) or, more simply, for comparison with the classification provided by the conventional cut-off point. On the contrary, if the estimates of the prevalence and the FOI are the only interest, this step can be skipped.

9.3.3 Smoothing the Posterior Mean of the Age-Dependent Mixture Probabilities

The hierarchical Bayesian mixture models are fitted using Markov Chain Monte Carlo (MCMC) methods, through Gibbs sampling (Diebolt and Robert, 1994; Frühwirth-Schnatter, 2007). In such a way, we can draw a sample for each parameter of interest from its posterior distribution. As a result, for each parameter we obtain the posterior density, which can be summarized by the posterior mean and the 95% credible interval (CI). However, since we used a probabilistic model for $\pi_j(a)$ at each age a, the posterior mean of the mixture probabilities by age, $\bar{\pi}_j(a)$, will show an irregular pattern characterized by many spikes. Therefore, in order to have a less spiky estimate of the prevalence, we smooth every MCMC replicate m of the prevalence, $m = 1,\ldots,M$, where M is the total number of MCMC saved replicates, using a generalized additive model (GAM, Hastie and Tibshirani, 1986) with logit link function:

$$\log\left(\frac{\pi_j(a)^m}{1-\pi_j(a)^m}\right) = \beta_0 + \beta_1 f(a), \tag{9.3.4}$$

where for the nonparametric function $f(a)$ we apply P-splines (Eilers and Marx, 1996) as smoother. So doing, the smoothed mixture probability is given by the mean of the smoothed probability's replicates per age over the M replicates.

9.4 Results

9.4.1 Age Patterns in Antibody Data

IgG antibody counts by age, with the corresponding smoothed average profiles, are reported in Figure 9.2. In both periods the average antibody levels initially decrease with age, reaching a minimum between 10 and 20 years, with a lag of approximately 5 years between the two periods, and then rise again, before plateauing or even slightly declining. The slightly larger prevalence in the youngest age groups in 2005–2006 may reflect the vaccination campaign, which mainly targeted children up to 10 years, but with special focus on young children. Average antibody levels in adults are lower in 2005–2006 than in 2003 and this could be due to a possible decay in the antibody counts, possibly due to the reduced boosting

from natural infection caused by the high vaccination coverage. However, after age 25 antibody counts remain quite high, with only a few cases below 2.5 $\log_{10}$ mUI/mL. Figure 9.3 reports the histogram of

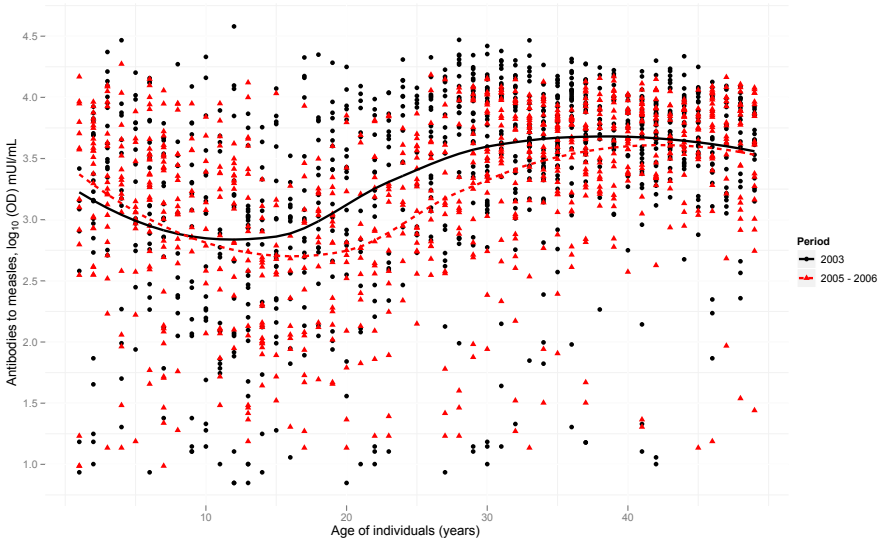


Figure 9.2: Scatterplot of antibodies IgG to measles (black dots for 2003, red triangles for 2005–2006) with over imposed the smoothed mean profiles for 2003 (black solid line) and for 2005–2006 (red dashed line).

antibody count data fitted by the optimal mixture models with age-independent weights; the figure shows already the main insight, that is, the presence of three (in 2003) and four (in 2005–2006) well identified subpopulations; the histogram also reports the two conventional cut-off points indicated by the assay's manufacturer: the cases lying at the left side of the lower cut-off are classified as susceptible, those lying at the right side of the upper cut-off are classified as immune, and those in between are considered as dubious. In Figure 9.4 we show a comparison of the age-dependent proportions seronegative $\hat{\pi}^S$ in the two periods obtained using the conventional cut-offs and by classification through the mixture models. In 2003, the proportions differ between the two methods mostly between 5 and 25 years, with the conventional cut-offs usually providing larger proportions susceptible, consistently with their higher specificity. However, this is not the case for 2005–2006, where the proportions seronegative estimated through the mixture tend to be larger than their conventional counterpart. Notwithstanding these differences, in both periods the two types of proportion susceptible are highly correlated, with linear correlations equal to 0.857 and 0.955 in 2003 and in 2005-2006, respectively.

9.4.2 Selection of the Optimal Number of Normal Components

The models in Eqs. (7.3.2) and (7.3.6) were fitted to data using MCMC methods (by JAGS software, version 3.1 (Plummer, 2003)), which are popular for Bayesian estimation of mixture models (Frühwirth-Schnatter, 2007). For each model parameter, we obtained its posterior density and we summarized it with

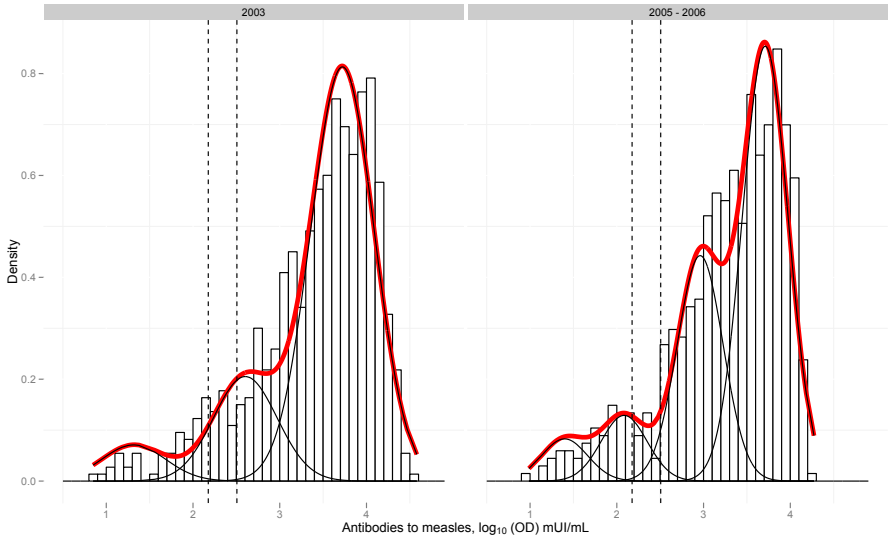


Figure 9.3: Histogram of antibodies IgG to measles in Tuscany in 2003 (left panel) and in 2005–2006 (right panel), fitted by the age-independent Normal mixture models with three (for 2003) and four (for 2005-2006) components and homogeneous variance. Conventional cut-off points are also reported (dashed lines).

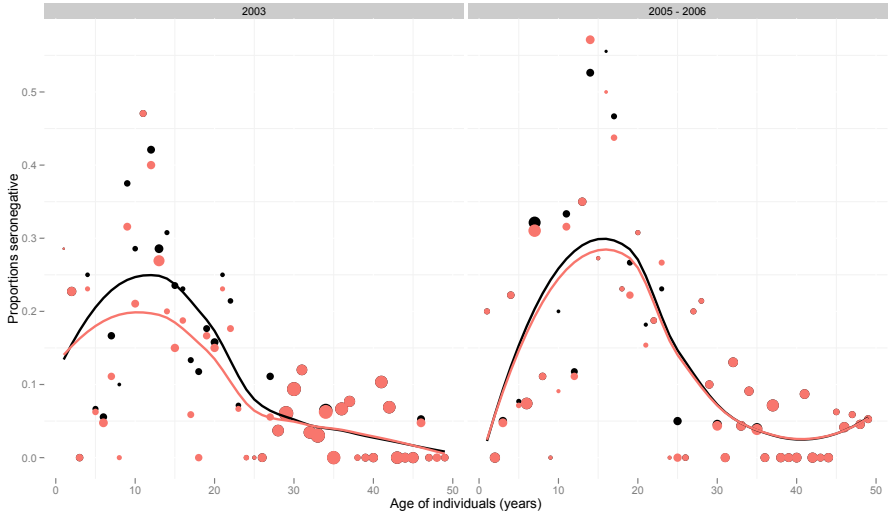


Figure 9.4: Comparison between the proportions seronegative by age in 2003 (left panel) and in 2005–2006 (right panel) given by the conventional cut-off point (black dots) and by classification through the age-dependent Normal mixture models (red dots), with over imposed the smooth mean profiles.

Table 9.3: PED for age-dependent mixture models with increasing number of components, either with homogeneous or heterogeneous variance of the components, based on further 20000 MCMC iterations.

Period	# Comp.	Variance	PED	Period	# Comp.	Variance	PED
2003	3	homogeneous	4937	2005–2006	3	homogeneous	5890
		heterogeneous	6862			heterogeneous	8300
	4	homogeneous	6095		4	homogeneous	4819
		heterogeneous	9864			heterogeneous	9200
	5	homogeneous	6435		5	homogeneous	5751
		heterogeneous	10492			heterogeneous	16235
	6	homogeneous	6371		6	homogeneous	5955
		heterogeneous	8649			heterogeneous	11475

Table 9.4: Posterior mean and 95% CI for the means μ_j and the standard deviations σ of the age-dependent four-component Normal mixture models for 2003 and 2005–2006.

Period	Par.	Post. Mean	95% CI	Period	Par.	Post. Mean	95% CI
2003	μ_1	1.530	(1.395, 1.642)	2005–2006	μ_1	1.730	(1.660, 1.800)
	μ_2	2.851	(2.748, 2.953)		μ_2	2.899	(2.832, 2.965)
	μ_3	3.779	(3.746, 3.812)		μ_3	3.639	(3.556, 3.695)
	-	-	-		μ_4	3.706	(3.658, 3.763)
	σ	0.350	(0.331, 0.371)		σ	0.308	(0.289, 0.331)

the posterior mean and the 95% credible interval (CI). To select the optimal number of components J, different age-dependent mixture models with a number of components ranging between three and six were estimated, under both the cases of homogeneous ("homoscedastic") and heterogeneous ("heteroscedastic") variance, and we compared their PED (Table 9.3). In particular, we discarded models with two components, because lacking the complexity necessary to accommodate for susceptible, naturally immune and vaccinate individuals. This approach is often followed in the literature about vaccine-preventable infections in post-vaccination regime (Gay et al., 2003; Vyse et al., 2006; Hardelid et al., 2008; Rota et al., 2008). In our case, two-components mixture models result to be completely inadequate to describe the histogram of the antibody count data. We found that the best model has three components in 2003 and four components in 2005–2006, both with homogeneous variance.

Note from Table 9.3 that heteroscedastic models are usually more penalized than homoscedastic ones (except for the models with two components), and that the penalty increases with the number of components. The corresponding posterior means (with 95% CI) and standard deviations of the "best" Normal mixtures for 2003 and 2005–2006 are reported in Table 9.4. We notice from Table 9.4 and Figure 9.3 that the means of the three- and the four- component models are quite similar (the 95% CI between the two periods are always overlapping), but for μ_1, whose posterior mean is significantly higher in 2005–2006.

9.4.3 Prevalence of the Different Components by Age

Following Section 9.3.2, we labelled the components and we classified each individual either as susceptible or immune. The first component is characterized by the lowest antibody counts and should represent the truly susceptible subjects, i.e., unvaccinated individuals who were never exposed to the infection.

Components 3 in 2003 and 3 and 4 in 2005-2006 are characterized by the highest antibody counts and thus represent people with a robust immune response, following either vaccination or natural infection.

Finally, the second component (in both 2003 and 2005–2006 surveys) is a clearly separate group between component 1 and components 3 and 4 and accounts mainly for persons who present neither a negative response nor a strong positive response.

As a consequence of this labeling, the practice of separating individuals into two groups would suggest to include in the immune group the components 2, 3, and 4:

$$\bar{Z}(a_i) = \begin{cases} 0, & k = 1 \in S, \\ 1, & k = 2,3 \in I_{2003}, \end{cases} \quad \text{and} \quad \bar{Z}(a_i) = \begin{cases} 0, & k = 1 \in S, \\ 1, & k = 2,3,4 \in I_{2005--2006}, \end{cases} \tag{9.4.5}$$

Our main results are summarized in Figure 9.5, where we compare the smoothed age-specific prevalences of susceptible individuals ($\bar{\pi}^S(a)$), i.e., those belonging to component 1 (dotted curves), and of immune individuals ($\bar{\pi}^I(a)$).

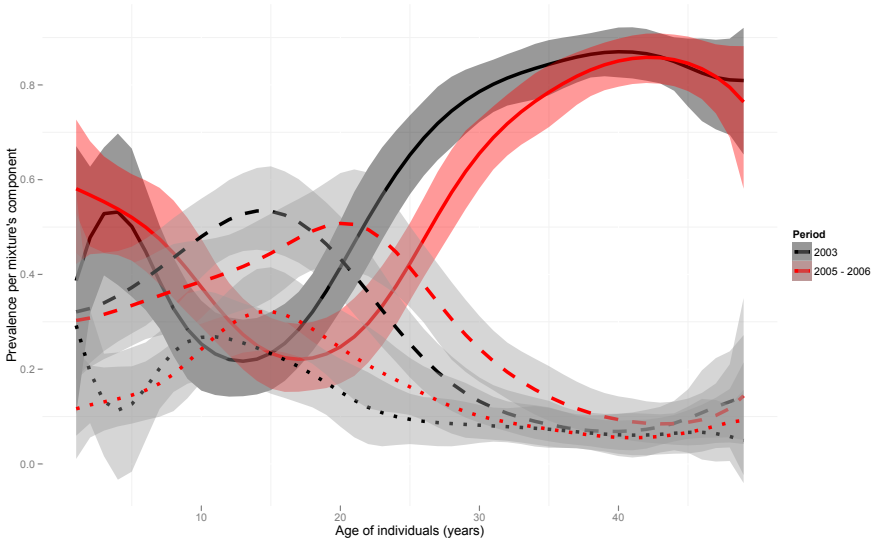


Figure 9.5: Comparison of the age-dependent prevalence of susceptible (component 1, dotted lines), "low responders" (component 2, dashed lines), and "high responders" (component 3 in 2003 and components 3 and 4 in 2005–2006, solid lines), for measles in Tuscany in 2003 (black lines) and 2005–2006 (red lines), with 95% pointwise CIs.

For clarity, we further split the latter prevalence into two parts, i.e., the "high responders" belonging

to components 3 and 4 (solid curves), and the "low responders" of component 2 (dashed curves), in view of their different age patterns (we recall that the sum of the three shown prevalence curves equals one in each age group).

We are thus able to show that component 3 in 2003 and components 3 and 4 in 2005-2006 ("high responders") account mainly for children aged 2–5 years old, i.e., cohorts with very high vaccine uptake, and persons older than 20, whose robust immunity also derives from natural infection and, perhaps, boosting, due to low vaccine uptake in the past. Their prevalence in 2003 reaches a first peak at age 2.5, which corresponds, including the observed vaccination delay ICONA (2008), to the completion of first dose immunization. Afterwards, it falls down to a minimum at age 13, mirroring the past history of increasing vaccine uptake and herd immunity. Then it rises again, until it finally flattens out. In 2005–2006, the prevalence of "high responders" has a peak at age 1.5, which is consistent with the decline in the age at first immunization and with the effort to decrease the corresponding delay. Then, it goes down reaching a minimum at around age 17 and finally rises up to a plateau after age 30 years. Both periods are characterized by high uncertainty after 45 years because of the very small sample sizes.

By comparing the two periods, we see that the prevalence of "high responders" in 2005–2006 is higher in the youngest age groups, up to age 10. This is consistent with the campaign targeted at children aged 6–14 conducted in Tuscany during 2004. Nonetheless, what is most striking from our analysis is the lower prevalence of "high responders" in the age-cohort 10–20 years. This trough corresponds to an increase in the prevalence of susceptible and "low responders" in this age group. Our analysis shows that both the first component (susceptible) and the second component ("low responders") mainly account for persons aged 10–20 in 2003 and 15–25 in 2005–2006. In particular the prevalence of susceptible in 2003 is estimated to peak at 22% at age 10, which is reflected in the peak at age 15 in 2005–2006, where the prevalence is even higher (about 25%). As concerns the "low responders", these represent a very substantial part of the teen-agers population, with peaks of prevalence above 50% at age 15 and at age 20, in 2003 and in 2005–2006, respectively.

9.5 Discussion

Measles serological studies conducted in Italian regions after the 2004 vaccination campaign reveal, using the conventional cut-off methodology, a worrying trough in teen age groups suggesting accumulation of susceptible at such ages (Bechini et al., 2010; Rota et al., 2008). To deep our understanding of this phenomenon, in this paper we used hierarchical Bayesian normal mixture models to estimate, directly from data (Gay et al., 2003; Gay, 1996), measles population prevalence and herd immunity in Tuscany in two distinct periods, 2003 and 2005–2006. Indeed, mixture models can give a better idea of the complexity behind the distribution of the immunity in the population, which in many cases cannot be reduced to the "simple cut-off" separation between susceptible and immune persons. Most of all, this is true in populations where the larger part of immunity is induced by vaccination. In this context, it becomes fundamental to study directly the antibodies with mixture model, given their ability in monitoring the age

pattern of the different components, which can account both for susceptibility and for different degrees of immunity. To estimate the population prevalence, we used a nonparametric model, which estimates a mixture probability for each age. Though paying some cost in terms of parsimony, our nonparametric approach allows a greater flexibility in data modeling compared to parametric models for $\pi(a)$ proposed by Hardelid et al. (2008). In particular it avoids to impose patterns to the data, e.g., the monotonicity restraints, which are hardly fulfilled under post-vaccination situations.

When dealing with post-vaccination data, we should be cautious with the interpretation of results, because we expect several mixture components to be necessary to accommodate the large variability of antibody data, due to the combination of susceptible, vaccinated and naturally immune individuals. For measles, natural immunity causes higher antibody titers than the vaccine (Krugman et al., 1965), especially for those who have received only one dose, as it has been the case for Italy. Therefore vaccinated individuals are likely to form a rather heterogeneous group, which includes a fraction of non-respondents to measles vaccination, estimated to be approximately 5% (Mossong et al., 1999), or respondent to vaccination with a low immune response. Using a conventional cut-off or in absence of appropriate modeling of this group (for instance, using a two-component mixture), these individuals would likely be classified either in the right tale of the susceptible group or in the left tail of the immune one. As a matter of fact, the problem of misclassification of observations following poor separation of components, due to the presence of vaccinated individuals, is a known issue of mixture models, as well of the conventional cut-off point, which is totally unable to distinguish naturally immune from vaccinated. For this reason, when using mixture models, we believe that model selection should not be based only on statistical selection criteria, but should also be supplemented by more articulated "interpretation" criteria, including *a priori* knowledge as well. This interpretation criterion is what, jointly with the poor graphic fit, led us to discard two-component mixtures, as it is often done in the literature.

We note that in this work we focussed only on normal mixture models, ignoring other possible distributions for the data. An alternative could have been to model the data with a skew-normal model, as we did in the two previous chapters. Using this distribution, with its ability to deal with skewness, we could have found that less components are needed to account for the immune response. Indeed, even though we did not perform any sensitivity analysis for the data distribution, we recognize that it would be quite interesting to see what would change in the description of the immunity levels using other distributions.

In any case, according to our mixture approach, the results show marked troughs in the prevalence of strong immune (i.e., the "high responders"), mirrored by a large prevalence of susceptible individuals and of weakly immune individuals (i.e., the "low responders"). As for susceptibility, the model accurately identifies the ages where pockets of susceptible exist, mostly between 10 and 20 years, and which therefore represent the target for further catch-up activities (Bechini et al., 2010). However, the unexpected result of our model is the large group of "low responders". The explanation of the large frequency of this group, located as well the susceptible in the age group 10–20, is less obvious and is probably due to the interplay of several factors. Surely these subjects belong to cohorts with a moderate vaccine uptake

(about 50–60%, but only first dose), but which also faced the decline in incidence due to the increasing uptake in subsequent years (see Table 9.2). The increase in herd immunity might have sharply reduced the effects of natural boosting of immunity, thereby reducing antibody titers (Mossong et al., 1999). Another factor could be the waning of measles vaccine protection, faster than the decaying that we would have with natural immunity, most of all in absence of a second dose of vaccination (Rota et al., 2008).

A possible caveat to the analysis regards the different geographical origin of the specimens in 2003 and in 2005–2006, as noted in Section 9.2.2. However, available Tuscan vaccination data show a good degree of homogeneity in measles coverages in the various Tuscan provinces prior to the serosurveys. In particular, they show that MMR coverages (which initiated in 1997) in the provinces of Lucca and Siena were almost overlapping with those registered in Florence (Bechini et al., 2006), and, moreover, pre-MMR coverages (before 1997) showed very mild differences among these provinces (given that we do not dispose of Tuscan pre-vaccination serological data, we can only conjecture that pre-vaccination epidemiology was homogeneous across Tuscan provinces). For these reasons, we consequently believe that the impact of the different geographical origin of the analyzed specimens on the sample representativeness at regional level should be minimal.

In conclusion, we believe that our analysis is effective in raising important questions about "low responders", namely, whether "low" might eventually become "at risk" due to insufficient immune memory. We feel that the current knowledge about immunity in highly vaccinated populations, especially when most individuals experienced only one dose, and the declined role of natural boosting, is still insufficient, and therefore careful monitoring of such phenomena is important. Unfortunately, what is currently preventing us from better interpreting post-vaccination antibody data is the lack of information about the actual difference between the antibody response in case of natural infection and in case of vaccination. From this viewpoint, serological data based on true survey studies capable to also collect information about past history of infection and vaccination could prove to be useful. Varicella, for which mass vaccination programs are not yet routinely used in Europe, is a good candidate for this.

Part III

Estimation of Transmission Parameters for Varicella in Europe

Chapter 10

Introduction: Modeling of Prevalence and Force of Infection for Varicella in Europe

10.1 Introduction

The third part of the thesis summarizes a part of the work carried out for a project supported by the European Center for Disease Control and Prevention (ECDC), entitled "Vaccine preventable disease modeling in the European Union and EEA/EFTA countries: Forecasting the effects of introducing a new vaccine in a national/regional program".

Varicella or chickenpox is a contagious disease caused by primary infection with varicella zoster virus (VZV), also known as human herpes type 3 type (HHV-3). Varicella is an airborne disease spread through coughing or sneezing and, more generally, close person-to-person contacts with infective subjects. This infection is usually benign in children, but can be more severe in elders, pregnant women, neonates, and immune-compromized subjects. After recovery from primary VZV infection, the virus becomes latent in the dorsal root ganglia an may reactivate at a later date, resulting in herpes zoster (HZ) or shingles, which is an important cause of morbidity. A live-attenuated varicella vaccine was first developed in 1974. Since 1995, the vaccine has been introduced among the recommended vaccination schedules for children in US. Currently, inclusion of this vaccine in routine pediatric immunization programs in Europe has been reported in Germany, Greece, in the Italian region of Sicily and in the Autonomous Community of Madrid in Spain (Bonanni et al., 2009).

The starting point of this part of the thesis (and of the ECDC project) is an explanatory analysis of age-specific serological data for varicella based on flexible statistical models. We use both parametric and non-parametric approaches. The purpose of the analysis is to "extract" optimal statistical descriptions of the age-patterns of acquisition of infection from the observed seroprofile, as typically summarized by a force of infection, and the implied "predicted" seroprofile, without any reference to some underlying

contact patterns. The usefulness of such analysis is twofold. Firstly, it offers optimal baselines against which to ground descriptions of seroprofiles based on recently acquired contact data, when these are available. Second, when contact data are not available, which is the case for most European countries, the approach provides the base for the use of the traditional indirect approaches to estimating contacts, e.g., the "WAIFW" approach (Anderson and May, 1991), or the proportional/preferred mixing approach (Hethcote, 1996).

10.2 Data

The data consist of cross-sectional serological samples by age from twelve European countries giving information about antibodies to VZV and current status infection, according to a conventional cut-off point given by the assay's manufacturer. The samples were collected either within the ESEN2 (Nardone and Miller, 2004; Nardone et al., 2007) or the POLYMOD frameworks (Mossong et al., 2008). Part of these data has been already analyzed using Bayesian mixture models in Chapter 7 and in Chapter 8. Table 10.1 reports information about the serum banks available, that is, the year of sample collection, the analyzed age range, the number of samples, and the commercial assay used to test data. The table is based on a selection of data after the removal of the equivocal cases, which cannot be classified either as positive or negative using the conventional cut-off, and of serological data for age below one, because of the maternal antibody protection still present. Figure 10.1 shows the observed prevalence per country

Table 10.1: Collection of main serum banks of participating countries and enzyme immunoassay used to test for antibodies to VZV. The reported number of samples in brackets does not include the equivocal cases.

Country	Framework	Year of collection	Age range	Sample size	Commercial assay
Belgium (BE)	ESEN2, POLYMOD	2001–2003	1–39	2432	Enzygnost
Finland (FI)	ESEN2, POLYMOD	1997–1998	1–60+	3113	Enzygnost
Germany (DE)	ESEN2	1995–1999	1–60+	4414	Enzygnost
Great Britain (GB)	ESEN2, POLYMOD	1996	1–20	2030	DiaMedix
Ireland (IE)	ESEN2	2003	1–60+	2430	DiaMedix
Israel (IL)	ESEN2	2000–2001	1–60+	1541	Enzygnost
Italy (IT)	POLYMOD	2003–2004	1–60+	2491	Enzygnost
Luxembourg (LU)	ESEN2	2000–2001	4–60+	2640	Enzygnost
Netherlands (NL)	ESEN2	1995–1996	1–60+	1967	Human
Poland (PL)	POLYMOD	2000–2004	1–19	1301	-
Slovakia (SK)	ESEN2	2002	1–60+	3515	Euroimmun
Spain (ES)	ESEN2	2002	2–39	3590	Enzygnost

and per age groups, the overall prevalence per country, and the sample size per country, respectively. For the age-specific prevalences, the main differences among countries are visible in the youngest age group, that is, 1–5 years: while in Belgium, Israel, Luxembourg, and the Netherlands the prevalence of varicella

is already higher than 50%, in Poland the prevalence is below 15%, while in Italy and in Slovakia it is somewhere in between, around 30%. For the age group 6–19, though in most countries prevalence lies above 90%, we notice the remarkable exception of Italy and Poland, which are only a little above 75%. Finally, in the older age group, the prevalence is always above 90%, and differences are less evident (no data are available for Poland and for Great Britain only data at age 20 are available).

10.3 Flexible Parametric and Nonparametric Models for Seroprevalence

In order to estimate the prevalence and the FOI from the serological data, given as current status data, we compared the results of three models, parametric and nonparametric. For the first class of models, we chose a model based on a piecewise-constant FOI (Anderson and May, 1991) and a constrained fractional polynomial (Royston and Altman, 1994). For the second class of models, we used a local polynomial (Fan and Gijbels, 1996). For all modeling approaches, the response, $Y(a_i)$, is the number of immune individuals in the sample at age a_i. We assume that

$$Y(a_i) \sim Bin(n(a_i), \pi(a_i)), \qquad i = 1, \ldots, n, \tag{10.3.1}$$

and

$$g(\pi(a)) = \eta(a), \tag{10.3.2}$$

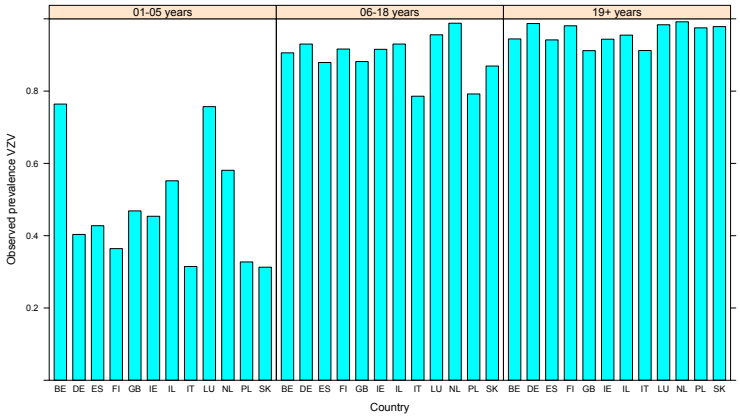
where $g(\cdot)$ is the link function and $\eta(a)$ is a linear predictor assumed to depend on the age of individuals.

All these flexible models can be used for comparison with models based on social contact data, providing thus a benchmark in terms of goodness-of-fit. Moreover, the estimates of the FOI can be used to estimate the transmission parameters of the WAIFW contact matrices, in order to derive both the basic reproduction number, R_0, and the critical threshold, p_c (Van Effelterre et al., 2009; Hens et al., 2012). This approach can still be useful to model VZV transmission in those countries that do not have social contact data at disposal.

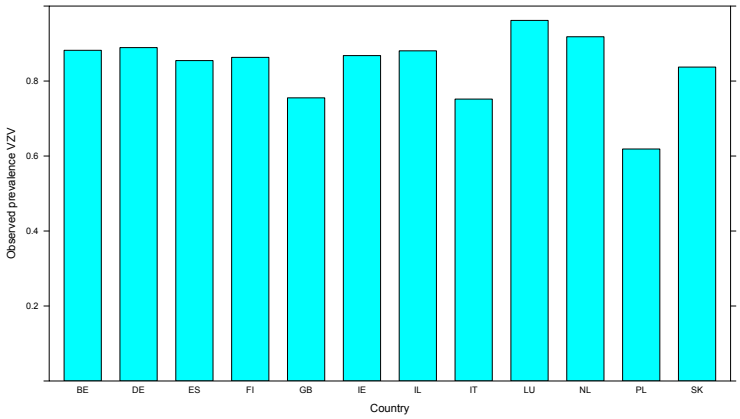
10.3.1 Model with Piecewise-Constant Force of Infection

For an elaborate discussion of this model, we refer to Section 7.3.2.2. In this chapter the model is used to estimate the prevalence and it will be used in the next chapter to estimate the prevalence for a given contact pattern. The model for the prevalence is given by

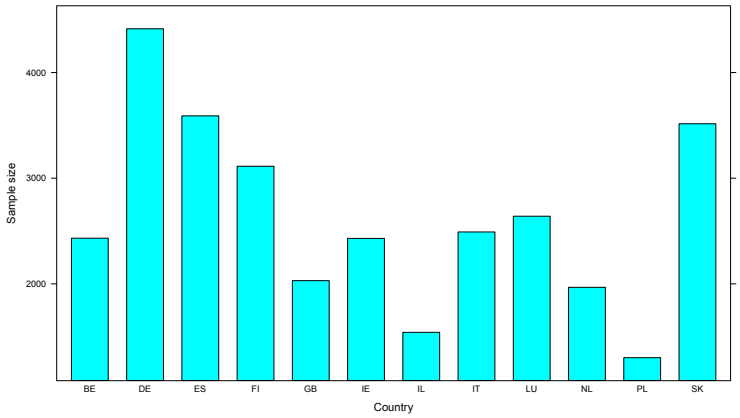
$$\pi(a) = 1 - \exp\left\{ -\sum_{j=1}^{J-1} \lambda_j (a_{j+1} - a_j) - \lambda_J (a - a_J) \right\}. \tag{10.3.3}$$



(a) Prevalence per age group



(b) Overall prevalence



(c) Sample size

Figure 10.1: From top to bottom: bar plot of the prevalence per country and per age group (“1–5 years”,“6–19 years”,“20+ years”); bar plot of the overall prevalence per country; bar plot of the sample size per country.

For the age groups, we used the following seven classes: 1–4, 5–9, 10–14, 15–19, 20–29, 30–49, 50+. For the estimation, we used an optimization algorithm that minimizes the binomial minus log-likelihood of the serological data, namely, a variable metric algorithm (Broyden-Fletcher-Goldfarb-Shanno method (BFGS)). Given that the FOI has to be a positive function, we imposed a positivity constraint on the estimates of the force of infection.

10.3.2 Fractional Polynomials

Royston and Altman (1994) proposed a family of curves, called "fractional polynomials" (FP), whose power terms are restricted to a small predefined set of numbers. The powers are selected in such a way that conventional polynomials are a subset of the family. An advantage of FP is that they have a greater flexibility than conventional polynomials and are straightforward to fit using standard methods.

We apply FP to the prevalence function and the model consists of a GLM, whose linear part is given by:

$$\eta(a;m,\mathbf{p}) = \beta_0 + \sum_{j=1}^{m} \beta_j H_j(a), \tag{10.3.4}$$

where the function $H_j(\cdot)$ is given by the Box-Tidwell's transformation (Box and Tidwell, 1962):

$$H_j(a) = \begin{cases} a^{p_j} & \text{if } p_j \neq p_{j-1}, p_j \neq 0, \\ \log(a) & \text{if } p_j = 0, \\ H_{j-1}(a)\log(a) & \text{if } p_j = p_{j-1}. \end{cases} \tag{10.3.5}$$

In order to find the best fitting model, it is worth comparing first- and second-order degrees models, that is, $m = 1,2$. For modeling data using fractional polynomials, the authors propose to determine the "best" value of m and of the power vector $\mathcal{P}$. The authors suggest to use the space $\mathcal{P} = \{-2,-1,-0.5,0,0.5,1,2,3\}$, which includes all conventional polynomials of degree less than or equal to m, because they assert that it is sufficiently rich to cover many practical cases adequately. More flexibility can be gained if we let p_j vary continuously between $\{-2,3\}$ (Shkedy et al., 2006). Concerning the choice of the degree m, once we have estimated the best model for each degree, we can decide to reject the model with degree m, in favor of model with degree $m+1$, if:

$$D(m,\tilde{\mathbf{p}}) - D(m+1,\tilde{\mathbf{p}}) > \chi^2_{2;0.90} = 4.61, \tag{10.3.6}$$

where $D(m,\tilde{\mathbf{p}})$ is the deviance of each FP model we want to test.

Order-constrained models were used for the prevalence, forcing it to be monotonically increasing, in order to guarantee that the FOI $\lambda(a)$ is always nonnegative. A model with logit link was used for the prevalence,

$$\log\left(\frac{\pi(a)}{1-\pi(a)}\right) = \eta(a), \tag{10.3.7}$$

and this implies that the force of infection $\lambda(a)$ is given by (Hens et al., 2012)

$$\lambda(a) = \eta\prime(a)\frac{e^{\eta(a)}}{1+e^{\eta(a)}}. \tag{10.3.8}$$

10.3.3 Local Polynomials

Local polynomial (LP) is a nonparametric method based on local approximations. It allows to fit simple models to localized subsets of data to describe their nonlinear pattern, without the specification of any parametric assumption. This means that the linear predictor $\eta(a)$ is approximated locally, in the neighborhood of a particular value a_0 of the age, by a specific function (Fan and Gijbels, 1996; Shkedy et al., 2003). The size of neighborhood of a_0 is controlled by a kernel function K_h, that assigns high weights to the age data points a_i close to a_0 and low or zero weights to farther data points. For this kernel function, we use a density function with mean 0 and variance h. The most common kernel functions are the Gaussian kernel, the tricube kernel and the Epanechnikov kernel. The variance h of the kernel controls the width of the "window" around the age value a_0. For this reason, it is called smoothing parameter. Along with the kernel function and the smoothing parameter, a third feature of a local polynomial is the degree of the function fitted in the neighborhood of the country. According to Shkedy et al. (2003), the degree $p = 2$ is the optimal degree when interest lies in the estimation of the FOI.

Hence, the local binomial likelihood function that we want to maximize in order to obtain the prevalence $\pi(a)$ is given by

$$L(a) = \sum_{i=1}^{n}[Y_i\log(\pi) + (n_i - Y_i)\log(1-\pi)]K_h(a_i - a_0), \tag{10.3.9}$$

where π is given by

$$\pi = g^{-1}(\eta(a_i - a_0)) \approx g^{-1}(\beta_0(a_0) + \beta_1(a_0)(a_i - a_0) + \beta_2(a_0)(a_i - a_0)^2), \tag{10.3.10}$$

following from a Taylor expansion of the age a_i in the neighborhood of a_0. The local estimate of the prevalence at age a_0 (Shkedy et al., 2003; Hens et al., 2012) is given by

$$\hat{\pi}(a_0) = g^{-1}(\hat{\beta}_0(a_0)). \tag{10.3.11}$$

The estimation of $g^{-1}(\hat{\beta}_0(a_0))$ must be repeated for each value of a_0. To obtain the FOI, the procedure is similar, since the local estimate of the FOI at age a_0 is

$$\hat{\lambda}(a_0) = \hat{\beta}_1(a_0)\delta(\hat{\beta}_0(a_0)), \tag{10.3.12}$$

where $\hat{\beta}_1(a_0)$ is the estimate of the local slope and $\delta(\hat{\beta}_0(a_0))$ is a function of the local intercept which is different for each link function (Shkedy et al., 2003).

Being the choice of the kernel function relatively unimportant and the function degree optimally set equal to 2, the most important parameter to choose is then the smoothing parameter h. A large h may result in over-smoothing, or miss important features in the data, while a small h may result in a fit that is too noisy. It may be desirable to vary h with the fitting point a_0. Model assessment can be performed in several ways, for instance, by using either a global criterion as the generalized cross validation (GCV), which estimates the averaged squared prediction error, or the information criteria, as the AIC or the BIC.

Finally, in order to account for the monotonicity of the prevalence function, we apply the pooled-adjacent-violators algorithm (PAVA) to the estimates of the prevalence, according to the principle "smooth, then constrain" (Hens et al., 2012).

10.4 Results

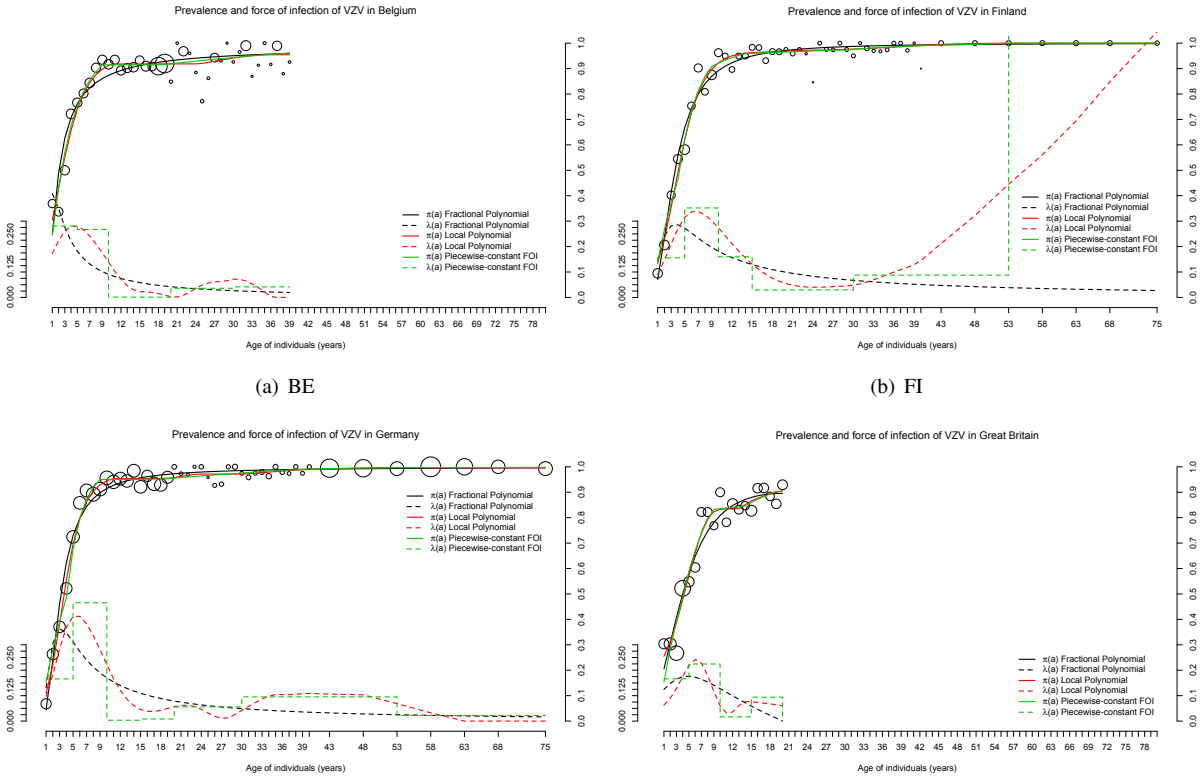
For each country, we fitted the three types of models considered. For the local polynomials, we selected the best bandwidth h taking the model with the smallest BIC. Table 10.2 reports, per country, the AIC and BIC for the models discussed in the previous section. Data and estimated models are shown in Figures 10.2 – 10.4.

Table 10.2: Comparison among three different models for $\pi(a)$: fractional polynomials (FP), local polynomials (LP) and piecewise-constant FOI models (PC).

	FP		LP		PC	
Country	AIC	BIC	AIC	BIC	AIC	BIC
BE	211.23	216.22	190.97	**200.09**	196.49	206.47
DE	191.60	197.15	177.58	**190.56**	186.73	199.68
ES	195.91	200.83	188.59	**197.01**	195.56	205.38
FI	152.20	157.75	147.12	**155.93**	154.99	167.94
GB	150.17	155.14	136.07	**141.59**	148.76	152.74
IE	147.34	153.63	135.17	**141.76**	138.98	147.79
IL	172.58	181.72	148.47	**159.09**	153.09	165.89
IT	199.92	205.47	188.15	**197.81**	187.93	200.88
LU	127.17	**132.52**	130.02	138.43	128.59	139.30
NL	117.43	126.68	105.60	119.92	104.60	**110.15**
PL	99.07	103.79	95.75	**101.05**	106.52	110.29
SK	178.17	187.32	170.01	**178.97**	173.02	185.83

In all the countries the prevalence is above 90% already at age 10 and the main peak of the FOI is usually found between age 5 and 10, as it appears from the piecewise-constant FOI model. Looking at the local polynomials, we notice that the peak tends to be very close to age 5, with a maximum around age 7 (Italy, Slovakia, and Spain). The peak is generally followed by a rapid decrease of the FOI (up to zero or very negligible values) in the age groups of the adolescents, who usually have few contacts with

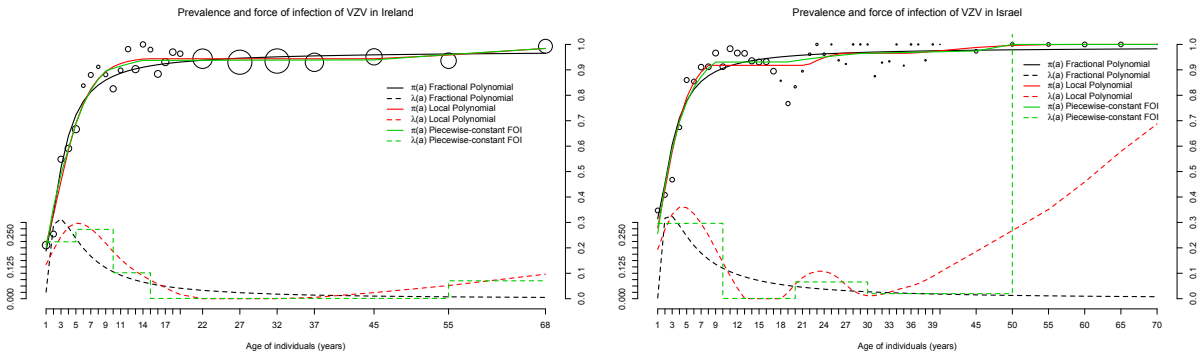
children and are too young to form a new family. From the local polynomials and the piecewise-constant FOI model, we find secondary peaks in the FOI in a number of countries. This is the case of Belgium, Germany, Great Britain, Israel, Luxembourg, and Spain, where a second peak in the FOI occurs between age 20 and 30. This age group accounts for young parents with young children, who might transmit the infection to their parents. Other secondary peaks in the FOI might happen, most of all as a consequence of the infection transmitted by children to parents and grandparents. The FOI yielded by the fractional polynomial has a similar shape for all the countries, because it is derived from the logit link function used for the model. The peak of the FOI occurs quite early in age, usually before age 5, and afterwards the FOI declines exponentially.



(a) BE (b) FI

(c) DE (d) GB

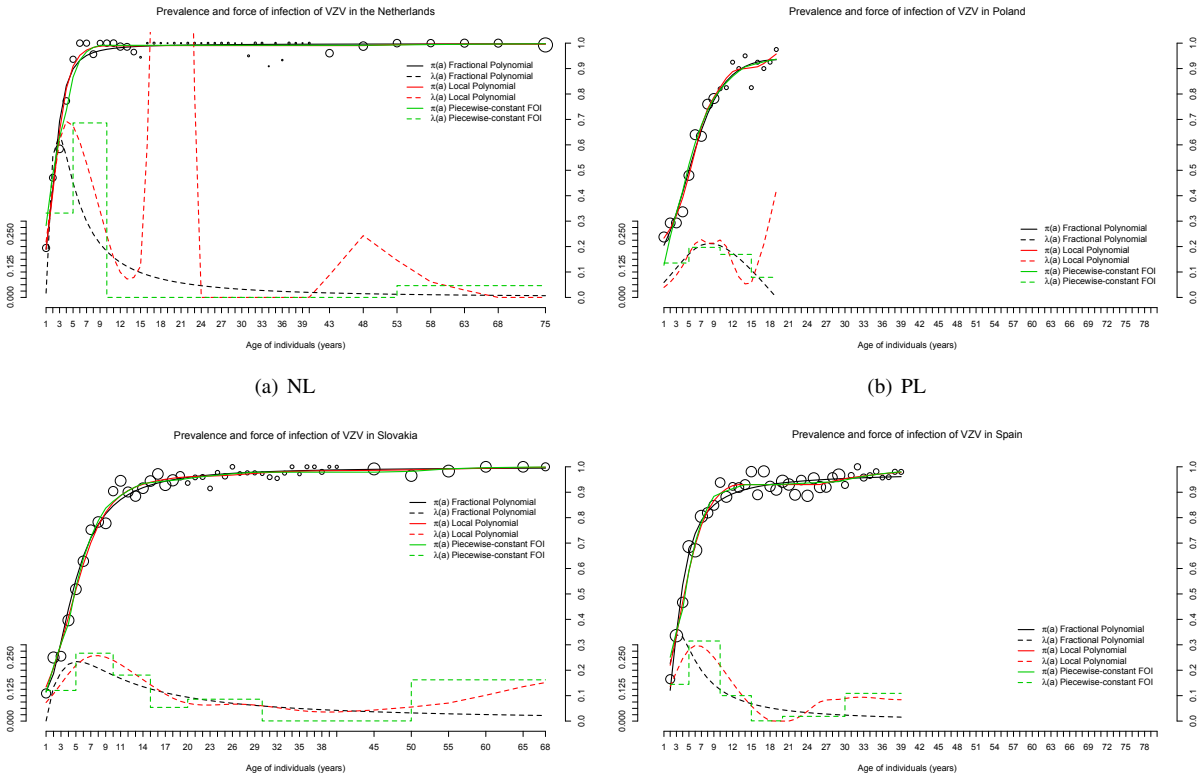
Figure 10.2: Comparison of the prevalence and FOI using fractional polynomials (FP), local polynomials (LP), and a model with piecewise constant force of infection (PC), for Belgium, Finland, Germany, and Great Britain.



(a) IE (b) IL

(c) IT (d) LU

Figure 10.3: Comparison of the prevalence and FOI using fractional polynomials (FP), local polynomials (LP), and a model with piecewise constant force of infection (PC), for Ireland, Israel, Italy, and Luxembourg.



(a) NL

(b) PL

(c) SK

(d) ES

Figure 10.4: Comparison of the prevalence and FOI using fractional polynomials (FP), local polynomials (LP), and a model with piecewise constant force of infection (PC), for the Netherlands, Poland, Slovakia, and Spain.

10.5 Discussion

The present chapter aimed to fit flexible statistical models for varicella European seroprevalence data. Varicella is a childhood disease and therefore the transmission of the infection occurs for the most part among children and, at a lesser extent, between children and other age groups.

All the fitted prevalence models reveal that the prevalence of immune reaches 90% already at age 10 in all the countries. Afterwards, the prevalence plateaus, slowly reaching almost 100%.

The results from the local polynomial and from the piecewise-constant FOI model agree in the prediction of a main peak of the FOI occurring at around age 7 or, more generally, in the age group [5–10). This group is formed by children attending primary school. They transmit the infection to each other as soon as they start the school and the pool of susceptible is rapidly emptied in the first years of school. In some countries (Belgium, Israel, Luxembourg, Netherlands), the peak of the FOI happens before, around age 5. This may be put in relation with the practice among parents of leaving very young children (age less than one year) in custody at daycare facilities.

After the peak, the FOI is usually characterized by a decrease almost down to zero, in the adolescent age groups, from age 10 to 20, when the subjects attend secondary schools and universities. These young subjects have usually very few contacts with young children and are still too young to form their own family, therefore they have less chances to enter in contact with the main transmitters of the infection.

In many countries, the FOI is characterized by secondary peaks, occurring after age 20 until age 30–35. We expect this rise of the FOI to be observed mainly in young parents living with young children, who transmit the infection to their parents. Further secondary peaks can be found in some countries, as a consequence of the transmission from children to parents and grandparents.

While the FOI estimates from the local polynomial and from the piecewise-constant FOI models usually agree, there are differences with the estimates from the fractional polynomial. The FOI function rises almost linearly in the very first age groups and reaches the main peak before age 5. After that, the FOI decreases with an exponential decline. Due to the mean structure of the model, it is not able to capture possible secondary peaks in the FOI, missing completely an important feature of varicella transmission.

A general issue with this analysis is that some of the FOIs estimated by the local polynomials and the piecewise-constant FOI models produce biologically nonsensible results, at least for some part of the age range. This is the case of Finland, Israel, and the Netherlands, where the FOI function skyrockets towards very huge values. The reason of this phenomenon is that the prevalence data in those specific age groups are poor and they can predict unrealistic big changes in prevalence and huge changes in the FOI. These problems occur usually with data for the elderly, even though in the Netherlands this is not case.

In conclusion, since we notice that a key feature of the transmission of varicella are the contacts within children and between children and adults, in Chapter 11 we will incorporate the information about contact data in the estimation of the prevalence and FOI. Even though we will not further use the FOI estimated in this chapter for the estimation of the transmission parameters, preferring rather a direct approach, which is based on the maximization of the log-likelihood of the binomial data, taking into

account the contact patterns, the results presented in Chapter 10 could be used as benchmark for the comparison with the estimates in Chapter 11, which are dependent on contact patterns.

Chapter 11

The Estimation of Contact Patterns Based Transmission Parameters of Varicella in Europe

11.1 Introduction

In Chapter 10, we have analyzed serological data for varicella in Europe, without making any assumption about the transmission of the infection, but rather describing the serology with flexible statistical models in order to estimate the prevalence and the FOI. In this chapter, we assume an age-structured SIR model for varicella and we use information about contact patterns between individuals in order to characterize the transmission process of the infection, by estimating fundamental parameters as the basic reproduction number R_0 and the critical threshold p_c. The first parameter, R_0, represents the expected number of secondary cases of infection caused by an infectious individual during his/her entire infectious period in a total susceptible population, while the second parameters, p_c, represents the minimal vaccination coverage that is necessary to eliminate the infection, under the assumption of a "perfect" vaccine, i.e., that is administered at birth, immunizing successfully in 100% of cases and conferring lifelong immunity. To calculate these quantities, it is necessary beforehand to estimate the transmission parameters associated with the contact matrices, which give information about the age-dependent contact patterns of individuals.

Countries with higher VZV transmission rates have usually lower rates of varicella complications and herpes zoster (HZ), as reported by Nardone et al. (2007). On the other hand, there are countries with lower VZV transmission rates, which could compensate with varicella vaccination. However, unless adequate vaccine coverage is achieved, namely, 85%–95% (WHO, 1998), the average age at infection is expected to increase, as well as the number of more severe cases. Moreover, there is the complication

of the unclear effects of varicella vaccination on HZ epidemiology (Brisson et al., 2003; Guzzetta et al., 2012). In Europe, this is causing the current paralysis to the introduction of mass vaccination against varicella. Therefore, a better understanding of VZV transmission is a key step for the development of better models for VZV transmission and reactivation to be used for the evaluation of the impact of vaccination programs. This can be achieved using mathematical modeling, that allows to test hypotheses about the epidemiology of an infection (Guzzetta et al., 2012), and to simulate the effects of different vaccination policies, e.g., no vaccination, one- or two-dose vaccination, etc. (Brisson et al., 2010; Van Hoek et al., 2011; Poletti et al., 2012).

A first analysis of varicella serological data from eleven European countries (Nardone and Miller, 2004) reveals a large variability in R_0 estimates among the countries. In their paper, Nardone et al. (2007) estimate the force of infection (FOI) for each country and then, using only one three-age-group WAIFW matrix (Anderson and May, 1991) to account for heterogeneity in mixing patterns (contrarily to the standard procedure, consisting in the comparison of different kinds of WAIFW matrices), they compute R_0 and p_c. They concluded that varicella transmission in Europe seems to be characterized by a neat clustering between high and low transmission countries, with the minimum found in Italy ($R_0 = 3.31$ and $p_c = 69\%$) and the maximum found in the Netherlands ($R_0 = 16.91$ and $p_c = 94\%$). We face, in particular, large variations in R_0, due either to differences in epidemiology among countries (higher and lower VZV transmission countries) or to high sensitivity of varicella to differences in contact patterns among individuals. More in general, we think that the main issue of the paper of Nardone et al. (2007) is their use of a coarse approach, consisting in the estimation of the FOI in only three age groups with a very assortative matrix. Therefore, it becomes fundamental to update their predictions on the base of more realistic and modern assumptions about contact patterns.

Measurement and understanding of contact patterns by age have seen remarkable advancements in recent years. Wallinga et al. (2006) used data on self-reported numbers of conversational contacts per person in order to estimate age-dependent transmission parameters for mumps in the Netherlands. Their approach consists in augmenting seroprevalence data with contact data arranged in age-dependent contact matrices, while assuming that the age distribution of the conversational contacts is proportional to the age distribution of the infected individuals, the so-called "social contact hypothesis". The POLYMOD study (Mossong et al., 2008) used the same approach of Wallinga et al. (2006), collecting contact data in eight European countries and classifying them, among others, for type and location. In this study, a contact was defined as either a skin-to-skin contact (a physical contact) or a two-way conversation with three or more words in the physical presence of another person without skin-to-skin contact (a nonphysical contact). Afterwards, other authors gave more detailed analyses of the available POLYMOD contact data (Hens et al., 2009a; Melegaro et al., 2011) or even presented the results of newly collected contact data (Hens et al., 2009b; Horby et al., 2011). In a similar manner, keeping the same proportionality assumption of infection data to contact data, alternative types of data on contact patterns were proposed. In particular, Zagheni et al. (2008) used time use data, while Iozzi et al. (2010) built contact matrices by counting contacts from a simulated artificial society that integrates routinely available socio-demographic data,

such as data on household composition or on school participation, with time use data.

Based on the availability of such contact matrices, considerable work specifically aimed to understand the type of contact patterns relevant for varicella transmission has been carried out. The major contributions in this area are represented by Goeyvaerts et al. (2010), Brisson et al. (2010), and Melegaro et al. (2011): Goeyvaerts et al. (2010) developed a series of models based on nested subsets of the total contact data and eventually focussed on the estimation of age-dependent transmission rates of varicella in Belgium; Brisson et al. (2010) developed a mathematical model for VZV and HZ infections and studied the effects of the introduction of a one-dose and a two-dose vaccination program, using information of contact patterns as estimated by POLYMOD; finally, Melegaro et al. (2011) determined which characteristics of the contact matrices had a stronger impact on the transmission rates of parvovirus B19 and varicella in five European countries.

This part of our work builds upon previous work done by Goeyvaerts et al. (2010), Melegaro et al. (2011), and Iozzi et al. (2010). However, our goal is different, since we aim to upgrade the work of Nardone et al. (2007) in order to provide a more detailed picture of the epidemiology of varicella in Europe in a pre-vaccination phase through a variety of approaches and data on contact patterns using robust statistical methods. Hence, our focus is more on the understanding of varicella transmission in Europe and on the uncertainty around its fundamental parameters, using data from several countries.

The chapter is structured as follows. In Section 11.2 we introduce the different types of age-specific contact matrices used in the analysis. In Section 11.3 we present the statistical methodology necessary for the estimation of the parameters of interest. In Section 11.4 we apply the methodology to varicella seroprevalence data from twelve European countries and we present the results about the uncertainty around the transmission parameters. Finally, in Section 11.5 we conclude with a discussion of the performances of the different contact matrices and of the main findings.

11.2 Social Contact Matrices

In this section we discuss alternative age-specific contact data: survey-collected contact data (Mossong et al., 2008); contact data from virtual populations (Fumanelli et al., 2012); WAIFW matrices empirically based on knowledge from the two previous types of contact data.

11.2.1 POLYMOD Matrices

POLYMOD contact data have been collected in a large multicountry population-based survey between 2005 and 2006. The following countries participated in the survey: Belgium (BE), Germany (DE), Finland (FI), Great Britain (GB), Italy (IT), Luxembourg (LU), Netherlands (NL), and Poland (PL). The participants were asked to fill in a diary with all the contacts had during a particular day. Contacts were defined either as skin-to-skin contacts or as two-way conversations with three or more words with the physical presence of the other person, but without skin-to-skin contact. For each contact, the participant

had to inform also on the age of the contacted person, the location of the contacts, the kind of contacts (physical or non-physical), the total duration of the contacts with the person, and the frequency of contacts with that person. Further details about the survey can be found in Mossong et al. (2008).

In order to use POLYMOD contact data to estimate transmission parameters, we arrange the data in contact matrices, $C_{ij} = \{c_{ij}\}$, with dimension 15×15, using fourteen 5-years age groups from 1 to 70 years and a final age for those individuals older than 70. Let y_{ij} be the observed number of contacts with individuals of age j reported per day by respondent of age i, with $i, j = 1, \ldots, J$. The mean number of contacts per day in the sample is defined as

$$m_{ij} = \frac{y_{ij}}{t_i}, \tag{11.2.1}$$

where t_i is the total number of respondents of age i. The total number of contacts per day in the population, μ_{ij}, is calculated by multiplying m_{ij} for the population size by age, w_i, obtained from demographic data,

$$\mu_{ij} = m_{ij} w_i. \tag{11.2.2}$$

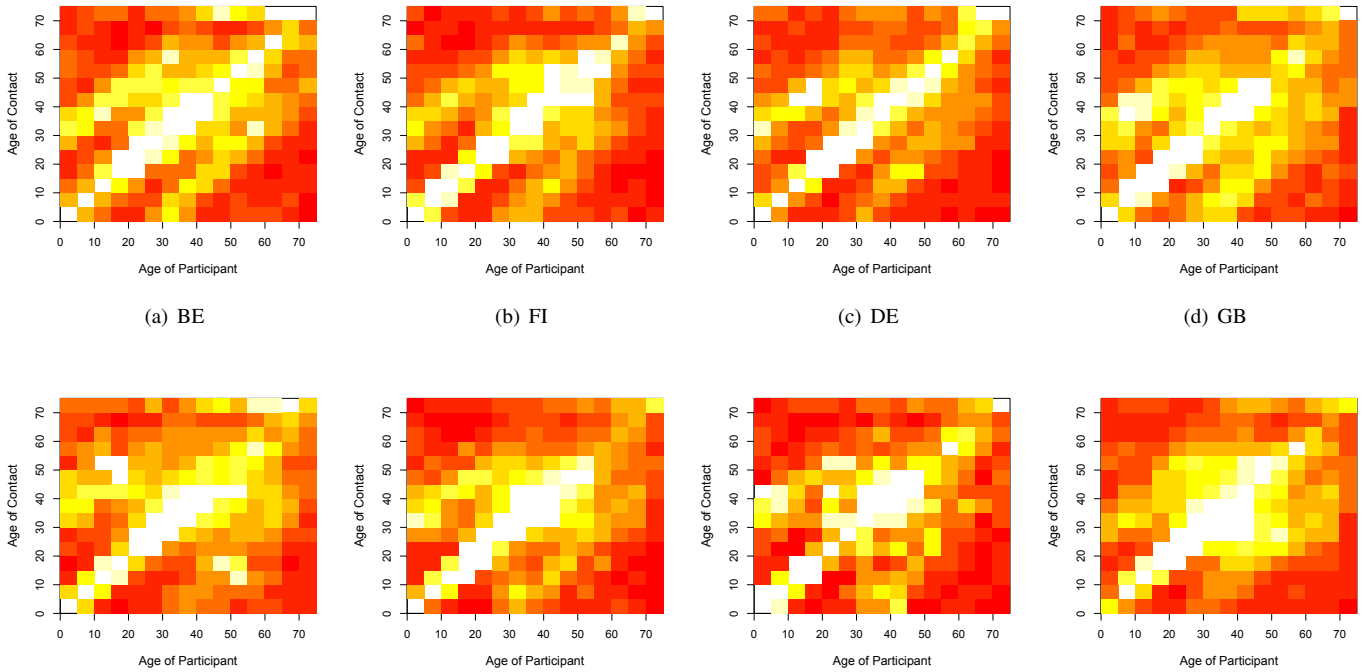
At this point, we have to take into account the reciprocal nature of contacts in the closed population, namely, $\mu_{ij} w_i = \mu_{ji} w_j$. Therefore, the total number of contacts per day in the population, after the correction for reciprocity is

$$\mu_{ij}^R = \frac{\mu_{ij} + \mu_{ji}}{2}. \tag{11.2.3}$$

Eventually, after the correction for reciprocity at the population level, we can calculate the adjusted mean number of contacts per day in the population, scaling μ_{ij}^R for the population by age, w_j:

$$c_{ij} = m_{ij}^R = \frac{\mu_{ij}^R}{w_j}. \tag{11.2.4}$$

We refer to the matrix C_{ij}, whose cells are denoted by c_{ij}, as the POLYMOD contact matrix. Figure 11.1 shows the POLYMOD contact matrices C_{ij} for all the eight countries. The main feature of the matrix is a high degree of age assortativeness, since the highest number of contacts is found among people in the same age group. However, this assortative mixing is attenuated by the presence of two wings, flanking the main diagonal, that represent the contacts between parents and children and grandparents and grandchildren. For the contacts in the remaining cells, it is reasonable to assume homogeneous mixing among individuals.



(a) BE (b) FI (c) DE (d) GB

(e) IT (f) LU (g) NL (h) PL

Figure 11.1: POLYMOD contact matrices for eight European countries. Each matrix presents the adjusted (corrected for reciprocity) mean number of contacts per day.

11.2.2 FBK Matrices

The POLYMOD matrices presented in the previous section require to conduct a survey in order to estimate the mean number of contacts per day. Such surveys are not available for all populations. An alternative approach to collecting data on contact patterns in a population survey is to construct a synthetic society by integrating available routine socio-demographic census data. Practically, researchers from the Fondazione Bruno Kessler (FBK, Trento, Italy) simulated, using an agent-based model, a virtual population of agents, of different ages, that interact with each other in different social settings: household, school, workplace and general community (Fumanelli et al., 2012). The main advantage of this method is that it is general and can be easily used for countries for which the necessary socio-demographic data are available.

The socio-demographic data used to build the virtual populations consist of descriptive statistics that were collected from the Statistical Office of the European Commission (Eurostat, 2011) and gave enough information to construct contact matrices for 26 European countries. These descriptive statistics included information on household size and composition, population age structure, rates of school attendance and employment/inactivity, school and workplace size, etc. Because of the lack of part of these socio-demographic data, Belgium, Malta, and Poland were excluded from the analysis.

In order to construct the overall country-specific contact matrix, setting-specific matrices were built and then joined in a linear combination. In this work, contacts are defined as having shared the same physical environment. Through an agent-based model, the socio-demographic information is used to create a simulated population, which is synthesized by the following considered environments: one for contacts within the households, that is, between children and parents (matrix H), one for contacts at school, between pupils and with teachers (matrix S), and one for contacts in the workplace, between colleagues (matrix W). A fourth environment is also considered, for contacts in the general community (matrix R). This environment is not described by the agent-based model, because of the absence of specific socio-demographic information, therefore homogeneous mixing is assumed.

For each setting, the respective matrix, containing relative frequencies of contacts between age groups, is calculated in the following way:

$$C_{ij}^K = p_i^K f_{ij}^K, \tag{11.2.5}$$

where p_i^K is the probability for the age group i to have at least one contact in the setting K, and f_{ij}^K is the frequency of contacts within setting K between individuals of age i and age j. Since not all the settings contribute in the same way to the infection transmission, the four matrices are combined together in a linear combination, with the weights representing the proportions of transmission in each setting. For this scope, the proportions for the transmission of influenza-like-illness (ILI) are used (Merler et al., 2011). Therefore, given that $K \in \{H,S,W,R\}$,the final contact matrix $C_{ij} = \{c_{ij}\}$ is obtained through the following linear combination:

$$C_{ij} = \sum_K \alpha_K C_{ij}^K = 0.30H_{ij} + 0.18S_{ij} + 0.19W_{ij} + 0.33R_{ij}. \tag{11.2.6}$$

Even though these coefficients α_K refer to the ILI, they have been already used to generate a synthetic contact matrix for Italy, which has been shown to fit properly varicella and parvovirus B19 data (Iozzi et al., 2010).

These FBK contact matrices have been validated against POLYMOD contact matrices, by jointly regressing their elements c_{ij} against those from POLYMOD matrices for all the countries. It has been found that they are linearly correlated ($R^2 = 0.72$), up to a scaling factor. As an example, Figure 11.2 shows a scatter plot of POLYMOD contact data against FBK contact data for Italy, from which it can be seen the linear relation existing between the two types of contact data. Moreover, the matrices have been

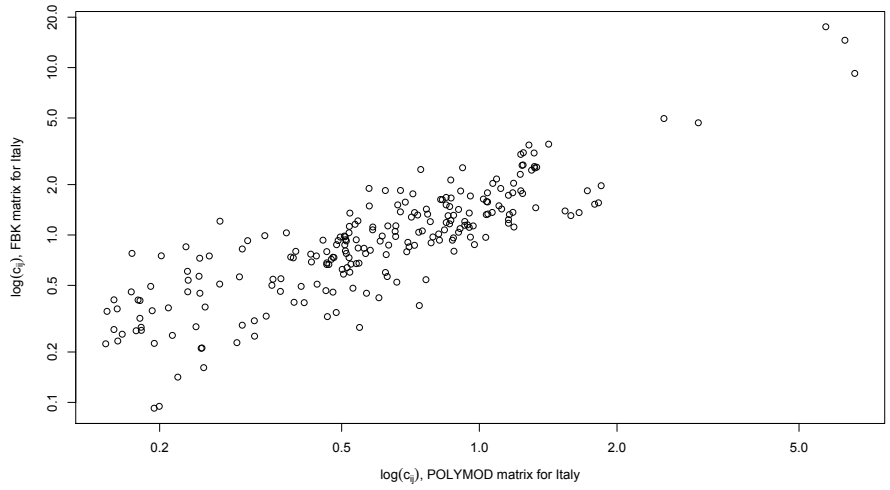
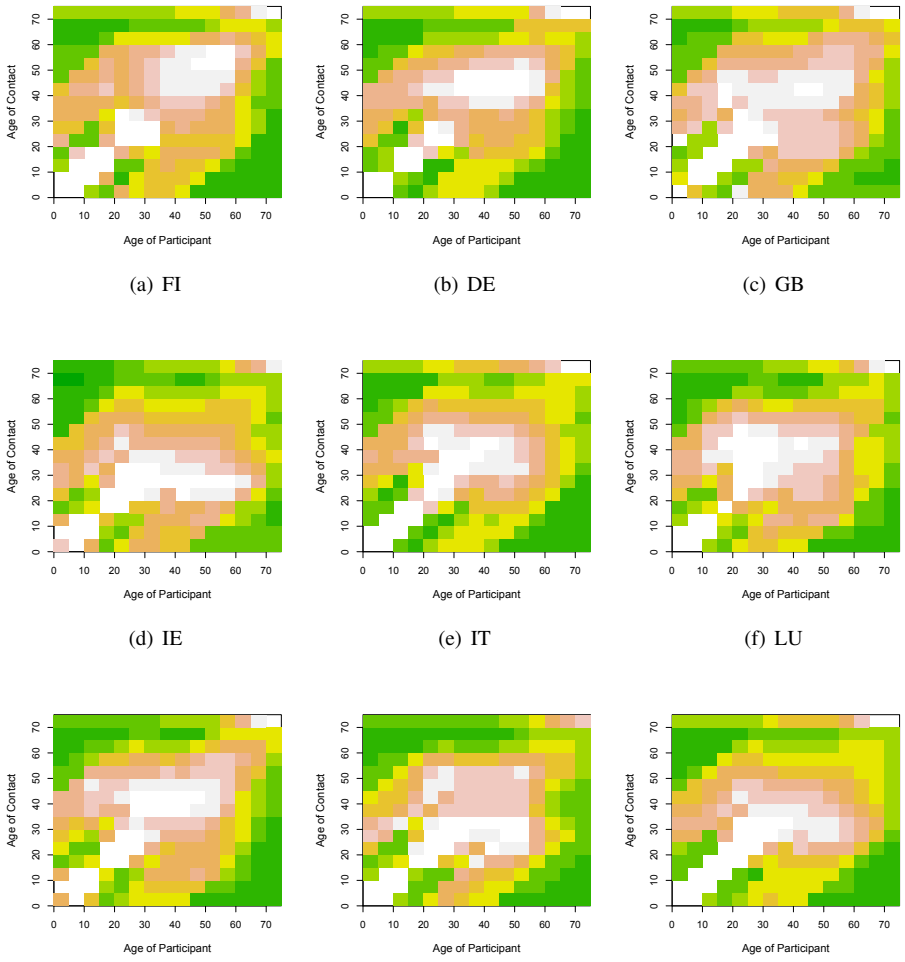


Figure 11.2: Scatter plot, on a logarithmic scale, of POLYMOD contact data (on the y axis) against FBK contact data (on the x axis) for Italy.

used in the analysis of influenza prevalence in UK and their performance has been compared to the one of POLYMOD contact matrices. The result is that FBK matrices provided an acceptable fit, quite similar to the one given by POLYMOD matrices. Note that, in contrast with POLYMOD data, FBK contact data are not affected by sampling error, since the contact matrices are a function of descriptive statistics, which, by definition, are without uncertainty. For this reason, the FBK contact data are smoother than POLYMOD contact data, because the latter are collected in a survey and thus are affected by sampling variability. However, this does not mean that FBK are smooth matrices in general, as it can be seen from Figure 11.3, mostly as a consequence of the mixing among children within classes.

As already mentioned, Figure 11.3 shows the overall FBK contact matrices for nine countries among those presented in Chapter 10 for the analysis of varicella serology. Similarly to POLYMOD matrices, FBK matrices present a high proportion of contacts in the main diagonal, accounting for assortative contacts by age, and two smaller diagonals, flanking the main one, accounting for the contacts within the households.



(a) FI (b) DE (c) GB

(d) IE (e) IT (f) LU

(g) NL (h) SK (i) ES

Figure 11.3: FBK contact matrices for nine European countries calculated according to Eq. (11.2.6).

11.2.3 Empirically Based WAIFW Matrices

For more than twenty years, the "Who Acquires Infection From Whom?" (WAIFW, Anderson and May, 1991) matrices have been the fundamental tool in order to model the transmission of infectious diseases accounting for heterogeneous mixing patterns among individuals. The characteristics of this technique is to provide, in a rather coarse manner, a structure for age-dependent contact patters, estimating transmission parameters from epidemiological data (seroprevalence or incidence data), either directly or indirectly. Van Effelterre et al. (2009), Ogunjimi et al. (2009), and Goeyvaerts et al. (2010), using data for varicella in Belgium, estimated WAIFW transmission parameters indirectly from seroprevalence data, given an estimate of the FOI at equilibrium. They used six different WAIFW matrices, with each matrix presenting a different degree of assortativeness among individuals, according to different theoretical assumptions about the mixing patterns of individuals. For identifiability reasons, the number of free parameters of the matrix is required to be equal to the number of considered age groups, therefore they chose to have six parameters for six age groups. On the other hand, Farrington et al. (2001) estimated WAIFW transmission parameters directly from seroprevalence data, maximizing the binomial likelihood of data with respect to the transmission parameters, under the constraint of positivity.

Although the recent availability of contact data, collected from population surveys, seemed to make the WAIFW approach, based rather on theoretical mixing patterns, quite obsolete, in this work, we propose a novel approach resurrecting the use of WAIFW matrices, taking advantage of the knowledge progress allowed by POLYMOD data. The idea is that the structure of the WAIFW matrix can be built upon the information derived by the collected or simulated contact data. This technique is a "cost-free" idea, since it only requires to arrange the structure of the matrix according to the shape of a contact matrix.

We consider the following two configurations for the WAIFW matrices, based on the transmission parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)$. They represent the "adequate" number of contacts in order to successfully transmit the infection. The transmission parameters are estimated following the direct approach of Farrington et al. (2001), based on the maximization of the log-likelihood of the seroprevalence data. The first WAIFW matrix, W1, shown in (11.2.7), is based on the following $J = 6$ age groups: 0–4, 5–9, 10–19, 20–29, 30–49, 50+. The first four parameters, from β_1 to β_4, account for assortativeness in the same age groups, the parameter β_5 for contacts between adults and children, e.g., in the family or at school with teachers, while β_6 is the parameter for the remaining contacts in the background. The second WAIFW matrix, W2, shown in (11.2.8), although it consists of fourteen 5-years age classes up to age 70, followed by a final class containing all the individuals older than 70, has only $J = 7$ transmission parameters β_j. In this way, the matrix is directly comparable with POLYMOD and FBK matrices, since they have the same number of age groups, but also is parsimonious, having only seven parameters. The first four parameters, from β_1 to β_4, account for assortative contacts up to age 20; the parameter β_5 stands for a broader assortative group for adults, between age 20 and 50, with contacts more typical on a work environment; finally, parameter β_6 accounts for contacts between adults and children and parameter β_7 for background contacts.

$$W1 = \begin{pmatrix} & 5 & 10 & 20 & 30 & 50 & 50+ \\ 5 & \beta_1 & \beta_6 & \beta_6 & \beta_5 & \beta_5 & \beta_5 \\ 10 & \beta_6 & \beta_2 & \beta_6 & \beta_6 & \beta_5 & \beta_5 \\ 20 & \beta_6 & \beta_6 & \beta_3 & \beta_6 & \beta_6 & \beta_5 \\ 30 & \beta_5 & \beta_6 & \beta_6 & \beta_4 & \beta_6 & \beta_6 \\ 50 & \beta_5 & \beta_5 & \beta_6 & \beta_6 & \beta_4 & \beta_6 \\ 50+ & \beta_5 & \beta_5 & \beta_5 & \beta_6 & \beta_6 & \beta_4 \end{pmatrix} \tag{11.2.7}$$

$$W2 = \begin{pmatrix} & 5 & 10 & 15 & 20 & 25 & 30 & 35 & 40 & 45 & 50 & 55 & 60 & 65 & 70 & 70+ \\ 5 & \beta_1 & \beta_7 & \beta_7 & \beta_7 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_7 & \beta_7 & \beta_7 & \beta_7 & \beta_7 & \beta_7 \\ 10 & \beta_7 & \beta_2 & \beta_7 & \beta_7 & \beta_7 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_7 & \beta_7 & \beta_7 & \beta_7 & \beta_7 \\ 15 & \beta_7 & \beta_7 & \beta_3 & \beta_7 & \beta_7 & \beta_7 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_7 & \beta_7 & \beta_7 & \beta_7 \\ 20 & \beta_7 & \beta_7 & \beta_7 & \beta_4 & \beta_7 & \beta_7 & \beta_7 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_7 & \beta_7 & \beta_7 \\ 25 & \beta_6 & \beta_7 & \beta_7 & \beta_7 & \beta_5 & \beta_5 & \beta_5 & \beta_5 & \beta_5 & \beta_5 & \beta_6 & \beta_6 & \beta_6 & \beta_7 & \beta_7 \\ 30 & \beta_6 & \beta_6 & \beta_7 & \beta_7 & \beta_5 & \beta_5 & \beta_5 & \beta_5 & \beta_5 & \beta_5 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_7 \\ 35 & \beta_6 & \beta_6 & \beta_6 & \beta_7 & \beta_5 & \beta_5 & \beta_5 & \beta_5 & \beta_5 & \beta_5 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 \\ 40 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_5 & \beta_5 & \beta_5 & \beta_5 & \beta_5 & \beta_5 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 \\ 45 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_5 & \beta_5 & \beta_5 & \beta_5 & \beta_5 & \beta_5 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 \\ 50 & \beta_7 & \beta_6 & \beta_6 & \beta_6 & \beta_5 & \beta_5 & \beta_5 & \beta_5 & \beta_5 & \beta_5 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 \\ 55 & \beta_7 & \beta_7 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 \\ 60 & \beta_7 & \beta_7 & \beta_7 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 \\ 65 & \beta_7 & \beta_7 & \beta_7 & \beta_7 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 \\ 70 & \beta_7 & \beta_7 & \beta_7 & \beta_7 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 \\ 70+ & \beta_7 & \beta_7 & \beta_7 & \beta_7 & \beta_7 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 \end{pmatrix} \tag{11.2.8}$$

These WAIFW matrices are compared with one of the classical matrices firstly presented by Anderson and May (1991), namely, a 6×6 matrices with five different parameters, from β_1 to β_5, accounting for assortativeness in the age groups and one final parameter, β_6, for the background transmission. This matrix, W3, is shown in (11.2.9).

$$W3 = \begin{pmatrix} & 5 & 10 & 20 & 30 & 50 & 50+ \\ 5 & \beta_1 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 \\ 10 & \beta_6 & \beta_2 & \beta_6 & \beta_6 & \beta_6 & \beta_6 \\ 20 & \beta_6 & \beta_6 & \beta_3 & \beta_6 & \beta_6 & \beta_6 \\ 30 & \beta_6 & \beta_6 & \beta_6 & \beta_4 & \beta_6 & \beta_6 \\ 50 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_5 & \beta_6 \\ 50+ & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_5 \end{pmatrix} \tag{11.2.9}$$

11.3 Statistical Methods

11.3.1 Estimating Transmission Rates, R_0 and p_c for POLYMOD and FBK Contact Matrices

We are interested in estimating the age-dependent VZV prevalence $\pi(a)$ and FOI $\lambda(a)$ for a discrete age-structured SIR model in endemic equilibrium. Assuming a constant population, type I mortality (death rate $\mu = 0$ until age at life expectancy L, afterwards everybody is considered dead), and discrete age groups, the SIR model for varicella is given by the following system of differential equations (Anderson and May, 1991):

$$\begin{cases} \dot{x}_i &= -\lambda_i x_i, \\ \dot{y}_i &= \lambda_i x_i - \nu y_i, \\ \dot{z}_i &= \nu y_i, \end{cases} \tag{11.3.10}$$

where x_i, y_i, and z_i are the proportions of susceptible, infectious, and immune (or recovered) individuals in the population at age a_i, respectively, λ_i is the age-dependent FOI, and ν is the recovery rate, equal to $1/D$, where D is here the average infectious period of varicella (taken equal to 7 days). Assuming a piecewise-constant model for the FOI λ_i within discrete age groups, the solution of the system in (11.3.10) is given by the following system of equations:

$$\begin{cases} x_i &= \exp\left(-\sum_{j=1}^{i} \lambda_j (a_j - a_{j-1})\right), \\ y_i &= D\frac{x_{i-1}-x_i}{a_i - a_{i-1}}, \\ z_i &= 1 - x_i - y_i. \end{cases} \tag{11.3.11}$$

According to the "social contact hypothesis" (Wallinga et al., 2006), the FOI acting on susceptible at age i at the endemic equilibrium is given by

$$\lambda_i = \sum_{j=1}^{J} q C_{ij} y_j, \tag{11.3.12}$$

where q is the transmission coefficient, $C_{ij} = \{c_{ij}\}$ is the contact matrix (either POLYMOD or FBK), containing the mean number of contacts per day between susceptible aged i and infectious aged j in the sample, after the correction for reciprocity, and y_j is the proportion of infectious aged j. The transmission coefficient q reflects the social contact hypothesis and is a scale factor which works as proportionality factor between the contact data and the infection data. However, more complex assumptions can be made. If the parameter depends on the age of susceptible, q_i, we talk about age-related "differential susceptibility", namely, the susceptibility of an individual to the infection changes with the age. If the parameter depends on the age of the infected, we have instead age-related "differential infectivity". Finally, the transmission parameter can depend on both the age of susceptible i and on the age of infected j (Goeyvaerts et al., 2010). In this way, we have a matrix of transmission rates, which behaves similarly to a WAIFW matrix.

Let z_k be the number of immune individuals of age k, $k = 1,\ldots,\max(\text{age})$, observed in a sample of size n_k. For the social contact hypothesis, in order to estimate the transmission parameter q, we need to to maximize the binomial log-likelihood of the observed serological data z_k (Wallinga et al., 2006; Goeyvaerts et al., 2010; Iozzi et al., 2010; Melegaro et al., 2011):

$$\ell(q) = \sum_{k=1}^{K} z_k \log(\pi_k(q)) + (n_k - z_k)\log(1-\pi_k(q)), \tag{11.3.13}$$

subject to the following conditions: (1) the chosen contact matrix $C_{ij} = \{c_{ij}\}$ and its age grouping, $j = 1,\ldots,J$; (2) the transmission parameters $q > 0$; (3) $R_0 > 1$, which is the necessary condition for the infection to reach the endemic state.

Finally, for the estimation of the basic reproduction number R_0, we calculate the dominant eigenvalue of the "next generation matrix" M_{ij}, given by

$$M_{ij} = DqC_{ij}, \tag{11.3.14}$$

where D is again the average infectious period of varicella. The critical threshold p_c is defined as $p_c = 1-(1/R_0)$ (Anderson and May, 1991).

In case more complex assumptions for the transmission parameters are used, the procedure to estimate R_0 is similar to the one described above.

11.3.2 Estimating Transmission Rates, R_0 and p_c for WAIFW Contact Matrices

Similarly to the methodology presented for POLYMOD and FBK contact matrices in Section 11.3.1, we can estimate the transmission parameters for the empirically based WAIFW matrices based on the maximization of the log-likelihood (Farrington et al., 2001). Consider a WAIFW matrix,

$$B_{ij} = \{\beta_j\} = qC_{ij}, \tag{11.3.15}$$

which implies that the transmission parameters β_j represent the mean number of "adequate" contacts to transmit the infection from infectious aged j to susceptible aged i. Therefore, we reparameterize Eq. (11.3.12) as

$$\lambda_i = \sum_{j=1}^{J} B_{ij} y_j, \tag{11.3.16}$$

and we maximize the log-likelihood of the seroprevalence data y_k with respect to $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_J)$, that is,

$$\ell(\boldsymbol{\beta}) = \sum_{k=1}^{K} y_k \log(\pi_k(\boldsymbol{\beta})) + (n_k - y_k)\log(1-\pi_k(\boldsymbol{\beta})). \tag{11.3.17}$$

Similar to Section 11.3.1, we restrict the transmission parameters β_j to be nonnengative. Eventually, the next generation matrix M_{ij}, from which we compute R_0 and p_c, is given by

$$M_{ij} = DB_{ij}. \tag{11.3.18}$$

11.3.3 Estimation of the Standard Errors

Standard errors and confidence intervals for the transmission parameters, R_0, and p_c are obtained using the bootstrap method. Different bootstrap algorithms were used for the different scenarios.

For the FBK matrices, the numbers in each cell c_{ij} actually represent the number of contacts between counted individuals aged i and individuals aged j in the synthetic population created using information provided by routine socio-demographic census data. Since these data are mostly descriptive census statistics, and the contact matrix is just of function of such descriptive statistics, the FBK contact matrix does not present any sampling variability (Fumanelli et al., 2012). This is reflected in the bootstrap procedure, thus we assume that the only source of uncertainty is the serological datum, thus we opt for a parametric bootstrap procedure, resampling the serological data, stratified by age, from a binomial density, with size n_i and probability $\hat{\pi}_i$, estimated with the piecewise-constant FOI model.

In contrast with the FBK matrices, the POLYMOD contact matrices are affected by sampling variability. According to a simulation study (Marangi, 2011), contact data appear to be the most important source of uncertainty, largely more than serological data. The author indeed showed that the total bootstrap standard error is essentially dominated by the bootstrap standard error of the contact component. As a consequence, we decided to adopt a approximated nonparametric bootstrap procedure, resampling the contact data only.

Finally, for the empirically based WAIFW matrices, we follow the same procedure of the FBK matrices, by resampling parametrically serological data stratified by age from a binomial density, with size n_i and probability $\hat{\pi}_i$, estimated with the piecewise-constant FOI model.

11.4 Results for Varicella Transmission in Europe

The matrices presented in the previous section were fitted to the serological data for VZV coming from ESEN2 and POLYMOD projects. For the contact matrices FBK and POLYMOD, we fitted two alternative models for the transmission parameter: a constant q, reflecting the social contact hypothesis, and a simple "differential susceptibility" model for q_i, consisting in a piecewise-constant age-dependent model within the age groups 5–9 and 10+.

Tables 11.1 – 11.2 present country-specific estimates of the transmission parameters, either $\boldsymbol{q}$ or $\boldsymbol{\beta}$, for all the matrices available for the specific country (note that FBK and POLYMOD matrices are not available for all twelve countries), the BIC, and the estimates with 95% percentile bootstrap (Davison and Hinkley, 1997) confidence intervals (CIs) for R_0 and p_c.

For what concerns the transmission parameters, we see that, depending on the country and on the specific contact matrix, it is preferable either the model with one q, favoring thus the standard social contact hypothesis of q as a proportionality factor, or the model with age-related differential susceptibility, with the transmission being higher either mostly among children than among adolescents and adults (see, for instance, Belgium, Great Britain, Ireland, Luxembourg, Netherlands, and Poland), even though in a few countries it happens the contrary (Germany, Finland and Slovakia). Large variations between the two q_i may indicate large differences in transmission between the two age groups and this thus produces changes in the FOI. Indeed, in those countries where we have an increase of the FOI in the older age groups are also those with a q_i larger for adolescents and adults.

Figure 11.4 displays, for each country, a comparison between the estimates of R_0 (with 95% CI) obtained from the matrices that we have presented and with the estimates of Nardone et al. (2007). The most evident detail emerging from Tables 11.1 – 11.2 and from Figure 11.4 is the large variability around R_0 and, consequently, p_c. We notice that FBK and POLYMOD generally lead to smaller R_0, while WAIFWs matrices tend to give higher values, due to the higher degree of assortativeness by age present in these matrices. We note that the estimates for FBK matrices have generally shorter 95% CIs than POLYMOD matrices, because the way they are built (contacts counting in synthetic populations) allows an intrinsic lower degree of uncertainty.

Moving to WAIFW matrices, the classical WAIFW of Anderson and May (1991), W3, is the one giving the largest value for R_0, because it is strongly assortative by age and only one parameter accommodate for the contacts between different age groups, for which we assume homogeneous mixing. The two new proposed WAIFWs, W1 and W2, built upon the observation of the shape of FBK and POLYMOD matrices, still yield larger values for R_0 than the contact matrices, but not so large as the classical WAIFW does. This happens because the high degree of assortativeness by age of the matrices is somehow mitigated by the use of two/three parameters for the contacts between different age groups (β_5 and β_6 in W1; and β_5, β_6, and β_7 in W2).

Some issues arose related to the estimation of the βs parameters of the WAIFW matrices. First, the piecewise-constant FOI λ_i is a linear combination of the βs and therefore, due to the parameterization of the model for the optimization problem, the configuration of the WAIFW matrices should be adjusted according to the age groups in the available data. Second, we encountered problems in the calculation of the percentile bootstrap CIs for R_0 for the WAIFW matrices. While the distribution of the bootstrap replicates for R_0 for FBK and POLYMOD results in a unimodal distribution (see, panels a and b in Figure 11.5), the distribution of the bootstrap replicates for the WAIFW matrices results in a bimodal one (see Figure 11.5, panel c and d). The reason of this phenomenon might be in the estimation procedure that we used, rather than in the configuration of the matrices. In consequence, we do not report the CIs for the WAIFW matrices in Tables 11.1 – 11.2.

As concerns goodness-of-fit, we could not find a matrix that fits well all the data for the all the country, but rather in every specific situation there is a different matrix that presents a smaller BIC.

In Figures 11.6 – 11.8, we give a graphical representation of the results. For each country, we display

the estimated prevalence (solid curve) and FOI (dashed curve) for all the available contact matrices. Also in this case, the peak of the FOI occurs in the age group 5–9, even though in some countries, like Belgium, the peak of FOI occurs before age 5, because of the widespread custom among families of having their babies guarded in the day-care centers.

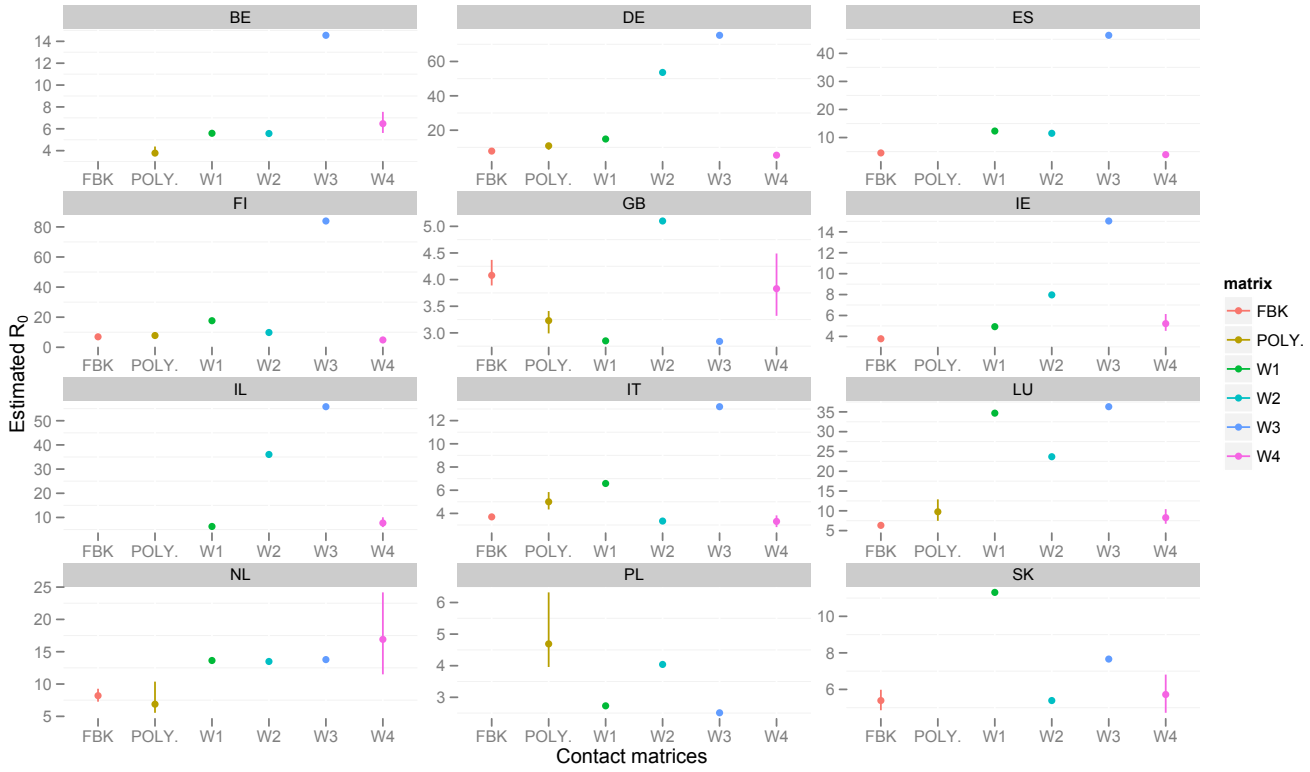
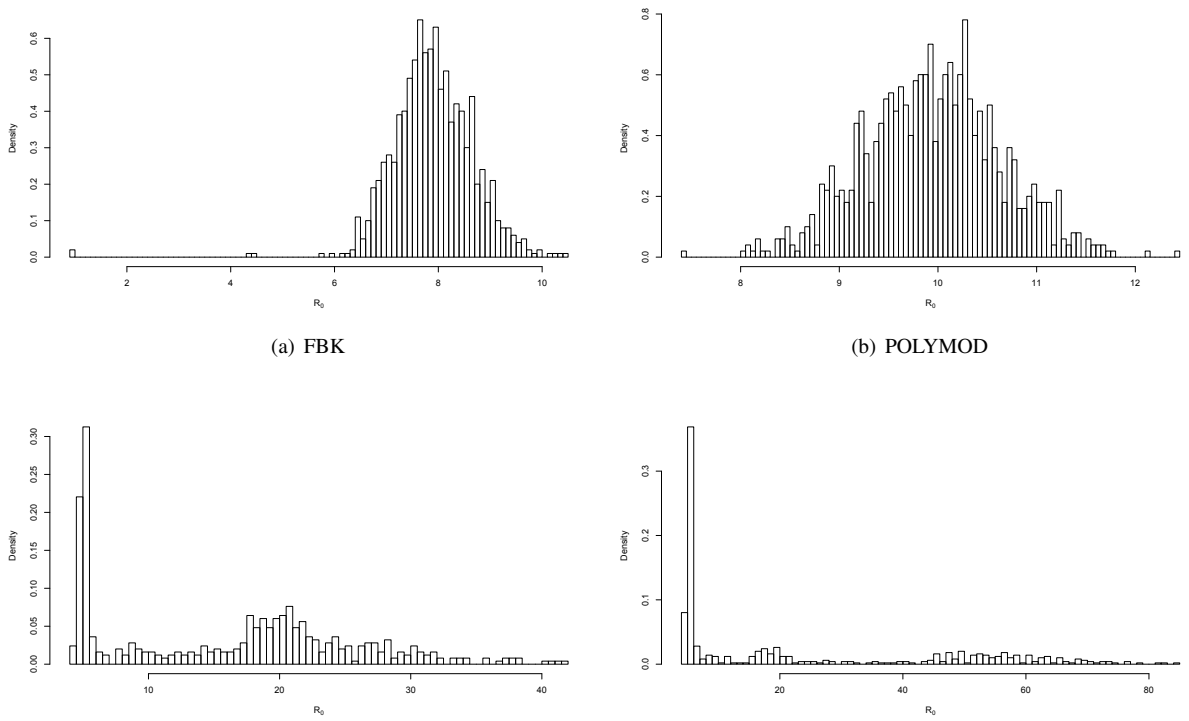


Figure 11.4: Comparison of the basic reproduction number R_0 per country, obtained using the proposed matrices. For comparison's sake, we displayed also the estimates obtained by Nardone et al. (2007) with the WAIFW matrix W4.



(a) FBK

(b) POLYMOD

(c) W1

(d) W2

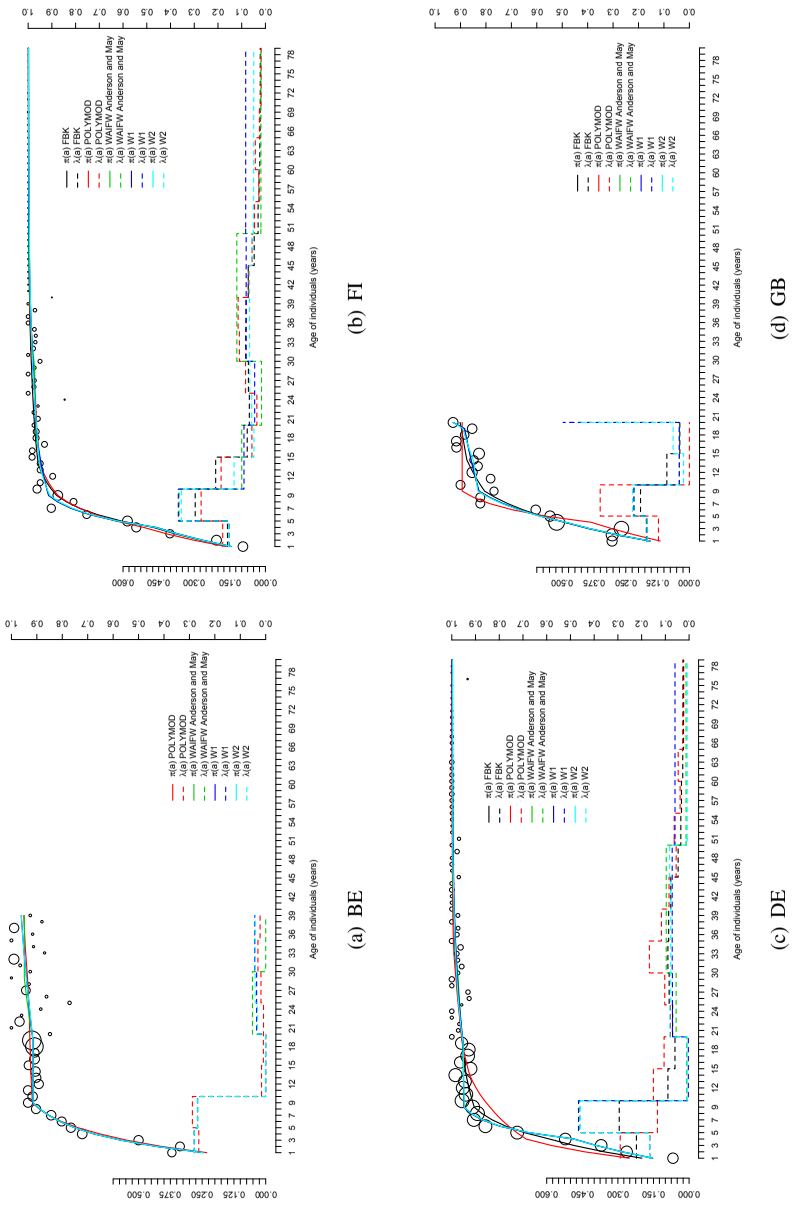
Figure 11.5: Distribution of bootstrap replicates of R_0, arising from a parametric bootstrap of the binomial data, under four contact matrices: FBK, POLYMOD, W1, and W2.

Table 11.1: Estimates of transmission parameters, R_0, and p_c (and respective 95% percentile bootstrap confidence intervals) for varicella in twelve European countries, according to different contact matrices: FBK, POLYMOD, data based WAIFW (W1 and W2), and theoretical WAIFW (W3, Anderson and May, 1991) matrices. Concerning the bootstrap CIs for the WAIFW matrices, see comments in Section 11.4.

Country	Matrix	BIC	q	q_i $[1,9)$	q_i $[10,\max(age))$	R_0	p_c
BE	POLYMOD	**194.83**	-	51.53 (39.54–64.92)	7.18 (0.00–14.52)	3.78 (3.54–4.38)	0.74 (0.72–0.77)
	W1	199.12	-	-	-	5.59	0.82
	W2	195.33	-	-	-	5.57	0.82
	W3	203.33	-	-	-	14.55	0.93
DE	FBK	224.65	-	21.92 (20.59–23.26)	33.81 (28.21–40.58)	7.85 (6.64–9.34)	0.87 (0.85–0.89)
	POLYMOD	319.12	-	42.46 (31.66–57.06)	70.96 (54.56–73.50)	10.92 (8.64–11.36)	0.91 (0.88–0.91)
	W1	213.18	-	-	-	14.87	0.93
	W2	**204.81**	-	-	-	53.61	0.98
	W3	208.66	-	-	-	75.22	0.99
ES	FBK	**190.88**	19.36 (18.56–20.19)	-	-	4.52 (4.33–4.71)	0.78 (0.77–0.79)
	W1	195.92	-	-	-	12.32	0.92
	W2	193.88	-	-	-	11.51	0.91
	W3	199.02	-	-	-	46.43	0.98
FI	FBK	**153.50**	28.97 (26.93–31.26)	-	-	6.94 (6.45–7.49)	0.86 (0.85–0.87)
	POLYMOD	158.07	-	26.74 (22.37–31.70)	35.11 (29.77–39.71)	7.80 (6.87–8.61)	0.87 (0.8–0.88)
	W1	160.40	-	-	-	17.64	0.94
	W2	165.01	-	-	-	9.80	0.90
	W3	164.70	-	-	-	83.93	0.99
GB	FBK	**147.60**	16.07 (15.32–16.83)	-	-	4.08 (3.89–4.27)	0.75 (0.74–0.77)
	POLYMOD	185.68	-	24.87 (21.25–28.86)	0.00 (0.00–7.36)	3.23 (2.99–3.41)	0.69 (0.67–0.71)
	W1	154.37	-	-	-	2.85	0.65
	W2	157.15	-	-	-	5.10	0.80
	W3	151.38	-	-	-	2.84	0.65
IE	FBK	151.73	-	31.90 (28.61–34.72)	1.53 (0.00–5.26)	3.77 (3.52–4.05)	0.73 (0.72–0.75)
	W1	156.35	-	-	-	4.93	0.80
	W2	161.69	-	-	-	7.97	0.87
	W3	**147.13**	-	-	-	15.04	0.93

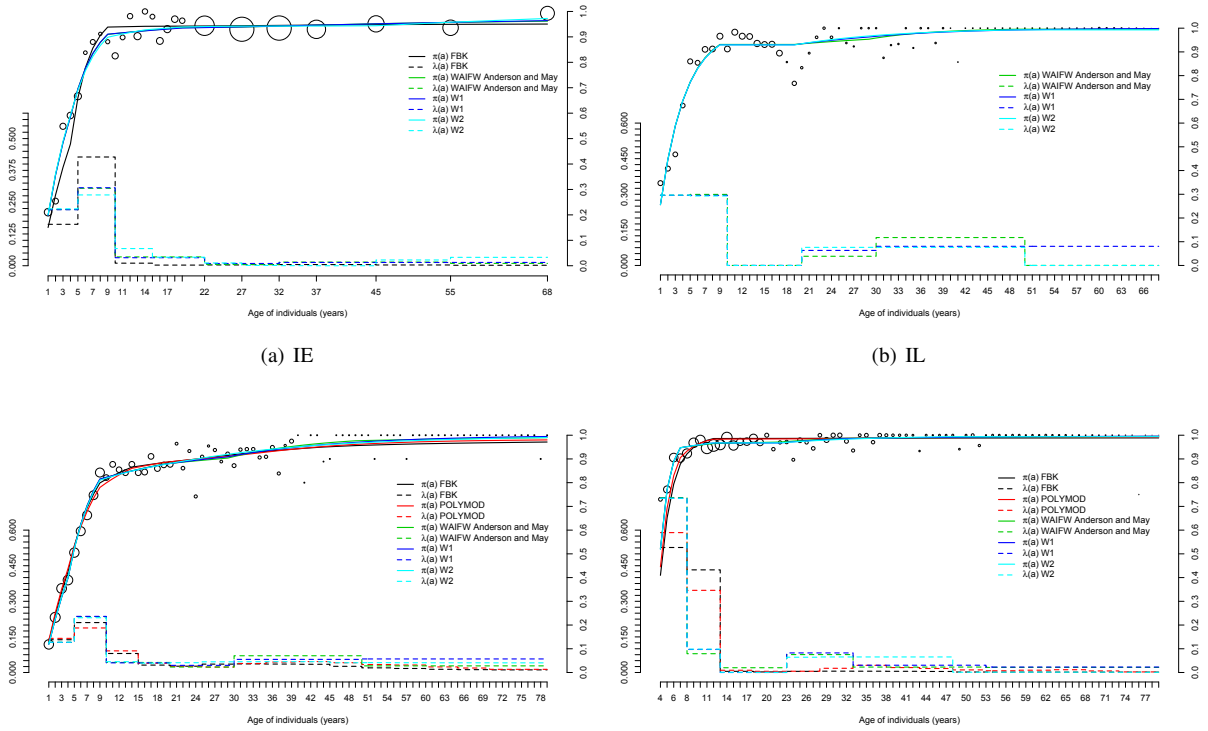
Table 11.2: Estimates of transmission parameters, R_0, and p_c (and respective 95% percentile bootstrap confidence intervals) for varicella in twelve European countries, according to different contact matrices: FBK, POLYMOD, data based WAIFW (W1 and W2), and theoretical WAIFW (W3, Anderson and May, 1991) matrices. Concerning the bootstrap CIs for the WAIFW matrices, see comments in Section 11.4.

Country	Matrix	BIC	q	q_i $[1,9)$	q_i $[10,\max(age))$	R_0	p_c
IL	W1	**157.14**	-	-	-	6.27	0.84
	W2	166.88	-	-	-	36.05	0.97
	W3	161.90	-	-	-	55.81	0.98
IT	FBK	204.53	15.67 (14.96–16.54)	-	-	3.70 (3.53–3.90)	0.73 (0.72–0.74)
	POLYMOD	**202.86**	11.94 (10.30–13.62)	-	-	5.00 (4.34–5.84)	0.80 (0.77–0.83)
	W1	210.70	-	-	-	6.58	0.85
	W2	216.60	-	-	-	3.34	0.70
	W3	215.38	-	-	-	13.20	0.92
LU	FBK	187.49	-	51.70 (47.35–55.83)	1.79 (0.00–11.77)	6.33 (5.89–7.00)	0.84 (0.83–0.86)
	POLYMOD	170.12	-	56.01 (43.38–73.60)	5.69 (0.00–9.22)	9.76 (7.47–12.90)	0.90 (0.87–0.92)
	W1	158.12	-	-	-	34.68	0.97
	W2	**157.67**	-	-	-	23.67	0.96
	W3	162.74	-	-	-	36.31	0.97
NL	FBK	**118.05**	-	71.28 (61.42–79.99)	0.00 (0.00–13.95)	8.20 (7.27–9.27)	0.88 (0.86–0.89)
	POLYMOD	147.03	-	27.20 (18.85–42.71)	13.23 (0.00–22.18)	6.88 (5.55–10.37)	0.85 (0.82–0.90)
	W1	135.45	-	-	-	13.64	0.93
	W2	139.80	-	-	-	13.49	0.93
	W3	135.50	-	-	-	13.78	0.93
PL	POLYMOD	152.69	-	21.92 (18.05–26.69)	13.85 (10.60–18.13)	4.69 (3.96–6.32)	0.79 (0.75–0.84)
	W1	116.52	-	-	-	2.73	0.63
	W2	**116.18**	-	-	-	4.04	0.75
	W3	116.52	-	-	-	2.51	0.60
SK	FBK	**173.56**	-	17.81 (16.49–19.19)	22.09 (19.49–24.96)	5.39 (4.86–5.98)	0.81 (0.79–0.83)
	W1	197.73	-	-	-	11.31	0.91
	W2	196.02	-	-	-	5.39	0.81
	W3	193.70	-	-	-	7.66	0.87



(a) BE (b) FI (c) DE (d) GB

Figure 11.6: Comparison of the prevalence (solid curve) and the piecewise-constant FOI (dashed curve) of varicella based on several contact matrices for Belgium, Finland, Germany, and Great Britain.



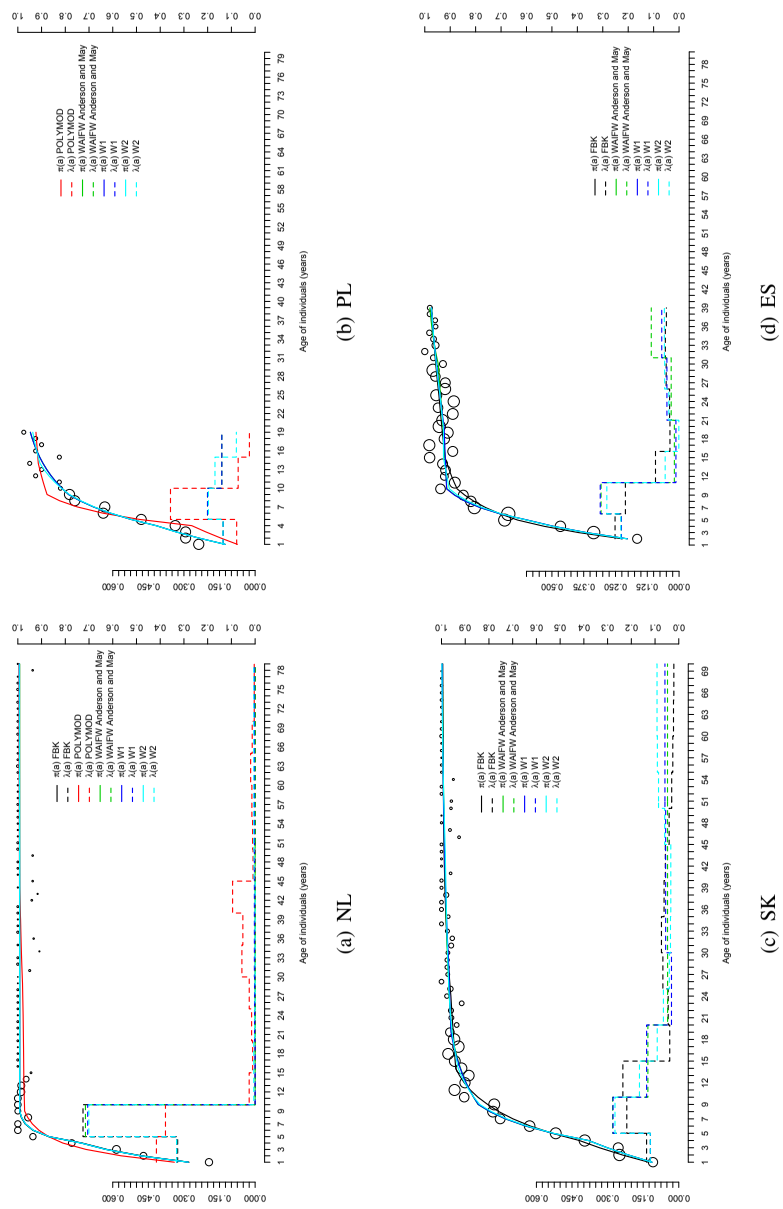
(a) IE

(b) IL

(c) IT

(d) LU

Figure 11.7: Comparison of the prevalence (solid curve) and the piecewise-constant FOI (dashed curve) of varicella based on several contact matrices for Ireland, Israel, Italy, and Luxembourg.



(a) NL (b) PL

(c) SK (d) ES

Figure 11.8: Comparison of the prevalence (solid curve) and the piecewise-constant FOI (dashed curve) of varicella based on several contact matrices for the Netherlands, Poland, Slovakia, and Spain.

11.5 Discussion

In this chapter we assessed the impact of different contact patterns on the transmission of varicella in twelve European countries. For this purpose we augmented the serological data with contact data arising from different sources, e.g., collected data from cross-sectional surveys in the general population (POLYMOD) and counted data in synthetic populations constructed using socio-demographic data (FBK), but we also used WAIFW matrices, reinvented by defining their structure based on the shape of the contact matrices FBK and POLYMOD. The main goal of this work were to show that we can realistically estimate the uncertainty around the basic reproduction number, R_0, related to the model assumption about the contact matrices. Moreover, we provided a framework for the estimation of R_0 for those countries without survey-collected contact data, using either FBK matrices or empirically based WAIFW matrices, or both.

Harmonized contact matrices, like FBK, resulted very important for modeling the serology of varicella. Not only they are in good agreement with POLYMOD matrices (they differ basically of a scale factor), but they present far more advantages than the latter ones. These matrices do not contain collected data from a cross-sectional survey, but rather data which are function of descriptive statistics provided by Eurostat. By definition, these statistics, which have been used to simulate a virtual population, are not affected by sampling error (Fumanelli et al., 2012). It follows that FBK matrices have an intrinsically lower degree of uncertainty than POLYMOD matrices, hence the associated parameters are characterized by less uncertainty, at least, not arising from sampling error. Indeed, when doing inference on the transmission parameters with bootstrap for POLYMOD matrices, we found that contact data are the dominant source of the uncertainty in the estimation (Marangi, 2011).

On the other hand, we recognize that the main advantage of POLYMOD data is the richness of information they encompass. While FBK matrices inform us only about the location of contacts, POLYMOD matrices incorporate much more information, such as type, duration and frequency of contacts.

Furthermore, we want to remark the good performance of the data-based WAIFW matrices: as far as we know, this is the first time that a WAIFW structure is proposed not based on some theoretical assumptions, but rather on data, in this case, POLYMOD and FBK data. These matrices, compared to the POLYMOD and FBK contact matrices, yield higher estimates of R_0 (but not as large as the classical WAIFW matrix proposed by Anderson and May (1991)), because of their higher degree of assortativeness. Nonetheless, these data-based WAIFW matrices are very easily obtained and are therefore very useful for those countries that do not have yet either the socio-demografic census data necessary to construct harmonized matrices or the possibility of organizing a survey similar to POLYMOD. Hence, the variety of data-based WAIFW matrices that can be built may be useful to get an idea of the uncertainty around the R_0. Nonetheless, the use of these matrices presents some difficulties regarding the estimation of their parameters. We have seen that it is possible to estimate the β parameters only for those age groups for which we have data at hand. Moreover, the bootstrap procedure failed to provide us with unimodal distribution for the parameters of interest, impeding the construction of CIs. Further research will be

necessary to find the reason of this phenomenon and solve it.

In conclusion, we found that the uncertainty around R_0 can be quite large and very dependent on the chosen contact matrix. It is thus important to be aware and take this uncertainty into account when developing a mathematical model for studying the transmission of varicella, because the assessment of the effect of the vaccination policies and the consequent cost-effective analyses strongly depend are strongly influenced by it.

Chapter 12

Discussion and Further Research

12.1 Statistical Modeling of HCV and HIV Infections Among Injecting Drug Users

In the first part of this thesis, we investigated the association between HCV and HIV infections in injecting drug users, which is one of the main risk group affected by hepatitis C as a consequence of their drug use habits, e.g., sharing syringes, that may lead to an exchange of infected blood between two persons. In Chapters 3 and 4 we focus on the association between the two infections using population prevalence data from Italy. Although these data cannot give us any information about the association between the infections at the individual level, their analysis with GLMMs allowed us to obtain a detailed picture of the evolution of the prevalence of HCV and HIV infections in the Italian regions between 1998 and 2007. We have shown that, at population level, the infections are positively correlated within the regions, therefore regions with a high prevalence of HCV infection tend to have also a high prevalence of HIV infection, and vice versa. This finding has an impact in terms of health policy decisions, since it strengthens the idea that interventions aimed to control the transmission of the two infections should not consider the two infections separately, but rather jointly.

Furthermore, we showed in Chapter 4 that the random intercept model is too simple to capture the whole variability in the data and, as a consequence, how to accommodate overdispersion, in addition to clustering, in multivariate binomial repeated data. Using a hierarchical Bayesian approach, it is possible to overcome the computational issues of the ML approach (Molenberghs et al., 2010), assuming complex structures for the different sets of random effects. The most complex model that we fitted considered time-dependent covariance matrices for the overdispersion random effects θ_{ijk}:

$$\begin{pmatrix} \theta_{ij1} \\ \theta_{ij2} \end{pmatrix} \sim MVN\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{D}_{\theta j}\right]. \tag{12.1.1}$$

This model assumes a unstructured relation of the covariance matrix $D_{\theta j}$with time. However, one could be interested in modeling this relation with time using a parametric model. A possibility is an exponential model that assumes that the additional correlation between HCV and HIV infections within the years, not explained by the correlation between the clustering random intercepts, either increases or decreases with time. The model is given by

$$\rho(\theta_{ij1}, \theta_{ij2}) = f(t) = \exp(\alpha t), \tag{12.1.2}$$

where α might be either positive or negative. Moreover, we could model the correlation between the overdispersion parameters for HCV and HIV according to information, collected at regional level, about the risk factors related to injecting drug use (percentage of sharing syringes or other paraphernalia, etc.) or about the results of interventions (percentage of supplies of clean drug injection equipment, opioid substitution and other forms of drug dependence treatments, antiviral treatments for HIV, health promotion, etc.).

A similar analysis has been presented in Chapter 5, where we modeled the co-infection with HCV and HIV in IDUs using either marginal and random-effects joint models for individual current status data. Using the marginal models, we could estimate the association measures (odds ratios, correlation) between the two infections in the individuals and investigate which risk factors have an impact on this association. Instead, using the random-effects models, we estimated the degree of individual heterogeneity in the acquisition of the infection. The results presented in Chapter 5 showed that a significant association between HCV and HIV infections within IDUs is related to a significant degree of individual heterogeneity in the acquisition of the infections (Farrington et al., 2001). Hence, for instance, the association between HCV and HIV infections in IDUs who reported ever sharing syringes is not significant because of a higher homogeneity in their behaviors and, therefore, in their acquisition of the infections.

Since these results were obtained using only current status data from Spain, it would be interesting to use the same statistical approaches on data from other countries, in order to confirm our findings. Besides, it would be worthy to assess the impact of other risk factors on the association between HCV and HIV infections, such as sharing other paraphernalia or unhygienic injecting, which were reported to have a strong impact on the prevalence of these drug-related infections (Mathëi et al., 2006). Another interesting point is to assess the impact of the intervention measures on the association between HCV and HIV infection. Indeed, we may expect that some interventions decrease the magnitude of the association between the two infections, making their transmission independent.

A fundamental issue for all these statistical problems regards the assessment of goodness-of-fit and the model selection. In our applications we used exclusively information criteria for model selection, in order to determine which is the model, given the data at hand, that better represents the compromise between goodness-of-fit (measured by the deviance) and parsimony in parameters (measured by the penalty). This decision has been motivated by the practical difficulty to assess the adequacy of the models to our data, given the high degree of complexity of the models used. For the models based on the ML estimation, presented in Chapter 3, we estimated the deviance (see Table 3.1) and we used the LRT for testing hypotheses about the covariance structures. However, this did not tell us a lot about the

goodness-of-fit, as displayed in Figure 4.1 in Chapter 4, where we showed the unsatisfactory fit of the basic correlated random-effects model (for what regards the fit of the individual profiles, but not for the fit of the marginal mean). This problematic is even more severe in the Bayesian framework, where most of the criteria presented in the literature are mainly focussed on the model selection rather than on the model fit, i.e., Bayes factors (Kass and Raftery, 1995), conditional predictive ordinate (CPO, Geisser and Eddy, 1979), DIC (Spiegelhalter et al., 2002), PED (Plummer, 2008), and the difference in posterior deviances (Aitkin, 2010). When dealing with random-effects models, it becomes difficult to assess the actual dimension of the model, therefore many of these methods fail because are not able to compute the correct penalty for the deviance (Celeux et al., 2006). The result is the odd behavior seen with the models that we fitted, where it seems that these information criteria find differences where there is none. Further research based on simulation studies will be therefore necessary in order to effectively evaluate the performance of these information criteria when dealing with random-effects models.

12.2 Bayesian Mixture Models for Antibody Titers

In the second part of this thesis, we used hierarchical Bayesian mixture models for the estimation of the prevalence and the force of infection from the continuous antibody titers. In Chapters 7 – 8 we applied these statistical models to antibodies to parvovirus B19 and VZV, two infections which are currently in a pre-vaccination status in Europe (except for a few areas, where it is routinely administered a vaccine against VZV). This means that we can assume the antibody titers to arise from a mixture of two components, one for susceptible and one for immune, whose weights (the probability to belong to such component) vary with the age.

Parvovirus B19 and VZV are both airborne infections that mainly affect children. Hence, sharing the same transmission route, the next generation of models should be multivariate mixture models. Figure 12.1 shows a joint scatter plot of the antibodies to parvovirus B19 and VZV in Belgium, with over imposed the respective cut-off points given by the assays' manufacturers. We notice that four groups can be identified: a small group of individuals who are susceptible to both infections, two groups of individuals who have experienced only one of the two infections (but very few individuals have only experienced an infection with parvovirus B19 and not with VZV), and a large group of individuals immune to both infections. The goal is therefore to estimate the joint prevalences and force of infection of parvovirus B19 and VZV infections using a bivariate mixture model.

The case of post-vaccination serology was discussed in Chapter 9, where we modeled the antibodies to measles in Tuscany using a multiple components hierarchical Bayesian mixture model, in order to estimate the prevalence, while accounting for different levels of immunity to the infection, following either natural infection or vaccination. We found that immune individuals can be categorized either as low or high responders to infection/vaccination. The existence of group of low responders is particularly important, because it could imply that some individuals might be at risk of becoming susceptible again, given the possibility that measles vaccination protection wane with time.

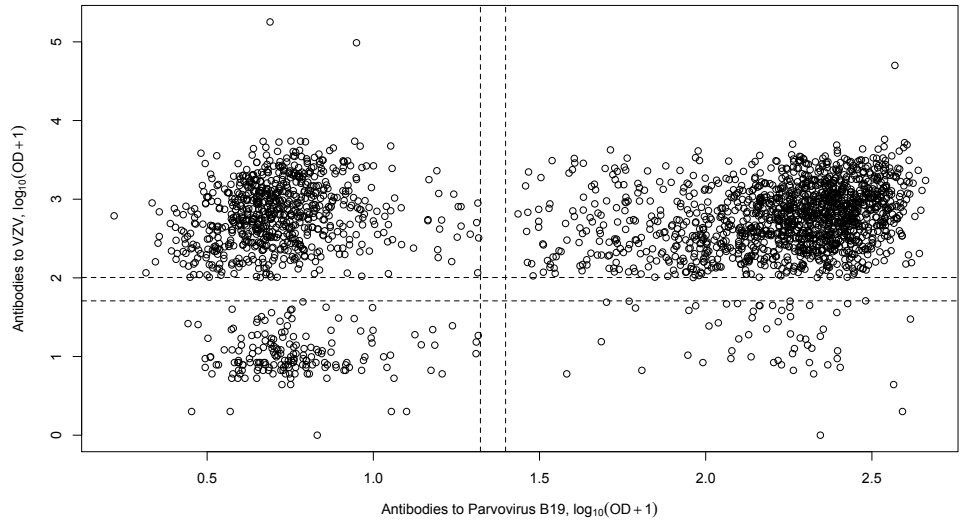


Figure 12.1: Joint scatter plot of antibodies IgG to parvovirus B19 (on the x axis) and to VZV (on the y axis).

We have shown that mixture models are descriptive methods which highlight the heterogeneity in the individual antibody response. Hence, it could be worthwhile to develop a mathematical model that explains the emerging of the different groups of individuals according to their response and the within-component heterogeneity. An hypothesis is that this between and within-component variability depends more on the individual heterogeneity rather than on factors like the age. In practice, this hypothesis can be investigated using an agent-based model, that is to say, a system of agent-specific differential equations, where the individual antibody response arise from a mixture model that describes the different subpopulations.

12.3 Estimation of Transmission Parameters for Varicella in Europe

Finally, in the third part of this thesis, we presented the work done in order to model the serology of varicella in Europe. This was accomplished in two different ways.

In Chapter 10, we did not assume any mathematical model for the transmission of varicella, but we rather used flexible statistical models for current status data in order to estimate the prevalence and the FOI of varicella in twelve European countries. The main features of varicella serology, captured by the statistical models, consist of a pronounced peak of the FOI in children aged from 5 to 10 years, as it is typical of close contact childhood infections. In this sense, children can be seen as the main vector of

the spread of this infections, both among themselves and towards adults. In effect, close to the first peak of the FOI, it can be found a second peak of the FOI, accounting for the infections occurring in parents and, at a lesser extent, in grandparents. Between the groups of children and of parents, we find another group, mainly composed by adolescents and young adults without family, who experience low or zero FOI, because of their lack of contacts with children.

The central analysis of the third part of the thesis was presented in Chapter 11, where we focussed on the transmission of the varicella taking into account the effect of contact patterns among individuals on the spread of the infection. In particular, we considered different contact matrices able to accommodate the heterogenous age-specific mixing patterns of individuals. Even though these matrices are constructed in different ways, they all share the same features: a strong assortativeness by age among children, as a result of school contacts; a more homogeneous mixing among adults, who meet in workplaces; a high frequency of contacts among parents and grandparents with children and grandchildren in the household. Using these matrices, we estimated the associated transmission parameters and we derived the basic reproduction number and the critical threshold for vaccination. In particular, we showed the high variability existing around the R_0 associated to the different contact matrices. Among the matrices, we remark the importance of the harmonized contact matrices based on synthetic populations and of the empirically based WAIFW matrices. Differently from the POLYMOD-type contact matrices, which require big surveys in order to collect the data, these two type of matrices can be constructed in an easier way. FBK-type contact matrices only need socio-demographic census data, while the WAIFW matrices we proposed only need to be defined looking at the social contact matrices already available.

My future interest in the field of contact matrices will be in the investigation of the impact of demography on the structure of contact matrices. The contact matrices presented in this part of the thesis are cross-sectional representations of the social contacts in the population, therefore we cannot assume that the observed mixing patterns remain constant in time. If this is true for Europe, whose population can be defined as stable (low fertility, replacement migration, postponing of life events, generalized aging), it is at a larger extent for the developing low-resource countries, which are characterized by an ongoing fertility transition and a faster than ever urbanization (which can have a dramatic impact on the family size). It becomes therefore important to construct, for these low-resource countries, social contact matrices that take into account the different locations where people live, being these either rural or urban areas, and the different household sizes and compositions.

Bibliography

Aceijas, C. and Rhodes, T. (2007). Global estimates of prevalence of hcv infection among injecting drug users. *International Journal of Drug Policy*, 18:352–358.

Agresti, A. (2002). *Categorical Data Analysis*. Wiley & Sons, New York, NY, 2nd edition edition.

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

Aitkin, M., Liu, C., and Chadwick, T. (2009). Bayesian model comparison and model averaging for small-area estimation. *Annals of Applied Statistics*, 3:199–221.

Alter, M. (2006). Epidemiology of viral hepatitis and HIV co-infection. *Journal of Hepatology*, 44:S6–S9.

Anderson, R. and May, R. (1991). *Infectious Diseases of humans, Dynamic and Control*. Oxford University Press, Oxford.

Anderson, R. M. and May, R. M. (1985). Vaccination and herd immunity to infectious diseases. *Nature*, 318:323–329.

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

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, 12:171–178.

Azzalini, A. (1986). Further results on a class of distributions which includes the normal ones. *Statistica*, 46:199–208.

Azzalini, A. and Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew-t distribution. *Journal of the Royal Statistical Society, Series B*, 65:367–389.

Baker, B. (2002). Natural history of hepatitis C. Technical report, National Institutes of Health (NIH), Bethesda, MD.

Barrio, G., de La Fuente, L., Toro, C., Brugal, M., Soriano, V., Gonzalez, F., Bravo, M., Vallejo, F., Silva, T., and Group, P. I. (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.

Baughman, A., Bisgard, K., Lynn, F., and Meade, B. (2006). Mixture model analysis for establishing a diagnostic cut-off point for pertussis antibody levels. *Statistics in Medicine*, 25:2994–3010.

Bayarri, M. J. and Castellanos, M. E. (2007). Bayesian Checking of the Second Levels of Hierarchical Models. *Statistical Science*, 22:322–343.

Bechini, A., Boccalini, S., Tiscione, E., Pesavento, G., Mannelli, F., Peruzzi, M., Rapi, S., Mercurio, S., and Bonanni, P. (2010). Progress towards measles and rubella elimination in tuscany, italy: the role of population seroepidemiological profile. *European Journal of Public Health*.

Bechini, A., Pesavento, G., Boccalini, S., Tiscione, E., Balocchini, E., Graziani, G., Pecori, L., Santini, M., Azzari, C., Peruzzi, M., Mannelli, F., Tomasi, A., Montomoli, E., Mazzoli, F., and Bonanni, P. (2006). Implementazione del piano per l'eliminazione del morbillo e della rosolia congenita in Toscana: progressi verso la seconda fase di controllo dell'infezione. *Bollettino Epidemiologico Nazionale (BEN)*, 19. (In Italian).

Becker, N. (1989). *Analysis of Infectious Disease Data*. Chapman & Hall, London.

Bollaerts, K., Aerts, M., Shkedy, Z., Faes, C., Van der Stede, Y., Beutels, P., and Hens, N. (2012). Estimating the Population Prevalence and Force of Infection Directly from Antibody Titers. *Statistical Modelling*, 12:441–462.

Bollepalli, S., Mathieson, K., Bay, C., Hillier, A., Post, J., van Thiel, D., and Nadir, A. (2007). Prevalence of risk factors for hepatitis C virus in HIV-infected and HIV/hepatitis C virus-coinfected patients. *Sexually Transmitted Diseases*, 34:367–370.

Bonanni, P., Bechini, A., Pesavento, G., Boccalini, S., Tiscione, E., Graziani, G., Santini, M., Balocchini, E., Pecori, L., and Peruzzi, M. (2005). Implementation of the Plan for Elimination of Measles and Congenital Rubella Infection in Tuscany: evidence of progress towards phase II of measles control. *Journal of Preventive Medicine and Hygiene*, 64:111–117.

Bonanni, P., Breuer, J., Gershon, A., Gershon, M., Hryniewicz, W., Papaevangelou, V., Rentier, B., Rümke, H., Sadzot-Delvaux, C., and Senterre, J. (2009). Varicella vaccination in Europe – taking the practical approach. *BMC Medicine*, 7:26.

Boncompagni, G., Incandela, L., Bechini, A., Giannini, D., C., C., Trezzi, M., Ciofi Degli Atti, M., Ansaldi, F., Valle, L., and Bonanni, P. (2006). Measles outbreak in grosseto, central italy. *Eurosurveillance*, 11:pii=3015.

Box, G. E. P. and Tidwell, P. W. (1962). Transformation of the independent variables. *Technometrics*, 4:531–550.

Breslow, N. (1984). Extra-Poisson variation in log-linear models. *Journal of the Royal Statistical Society, Series C*, 33:38–44.

Breslow, N. and Clayton, D. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88:9–25.

Brisson, M., Edmunds, W. J., and Gay, N. (2003). Varicella vaccination: Impact of vaccine efficacy on the epidemiology of VZV. *Journal of Medical Virology*, 70:S31–S37.

Brisson, M., Melkonyan, G., Drolet, M., Serres, G. D., Thibeault, R., and Wals, P. D. (2010). Modeling the impact of one- and two-dose varicella vaccination on the epidemiology of varicella and zoster. *Vaccine*, 28:3385–3397.

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

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

CDC (2005). Parvovirus B19 Infection and Pregnancy. `http://miscarriage.about.com/gi/o.htm?zi=1/XJ&zTi=1&sdn=miscarriage&cdn=health&tm=8&f=00&tt=3&bt=0&bts=0&zu=http%3A//www.cdc.gov/ncidod/dvrd/revb/respiratory/B19%26preg.htm`. Last Accessed 4 sep 2012.

CDC (2011). Increased transmission and outbreaks of measles–European Region, 2011. *MMWR. Morbidity and mortality weekly report*, 60:1605–1610.

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

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

Clayton, D. (1978). A model for association in bivariate life tables and its application in epidemiological studies. *Biometrika*, 65:141–151.

Clayton, D. (1996). Generalized linear mixed models. In W.R., G., S., R., and D.J, S., editors, *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.

Congdon, P. (2003). *Applied Bayesian Modelling*. Wiley & Sons, Chichester.

Corcoran, A. and Doyle, S. (2004). Advances in the biology, diagnosis and host-pathogen interactions of Parvovirus B19. *Journal of Medical Microbiology*, 53:459–475.

Coutinho, F., Massad, E., Lopez, L., , and Burattini, M. (1999). Modelling heterogeneities in individual frailties in epidemic models. *Mathematical and Computer Modelling*, 30:97–115.

Crofts, N., Dore, G., and Locarnini, S., editors (2001). *Hepatitis C: An Australian Perspective*. IP Communications, East Hawthorn, Vic.

Dale, J. (1986). Global cross-ratio models for bivariate, discrete, ordered responses. *Biometrics*, 42:721–727.

Davison, A. and Hinkley, D. (1997). *Bootstrap Methods and Their Application*. Cambridge University Press.

Del Fava, E., Kasim, A., Usman, M., Shkedy, Z., Hens, N., Aerts, M., Bollaerts, K., Scalia Tomba, G., Vickerman, P., Sutton, A., Wiessing, L., and Kretzschmar, M. (2011a). Joint modeling of HCV and HIV infections among injecting drug users in Italy using repeated cross-sectional prevalence data. *Statistical Communications in Infectious Diseases*, 3:1.

Del Fava, E., Shkedy, Z., Hens, N., Aerts, M., Suligoi, B., Camoni, L., Vallejo, F., Wiessing, L., and Kretzschmar, M. (2011b). 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*, 3:3.

Diebolt, J. and Robert, C. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, Series B*, 56:363–375.

Dumchev, K., Soldyshev, R., Qian, H.-Z., Zezyulin, A., Chandler, S., Slobodyanyuk, P., Moroz, L., and Schumacher, J. (2009). HIV and hepatitis C virus infections among hanka injection drug users in central Ukraine: a cross-sectional survey. *Harm Reduction Journal*, 6:23.

Eilers, P. H. and Marx, B. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11:89–102.

EMCDDA (2010). Annual report 2010: The state of the drugs problem in europe. Technical report, European Monitoring Centre for Drugs and Drug Addiction, Luxembourg: Publications Office of the European Union.

Eurostat (2011). http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/search_database.

Evans, R. and Erlandson, K. (2004). Robust Bayesian prediction of subject disease status and population prevalence using several similar diagnostic tests. *Statistics in Medicine*, 23:2227–2236.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications*. Chapman & Hall, London.

Farrington, C. (1990). Modeling forces of infection for measles, mumps and rubella. *Statistics in Medicine*, 9:953–967.

Farrington, C., Kanaan, M., and Gay, N. (2001). Estimation of the basic reproduction number for infectious diseases from age-stratified serological survey data (with discussion). *Journal of the Royal Statistical Society, Series C*, 50:251–292.

Farrington, C. P., Unkel, S., and Anaya-Izquierdo, K. (2012). The relative frailty variance and shared frailty models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 74:673–696.

Frühwirth-Schnatter, S. (2007). *Finite Mixture and Markov Switching Models*. Springer, Berlin.

Frühwirth-Schnatter, S. and Pyne, S. (2010). Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions. *Biostatistics*, 11:317–336.

Fumanelli, L., Ajelli, M., Manfredi, P., Vespignani, A., and Merler, S. (2012). Inferring the Structure of Social Contacts from Demographic Data in the Analysis of Infectious Diseases Spread. *PLoS computational biology*, 8:e1002673.

Gay, N. (1996). Analysis of serological surveys using mixture models: application to a survey of Parvovirus B19. *Statistics in Medicine*, 15:1567–1573.

Gay, N., Vyse, A., Enquselassie, F., Nigatu, W., and Nokes, D. (2003). Improving sensitivity of oral fluid testing in IgG prevalence studies: application of mixture models to a rubella antibody survey. *Epidemiology and Infection*, 130:285–291.

Geisser, S. and Eddy, W. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, 74:153–160.

Gelfand, A. and Kuo, L. (1991). Nonparametric Bayesian bioassay including ordered polytomous response. *Biometrika*, 78:657–666.

Gelfand, A., Sahu, S., and Carlin, B. (1996). *Efficient parametrizations for generalised linear mixed models (with discussion)*, pages 165–180. Oxford University Press.

Gelfand, A., Smith, A., and Lee, T.-M. (1992). Bayesian analysis of constrained parameter and truncated data problems. *Journal of the American Statistical Association*, 87:523–532.

Gelfand, A. E. (2003). Some comments on model criticism. In Green, P. J., Hjort, N. L., and Richardson, S., editors, *Highly Structured Stochastic Systems*. Oxford University Press, Oxford.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2004). *Bayesian Data Analysis*. Chapman & Hall, London, 2nd edition edition.

Gelman, A., Meng, X. L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6:733–759.

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

George, E. and McCulloch, R. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88:881–889.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In Bernado, J., Berger, J., Dawid, A., and Smith, A., editors, *Bayesian Statistics 4*. Clarendon Press, Oxford.

Gilks, W., Richardson, S., and Spiegelhalter, D. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.

Goeyvaerts, N., Hens, N., Aerts, M., and Beutels, P. (2011). Model structure analysis to estimate basic immunological processes and maternal risk for parvovirus B19. *Biostatistics*, 12:283–302.

Goeyvaerts, N., Hens, N., Ogunjimi, B., Aerts, M., Shkedy, Z., van Damme, P., and Beutels, P. (2010). Estimating infectious disease parameters from data on social contacts and serological status. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 59:255–277.

Greiner, M., Bhat, T., Patzelt, R., Kakaire, D., Schares, G., Dietz, E., Böhning, D., Zessin, K., and Mehlitz, D. (1997). Impact of biological factors on the interpretation of bovine trypanosomosis serology. *Preventive Veterinary Medicine*, 30:61–73.

Greiner, M., Franke, C., Böhning, D., and Schlattmann, P. (1994). Construction of an intrinsic cut-off value for the sero-epidemiological study of *Trypanosoma evansi* infections in a canine population in Brazil: a new approach towards an unbiased estimation of prevalence. *Acta Tropica*, 56:97–109.

Grün, B. and Leisch, F. (2008). Finite mixtures of generalized linear regression models. In Shalabh and Heumann, C., editors, *Recent Advances in Linear Models and Related Areas*, pages 205–230. Springer.

Guzzetta, G., Poletti, P., Del Fava, E., Ajelli, M., Scalia Tomba, G., Merler, S., and Manfredi, P. (2012). Hope-Simpson's progressive immunity hypothesis may explain Herpes Zoster incidence data. *American Journal of Epidemiology*, 0:00–00.

Hagan, H. and Des Jarlais, D. (2000). HIV and HCV infection among injecting drug users. *Mount Sinai Journal of Medicine*, 67:423–428.

Hardelid, P., Williams, D., Dezateux, C., Tookey, P., Peckham, C., Cubitt, W., and Cortina-Borja, M. (2008). Analysis of rubella antibody distribution from newborn dried blood spots using finite mixture models. *Epidemiology and Infection*, 136:1698–706.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, 1(3):297–318.

Heidelberger, P. and Welch, P. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, 31(1109–1144).

Hens, N., Aerts, M., Shkedy, Z., Theeten, H., van Damme, P., and Beutels, P. (2008). Modelling multisera data: the estimation of new joint and conditional epidemiological parameters. *Statistics in Medicine*, 27:2651–2664.

Hens, N., Ayele, G. M., Goeyvaerts, N., Aerts, M., Mossong, J., Edmunds, W. J., and Beutels, P. (2009a). Estimating the impact of school closure on social mixing behaviour and the transmission of close contact infections in eight European countries. *BMC Infectious Diseases*, 9:187.

Hens, N., Goeyvaerts, N., Aerts, M., Shkedy, Z., van Damme, P., and Beutels, P. (2009b). Mining social mixing patterns for infectious disease models based on a two-day population survey in Belgium. *BMC Infectious Diseases*, 9:5.

Hens, N., Shkedy, Z., Aerts, M., Faes, C., Van Damme, P., and Beutels, P. (2012). *Modeling Infectious Disease Parameters Based on Serological and Social Contact Data*. Springer.

Hens, N., Wienke, A., Aerts, M., and Molenberghs, G. (2009c). 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.

Hethcote, H. (1996). Modelling heterogeneous mixing in infectious diseases dynamics. In Isham, V. and Medley, G., editors, *Models for Infectious Human Diseases: Their Structure and Relation to Data*. Cambridge University Press.

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

Hope, V., Judd, A., Hickman, M., Sutton, A., Stimson, G., Parry, J., and Gill, O. (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.

Horby, P., Thai, P. Q., Hens, N., Yen, N. T. T., Mai, L. Q., Thoang, D. D., Linh, N. M., Huong, N. T., Alexander, N., Edmunds, W. J., Duong, T. N., Fox, A., and Hien, N. T. (2011). Social Contact Patterns in Vietnam and Implications for the Control of Infectious Diseases. *PLoS ONE*, 6:e16965.

ICONA (2003). ICONA 2003: Indagine nazionale sulla copertura vaccinale infantile. Technical Report 07/37, Istituto Superiore di Sanità, Roma.

ICONA (2008). ICONA 2008: Indagine nazionale sulla copertura vaccinale infantile. Technical Report 09/29, Istituto Superiore di Sanità, Roma.

Iozzi, F., Trusiano, F., Chinazzi, M., Billari, F., Zagheni, E., Merler, S., Ajelli, M., Del Fava, E., and Manfredi, P. (2010). Little Italy: an agent-based approach to the estimation of contact patterns- fitting predicted matrices to serological data. *PLoS computational biology*, 6:e1001021.

Jewell, N. and van der Laan, M. (2004). Current status data: Review, recent developments and open problems. In N., B. and C.R., R., editors, *Advances in Survival Analysis*, pages 625–643. Elsevier, North Holland.

Jewell, N. and Van Der Laan, M. J. (2002). Case-control current status data. Technical Report 116, University of California Berkeley Division of Biostatistics, Berkeley.

Kasim, A., Shkedy, Z., and Kato, B. (2012). Estimation and inference under simple order restrictions: hierarchical Bayesian approach. In Dan, L., editor, *Modeling Dose-response Microarray Data in Early Drug Development Experiments Using R*. Springer-Verlag, Berlin.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90:773–795.

Keiding, N. (1991). Age-specific incidence and prevalence: a statistical perspective. *Journal of the Royal Statistical Society, series A*, 154:371–412.

Keiding, N., Begtrup, K., Scheike, T., and Hasibeder, G. (1996). Estimation from current status data in continuous time. *Lifetime Data Analysis*, 2:119–129.

Kremer, J. R., Schneider, F., and Muller, C. P. (2006). Waning antibodies in measles and rubella vaccinees–a longitudinal study. *Vaccine*, 24:2594–2601.

Kretzschmar, M. and Wiessing, L. (2008). New challenges for mathematical and statistical modeling of HIV and hepatitis C virus in injecting drug users. *AIDS*, 22:1527–1537.

Krugman, S., Giles, J., Friedman, H., and Stone, S. (1965). Studies on immunity to measles. *Journal of Pediatrics*, 66:471–488.

Lalkhen, A. G. and McCluskey, A. (2008). Clinical tests: sensitivity and specificity. *Continuing Education in Anaesthesia, Critical Care & Pain*, 8:221–223.

Lehmann, H., von Landenberg, P., and Modrow, S. (2003). Parvovirus B19 infection and autoimmune disease. *Autoimmunity Reviews*, 2:218–223.

Lucidarme, D., Bruandet, A., Ilef, D., Harbonnier, J., Jacob, C., Decoster, A., Delamare, C., Cyran, C., Van Hoenacker, A., Frémaux, D., Josse, P., Emmanuelli, J., Le Strat, Y., Desenclos, J., and Filoche, B. (2004). Incidence and risk factors of HCV and HIV infections in a cohort of intravenous drug users in the north and east of France. *Epidemiology and Infection*, 132:699–708.

Marangi, L. (2011). *Contatti sociali, stima ed inferenza su parametri di infezioni a trasmissione diretta. Il caso della varicella.* PhD thesis, University of Florence, Italy.

Mathëi, C., Buntinx, F., and Van Damme, P. (2002). Seroprevalence of hepatitis C markers among intravenous drug users in western European countries: a systematic review. *Journal of Viral Hepatitis*, 9:157–173.

Mathëi, C., Shkedy, Z., Denis, B., Kabali, C., Aerts, M., Molenberghs, G., van Damme, P., and Buntinx, D. (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.

Mathers, B., Degenhardt, L., Phillips, B., Wiessing, L., Hickman, M., Strathdee, S., Wodak, A., Panda, S., Tyndall, M., and Toufik, A. (2008). Global epidemiology of injecting drug use and HIV among people who inject drugs: a systematic review. *Lancet*, 372:1733–1745.

McCulloch, C. and Searle, S. (2001). *Generalized, Linear and Mixed Models*. Wiley & Sons, New York, NY.

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

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

Melegaro, A., Jit, M., Gay, N., Zagheni, E., and Edmunds, W. (2011). What types of contacts are important for the spread of infections?: using contact survey data to explore European mixing patterns. *Epidemics*, 3:143–151.

Merler, S., Ajelli, M., Pugliese, A., and Ferguson, N. M. (2011). Determinants of the Spatiotemporal Dynamics of the 2009 H1N1 Pandemic in Europe: Implications for Real-Time Modelling. *PLoS computational biology*, 7:e1002205.

Miller, C., Wood, E., Spittal, P., Li, K., Frankish, J., Braitstein, P., Montaner, J., and Schechter, M. (2004). The future face of coinfection: prevalence and incidence of HIV and hepatitis C virus coinfection among young injection drug users. *Journal of Acquired Immune Deficiency Syndromes*, 36:743–749.

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

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

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

Morimune, K. (1979). Comparison of normal and logistic models in the bivariate dichotomous analysis. *Econometrica*, 47:957–975.

Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., Massari, M., Salmaso, S., G., S. T., Wallinga, J., Heijne, J., Sadkowska-Todys, M., Rosinska, M., and Edmunds, W. (2008). Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Medicine*, 5:e74.

Mossong, J., Nokes, D., Edmunds, W., Cox, M., Ratnam, S., and Muller, C. (1999). Modeling the impact of subclinical measles transmission in vaccinated populations with waning immunity. *American Journal of Epidemiology*, 150:1238–1249.

Nardone, A., de Ory, F., Carton, M., Cohen, D., van Damme, P., Davidkin, I., Rota, M., de Melker, H., Mossong, J., Slacikova, M., Tischer, A., Andrews, N., Berbers, G., Gabutti, G., Gay, N., Jones, L., Jokinen, S., Kafatos, G., de Aragon, M., Schneider, F., Smetana, Z., Vargova, B., Vranckx, R., and Miller, E. (2007). The comparative sero-epidemiology of varicella zoster virus in 11 countries in the European region. *Vaccine*, 25:7866–7872.

Nardone, A. and Miller, E. (2004). Serological surveillance of rubella in Europe: European sero-epidemiology network (ESEN2). *Eurosurveillance*, 9:5–7.

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.

Nielsen, S., Toft, N., Jørgensen, E., and Bibby, B. (2007). Bayesian mixture models for within-herd prevalence estimates of bovine paratuberculosis based on a continuous ELISA response. *Preventive veterinary medicine*, 81:290–305.

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

Ntzoufras, I. (2002). Gibbs variable selection using BUGS. *Journal of Statistical Software*, 7(7).

Ødegård , J., Madsen, P., Gianola, D., Klemetsdal, G., Jensen, J., Heringstad, B., and Korsgaard, I. (2005). A Bayesian threshold-tormal mixture model for analysis of a continuous mastitis-related trait. *Journal of Dairy Science*, 88:2652–2659.

Ogunjimi, B., Hens, N., Goeyvaerts, N., Aerts, M., Van Damme, P., and Beutels, P. (2009). Using empirical social contact data to model person to person infectious disease transmission: an illustration for varicella. *Mathematical Biosciences*, 218:80–87.

O'Hara, R. and Sillanpää, M. (2009). A Review of Bayesian Variable Selection Methods: What, How and Which. *Bayesian Analysis*, 4:85–118.

Pallás, J., Farinas-Álvarez, C., Prieto, D., and Delgado-Rodríguez, M. (1999). Coinfections by HIV, hepatitis B and hepatitis C in imprisoned injecting drug users. *European Journal of Epidemiology*, 15:699–704.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics*, 57:120–125.

Perz, J., Armstrong, G., Farrington, L., Hutin, Y., and Bell, B. (2006). The contributions of hepatitis B virus and hepatitis C virus infections to cirrhosis and primary liver cancer worldwide. *Journal of Hepatology*, 45:529–538.

Piantadosi, S., Byar, D. P., and Green, S. B. (1988). The ecological fallacy. *American Journal of Epidemiology*, 127:893–904.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, Vienna.

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

Plummer, M. (2011). *JAGS Version 3.1. 0 user manual*.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). Coda: Convergence diagnosis and output analysis for mcmc. *R News*, 6:7–11.

Poletti, P., Melegaro, A., Ajelli, M., Del Fava, E., Guzzetta, G., Faustini, L., Scalia Tomba, G., Lopalco, P., Rizzo, C., Merler, S., and Manfredi, P. (2012). Perspectives on the impact of varicella immunization on herpes zoster. A model-based evaluation from three European countries. *BMC Medicine*, submitted for publication.

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

Rahimi-Movaghar, A., Razaghi, E., Sahimi-Izadian, E., and Amin-Esmaeili, M. (2010). HIV, hepatitis C virus, and hepatitis B virus co-infections among injecting drug users in Tehran, Iran. *International Journal of Infectious Diseases*, 14:e28–e33.

Rhodes, T., Platt, L., Judd, S., Mikhailova, L., Sarang, A., Wallis, N., Alpatova, T., Hickman, M., and Parry, J. (2005). Hepatitis C virus infection, HIV co-infection, and associated risk among injecting drug users in Togliatti, Russia. *International Journal of STD & AIDS*, 16:749–754.

Roberts, G. and Sahu, S. (1997). Updating schemes, correlation structure, blocking and parameterisation for the gibbs sampler. *Journal of the Royal Statistical Society, Series B*, 59:291–317.

Roca, B., Suarez, I., Gonzalez, J., Garrido, M., de la Fuente, B., Teira, R., Geijo, P., Cosin, J., Perez-Cortes, S., Galindo, M., Lozano, F., Domingo, P., Viciana, P., Ribera, E., Vergara, A., and Sánchez, T. (2003). Hepatitis C virus and human immunodeficiency virus coinfection in Spain. *Journal of Infection*, 47:117–124.

Rockstroch, J. and Spengler, U. (2004). HIV and hepatitis C virus co-infection. *The Lancet Infectious Diseases*, 4:434–444.

Rota, M., Massari, M., Gabutti, G., Guido, M., De Donno, A., and Ciofi degli Atti, M. (2008). Measles serological survey in the Italian population: Interpretation of results using mixture model. *Vaccine*, 26:4403–4409.

Royston, P. and Altman, D. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *Journal of the Royal Statistical Society, Series C*, 43:429–467.

Servey, J., Reamy, B., and Hodge, J. (2007). Clinical presentations of Parvovirus B19 infection. *American Family Physician*, 75:373–376.

Shepard, C., Finelli, L., and Alter, M. (2005). Global epidemiology of hepatitis C virus infection. *The Lancet Infectious Diseases*, 5:558–567.

Sherman, K., Rouster, S., Chung, R., and Rajicic, N. (2002). Hepatitis C virus prevalence among patients infected with human immunodeficiency virus: a cross-sectional analysis of the US Adult AIDS Clinical Trials Group. *Clinical Infectious Diseases*, 34:831–837.

Shkedy, Z., Aerts, M., Molenberghs, G., Beutels, P., and van Damme, P. (2003). Modeling forces of infection by using monotone local polynomials. *Journal of the Royal Statistical Society, Series C*, 52:469–486.

Shkedy, Z., Aerts, M., Molenberghs, G., Beutels, P., and van Damme, P. (2006). Modelling age-dependent force of infection from prevalence data using fractional polynomials. *Statistics in Medicine*, 25:1577–1591.

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

Spiegelhalter, D., Best, N., Carlin, B., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, 64:583–640.

Steinbakk, G. H. and Storvik, G. O. (2009). Posterior Predictive p-values in Bayesian Hierarchical Models. *Scandinavian Journal of Statistics*, 36:320–336.

Strickland, G. T., El-Kamary, S. S., Klenerman, P., and Nicosia, A. (2008). Hepatitis C vaccine: supply and demand. *The Lancet Infectious Diseases*, 8:379–386.

Su, Y.-S. and Yajima, M. (2012). *R2jags: A Package for Running jags from R*.

Sutton, A., Gay, N., Edmunds, W., Hope, V., Gill, O., 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, A., Hope, V., Mathëi, C., Mravcik, V., Sebakova, H., Vallejo, F., Suligoi, B., Brugal, M., 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.

Tolfvenstam, T., Papadogiannakis, N., Norbeck, O., Petersson, K., and Broliden, K. (2001). Frequency of human parvovirus B19 in intrauterine fetal death. *Lancet*, 357:1494–1497.

UNAIDS (2008). Report on the global HIV/AIDS epidemic 2008: executive summary. Technical report, Joint United Nations Programme on HIV/AIDS, Geneva.

UNAIDS (2009). AIDS epidemic update: November 2009. Technical report, Joint United Nations Programme on HIV/AIDS, Geneva.

van de Laar, T., Paxton, W., Zorgdrager, F., Cornelissen, M., and de Vries, H. (2011). Sexual transmission of hepatitis C virus in human immunodeficiency virus-negative men who have sex with men: a series of case reports. *Sexually Transmitted Diseases*, 38:102–104.

Van Effelterre, T., Shkedy, Z., Aerts, M., Molenberghs, G., Van Damme, P., and Beutels, P. (2009). Contact patterns and their implied basic reproductive numbers: an illustration for varicella-zoster virus. *Epidemiology and infection*, 137:48–57.

van Hoek, A. J., Melegaro, A., Zagheni, E., Edmunds, W. J., and Gay, N. (2011). Modelling the impact of a combined varicella and zoster vaccination programme on the epidemiology of varicella zoster virus in England. *Vaccine*, 29:2411–2420.

Vandelli, C., Renzo, F., Romanò, L., Tisminetzky, S., De Palma, M., Stroffolini, T., Ventura, E., and Zanetti, A. (2004). Lack of evidence of sexual transmission of hepatitis C among monogamous couples: results of a 10-year prospective follow-up study. *The American Journal of Gastroenterology*, 99:855–859.

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

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.

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 Diseases*, 85:ii23–ii30.

Vyse, A., Gay, N., Hesketh, L., Morgan-Capner, P., and Miller, E. (2004). Seroprevalence of antibody to varicella zoster virus in England and Wales in children and young adults. *Epidemiology and Infection*, 132:1129–1134.

Vyse, A., Gay, N., Hesketh, L., Pebody, R., Morgan-Capner, P., and Miller, E. (2006). Interpreting serological surveys using mixture models: the seroepidemiology of measles, mumps and rubella in England and Wales at the beginning of the 21st century. *Epidemiology and Infection*, 134:1303–1312.

Wallinga, J., Teunis, P., and Kretzschmar, M. (2006). Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents. *American journal of epidemiology*, 164:936–944.

WHO (1998). The who position paper on varicella vaccines. *The Weekly Epidemiological Record*, 73:241–248.

WHO (1999). Hepatitis C - global prevalence (update). *The Weekly Epidemiological Record*, 74:425–427.

WHO (2009). Measles vaccines: Who position paper. *The Weekly Epidemiological Record*, 84:349–360.

WHO (2011). Fourth meeting of the Global Polio Eradication Initiative's Independent Monitoring Board. *Relevé épidémiologique hebdomadaire / Section d'hygiène du Secrétariat de la Société des Nations = Weekly epidemiological record / Health Section of the Secretariat of the League of Nations*, 86:557–558.

Williams, J. R., Manfredi, P., Butler, A. R., Ciofi degli Atti, M., and Salmaso, S. (2003). Heterogeneity in regional notification patterns and its impact on aggregate national case notification data: the example of measles in Italy. *BMC Public Health*, 3:23.

Yee, T. (2010). The VGAM package for categorical data analysis. *Journal of Statistical Software*, 32:1–34.

Zagheni, E., Billari, F. C., Manfredi, P., Melegaro, A., Mossong, J., and Edmunds, W. J. (2008). Using Time-Use Data to Parameterize Models for the Spread of Close-Contact Infectious Diseases. *American journal of epidemiology*, 168:1082–1090.

Zhao, Y., Staudenmayer, J., Coull, B. A., and Wand, M. P. (2006). General design Bayesian generalized linear mixed models. *Statistical Science*, 21:35–51.

Zocratto, K., Caiaffa, W., Proietti, F., Carneiro-Proietti, A., Mingoti, S., and Ribeiro, G. (2006). HCV and HIV infection and co-infection: injecting drug use and sexual behavior, AjUDE-Brasil I Project. *Cadernos de Saúde Pública*, 22:839–848.

Hoofdstuk 13

Samenvatting

Dit proefschrift bestaat uit drie delen. Het eerste deel is gewijd aan statistische modellen van HCV en HIV-infecties onder injecterende drugsgebruikers, het tweede deel aan hiërarchische Bayesiaanse mengselmodellen toegepast op antilichamen in serologische stalen, en het derde deel aan de epidemiologie van varicella in Europa, met de nadruk op de schatting van de overdrachtsintensiteiten.

In het eerste deel van het proefschrift wordt de statistische modellering van het hepatitis C-virus (HCV) en het humaan immunodeficiëntie virus (HIV) infecties onder injecterende drugsgebruikers (ID's) in Europa behandeld. De keuze van deze twee infecties wordt gemotiveerd door de hoge frequentie van ID's die besmet zijn met beide virussen. De beschikbare gegevens bestaan uit bevolkingsgegevens en individuele gegevens. Het eerste type van gegevens kan worden gebruikt om de associatie tussen de infecties in de bevolking te beoordelen, hoewel dit geen associatie binnen individuen impliceert. Deze laatste associatie kan in plaats daarvan worden beoordeeld met behulp van individuele gegevens, die gebruikt kunnen worden om de co-infectie met beide virussen te onderzoeken. Na een introductie van het onderwerp (Hoofdstuk 2), willen we in Hoofdstuk 3 en Hoofdstuk 4 de correlatie tussen HCV en HIV-infecties op populatieniveau schatten en daarom modelleren we de bevolkingsgegevens van Italië met gezamenlijke modellen voor herhaalde binomiale gegevens. Deze modellen kunnen rekening houden met zowel clustering (Hoofdstuk 3), omdat er associatie is tussen de observaties van dezelfde regio, als overdispersie (Hoofdstuk 4), met name de overmatige variabiliteit in de gegevens, die niet door de binomiale verdeling wordt verklaard. Dit doel wordt bereikt door middel van verschillende verzamelingen van random effecten, één voor clustering en één voor overdispersie. Voor elke verzameling worden verschillende covariantiematrices getest. Voor de schatting gebruiken we zowel maximum-likelihood methoden als een Bayesiaanse benadering door middel van Monte Carlo Markov Chain (MCMC) methoden. Tenslotte modelleren we in Hoofdstuk 5 de individuele gegevens om de co-infectie met HCV en HIV-virus te bestuderen. Met behulp van marginale modellen (ALR, BDM, BPM), kunnen we onderzoeken welke risicofactoren de associatie (odds ratio, correlatie) tussen de twee infecties beïnvloeden.

Daarnaast, gebruiken we random effecten modellen (GLMM, shared gamma frailty modellen) om het niveau van individuele heterogeniteit in het verwerven van de infecties te schatten.

Het tweede deel van het proefschrift is gewijd aan de hiërarchische Bayesiaanse mengselmodellen voor antilichamen om direct de prevalentie en de infectiedruk te schatten, zonder gebruik te maken van een conventionele grenswaarde. Zo gebruiken wij de gehele informatie bevat in de antilichaamgegevens, zonder dat we ze tot binaire gegevens reduceren, zoals het geval is bij de standaardbenadering gebaseerd op de grenswaarde. Bovendien laten mengselmodellen toe om de prevalentie correct te schatten als het leeftijd-specifiek mengselgewicht van de immune component, zonder het probleem van onderschatting dat typisch is voor de grenswaardebenadering. De laatste vermelde benadering is ontworpen om te werken voor een goede diagnostiek en niet voor serologisch onderzoek. Na een introductie van het onderwerp (Hoofdstuk 6), presenteren we in Hoofdstuk 7 en in Hoofdstuk 8 een toepassing op parvovirus B19 in België en in Italië. Voor deze infectie is er geen vaccinatie. Daarom kunnen we individuen onderverdelen in vatbaren, zonder bewijzen van eerdere infectie, en in immunen, met bewijzen van eerdere infectie. In Hoofdstuk 7 presenteren we een twee-componenten hiërarchisch Bayesiaans mengselmodel, dat we gebruiken om de leeftijdsafhankelijke prevalentie en de infectiedruk te schatten. We testen een aantal aannames voor de verdeling van de gegevens (symmetrische en scheve verdelingen, zoals de normale en de scheef-normale verdeling) en voor het model van de leeftijdsspecifieke prevalentie en de infectiedruk (log-logistisch model, stuksgewijs constant model voor de infectiedruk en niet-parametrisch model). Het model wordt geschat door middel van MCMC methoden. In Hoofdstuk 8 breiden we het model in het vorige hoofdstuk uit en versoepelen we de aanname van constante gemiddelde van de immune component. Hiervoor, definiëren we een stuksgewijs constant model voor het gemiddelde gehalte aan antilichamen van de immunen in de gekozen leeftijdsgroepen en dan kunnen we gebruik maken van een Bayesiaanse variabele selectie (BVS) benadering om het beste model te kiezen. De epidemiologische idee achter deze aanname is dat veranderingen in de antilichaamniveaus (als gevolg van afnemende immuniteit of versterkte immuniteit) kunnen leiden tot veranderingen in de prevalentie en in de infectiedruk. Vervolgens geeft Hoofdstuk 9 een toepassing op mazelen in Toscane (Italië). Mazelen is een infectie waarvoor gevaccineerd wordt. Dit betekent dat immuniteit kan verkregen worden door vaccinatie of door een natuurlijke infectie. Immuniteit na vaccinatie kan echter onvolledig zijn, meestal na één enkele dosis (in plaats van twee), maar ook vanwege afnemende immuniteit. Daarom maken we gebruik van een Bayesiaans mengselmodel met meerdere componenten, dat een niet-parametrisch model voor de prevalentie gebruikt. Met behulp van dit model kunnen we de serologie van mazelen in Toscane in 2003 en in 2005-2006 vergelijken, namelijk voor en na een belangrijke vaccinatiecampagne. We vonden een grote pool van vatbaren en zwakke immunen na vaccinatie in de leeftijdsklasse tussen 10 en 20 jaar oud. Specifieke vaccinatiecampagnes moeten dus op deze groepen gericht worden.

Tenslotte wordt het derde deel van het proefschrift gewijd aan de modellering van varicella in Europa en aan de beoordeling van de onzekerheid over de overdrachtsintensiteiten van de infectie, zoals het basisreproductiegetal. In Hoofdstuk 10 geven we een overzicht van de beschikbare serologische gegevens voor varicella, afkomstig uit twaalf Europese landen en we modelleren ze met flexibele leeftijdsspecifieke

modellen voor de prevalentie en de infectiedruk. In het laatste hoofdstuk hebben we ons gericht op de schatting van het basisreproductiegetal voor varicella, het gebruik van verschillende contactstructuren om rekening te houden met heterogene vermenging tussen individuen: sociale contactgegevens verzameld met een multinationaal cross-sectioneel onderzoek (POLYMOD), sociale contactgegevens geteld in een synthetische populatie op basis van socio-demografische volkstellingsgegevens (FBK) en empirische gebaseerde WAIFW matrices, die hier geïntroduceerd worden. We vonden dat er een grote variabiliteit van het basisreproductiegetal is, die te wijten is aan de mate van overdracht (hoog of laag) en de gebruikte contactmatrices. Dit werk is belangrijk omdat het, rekening houdend met de onzekerheid in de overdrachtsintensiteiten, de elementen biedt om een solide wiskundig model te bouwen om de effecten van de vaccinatie voor varicella op de overdracht van varicella en herpes zoster te bestuderen.