2014 | Faculty of Sciences

DOCTORAL DISSERTATION

Mathematical and statistical models applied to HIV and Hepatitis C co-infection and to nosocomial infections

Doctoral dissertation submitted to obtain the degree of
doctor of Science: Statistics, to be defended by

Amparo Yovanna Castro Sanchez

Promoter: Prof. Dr Marc Aerts | UHasselt / tUL
Co-promoter: Prof. Dr Niel Hens | UHasselt / tUL

D/2014/2451/68

Interuniversity Institute for Biostatistics
and statistical Bioinformatics

Acknowledgements

Many people helped me to accomplish this project, some of them from the academical point of view and others from an emotional point of view. For both I would like to express my sincere gratitude, to my supervisor and co-supervisor Marc Aerts and Niel Hens, respectively for their patience, encouragement, support, valuable help and comments. To Ziv Shkedy, thank you for include me in the HCV - HIV team, for his wonderful ideas, his help with the programs, and his infinite patience to hear all my work worries.

Thanks to my coauthors for their comments, suggestions and helpful discussions. My appreciation also goes to the jury members for their suggestions, questions and comments. Thank you very much to Tom Cattaert who wrote initial versions of the Matlab programs for the frailty models and dedicated time to explain me how do they works and how I could modify them; and to Steven Abrams for the insightful discussions regarding the solution of the mathematical models.

I want to thank to all the people who encourage me to finish this project during the past years, thanks God the list is really long the problem is my very fragile memory; so I will just mention some of them: Annemie, Sofie & family, Opa (another angel) and Oma (all mijn Belgisch familie), and Javier.

To a very international group of friends my sincere appreciation for taking care of the fun part: Emanuele, Greet, Hong, Tanya & family, Ambily, Ariel, Pia, Bea, Caro, Consuelo, Donato, Eva & Jimmy, Martin, Filippo, Fortunato, Kim, Shreosi, and others. You constantly remind me what life-work balance means. I learned a lot about music, food, cooking, wonderful places, and in general international culture, thanks to all of you guys.

Thank you to the staff of Censtat for the wonderful experience, to Martine Machiels, Marc Toelen and Karin Daniëls for their support to the international students. To Tanya Spasibo for all preparations for the PhD defense day, including the last minute translation. Thanks to those who join that team effort. Bedankt to Greet and molte grazie to Emanuele for their helpful comments. Thank you to my Roche colleagues and managers for their support in the past two years.

My gratitude also goes to the individuals who took part of the studies: Amsterdam Cohort Studies, Itinere Project, Vedette Project, and Probiotics Project. Without their willingness to participate on the studies this work would not have been possible. I gratefully acknowledge the contribution of the people and institutions involved in the participant recruitment, follow up, and data management. My deepest appreciation to Marieke, Fernando, Fabrizio, Lucas and Mirjam for kindly provided me with the data of the studies.

I would like to thank Hasselt University for the financial support during my PhD with the Bijzonder Onderzoekfonds (BOF) grant and to the Vlaamse Interuniversitaire Raad (VLIR-UOS) for the financial support during of my master.

También quisiera agradecer a mi amado esposo por el apoyo emocional incluso en los momentos más angustiosos del doctorado, sin tí no hubiera sido posible culminar este proyecto. Amor, gracias por darle prioridad a mis objetivos profesionales, no muchas parejas están dispuestas a dejarlo todo para seguir el sueño de uno solo. Gracias por tu comprensión y tu dedicación para cuidar de nuestra pequeña princesa cuando he estado tan ocupada. Gracias por la linda familia que tenemos y por tu paciencia infinita. Sulay y tú fueron la motivación para terminar este trabajo. Sulay desde que llegaste a mi vida has sido mi mejor maestra, gracias preciosa.

A mi familia cercana: Amanda, Mile y Juan yo también los amo mucho, gracias por apoyarme en la distancia y por entender mis ausencias para poder terminar la tesis. A mi familia lejana, a mis amigos y colegas en Colombia gracias por el ánimo y el interés en mi trabajo en especial a Liliana, Edith, Angélica, Adriana y Alejo. A quienes desde el cielo me apoyan en especial a mi excelente madre Amanda quien solía decir que la educación era la mejor herencia para un hijo, gracias por tu buen ejemplo, por tus sabios consejos, por tu oido atento y por todos los sacrificios que hiciste para educarnos. Mami te he extrañado mucho desde tu reciente partida...como me hubiera gustado que estuvieras conmigo en este día. Finalmente gracias a Dios por la inmensa ayuda.

Y como diría Cerati a todos: Gracias Totales!

December, 2014

Contents

List of Abbreviations

ACS	Amsterdam Cohort Studies
AF	Acceleration Factor
AFT	Accelerated Failure Time Models
AIC	Akaike's Information Criterion
ARE	Ampicillin-resistant *Enterococcus faecium*
CI	Confidence Interval
ESE	Empirical Standar Error
FOI	Force of Infection
HBV	Hepatitis B Virus
HCV	Hepatitis C Virus
HIV	Human Immunodeficiency Virus
IDU	Injecting Drug User
ML	Maximum Likelihood
MSE	Mean Squared Error
NPMLE	Nonparametric Maximum Likelihood Estimator
ODE	Ordinary Differential Equation
SD	Standard Deviation
SE	Standard Error

Chapter 1

Introduction

From an infectious disease perspective, statistical models are a flexible approach to describe association between the variables, expressing the relation with a functional form. As a result, these models allow the identification of risk factors although there is not explicit causality.

In contrast, mathematical models provide a representation of how disease burden is established, they can be used to predict prevalence and incidence of disease even beyond range of data (Garnett et al.; 2011).

In this thesis we propose statistical and mathematical models to HIV and hepatitis C co-infection and to hospital-acquired (nosocomial) infections. Chapter 2 provides concepts and methods to analyse survival data; additionally we describe the mathematical models for infectious disease transmission. Both parts constitute the theoretical bases for the models included in this thesis.

In Chapter 3 we apply survival analysis models to quantify the effect of probiotics and antibiotics on nosocomial infection. The dataset comes from a clinical trial performed in a university hospital in the Netherlands. Here we focus on interval-censored data with time-dependent covariates.

The statistical models presented in Chapters 4 and 5 rely also on survival analysis methodology and are applied to cohort study on injecting drug users collected in the Netherlands.

In Chapter 6 we describe basic transmission models for Hepatitis C virus (HCV) and HIV separately, and we introduce a joint mathematical model accounting for both HIV and HCV co-infection attributed to sharing syringes and other paraphernalia. In Chapter 7 we describe a procedure to assess the joint mathematical model from a statistical perspective. The model is calibrated using a longitudinal study of heroin users in Italy and a cross-sectional study of young adult heroin users and Injecting Drug Users in Spain.

In this chapter we describe the importance of study HIV and hepatitis C co-infection as well as some preventive measures for nosocomial infections. Subsequently, we describe the data sources used in this thesis.

1.1 HIV and hepatitis C co-infection

The Acquired Immune Deficiency Syndrome (AIDS) was defined in 1982 after the occurrence of deaths related with a reduced number of helper T cells. Two years later, the causing virus of AIDS was isolated and named Human Immunodeficiency Virus (HIV) which made the development of a blood sample based diagnostic test possible. Afterwards, clinical studies helped to describe the history of HIV infection, providing information about incubation time. Then, the antiretroviral therapy was introduced as mechanism to combat the progress of the virus. Despite the huge advances in the last decades, AIDS is still an enormous issue in public health. According to UNAIDS in 2010, 34 million people lived with HIV, 2.5 million people were newly infected, and more than two million died due to AIDS-related diseases (UNAIDS; 2010).

On the other hand, hepatitis C is a viral infection of the liver whose virus HCV was identified in 1989. The virus, which is spread by direct contact with infected blood, is one of the major causes of hepatitis and chronic liver diseases such as cirrhosis and liver cancer. The screening procedures for blood products have reduced the number of infections. However, according to estimates of the World Health Organization, 150 million people are infected around the world, and between 3 and 4 million are newly infected each year (World Health Organization (WHO); 2011).

There have been defined two disease stages: acute hepatitis C and chronic hepatitis C. The first one lasts around six months, is mostly asymptomatic and leads to chronic infection in around 80% of the cases. Chronic hepatitis C can last up to 20 years, spontaneous clearance of the virus is rare, and increases the risk of cirrhosis, hepatic decompensation and liver cancer. There is no vaccine against hepatitis C partly because the virus mutates very easily. As re-infection can occur, the role of a potential vaccine may be to prevent the progression of acute hepatitis C to chronic infection (Wasmuth; 2010).

Co-infection between HCV and HIV often occur. A third of HIV infected individuals in Europe and USA are co-infected with HCV and it is known that HIV accelerates the development of liver disease related with HCV and reduces the chance of spontaneous clearance and possibly increases infectivity (Rockstroh and Spengler; 2004). The majority of the co-infected people are Injecting Drug Users (IDUs), so focusing on this high-risk population could give us insights about the transmission of HIV/HCV co-infection, how to formulate interventions and how the treatment of acute or chronic infections can affect the prevalence.

Hepatitis C virus is usually acquired rapidly after having started with injecting drugs. In 2011, between 60-80% of the IDUs from 25 countries had a positive anti-HCV test result. Whereas in 12 other countries a higher seroprevalence has been reported (Nelson et al.; 2011). IDUs are also at higher risk to acquire HIV, in some countries the HIV prevalence can reach up to 20% (Aceijas et al.; 2004). The main transmission route for both viruses in this population is sharing injecting equipment (Mathei et al.; 2006).

Based on mathematical models for HCV and HIV co-infection, HCV prevalence has been proposed as an indicator of HIV among IDUs (Vickerman et al.; 2010). Then De Vos et al. (2012) shows the existence of a threshold HCV equilibrium value below which HIV cannot establish itself. (See Chapter 2).

Below we describe three datasets where the study populations are injecting drug users. The projects were carried out in the Netherlands, Italy and Spain. The studies in the Netherlands and Italy were epidemiological studies aiming to assess the impact of interventions.

1.2 Studies about injecting drug users

1.2.1 Amsterdam Cohort Studies

The Amsterdam Cohort Studies (ACS) is a collaboration of Amsterdam Health Service, Academic Medical Center of the University of Amsterdam, Sanquin Blood Supply Foundation, and the University Medical Center Utrecht. ACS is part of the Netherlands HIV Monitoring Foundation and is financially supported by the Netherlands National Institute for Public Health and the Environment.

The prospective cohort study initiated in 1985 to investigate the prevalence, incidence, and risk factors of HIV infections and other blood-borne and/or sexually transmitted diseases, as well as the effects of intervention. Participation in the ACS is voluntary, and informed consent is obtained for every individual at intake. ACS participants visit the Amsterdam Health Service every 4-6 months. They complete a standardized questionnaire about their health, risk behaviour, and sociodemographic situation. Questions at ACS entry refer to the 6 months preceding the visit; questions at follow-up refer to the interim since the preceding visit. Blood is drawn during each visit for laboratory testing and storage. Until 2010, 1,657 injecting drug users were included in the ACS. The recruitment for the drug users was via methadone programs, via a sexually transmitted diseases clinic for drug using sex workers and by word of mouth.

The ACS database up to 2005 contains information on 1,206 IDUs of whom 254 lack information about their HCV serostatus since only those with at least two study visits have been tested for HCV (Van den Berg et al.; 2007a,b). There were 3, 12 and 2 individuals having zero or negative time to infection for HIV only, HCV only and both HIV and HCV, respectively. Zero time to infection implies that the year of first injection coincides with the year of the first positive result, whereas negative time to injection refers to individuals who had positive results before becoming IDUs. Table 1.1 shows the HIV and HCV serostatus for the remaining 935 individuals.

In Chapter 4 we included only the individuals who entered negative for HCV, totalling 165 individuals (58 seroconverters and 107 who remained negative). On the other hand, in Chapter 5 we consider all the 935 individuals, Table 1.2 and Figure 1.1 present the descriptive statistics for this group of individuals.

From table 1.2, 61.3% of the individuals were males; 41.6% stated sharing syringes at least once during the follow up period. Concerning the frequency of injection at first visit, 23.2% did not inject recently, 30.5% reported using drugs more than once a day and 29.4% used drugs between 2-6 days per week. The most com-

Table 1.1: Amsterdam Cohort Studies dataset. Number of patients according to their serostatus for HIV and HCV.

HIV status	HCV status			Total
	Negative at the end of follow up	Positive before entry	Seroconverter during the study	
Negative at the end of follow up	104	456	45	605
Positive before entry	0	240	1	241
Seroconverter during the study	3	74	58	89
Total	107	770	58	935

mon drug was a combination of cocaine and heroin: 42.2%; followed by heroin and cocaine use alone with 13.7% and 9.8%, respectively. The year in which individuals start to inject drugs was highly variable (45.6% between 1962-1980; 40.4% between 1981-1990; and 14% between 1991-2002). The average age of first injection was 22.4 years (SE 6.4 years), whereas the mean age at first visit was 31.6 years (SE 6.5 years).

Figure 1.1 shows the time to infection for both viruses. A large percentage of individual (48.8%) get infected with HCV but remain negative for HIV. For the individuals who become infected with any of the viruses the exact time to infection is unknown, but partial information is available thanks to the regular visits (check ups). Then the infection occurs between the last negative result and the first positive result, this is known as case II interval-censored observation. To simplify the graph we represent those observations using the mid-point imputation.

Table 1.2: Amsterdam Cohort Studies dataset. Descriptive statistics for all the IDUs

Individuals (n=935)	n	(%)	n	(%)
HCV serostatus				
Negative	107	11.44		
Positive	828	88.56		
HIV serostatus				
Negative	605	64.71		
Positive	330	35.29		
Sharing syringes				
No	542	57.97		
Yes	389	41.60		
Unknown	4	0.43		
Year first injection				
1962- 1980	426	45.56		
1981- 1990	378	40.43		
1991- 2002	131	14.01		
Gender				
Male	573	61.28		
Female	362	38.72		
	First follow up visit		Last follow up visit	
Frequency of injection				
No recent injections	217	23.21	503	53.8
More once per day	285	30.48	99	10.59
Once daily	45	4.81	14	1.50
2-6 days per week	275	29.41	129	13.8
Once a week	23	2.46	23	2.46
2-3 days per month	24	2.57	42	4.49
One day a month	10	1.07	14	1.50
Less than one day a month	52	5.56	77	8.24
Unknown frequency of injection[1]	4	0.43	34	3.64
Drug of injection				
No recent injections	217	23.21	503	53.80
Heroin	128	13.69	91	9.73
Cocaine	92	9.84	53	5.67
Cocaine and heroin	395	42.25	215	22.99
Amphetamine	48	5.13	21	2.25
Methadone	25	2.67	17	1.82
Unknown drug of injection[2]	30	3.21	35	3.74
	Mean	Std. Dev.		
Duration of injection at first visit (years)	9.13	6.62		
Duration of injection at last visit (years)	17.29	8.27		

[1] Unknown frequency of injection was not associated with any of the covariates

[2] Unknown drug of injection at first follow up visit was associated sharing syringes and year of first injection

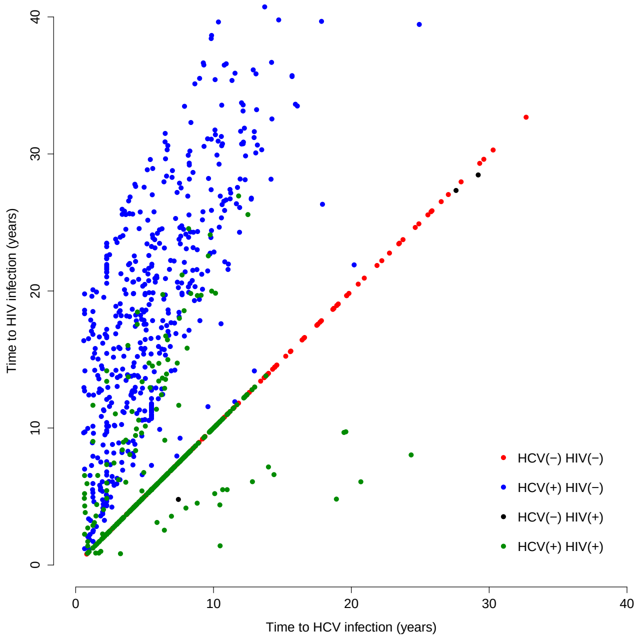


Figure 1.1: Amsterdam Cohort Studies. Scatterplot of the time to HCV infection vs time to HIV infection. The red dots represent the individuals who were not infected with any of the viruses during the follow up time (right censored observations). The green dots correspond to the individuals infected with both viruses. The blue dots represent the individuals infected with HCV but negative for HIV. For the infected individuals we use the mid-point imputation of the interval between the last negative result and the first positive result.

1.2.2 Vedette study

In Chapter 7 we used the Vedette study as one of the illustrative examples. The dataset comes from a longitudinal study of heroin users in the Piedmont region, Italy. All individuals were followed during 18 months from September 1998 to March 2001. Clinical history and personal information were collected at entry (Bargagli et al.; 2006; Davoli et al.; 2007; Salamina et al.; 2010). The main goal of the study was to evaluate the effectiveness of treatments provided by the National Health Services. In total, the study was based on 115 drug treatment centers and included 10,454 heroin users in 13 Italian regions. Antibody levels were determined for HIV, HCV and Hepatitis B.

For the analyses presented here, the individuals with missing serostatus for both infections and the duration of injection are not included. Table 1.3 and Figure 1.2 provide a description of the variables included in the dataset. Figure 1.2 presents the joint and the marginal prevalence according to the length of injecting career, the proportion of individuals who remain negative for both infections (p_{00t}) diminishes sharply in the first five years of injection, whereas the individuals that become positive only for HIV (p_{01t}) remains very low (almost constant). In fact, most of the individuals positive for HIV were also co-infected with HCV (p_{11t}). For the Vedette study the exact time to infection is unknown, as for the ACS dataset. However, here we have a unique evaluation of the serostatus instead of a long follow up. This is known as case I interval-censored data.

Table 1.3: Vedette IDU dataset. Descriptive statistics

Individuals (n=1,846)	n	(%)
HCV serostatus		
Negative	343	18.58
Positive	1,372	74.32
Unknown HCV serostatus[1]	131	7.10
HIV serostatus		
Negative	1,426	77.25
Positive	125	6.77
Unknown HIV serostatus[2]	295	15.98
HBV serostatus		
Negative	802	43.45
Positive	854	46.26
Unknown HBV serostatus[3]	190	10.29
Sharing syringes within the six months before the interview		
No	1,625	88.03
Yes	144	7.80
Unknown sharing syringes status[4]	77	4.17
Sharing other paraphernalia within the six months before the interview		
No	1,452	78.66
Yes	311	16.85
Unknown sharing paraphernalia status[5]	83	4.50
Gender		
Male	1,500	81.26
Female	346	18.74
	Mean	Std. Dev.
Age at first injection (years)	21.28	6.62
Duration of injection (years)	9.36	5.04

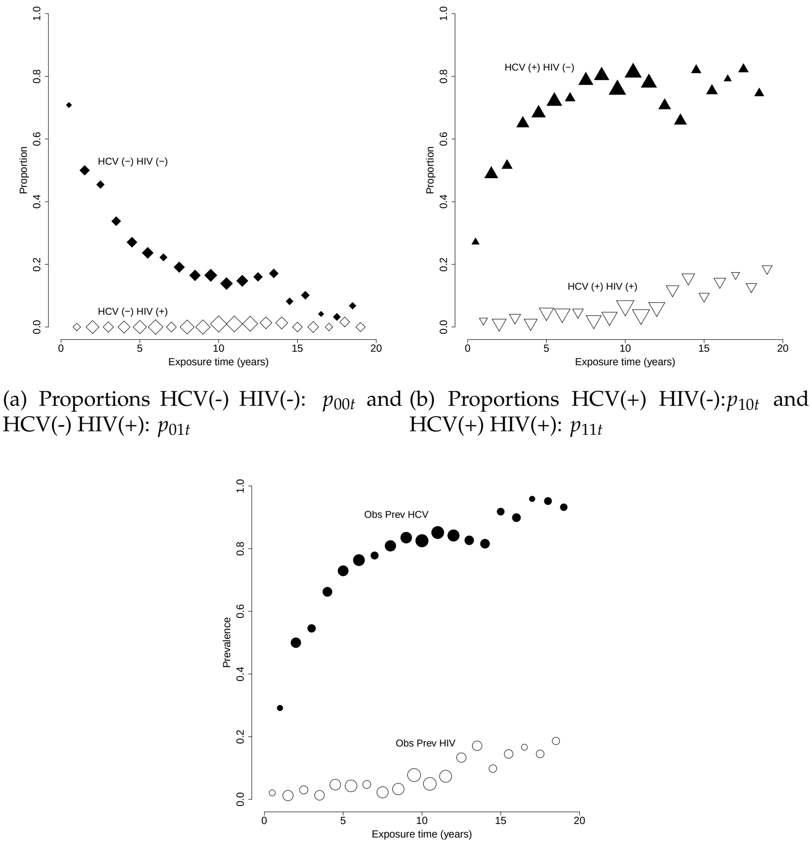
[1] Unknown HCV serostatus was not associated with any of the covariates

[2] Unknown HIV serostatus was associated with HBV serostatus

[3] Unknown HBV serostatus was associated with HCV serostatus

[4] Unknown sharing syringes status was associated with HCV serostatus, and sharing other paraphernalia

[5] Unknown sharing other paraphernalia status was associated with HCV sharing syringes



(a) Proportions HCV(-) HIV(-): p_{00t} and HCV(-) HIV(+): p_{01t}

(b) Proportions HCV(+) HIV(-):p_{10t} and HCV(+) HIV(+): p_{11t}

(c) Prevalences for HCV ($p_{r.t}$) and HIV ($p_{.st}$)

Figure 1.2: Vedette IDU dataset. Observed proportions: joint and marginal prevalence according to the length of injecting career. The size of the symbols is proportional to the observed number of individuals at each exposure time.

1.2.3 Itinere study

The dataset comes from a study of young adult heroin users and IDUs in Spain. All individuals were tested for both HCV and HIV between 2001 and 2003. Additionally, information about the length of the injecting career, the frequency of injecting and sharing syringes was also collected (De La Fuente et al.; 2006). The main goals were to monitor the health impact of drug use and to identify related factors. The study was based on street recruitment, referred by other participants or by non-participants (either drug users or ex-drug users).

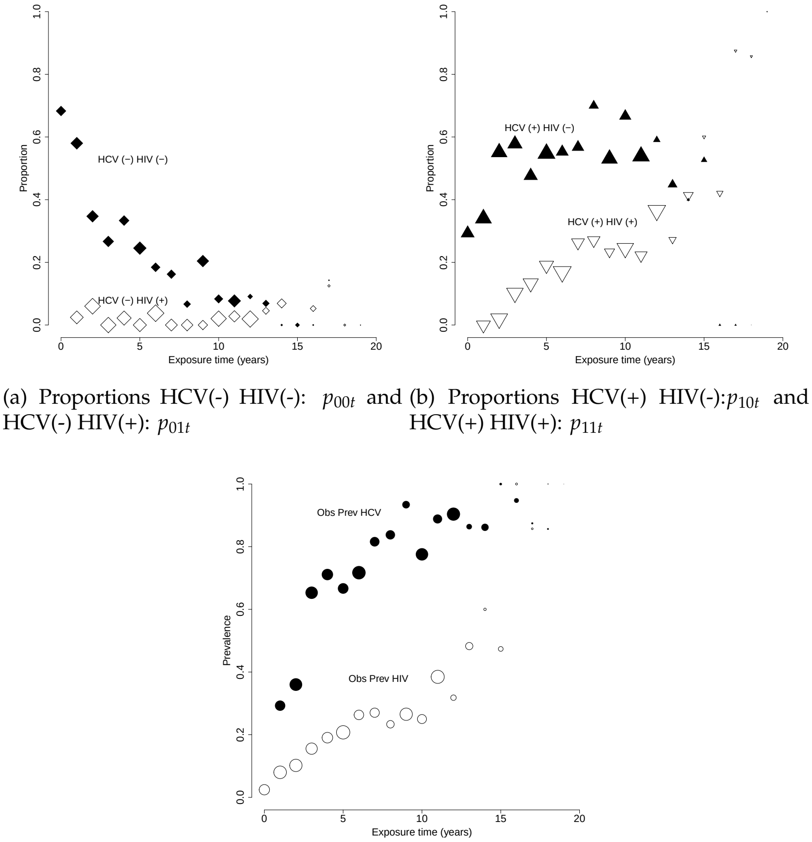
Table 1.4: Itinere IDU dataset. Descriptive statistics

Individuals (n=619)	n	(%)
HCV serostatus		
Negative	165	26.66
Positive	454	74.34
HIV serostatus		
Negative	462	74.64
Positive	157	25.36
HBV serostatus		
Negative	482	77.87
Positive	137	22.13
Sharing syringes within the 12 months before the interview		
No	476	76.89
Yes	114	18.42
Unknown sharing syringes status[1]	29	4.68
Gender		
Male	459	74.15
Female	160	25.85
	Mean	Std. Dev.
Age at first injection (years)	19.37	3.80
Duration of injection (years)	6.72	4.52

[1] Unknown sharing syringes status in the past 12 monts was not associated with any of the covariates

For the analyses presented here, the individuals with missing serostatus for both infections and the duration of injection are not included.

Table 1.4 provides a description of the variables included in the dataset. The HIV prevalences for the Itinere dataset are larger than those for the Vedette dataset (see Figures 1.2 and 1.3). In fact, the proportion of individuals positive for both viruses is higher. This difference maybe attributed to the characteristics of the study populations, for the Vedette data the individuals attended drug treatment centers whereas the individuals that took part in the Itinere project were mainly street users.



(a) Proportions HCV(-) HIV(-): p_{00t} and HCV(-) HIV(+): p_{01t}

(b) Proportions HCV(+) HIV(-):p_{10t} and HCV(+) HIV(+): p_{11t}

(c) Prevalences for HCV ($p_{r.t}$) and HIV ($p_{.st}$)

Figure 1.3: Itinere IDU dataset. Observed proportions: joint and marginal prevalence according to the length of injecting career. The size of the symbols is proportional to the observed number of individuals at each exposure time.

1.3 Nosocomial infections: preventive measures

Antibiotics are used to treat infections caused by bacteria; however many bacteria have become resistant to antibiotics. Hospital-acquired infections caused by antibiotic-resistant bacteria are associated with longer hospitalization time, and with higher morbidity and morality compared with infections caused by antibiotic-susceptible bacteria (Davey et al.; 2005). Additionally, acquired antibiotic resistance seriously limits the therapeutic options to treat the patients when infections occur, increasing the clinical treatment failure and the mortality (Brown et al.; 2006).

In the United States more than ten percent of the hospital-acquired infections are attributed to *Enterococcus* species. The antibiotic resistance of *Enterococcus* has been widely documented. In 1972 the antibiotic vancomicin was used for the first time and only 15 years later vancomycin-resistant enterococci were observed. Among the *Enterococcus* species, *E. faecium* poses the higher antibiotic resistance threat (Fisher and Phillips; 2009).

Over the past years, despite of a huge effort to improve how antibiotics are prescribed by physicians in hospitals, is estimated that half of the use of antibiotics is inappropriate (Davey et al.; 2005). Therefore additional strategies should be implemented to reduce the impact of antibiotic resistance. One strategy is the use of probiotics to prevent infection. Even though some studies point the beneficial impact of probiotics to maintain and restore the intestinal flora (Hickson et al.; 2007; DSouza et al.; 2002), the evidence of efficacy of probiotics in infection prevention needs further study (Oudhuis et al.; 2011).

UMCU probiotics study

The University Medical Center Utrecht (UMCU) designed a cohort study to quantify the effects of probiotics and antibiotics on acquisition of ampicillin-resistant *Enterococcus faecium* (ARE) in patients admitted to two hospital wards with documented high prevalence of intestinal ARE carriage (de Regt et al.; 2008).

Of the 530 included patients, 436 patients were at risk for acquisition of ARE (236 females 54.1%). Their mean age at admission was 62.6 years (Std. Dev. 18.12 years) and their average length of hospital stay was 12 days (Std. Dev. 8.7 days). Of these 436 patients, 111 (25.5%) received probiotics during at least one day and 207 (47%) were treated with antibiotics.

1.4 Outline of the thesis

This thesis aims to develop statistical and mathematical models for HIV and hepatitis C virus (HCV) co-infection in the context of Injecting Drug Users (IDUs) as well as for nosocomial infections. The models applied in this thesis take into account the objectives of the study and the type of data.

The Amsterdam Cohort Studies and the UMCU probiotics are cohort studies that provide information about time to event, in the first case about time to HIV and HCV infection and in the second case about time to ARE acquisition. Considering the nature of both datasets, survival analysis models are suitable to describe the force of infection and to identify risk factors associated with time to event. In Chapter 2 we first describe the methods to analyse survival data ranging from completely non-parametric to fully parametric methods. Since the time to event is interval-censored for both datasets, we also include a literature review on this topic.

The Vedette and the Itinere projects also provide information about time to infection, but they are not follow up studies. For both studies the serostatus of the participants is given at a specific time point. Here, the main goal is to model the transmission process using a mathematical model. Therefore, Chapter 2 includes concepts about mathematical models reviewing the basic SIR model and a review of transmission models for HCV and HIV in the context of injecting drug users.

Chapter 3 presents survival analysis methods to account for case II interval-censored data including both fixed and time-dependent covariates. We focus on the effects of covariates since the main goal is to measure the impact of probiotics and antibiotics on ARE-colonization.

Chapter 4 includes: i) the estimation of the force of infection for HCV applying the concepts and methods of survival analysis within the interval-censoring framework; ii) the impact of risk factors such as frequency of injection, drug injected, sharing of syringes and time of first injection on the time to HCV infection. We used data from the Amsterdam Cohort Studies collected in The Netherlands, focusing on those individuals who were HCV negative upon entry into the study. Previous estimates of the force of infection for HCV in IDU were based on cross-sectional data (Del Fava et al.; 2011; Mathei et al.; 2006; Namata; 2008; Platt et al.; 2009; Sutton et al.; 2006, 2008). Here we use a large cohort study with more than 25 years of follow up.

The analyses presented in Chapter 4 contributed to the work of the "European Study Group for Mathematical Modelling and Epidemiological Analysis of Drug-Related Infectious Diseases", coordinated by European Monitoring Centre for Drugs and Drug Addiction (EMCDDA) and the Center for Infectious Disease Control (RIVM) with funding from World Health Organization/Europe and the government of The Netherlands.

Chapter 5 describes the frailty models as an alternative to analyse multivariate survival data. Here the excess of risk of an event in a cluster (or an individual) is represented by a random effect: the frailty term. The model is defined as shared frailty if a common random effect is assumed for all the members within a cluster. On the other hand, if the random effect is specific to an individual within the cluster then is a correlated frailty model, since the correlation between the random effects may be assumed.

In infectious disease epidemiology, the shared frailty has been applied before

(Farrington et al.; 2001; Sutton et al.; 2006, 2008). Then, Cattaert (2008) applies several frailty models to seroprevalence data on mumps and rubella and to parvo and varicella data. After, Hens et al. (2009) studied the behaviour of the bivariate-correlated gamma frailty model for case I interval-censored data (current status data) and compared the correlated with the shared frailty model using cross-sectional data on hepatitis A and B.

Chapter 5 builds on the work of Cattaert (2008) and Hens et al. (2009) considering exact time to event, right censored and case II interval-censored data. First, we describe the gamma frailty model and the estimation procedure, then we apply several frailty models to the Amsterdam Cohort Studies data. Finally, we perform two simulation studies, the first one is to assess the performance of the correlated frailty model in presence of interval-censored data and the second one is to evaluate how different frailty variances impact the estimation of the parameters in the model.

On Chapter 6 we move to mathematical models, focussing on the transmission. We first propose basic mathematical models for HCV and HIV, for some particular cases we found the solution of the system of the equation. Then we combine the two basic models into a joint model accounting for HIV/HCV co-infection. The mathematical models are proposed in the contex of injection drug users.

Chapter 7 presents a statistical concepts and methods are used to assess the joint model from a statistical perspective, in order to get further insights in: i) the comparison and selection of optional model components, ii) the unknown values of the numerous model parameters, iii) the parameters to which the model is most 'sensitive' and iv) the combinations or patterns of values in the high-dimensional parameter space which are most supported by the data. Data from heroin users in Italy and Spain are used to illustrate the application of the proposed joint model and its statistical assessment. The model assessment of the joint transmission model includes the estimation of the parameters or the calibration of the model to data, the quantification of model uncertainty and model selection, the assessment of the statistical variability, and analyses of the model parameters in the high-dimensional parameter space. Finally, we close with some conclusions and further research.

Chapter 2

Concepts of statistical and mathematical models

2.1 Statistical models

Models are simplified representations of reality and are used in many areas of science, finance and industry. When a model includes a probabilistic component is called a statistical model (Lindsey; 2007). Statistical modelling has been a very active area of research, taking into account the nature of the outcome variable and explanatory variables. We may consider simple linear regression model when the outcome variable follows a normal distribution and the interest is to assess the impact of a covariate. If the outcome variable does not follow a normal distribution the generalized linear models are an option. Here, the generalization has two aspects: the outcome may follow a distribution of the exponential family and the models include a transformation of the mean. Lindsey (2007) presents an interesting classification of the outcome variable: i) measurements that can take positive or negative values, ii) measurements strictly positive, and iii) number of occurrences of one or more kinds of events. Additionally, the random component corresponds to the distribution of the outcome variable. The time to some event is of special interest among the outcomes in the second type. The Amsterdam Cohort Studies and the Probiotics study provide time to event data. Therefore, in the following section we describe those models in detail.

2.2 Survival analysis: concepts and methods

In survival analysis an individual is followed over time for the occurrence of a specific event (recovery, death, infection). The main outcome is the time to the event and typically we are interested on estimating the time until the event happens (survival function) or the event risk (hazard function) and assessing the impact of covariates.

One of the distinctive features of the time to event data is censoring, which occurs whenever the exact time to event is unknown (but there is some level of information). The most common type of censoring is right-censoring where by the end of the observation period the event of interest has not yet occurred (the time to event is larger than the censoring time). Left-censoring is the censoring type where the event occurred before the first observation time. Sometimes the event is known to have occurred within two observation times but the exact time is unknown. The time to event is then said to be interval-censored with right and left censoring as special cases.

A second characteristic of survival data is truncation, that is only those individuals whose event time lies within a certain observational window are observed. In Chapter 4 we provide further details about left truncation and how it can be taken into account in an analysis.

The event time of an individual can be represented by a nonnegative random variable T with a cumulative distribution function $F(t) = P(T \leq t)$. The complement of $F(t)$ is the survival function $S(t) = 1 - F(t)$, which is the probability that the individual survives beyond time t. The hazard function $\lambda(t)$ also known as the force of infection, or the intensity function, describes the instantaneous probability of the event, conditional on having survived up to time t is defined as:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)} = -\frac{d}{dt} \ln S(t). \tag{2.1}$$

In fact, the definition given by (Klein and Moeschberger; 2003) is:

$$\lambda(t) = \frac{f(t)}{S(t)} = \frac{f(t)}{1 - F(t)} \tag{2.2}$$

The cumulative hazard function is given by:

$$\Lambda(t) = \int_0^t \lambda(u) du = -\ln S(t) \tag{2.3}$$

Throughout the present work we assume independent censoring: the censoring mechanism is independent of the time to event.

2.2.1 Nonparametric estimation of the survival and the hazard functions

The survival function can be estimated without any a priori assumption regarding the distribution of the time to event using the estimator Kaplan-Meier also known as the Product-Limit estimator by Kaplan and Meier (1958) for right-censored data. The estimator is a step function with jumps at the observed event times; the size of the jumps depends on the number of events observed at each event time and the previous censored observations. Another option is to estimate the cumulative hazard function using the Nelson-Aalen estimator, originally proposed by Nelson (1972) and rediscovered by Aalen (1978), which has better small-sample-sized properties than the Product-Limit estimator.

For interval-censored data Turnbull (1976) extended and generalized previous results from Peto and Lee (1973). Here, the survival function estimator is also known as Nonparametric Maximum Likelihood Estimator. It is an iterative procedure, first a grid of time points is defined and an initial estimate of the survival function for the grid should be provided, then the algorithm is as follows: i) computes the probability of an event, ii) estimate the number of events, iii) computed the estimated number of people at risk and iv) computed the updated Product-Limit estimator using the estimated data from ii) and iii). The process is repeated until the old and the updated survival function are close.

The most common algorithm to obtain the NPMLE is the self-consistency algorithm proposed by Turnbull (1976). The interval-censored data is treated as incomplete data and the Expectation-Maximization (EM) algorithm (Dempster et al.; 1977) is applied to take these incomplete data into account. The drawback is that it can be very slow for large sample sizes and the convergence is not guaranteed (Gómez et al.; 2009). More efficient proposals are the Iterative Convex Minorant (ICM) proposed in 1992 and the EM-ICM proposed in 1998; another option is to use Project Gradient Methods (PGM).

Comparing survival curves

If the interest is in testing the hypothesis comparing either survival or hazard functions there are several tests proposed for right-censored data. Among them are: the log-rank test, the Gehan, the Tarone-Ware, the Peto and the Fleming-Harrington test. The last one is in fact a general class of tests that includes the Log-rank test as special case. Klein and Moeschberger (2003, chap. 7) provides a detailed description of the tests to compare survival curves.

For interval-censored data there are two type of tests to compare survival curves: rank-based and survival-based tests. The first class is based on weighted differences between the estimated hazard functions and it is appropriate to detect ordered hazard alternatives but unsuitable for crossing hazards. Whereas the second focuses on the estimated survival functions and is applicable for ordered survival functions but inappropriate for crossing survival functions. For a detailed description of the tests we refer the paper by Lesaffre et al. (2005) and the book of Sun (2006). Among the more recent proposals are: a family of tests that extends the Fleming-Harrington test described in the paper by Gómez et al. (2009) and a weighted logrank test proposed by Fay and Shaw (2010).

The hypothesis testing procedures do not provide any parameter to quantify the effect of each covariate; they only produce information regarding their significance. When the main interest is to quantify the impact of several factors, or to predict the time to event, then regression modelling techniques are more appropriate. Those include the well-known Cox proportional hazard models (semi-parametric) and the Accelerated Failure Time (AFT) models (parametric) that we describe below.

2.2.2 Semi-parametric regression - Cox model

The Cox model is the most common method to analyse right-censored data. Here covariates can be fixed or time-dependent. The hazard function is modelled as a product of two factors: the baseline hazard that is left unspecified (nonparametric part) and a factor that characterizes how the baseline hazard function changes as a function of subject covariates (parametric part). The estimation of the model is done using partial likelihood and was proposed by Cox (1972).

$$\lambda(t) = \lambda_0(t)\exp(\boldsymbol{\gamma}'\boldsymbol{X}), \tag{2.4}$$

where $\boldsymbol{\gamma}' = (\gamma_1, \gamma_2, \ldots \gamma_p)$ is a vector of regression coefficients, $\boldsymbol{X} = (x_1, x_2, \ldots, x_p)'$ is a p-variate vector of covariates. The Cox model is also known as the proportional hazard models, in fact the ratio between the hazard of two different individuals is constant over time.

Note that there are some parametric options of proportional hazard model such as the Weibull distribution, where the baseline hazard is fully specified. Additionally, there are also some semiparametric AFT models where the error distribution is unknown.

Semi-parametric regression of interval-censored data

There are still unresolved issues to extend the Cox model to interval-censored data, and so far there is no unified approach. Next we describe some approaches have been followed grouping them according to the methods used.

Some proposals are based on the EM algorithm and an approximation to the likelihood function (see for instance Finkelstein (1986) and Goetghebeur and Ryan (2000)), the last one reduces to the Cox proportional hazards model in absence of interval-censored data.

Some other authors relied on multiple imputation of the unobserved survival times. Here we have the rank-based method proposed by Satten (1996), which combines Monte Carlo simulations with the EM algorithm and fits Cox models for complete-data setting. Two years later, Satten et al. (1998) used a parametric model of the baseline hazard function to impute the exact time in case of interval and right censored observations. Whereas, Goggins et al. (1998) modified the initial propose of Satten (1996) using Montecarlo simulations only in the E-step. Then Pan (2000) proposed a multiple imputation approach based on Breslow's estimate of the survivorship function. The drawback of these methods is that they are computationally demanding.

Other proposals include smoothing of the baseline hazard: Kooperberg and Clarkson (1997) used smoothed parametric splines; Betensky et al. (1999) proposed the use of local likelihood smoothing focusing on the estimation of the baseline hazard without considering covariates. In Betensky et al. (2002) extended the local likelihood estimation procedure to include covariates making minimal assumptions about the hazard. However, the method requires manual entry of the bandwidth which determines the amount of smoothing in the hazard function. Additionally, the analytical standard errors were not derived; hence, bootstrap standard errors may be calculated; however, in this setting are quite computationally demanding. Cai and Betensky (2003) presented a linear spline model assuming a log linear spline mixed model for the baseline hazard, and the Cox proportional hazards for the covariate effect. Finally, the proposal of Sun (2006) is to restrict the baseline hazard using non decreasing and continuous piecewise linear functions. An interesting feature of these models is that predictive smooth survival and the hazard functions are available as a results of the model fit.

It is worth to mention the paper by Zhang and Davidian (2008) proposing a general framework for regression analysis under different censoring settings assuming that the elements of the density of the survival time can be approximated by a "semi-nonparametric" (SNP) density estimator.

2.2.3 Accelerated failure time models

One could use a specific shape for the hazard function that includes covariates following a fully parametric approach. One option is the Accelerated Failure Time (AFT) model, which assumes that the log-transformed time to event is a linear function of the predictor variables (similar to a classical linear regression model).

$$T = \ln(Y) = \mu + \boldsymbol{\gamma' X} + \sigma W, \tag{2.5}$$

where $\boldsymbol{\gamma'} = (\gamma_1, \gamma_2, \dots \gamma_p)$ is a vector of regression coefficients, $\boldsymbol{X}$ is a vector of covariates and W is an error term, assumed to follow a certain distribution. This type of models will be further implemented in Chapters 3 and 4.

Under the accelerated failure time model 2.5, the hazard function for an individual with covariate X is related to a baseline hazard rate λ_{0HCV} as follows:

$$\lambda(t, X) = \exp(-\gamma' X)\lambda_0[t \exp(-\gamma' X)] \tag{2.6}$$

The factor $\exp(-\gamma' X)$ is called the acceleration factor, which reflects the expansion or the contraction of survival time as a function of the covariates. Table 2.1 shows the different distributions and the corresponding hazard functions as considered in Chapter 4.

The AFT model can be directly used in presence of interval-censored data, applying maximum likelihood methods to estimate parameters of the baseline and regression coefficients. The construction of the likelihood considers what information provides each individual. That means, if for an individual the exact time to event is observed, he/she provides information on the probability that the event occurs at that specific time $f(t)$, which is approximately equal to the density function. On the other hand, for an individual with right censored observation we only know the event occurs after the censoring time, then the information he/she provides corresponds to the survival function evaluated at censoring time $S(C_r)$. For a left-censored observation we only know the event has already occurred, so his/her contribution is in terms of the cumulative distribution function $F(C_l) = 1 - S(C_l)$. Finally, if the observation is interval-censored we only know the event time is in the interval $S(L_j) - S(R_j)$.

Table 2.1: Force of infection and survival functions for the different parametric distributions

Distribution	Force of Infection (FOI) $\lambda(t)$	Survival Function $S(t)$
Weibull $\alpha, \beta > 0, t \geq 0$	$\alpha\beta t^{\beta-1}$	$\exp[-\alpha t^{\beta}]$
Gompertz $\alpha, \beta > 0, t \geq 0$	$\alpha \exp(\beta t)$	$\exp\left(-\frac{\alpha}{\beta}[\exp(\beta t) - 1]\right)$
Log-normal $-\infty < \alpha < \infty, \beta > 0, t \geq 0$	$\frac{\phi[(\ln t-\alpha)/\beta]}{\beta t\Phi[-(\ln t-\alpha)/\beta]}$	$1 - \Phi\left[\frac{\ln t-\alpha}{\beta}\right]$
Log-logistic $\alpha, \beta > 0, t \geq 0$	$\frac{\alpha\beta t^{\beta-1}}{1+\alpha t^{\beta}}$	$\frac{1}{1+\alpha t^{\beta}}$
Generalized Gamma $\alpha, \beta, \delta > 0, t \geq 0$	$\frac{f(t)}{S(t)}$	$1 - I^{*}[\delta t^{\alpha}, \beta]$

I^{*} denotes the incomplete gamma function

The likelihood function is then the product of all the individual contributions.

An extension of the parametric regressions models to account for both fixed and time-dependent covariates was proposed by Sparling et al. (2006). The method is based on a general form of the hazard function. A detailed description of the method and an application is shown in Chapter 3.

2.2.4 Other methods to analyse interval-censored data

Other regression models for time to event data are: i) the additive risk model where the hazard function at time t is given by the sum of the baseline hazard and a linear combination of the covariates (more details in Klein and Moeschberger (2003, chap. 10)); ii) weakly parametric methods (named due to their flexibility). Essentially, the baseline hazard is estimated using set of parameters that may change during the observation period. These models can be additive or multiplicative Farrington (1996).

A similar approach may be to consider an arbitrarily interval-censored problem as a binary outcome regression problem with a complementary log-log as link function, a detailed description of the method is provided in Hosmer et al. (2008). In this case the outcome variable can be zero if the individual survives the interval and one if the event is observed during the interval for that specific individual. This is also a semi-parametric approach since the baseline hazard is left unspecified. The estimation of the model can be done via maximum likelihood. If a logistic distribution is used as a link function the model uses the proportional odds assumption, if a complementary log-log is used as a link function this is also a proportional hazards model.

2.3 Frailty models

In survival analysis, frailty models account for correlation between (survival) times and deal with the problem of heterogeneity due to unobserved covariates. For instance, event times from individuals who have common characteristics (siblings, married couples, and so on), or times to ocurrence of different diseases within the same subject.

Assuming a proportional hazards model with frailties, the conditional hazard function of the event is given by:

$$\lambda(t|\boldsymbol{\omega}) \quad = \quad \lambda_0(t)\exp\left(\gamma'\boldsymbol{X} + \boldsymbol{\omega}\right)$$

where γ is the vector of regression coefficients, $\boldsymbol{X}$ denotes the matrix of covariates, and $\boldsymbol{\omega}$ denotes the random effect, which follows a certain distribution f_Ω. This model can be rewritten as follows:

$$\lambda(t|\boldsymbol{\omega}) \quad = \quad \lambda_0(t)\exp\left(\gamma'\boldsymbol{X}\right)\boldsymbol{Z} \tag{2.7}$$

where $Z = \exp(\omega)$ is called the frailty which follows a distribution f_Z.

Different choices are possible for the frailty distribution. However, the distribution is often chosen out of mathematical convenience. Here we summarize some of the options described by Wienke (2011):

- Discrete frailty, where the population is divided in a given number of subgroups, each of them with different risk of an event.
- Gamma frailty, this is one of the most common frailty distributions due to its computational and analytical advantages. As we see in this chapter, thanks

to the simplicity of the Laplace transform, it is posible to derive closed form expressions for the hazard function and the unconditional survival function. However, there is no "biological" reason to prefer this distribution.

- Positive stable frailty. If the normalized sum of n independent random variables has the same distribution than the original variables, then the distribution is called positive stable. The Laplace transform has a simple form, however, there is no closed form expression for the survival function of a random variable with a positive stable distribution.

- Inverse gaussian frailty. It was proposed as an alternative to the gamma frailty model, having closed form expressions for the conditional survival and the hazard function. The power variance function distribution is a three parameter family and includes gamma, inverse gaussian and positive stable distribution.

- Compound Poisson Frailty. It is based on the sum of a Poisson distributed number of independent and identically gamma distributed random variables. There are closed form expressions for the Laplace transform, the marginal survival and hazard function. One interesting property is that it includes a subgroup of zero frailty, with no event.

- Log-normal frailty. Since there is no explicit form of the unconditional likelihood, the estimation strategies for this model rely on numerical integration.

- Univariate frailty cure model. Here we assume that a proportion of people in the population will not experience the event of interest (cured fraction).

In Chapter 5 we restrict to gamma frailty models thanks to its mathematical properties.

In the infectious diseases context, statistical models are very flexible and are useful for identifying risk factors for the infection at hand; however, they do not focus on the transmission process of the viruses (Garnett et al.; 2011). On the other hand, the mathematical models described in Section provide a mechanistic representation of the disease spread.

2.4 Mathematical models for infectious disease transmission

Infection is an invasion of one host organism by smaller disease-causing organism (pathogen), constitutes an ubiquitous phenomenon. There is a variety of ways in which transmission can occur, for instance, through air, from direct or indirect contact with an infected person, and through contaminated food or water among others. There are also many infectious agents: microparasites (virus, bacteria and protozoa) and macroparasites (Helminths and arthorpods). Understanding the dynamic of a infectious disease is key to provide a sensible model.

A considerable amount of literature has been published on mathematical models for infectious disease transmission. Hens et al. (2012) provide an overview of the early contributions on the topic: dÁlembert (1761) is the oldest of those; then Bernoulli (1766) presented a differential equations model of smallpox dividing the population in susceptible and immune (Dietz and Heesterbeek; 2002).

After the papers of Ross (1916) and Ross and Hudson (1917) outlined the deterministic theory for the spread of epidemics in three papers. Then, Kermack and McKendrick (1927, 1932, 1933) studied a deterministic model for a closed population with susceptibles, infectives and removals (those who had died, or had recovered or were immune). The stochastic equivalent of the Kermack and McKendrick model was proposed by Bartlett (1949) (Gani; 2010).

Over the years, the transmission of many infectious diseases such as: smallpox, rubella, malaria, dengue fever, HIV and Hepatis A among others have been an important research topic, with increasing research activity over the last years (Hens et al.; 2012). The transmission models have been extremely helpful in the control of infections and in assessing the effectiveness of vaccination and intervention strategies.

In the next section we present some concepts about the basic SIR model as described in Hens et al. (2012). Subsequently, we describe some transmission models on HCV and HIV for injecting drug users.

2.4.1 Compartmental models

Let us consider a transmission model with three compartments: susceptibles (S), infected (I), immune or recovered (R), we assume all the individuals are born into the susceptble class, then the age of the individual constitutes the exposure time. After infection, the individuals move to the infected class and after clearing the infection

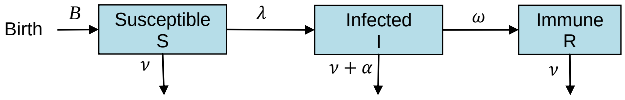


Figure 2.1: Flow diagram for the SIR model. S: susceptible, I: infected, R: recovered/immune

the individuals move to the rcovered/immune class. We assume all the individuals gain lifelong immunity after recovery and do not participate anymore in the transmission process. The SIR model is widely used to model many viral infections in childhood.

Although the exposure time is the age, the transmission parameters may depend on the calendar time. In this case the SIR model can be described by a following set of partial differential equations:

$$\begin{aligned}\frac{\partial S(a,t)}{\partial(a)}+\frac{\partial S(a,t)}{\partial(t)} &= -(\lambda(a,t)+\nu(a,t))S(a,t),\\ \frac{\partial I(a,t)}{\partial(a)}+\frac{\partial I(a,t)}{\partial(t)} &= \lambda(a,t)S(a,t)-(\omega(a,t)+\alpha(a,t)+\nu(a,t))I(a,t),\\ \frac{\partial R(a,t)}{\partial(a)}+\frac{\partial R(a,t)}{\partial(t)} &= \omega(a,t)I(a,t)-\nu(a,t)R(a,t), \qquad (2.8)\end{aligned}$$

where $S(a,t)$, $I(a,t)$, and $R(a,t)$ are the age- and time-specific number of susceptibles, infected, and recovered respectively. There are some boundary conditions given by the assumptions of the model: $S(0,t) = B(t)$ the number of births at time t in the population, $I(0,t) = R(0,t) = 0$, discarding vertical transmission of the infection. Additionally, $N(a,0)$ denotes the age-specific population size at time zero, whereas $\nu(a,t)$ and $\alpha(a,t)$ denote the natural and disease-related death rate, respectively. The force of infection $\lambda(a,t)$ is the rate at which individuals are infected and $\omega(a,t)$ is the recovery rate.

The mathematical models referred in Chapters 5 and 7 are static models, that is, assume time homogeneity. In the manuscript we consider the time since the individual starts to inject drugs as the exposure time. Furthermore, we consider a dynamic transmission model where the force of infection depends on the number of infected individuals in the population (Vynnycky and White; 2010).

As a general framework, the next section describes some transmission model of hepatitis C in injecting drug users proposed by Kretzschmar and Wiessing (2004) to provide some concepts and to introduce the notation used in Chapters 6 and 7. Then, we also present transmission models accounting for HIV and HCV co-infection proposed by Vickerman et al. (2008) and De Vos et al. (2012).

2.4.2 Mathematical models for HCV, and HIV/HCV co-infection on IDUs

To date various models have been developed and introduced with at least one of the following objectives: i) to estimate the prevalence and incidence of HCV and/or HIV, ii) to quantify the disease burden of the two viruses and iii) to assess the impact of the treatment. This section provides an overview of some transmission models which have been proposed.

2.4.2.1 HCV transmission models on IDUs

Based on the natural course of the infection, the model considers three disease stages: acute infection, chronic carrier and recovered. At this point there was a lot of uncertainty about secondary infections so the authors did not consider them in the model.

S_{HCV} denotes the number of susceptibles, I_{HCV} the number of acute infectious individuals, CC_{HCV} the number of chronic carriers and R_{HCV} the number of recovered. The recruitment rate is determined by B and the exit rate is denoted by ν. The force of infection depends on the rate of borrowing injecting equipment κ and the transmission rate. If a susceptible individual borrows equipment from a someone with acute infection the transmission rate is denoted by b_I, and is denoted by b_{CC} if the infectious individual is a chronic carrier. N denotes the total population size. A person with primary acute infection leaves that state with rate ω_I, with a fraction ψ of becoming chronic carriers and the remaining $1 - \psi$ recovering completely. The chronic carriers can still recover at a rate ω_{CC}. The model can be represented by the diagram presented in Figure 2.2.

The following set of differential equations can be used to describe the model:

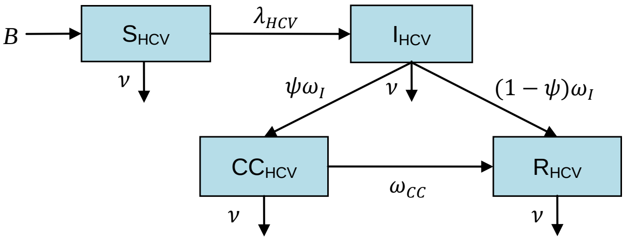


Figure 2.2: Flow diagram of the mathematical model for HCV. S_{HCV}: susceptible HCV, I_{HCV}: acute HCV infected, CC_{HCV}: chronic HCV carrier R_{HCV}: recovered/immune

$$\begin{aligned}
\frac{dS_{HCV}(t)}{d(t)} &= B - \lambda_{HCV}(t)S_{HCV}(t) - \nu S_{HCV}(t), \\
\frac{dI_{HCV}(t)}{d(t)} &= \lambda_{HCV}(t)S_{HCV}(t) - \omega_I I_{HCV}(t) - \nu I_{HCV}(t), \\
\frac{dCC_{HCV}(t)}{d(t)} &= \psi\omega_I I_{HCV}(t) - \omega_{CC}CC_{HCV}(t) - \nu CC_{HCV}(t), \\
\frac{dR_{HCV}(t)}{d(t)} &= (1-\psi)\omega_I I_{HCV}(t) + \omega_{CC}CC_{HCV}(t) - \nu R_{HCV}(t), \qquad (2.9)
\end{aligned}$$

with the force of infection given by:

$$\lambda_{HCV}(t) = \kappa\left(b_I\frac{I_{HCV}(t)}{N(t)} + b_{CC}\frac{CC_{HCV}(t)}{N(t)}\right). \qquad (2.10)$$

To account for the heterogeneous behaviour of the injecting drug users, the model can be extended assuming that there are two subgroups in the population. One of the subgroups with a high average rate of needle sharing and one with a low rate. The authors assume that the subgroups differ in their behaviour but not in the disease-specific parameters, and all the people entering to specific risk group will remain there during their entire injecting career. The time at risk is actually the duration of injection reflecting the exposure time of the individuals during their injecting career.

The HCV transmission model presented in Chapter 7 is an extended version of model proposed by Kretzschmar and Wiessing (2004). The joint model accounts for multiple HCV infections and distinguishes between acute, chronic infected and susceptible individuals who spontaneously clear the virus. The HCV transmission model in Chapter 5 also considers multiple HCV infections; however, it does not include recruitment and exit rates.

Other models HCV include: Hutchinson et al. (2005) who estimated the current and future burden of hepatitis C in Scotland; the authors developed a transmission model including in the disease stages chronic HCV (mild and moderate), compensated and decompensated cirrhosis, hepatocellular carcinoma and as treatment alternative the liver transplantation for a proportion of patients suffering form decompensated cirrhosis. Hutchinson et al. (2006a) considers an stochastic model with two sequential acute infectious stages (not infectious and infectious), recovery and chronic infection.

Then, Vickerman et al. (2007) developed a mathematical model to explore the impact of strategies to decrease syringe sharing in London. The model is an adaptation of the model proposed by Kretzschmar and Wiessing (2004), modified to allow two kinds of acute infection, one leading to a chronic HCV infection and the other to spontaneous clearance of the virus.

In Australia, mathematical models have been proposed to estimate HCV incidence and prevalence (Law et al.; 2003) and to assess the economic impact of the treatment uptake (NCHECR; 2010). Other proposals account by hepatitis C treatment for injecting drug users are: Martin et al. (2011b), Martin et al. (2011a), and among others.

2.4.2.2 HCV and HIV co-infection models

Vickerman et al. (2008) developed a mathematical model for HIV and HCV co-infection aimed at assessing the cost-effectiveness of needle and syringe programmes. The Figures 2.3 and 2.4 represent the model.

A similar model was used to explore the hypothesis of a low prevalence of HCV despite the high rates of sharing needles/syringes, and to project future HIV/HCV co-infection while assessing the impact of interventions (Vickerman et al.; 2009). Recently, Vickerman et al. (2011) used a mathematical model to understand the trends in HIV and HCV prevalences, determining epidemiological profiles.

De Vos et al. (2012) developed a mathematical model to investigate the relationship between the prevalences and the heterogeneity of injecting risk behavior. The authors found that there is threshold HCV prevalence at which HIV can invade into an IDU population. Their results agreed with previous results from Vickerman et al. (2010).

The development of a mathematical model should consider an assessment of un-

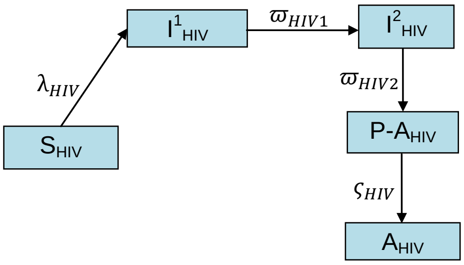


Figure 2.3: Flow diagram of the mathematical model for HIV Vickerman et al. (2008). S_{HIV}: susceptible HIV, I^1_{HIV}: Infected in a high viraemia stage, I^2_{HIV}: Infected in a low viraemia stage, $P - A_{HIV}$: Pre-AIDS, and A_{HIV}: AIDS

.

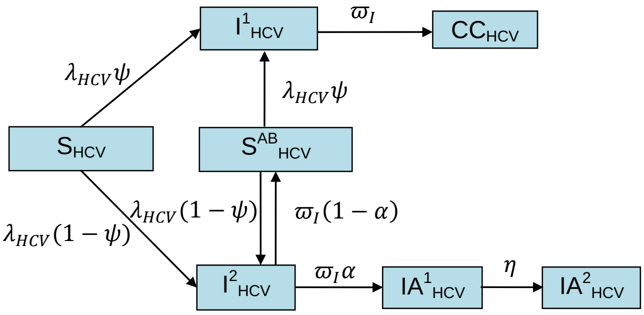


Figure 2.4: Flow diagram of the mathematical model for HCV Vickerman et al. (2008). S_{HCV}: susceptible HCV, I^1_{HCV}: acute HCV infected who afterwards become chronic carriers, CC_{HCV}: chronic HCV carrier, I^2_{HCV}: acute HCV infected who may spontaneously clear the virus, S^{AB}_{HCV}: susceptible individuals who spontaneously clear infection, IA^1_{HCV}: Immune with antibodies, and IA^2_{HCV}: Immune without antibodies

certainty. Below we present the theoretical framework we follow to assess model uncertainty and calibration.

2.5 Model uncertainty and calibration in infectious disease models

Several authors have pointed to the importance of rigorous sensitivity analysis to model the dynamics of a given infectious disease (Bilcke et al.; 2011; Garnett et al.; 2011; Jit and Brisson; 2011; Okais et al.; 2010; Vanni et al.; 2011). Vanni et al. (2011) and Bilcke et al. (2011) provide a methodological framework to account for different sources of uncertainty and to calibrate the model to observational data. Bilcke et al. (2011) classifies the uncertainty in decision analytic-models in three components: methodological (which normative modelling approach should be used?), structural (what structural aspects should be incorporated to capture the relevant characteristics of the disease?), and parameter uncertainty (what is the true value of each model parameter?). Jit and Brisson (2011) adds the model uncertainty to refer to variations due to different categorical choices that cannot be readily parameterized (e.g. a choice between static and dinamic models).

In the HIV/HCV model presented in Chapter 6 we account by structural uncertainty because we consider two definitions for the force of infection and parameter uncertainty.

Vanni et al. (2011) proposed a seven steps approach to guide the calibration process of a mathematical model:

- Select the parameters should be varied in the calibration process
- Select the calibration targets. It refers to the data used to calibrate the model based on the objectives of the study.
- Define a goodnes of fit measure. It allows to assess how close is the model to the data
- Select a parameter search strategy.
- Determine acceptable goodness of fit sets. It is directly related with the convergence criteria
- Determine the stopping rule to ends the calibration process.

- Integrate the calibration results and the economical parameters. This is useful to measure intervention strategies.

Following Vanni et al. (2011) on Chapter 7 we calibrate HIV/HCV model (presented in Chapter 6) using statistical methods and concepts. The seven steps we follow are: i) we did not discard any parameter in the calibration process; ii) our target is to reproduce the trends observed on a regional data from Italy and Spain; iii) our goodness of fit measure is a multinomial likelihood considering the serostatus of the individuals and the duration of injection; iv) we use the latin hypercube sampling as parameter search strategy; v) the acceptance criterion was based on the observed percentiles of the multinomial likelihood; vi) as stopping rule we consider 500,000 parameter sets; the last point was not considered.

Additionally we assess the statistical variability using bootstrap and perform analyses of the model parameters in the high-dimensional parameter space.

2.6 Statistical and mathematical models to analyse cross-sectional data

The force of infection can be estimated from cross-sectional seroprevalence surveys. In such survey, taken at a certain calendar time, each participant is tested for the presence of infection-specific antibodies, a marker for a past infection and thus constituting current status data on past infection. The participant age is usually considered as the time at risk.

Hens et al. (2010) presented an overview on estimating the force of infection from cross-sectional data. Key-contributions of the 20-th century on the topic are attributed to: Muench (1934), Wilson and Worcester (1941) (with constant force of infection), Griffiths (1974) (with linear force of infection), Grenfell and Anderson (1985) (polynomial), Farrington (1990) (non-linear) and Keiding (1991) (non-parametric estimation).

At Hasselt University a lot of work has been done in the area of modelling infectious diseases. Shkedy et al. (2003, 2006); Namata et al. (2007) proposed non-parametric, semiparametric and parametric methods to model seroprevalence data in the generalized linear mixed model framework by using local polynomials, fractional polynomials and penalized splines. Additional contributions have been included in the book of Hens et al. (2012).

In the IDU setting serological data constitutes one valuable source of information for understanding HIV and HCV epidemiology. Serological data comprises the serostatus for each individual and self-reported data on the duration of injection can be considered as a measurement of the time at risk. Several studies have addressed modelling the force of infection and co-infection for HCV and HIV in IDU populations based on serological data (Sutton et al.; 2008; Platt et al.; 2009; Del Fava et al.; 2011). In these models, the force of infection has been estimated as a function of the exposure time and a term reflecting the individual heterogeneity in the acquisition of the virus; a frailty term.

Platt et al. (2009) use a piecewise constant model and consider the serostatus for HIV and HCV. Sutton et al. (2008) use a random effect model considering three viruses: HIV, HCV and HBV. Where the proportion of individuals infected is a function of the cumulative force of infection. The study evidences the individual heterogeneity of force of infection estimates within the overall IDU population. Del Fava et al. (2011) use marginal models to estimate the association measures between HCV and HIV infections. The authors also consider a risk factor analysis and some random effect models to take into account the individual heterogeneity in the acquisition of the infections.

The statistical models are very flexible and are useful for identifying risk factors for the infection at hand; however, they do not focus on the transmission process of the viruses (Garnett et al.; 2011). On the other hand, the mathematical models described in section 2.4 provide a mechanistic representation of the disease spread.

The models proposed by: Vickerman et al. (2007, 2011) are some examples of mathematical models calibrated to cross-sectional surveys. Whereas (Law et al.; 2003; Vickerman et al.; 2008; Hutchinson et al.; 2005) consider several sources of information simultaneously for the calibration process.

Chapter 3

Effect of probiotics on acquisition of multiresistant Enterococci: survival analysis with interval-censored data and time-dependent covariates

As indicated in Chapter 1, antibiotic resistance has a considerable impact on morbidity and mortality of the hospitalized patients. Although new generations of antibiotics with fewer side effects have been developed, the incidence of antibiotic associated diarrhea ranges between 3.2 and 29 per hundred hospitalized patients (Gupta and Garg; 2009). Some studies suggest that probiotics may help to maintain the integrity of the intestinal flora and that they augment restoration of integrity after disruption (Hickson et al.; 2007; DSouza et al.; 2002). However, there is need of evidence to support the use of probiotics to prevent infections as it has been pointed by Oudhuis et al. (2011). Well-designed clinical trials are the answer to quantify the impact of probiotics to prevent infections.

In the University Medical Center Utrecht proportions of ampicillin-resistant *Enterococcus faecium* (ARE) increased in the last 15 years, from 2% in 1994 tot 50% in 2008. A prospective cohort was designed to quantify the effects of probiotics and antibiotics on acquisition of ARE-colonization in patients admitted to two hospital wards with, previously documented, high prevalence of intestinal ARE carriage (de Regt et al.; 2008).

Acquired antibiotic resistance seriously limits therapeutic options when infections occur, increasing the risk of treatment failure (Brown et al.; 2006). Typically, infections with multiresistant nosocomial pathogens such as ARE are preceded by colonization. Since colonization is asymptomatic, colonized patients serve as silent reservoirs with a high propensity of unnoticed spread to the other patients.

The data used in this chapter was kindly provided by M. de Regt. This chapter constitutes an extension of the previous analyses presented on de Regt et al. (2010). Here we account for the interval-censored observations and time dependent covariates. We first estimate the survival curve (the probability that the time to event does not occur before a specific point in time) without any distributional assumption (nonparametrically). Secondly, we present several methods to compare survival curves between groups. Next, we introduce semi-parametric, parametric regression methods accounting by interval-censored data. The parametric regression technique applied here also accounts for time-dependent covariates.

The study was performed in the gastroenterology/nephrology and geriatric ward, of the University Medical Center Utrecht between June 2007 and March 2008. The patients with expected length of stay longer than two days were screened for ARE colonization by obtaining perianal swabs, the screenings took place within 48 hours after admission, twice weekly and within 48 hours before discharge. In each ward there were two periods one without the intervention, being the control period, and one with intervention in which all the recruited patients were offered probiotics twice daily during their entire stay on the wards.

The study collected information about 530 patients, 94 (18%) were ARE colonized at admission and were not included in this analyses. Of remaining 436 noncolonized patients at admission, 92 acquired are colonization. For the purpose of this analysis, we only include the 436 patients. The data includes: i) demographic characteristics of the patienst such as gender and age; ii) variables related to the hospitalization such as admision and discharge dates, ward, date of last negative result for ARE colonization, date of first positive result for ARE colonization, probiotics use, and first date and last date of probiotics intake.

In the following sections, we discuss and illustrate three different approaches to analyze interval-censored data. First, we do not make any assumption regarding the shape of the survival function nor the relationship between the covariates and the time to acquisition (nonparametric). Next, we assume a specific type of relationship between the hazard function and the covariates but not regarding the shape of the hazard function (semi-parametric). Finally, we assume a distribution for the survival function and for the covariates (parametric).

3.1 Non-parametric analysis of the survival

Initially, we estimate the survival curve without any a priori assumption regarding the distribution of the time to ARE acquisition. We use the Nonparametric Maximum likelihood Estimator Turnbull (1976) described in chapter 2. To estimate the survival curve the self consistency algorithm available in the R packages 'Interval' and 'Icens' is used. The estimated curve provides the probability that a study subject survives past a specified time. Other two algorithms that can be used are: hybrid Expectation Maximization-Iterative Convex Minorant estimator and Project Gradient Methods.

Secondly, we distinguish between subgroups as defined by several categorical covariate: probiotics use and antibiotics use. Originally, age at admission is continuous, yet in this section we dichotomize the variable as: 60 years or younger vs older than 60 years. After, we describe some formal tests to compare the survival curves providing information about the significance of the factors considered.

There are two types of tests: rank-based and survival-based tests. The rank-based tests consider a weighted difference of the estimated hazard functions in the different groups. The survival-based tests rely on the weighted differences of the estimated survival functions. Only a few of these tests have been made available in existing or implemented in new software. In R, the packages 'Interval' and 'glrt' can be used to perform some of the tests. In SAS there is a macro that allows the comparison of two survival curves accounting by interval-censored data (details are shown in Chang (2006)).

We focus on four tests: in the first one the scores are associated with the grouped proportional hazards model of Finkelstein (1986); the second corresponds to a generalization of the Wilcoxon Mann Whitney scores; and the last two are generalized tests proposed by Zhao and Sun (2004) and Zhao et al. (2008).

3.2 Semi-parametric regression for interval-censored data

The Cox model is the most commonly used method to analyze right-censored data and is available in most statistical software packages. The Cox model provides an estimate of the hazard ratio (a ratio of the hazards rates in the treated versus the control group) and its confidence interval. The hazard for an individual is modelled as a product of two factors: the baseline hazard that is left unspecified (nonparametric part) and a factor that characterizes how the baseline hazard function changes as a function of subject covariates (parametric part).

For interval-censored data a unified approach is still lacking, since, the estimation of the unknown baseline hazard is challenging. A more detailed overview was presented on Chapter 2. Partly due to the great computational burden required to implement any of those methods and due to the lack of a unified approach, only the proposal made by Pan (2000) has been implemented in R, more specifically, in the package 'intcox' albeit that the package does not provide standard errors for the estimated regression parameters. Consequently, the authors of the package suggest using a nonparametric bootstrap (Davison and Hinkley; 1997).

3.3 Parametric regression models for interval-censored data with time-dependent covariates

In the parametric regression models, we assume a distribution for the time to event outcome. In the previous section the underlying assumption is that the effect of the covariates is proportional (multiplicative) with respect to the hazard, whereas in the AFT model the effect of the covariates is proportional (multiplicative) with respect to the survival time.

To assess the impact of the multiple time dependent covariates, we use an approach proposed by Sparling et al. (2006). In this framework a parametric model for the survival curve is assumed and time dependent covariates are also taken into account. The approach is based on a smart parameterization of the hazard function. And some of the well know distributions of the AFT model are special cases.

The general form of the hazard function is given by:

$$\lambda(t) = \frac{\alpha\beta t^{\alpha-1}}{[1+\beta t^{\alpha}]^{\kappa}}. \tag{3.1}$$

($\alpha, \beta > 0$; κ is any real number). For specific values of κ, we can appreciate similarities with some of the distributions presented in Table 2.1. If $\kappa = 0$, $\lambda(t)$ is the hazard function for the Weibull distribution, when $\kappa = 1$, the family yields the log-logistic distribution.

Let $x_i = (x_{i1}, \ldots, x_{ip})'$ be a vector of p fixed covariates for the i-th subject. Additionally the time dependent covariates are updated at a sequence of updated times $\tau_{i0}, \ldots, \tau_{ik_i}$, where τ_{i0} is the time at which a subject enters follow up. The set of update times τ_{ij} may differ among subjects. At the jth update time of the ith subject τ_{ij}, let $y_{ij} = (y_{ij1}, \ldots, y_{ijq})'$ denote the vector of q time dependent covariates values that are updated at that time.

Let γ and η be the coefficient vector for the fixed and time dependent covariates x_i and y_{ij}, respectively, so that: $\gamma' x_i = \gamma_1 x_{i1} + \ldots + \gamma_p x_{ip}$ and $\eta' y_{ij} = \eta_1 y_{ij1} + \ldots + \eta_q y_{ijq}$

Let θ be the intercept of the model. $\beta_{ij} = \exp(\theta + \gamma' x_i + \eta' y_{ij})$, is a rate parameter conditional on the covariate values at update time τ_{ij}. For subject i, conditional on covariates measured at baseline at time τ_{ij}, this hazard can be expressed as:

$$\lambda(\tau_{ij}|z_i, y_{ij}) = \frac{\alpha\beta_{ij}\tau_{ij}^{\alpha-1}}{[1+\beta_{ij}\tau_{ij}^{\alpha}]^{\kappa}} \tag{3.2}$$

In the study there are several time dependent covariates: probiotics use which is one if the patient took the probiotics an specific day and zero otherwise, colonization pressure defined as the proportion of colonized patients in a ward, antibiotics intake as dummy variable being one when the patient took antibiotics and zero otherwise, and as a fixed covariate we consider age at admission and ward. We also consider the interaction between antibiotics and probiotics.

3.4 Application to the UMCU probiotics dataset

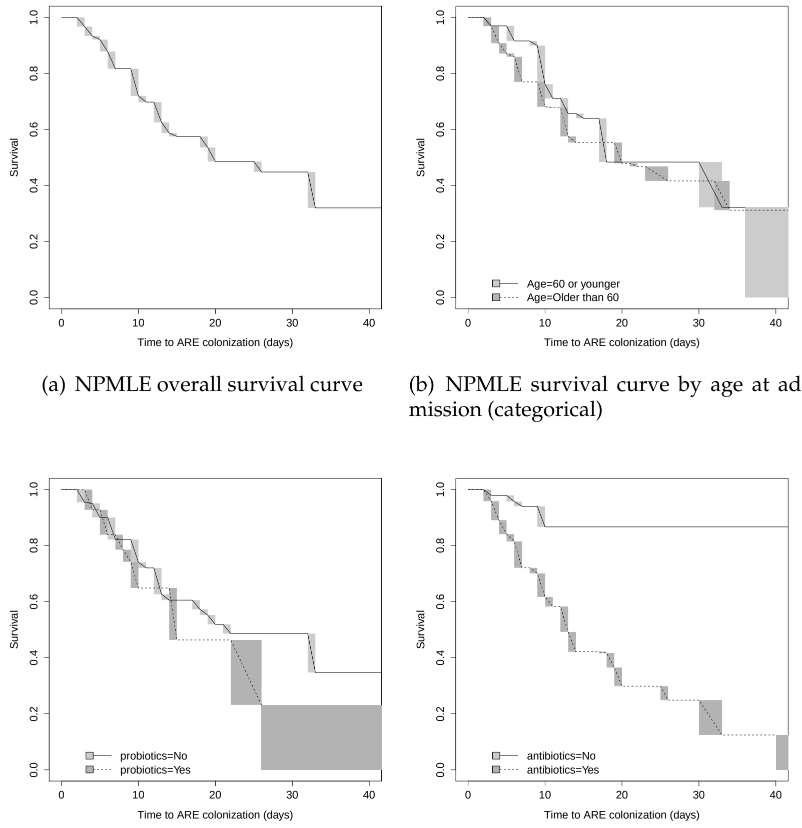
In this section we present the results of the analysis of the UMCU dataset. Initially we present the nonparametric estimates of the survival function without any covariates and then considering important covariates such as probiotics use, antibiotics use and age of admision. Initially, we ignore the time-dependence nature of some of the covariates such as probiotics and antibiotics use and colonization pressure.

Next, we present the results of a semi-parametric regression model proposed by Pan (2000) and parametric survival model for interval-censored accounting by time-dependent covariates following the approach proposed by Sparling et al. (2006).

3.4.1 Nonparametric estimation of the survival functions

Initially, we estimate the survival curve without any a priori assumption regarding the distribution of the time to ARE acquisition. Figure 3.1 shows the nonparametric maximum likelihood estimates (NPMLE) of the survival curves, based on the self consistency algorithm, first without covariates (Figure 1a), and then with covariates Figures (1b-1d). The estimated curve (Figure 1a) provides the probability that a study subject do not get infected after a specified period time (expressed in days), e.g. the probability that a subject is free of colonization after one month in the hospital is 0.32.

Next, we distinguish between subgroups as defined by the categorical covariates: Figure 1b for age at admission; Figure 1c for probiotics use and Figure 1d for antibiotics use. Originally, age at admission is continuous, yet in this section we dichotomize the variable as: 60 years or younger vs older than 60 years. Based on Figure 1b, the effect of age at admission seems more pronounced in the first 16 days, with a longer time to acquisition for patients of 60 years or younger. Based on Figure 1c there is no consistent impact of probiotics use on the survival function: the two curves are very close to each other during the first 15 days, whereas the estimated survival curve of those who were not treated with probiotics is higher from day 16 to day 40. On the other hand, the impact of antibiotics use is quite clear; resulting in a higher curve for the group without antibiotics (Figure 1d). The patients who took antibiotics could have been exposed to an ARE therefore we see that the NPMLE for time to colonization decreases in this group. For estimation purposes, probiotic and antibiotics use was dichotomized in 'Yes' for those who took probiotics/antibiotics at least once and 'No' otherwise.



(a) NPMLE overall survival curve

(b) NPMLE survival curve by age at admission (categorical)

(c) NPMLE survival curve according to probiotics use

(d) NPMLE survival curve by antibiotics

Figure 3.1: UMCU probiotics dataset. Nonparametric maximum likelihood estimates of the survival functions.

We obtain similar results using two other algorithms to estimate the NPMLE: the hybrid Expectation Maximization-Iterative Convex Minorant estimator and Project Gradient Methods (results are not shown).

After comparing the NPMLE, we perform some non-parametric tests to compare survival curves according to probiotics, antibiotics and age at admission.

In order to compare survival curves for interval-censored data several tests have been proposed. For a detailed description we refer to (Lesaffre et al.; 2005) and Sun (2006). Among the more recent proposals are: a family of tests that extends the Fleming-Harrington test described in the paper by Gómez et al. (2009) and a weighted logrank test proposed by Fay and Shaw (2010).

The tests can be classified into rank-based or survival-based tests. The rank-based tests consider a weighted difference between the estimated hazard functions in the different groups. The survival-based tests rely on weighted differences between the estimated survival functions.

Despite the considerable amount of tests available for interval-censored data, a detailed comparison including all the most recent proposals has not been performed. However, some comparisons including the performance of certain tests have been reported in Fay (1999) and Huang et al. (2008).

Deriving the asymptotic behaviour of the tests in an interval censored setting is more complex, therefore some authors use asymptotic methods with the observed Fisher's information, and some rely on resampling methods including permutations, multiple imputation (Huang et al.; 2008) or bootstrap.

Furthermore, few of the tests have been directly implemented in software; in R the packages 'interval' and 'glrt' compute some of the tests. Additionally, Gómez et al. (2009) shows a detailed description in order to use the family of tests developed by Gomez and Oller (2008).

In SAS the macro

It is worth to mention the time dependence nature of the covariates of the covariates in the study is not taken into account under any of the approaches formulated before.

We obtain consistent results with all the tests considered. Probiotics use was found to be non significant, whereas antibiotics use was (Table 3.1). The results for age of admission are not that clear, the p-values are between 5% and 10%. As we will see in the next section the age of admission has an indirect effect on the outcome. In fact the NPMLE shows that the survival curves cross after 20 days, this is certainly a challenge, since the tests are more suitable to detect ordered survival curves.

The nonparametric tests presented above allow comparisons for each categorical

Table 3.1: UMCU probiotics dataset. Comparison of the survival functions according to: probiotics, antibiotics and age at admission using different scores.

	Finkelstein scores Finkelstein (1986)		Generalized Wilcoxon Mann-Withney scores		Generalized test Zhao and Sun (2004)		Generalized test Zhao et al. (2008)	
Covariate	Z	P-value	Z	P-value	χ^2	P-value	χ^2	P-value
Age at admission (categorical)	2.876	0.0899	-1.893	0.0583	2.915	0.0877	2.750	0.0972
Probiotics	-0.917	0.3589	-0.829	0.4070	0.816	0.3663	0.844	0.3584
Antibiotics	-6.280	<0.0001	-6.042	<0.0001	39.744	<0.0001	39.528	<0.0001

factor one at a time; moreover, the joined effect of two (or more) categorical covariates can be studied by defining a new variable with levels being all combinations of the original covariates. However, this procedure might produce subgroups with a very low number of observations or even empty subgroups. These tests can also be used to compare subgroups of continuous covariates after categorization (as previously shown for age at admission).

The tests previously described do not provide any parameter quantifying the effect of each covariate; they only provide information regarding their significance. When the main interest is to quantify the impact of several factors, or to predict the time to event, then regression modelling techniques are more appropriate. Those include the extensions of the well-known Cox proportional hazard models (semi-parametric) and the Accelerated Failure Time (AFT) models (parametric) that we describe below.

Furthermore, note that the time dependent nature of the covariates in the study has not been taken into account using the nonparametric approach. In the following sections, we also point to some methods that can be applied in the presence of time dependent covariates.

3.4.2 Semi-parametric model for interval-censored data

We applied the method proposed by Pan (2000) and calculated the bootstrap confidence intervals using nonparametric bootstrap (Davison and Hinkley; 1997). Based on the results we conclude that age at admission and probiotics use are not significant. Antibiotics use is found to be a significant risk factor with the hazard of ARE colonization being five times higher for those who were treated with antibiotics as compared to those who did not receive antibiotics (3.2). Notice that the interest is to reject the null hypothesis that the hazard ratio is equal to one, implying that there is no impact of the variable on the hazard function.

3.4.3 Parametric models for interval-censored data considering time-dependent covariates

Now we assume a particular distribution for the time to event outcome. The most often applied method is the Accelerated Failure Time (AFT) Model. In the previous section the underlying assumption is that the effect of the covariates is proportional (multiplicative) with respect to the hazard, whereas in the AFT Model the effect of the covariates is proportional (multiplicative) with respect to the survival time.

Table 3.2: UMCU probiotics dataset. Parameter estimates and standard errors for the semi-parametric regression model

Parameter	Estimate (SE)	P-value	Hazard ratio	95% CI Hazard ratio
Age at admission	0.007 (0.0060)	0.1155	1.007	(0.996 ; 1.020)
Probiotics	0.259 (0.232)	0.1317	1.296	(0.824 ; 1.994)
Antibiotics	1.637 (0.352)	<0.0001	5.142	(2.967 ; 11.977)

The AFT model assumes that the log-transformed time to event is a linear function of the predictor variables (similar to a classical linear regression model). One common choice for the distribution is the Weibull distribution, implying that the error term follows an extreme value distribution and that the proportional hazards property (as in the Cox Proportional hazard model) is satisfied. Other popular distributions are the log-logistic and the log-normal distribution. To select the model that fits the data best, likelihood ratio tests can be selected for the nested models and selection criteria like the AIC for the non-nested models. Sparling et al. (2006) propose a model that accounts for interval-censored data considering time-dependent covariates. The parameterization proposed by Sparling et al. (2006) described in Section 3.3 has the Weibull and the log-logistic distributions as particular cases.

Firstly, we consider several models with one time-dependent covariate each (results are not shown), from them only age at admission and antibiotics use were significant. The remaining time-dependent covariates probiotics use, isolation and colonization pressure were not significant. Subsequently we consider a full model with the following terms: as main effects we include antibiotics use, probiotics use, isolation and colonization pressure; and three interaction terms probiotics-age, probiotics-antibiotics and antibiotics age. The three interaction terms as well as the main effects for colonization pressure and isolation were sequentially removed from the model following a backward selection procedure. We also assumed the shape parameter κ equals zero taking into account the correspondent p-value. Hence, we assume a Weibull distribution as a parametric form for the model. The estimates for the reduced model are shown in Table 3.3.

Based on the results of Table 3.3, those individuals who take antibiotics have higher risk of developing ARE than those who do not. Besides we observe a slight higher risk of ARE in the individuals that took probiotics. The results are counterintuitive because the purpose of the probiotics administration is to protect again ARE

Table 3.3: UMCU probiotics dataset. Estimates for the parametric reduced model with age as fixed effect and probiotics and antibiotics use as time-dependent covariates.

Parameter	Estimate (SE)	P-value	Acceleration Factor (AF)	95% CI for AF
Intercept	-6.239 (0.561)	<0.0001		
Age at admission	0.009 (0.006)	0.1053	1.009	(0.998 ; 1.021)
Probiotics use: yes	0.549 (0.268)	0.0404	1.732	(1.024 ; 2.929)
Probiotics use: no	0.000		1.000	
Antibiotics use: yes	2.150 (0.268)	<0.0001	8.588	(5.075 ; 14.534)
Antibiotics use: no	0.000		1.000	
Scale	1.302 (0.128)			

colonization. Besides, probiotics use appears to be significant whereas the previous tests did not show any significant result.

In contrast, age at admission showed some borderline significant results in the non-parametric tests, when the variable was dichotomized, but does not appear significant in the model. As we pointed out before, this may be attributed to the fact that the age effect is changing during the analysis period. Therefore, we focus on the marginal distribution of the age at admission by probiotics use.

Figure 3.2 shows some difference in the distribution of age according to probiotics intake, the peak of the density function for this group is around 80 years old, whereas the peak for the group without probiotics is around 60 years. Meaning that the individuals who took probiotics where in general older than those who did not. As we can expect, older people run a higher risk of ARE-acquisition compared to younger individuals.

We categorize the age variable because we do not think the relationship between time to colonization and age is linear. We define four groups: younger than 41 years old, between 41 and 60 years, between 61 and 80 years and older than 80 years.

In fact, the results considering a model with the three age categories shown in Table 3.4 reveal that the impact of probiotics use was not significant in this study; additionally, there is no significance difference in the time to ARE infection for the patients in older age groups, the difference is only significant between the first category (younger than 41) and the baseline (older than 80 years old).

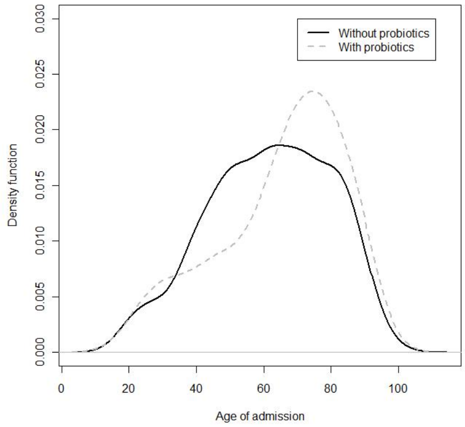


Figure 3.2: UMCU probiotics dataset. Estimated density function of age at admission according to the probiotics intake.

3.5 Concluding remarks

The study was performed in a hospital with documented high prevalence of intestinal ARE carriage, in this setting we did not find significant impact of daily probiotics intake on the reduction of the time to ARE acquisition. In the same sense, a recent meta-analysis by Hempel et al. (2012) mentions that most of the trials did not show a statistically significant advantage of probiotics use and a review made by (Oudhuis et al.; 2011) shows conflicting results regarding the effects of probiotics on infection rates.

When we compare distribution of the age of admission for the two groups: with and without probiotics we notice the patients who receive probiotics tend to be older than those who did not receive them. This may indicate a selection bias which can clear affect the results agains probiotics use.

Therefore, more research is needed to demonstrate which are the most effective probiotics and to target the patients according to the type of antibiotics they take (Hempel et al.; 2012). Another aspect that should be taken into account is the clinical condition of the patient (Oudhuis et al.; 2011).

As a sensitivity analysis we assess the impact of the interval-censoring on the results. We compare the model estimates from Table 3.3 with imputed time to event. The imputation approach is a common naive way to deal with interval-censored

Table 3.4: UMCU probiotics dataset. Estimates for the parametric reduced model with age as fixed effect categorized in three groups and probiotics and antibiotics as time-dependent covariates.

Parameter	Estimate (SE)	P-value	Acceleration Factor (AF)	95% CI for AF
Intercept	-5.606 (0.449)	<0.0001		
Age at admission $\leq$40	-1.088 (0.489)	0.0260	0.337	(0.129 ; 0.878)
Age at admission 41-60	0.011 (0.290)	0.9705	1.011	(0.572 ; 1.786)
Age at admission 61-80	0.090 (0.253)	0.7213	1.095	(0.666 ; 1.799)
Probiotics use: yes	0.520 (0.268)	0.0525	1.682	(0.994 ; 2.846)
Probiotics use: no	0.000		1.000	
Antibiotics use: yes	2.173 (0.268)	<0.0001	8.789	(5.192 ; 14.879)
Antibiotics use: no	0.000		1.000	
Scale	1.319 (0.130)			

data, here the time to event is imputed using a left, right or midpoint value of the interval after which standard methods to analyse the data can be applied. It has been shown that this approach can lead to biased and misleading results (Lindsey and Ryan; 1998). Additional analyses carried out shown that for this specific data the impact of interval-censored data is negligible.

However, in general ignoring interval-censored data and naively imputing the exact time to event is not recommended. Several methods can be implemented depending on the objectives of the study. In this Chapter, we presented a brief overview of methods and available software, with an illustration in the microbiological context. We did not attempt to present an exhaustive review, excluding topics such as clustering and correlated data, current status data as a particular case of interval-censored, or casting methods in a bayesian framework. Chapter 5 describes some methods to account for clustering and current status data.

We also did not mention flexible modelling, such as smoothing splines for continuous covariates or mixture distributions for the error terms, which can be applied. For the weakly parametric methods, the 'Epi' package from R, can be used to fit additive and multiplicative models as proposed by Farrington (1996).

We present three algorithms to obtain a Nonparametric Maximum Likelihood Estimator, as well as some options to compare survival curves considering a categorical covariate. To test for continuous covariates or more than one covariate simultaneously parametric regression methods are straightforward to apply and, importantly, time dependent covariates can be taken into account albeit with some software limitations. In this framework the inferences depend upon the assumption of the model, which is difficult to assess in this setting. Therefore the approach is parametric in nature, in contrast to the nonparametric estimation (NPMLE).

Interval-censored data is very common when the event of interest can only be monitored at specific time points, yet many of the proposed methods have not been implemented in any statistical software. The lack of software is notorious in the extensions of the Cox proportional hazard models, the most popular method for right censored data, partly due to the lack of a unified approach in this setting in combination with the large computational efforts that are needed. Finally, for right censored settings the assumption of proportional hazards in the Cox model can be tested by visual methods as well as formal tests based on the residuals of the models or the inclusion of time dependent indices (Therneau and Grambsch; 2000). For interval-censored data the proportionality may only be assessed using graphical techniques.

All the methods discussed here assumed that the censoring mechanism is independent of the time to event (independent interval censoring). To deal with dependent or informative interval censoring we refer to Sun (2006).

Chapter 4

The estimation of the force of infection for HCV among injecting drug users using interval-censored data

As has been mentioned in Chapter 1, hepatitis C virus is a clear treat for public health. In developed countries, 90% of the persons with chronic HCV infection are current or ever injecting drug users, or have a history of transfusion with unscreened blood. In addition, IDUs are also at high risk to acquire HIV and other infectious diseases due to common transmission routes such as sharing syringes or other injecting paraphernalia.

There is a vast amount of literature dealing with the estimation of the infection hazard, often referred to as the force of infection (FOI), from cross-sectional seroprevalence surveys (see Hens et al. (2010), for an overview). In such a survey, taken at a specific calendar time, each participant is tested for the presence of infection-specific antibodies, a marker for past infection and thus constituting current status data on past infection. In general, the participant's age is considered the time at risk. However, in the setting of IDUs, the cross-sectional sample has information on the serostatus of each individual and the self-reported duration of injection is usually considered as a more precise measurement of the time at risk. A quintessential assumption in the estimation of the FOI from cross-sectionally collected seroprevalence data is the assumption of time homogeneity, i.e. assumes that the FOI is invariant with respect to calendar time. This assumption can be relaxed when either a cohort study or repeated cross-sectional studies are available.

The Amsterdam Cohort Studies is a prospective cohort study that test participants' blood for infections at each follow-up visit. Therefore, the exact time to event is unknown but the time interval in which the infection occurs is known. In survival analysis, data of this type are known as type II interval-censored data, whereas current status data constitute type I interval-censored data (Sun; 2006). In the literature, several authors ignored the interval and imputed the time to event T using the left, right or midpoint value of the interval after which they applied standard time to event techniques to analyse the data. It has been shown that this approach can lead to biased and misleading results (Lindsey and Ryan; 1998); e.g. the right endpoint imputation yields inflated estimates of the risk Dorey et al. (1993). Therefore appropriate techniques have to be used (Sun; 2006).

The estimation of the force of infection for the IDU population was based previously on cross-sectional data (Del Fava et al.; 2011; Mathei et al.; 2006; Namata; 2008; Platt et al.; 2009; Sutton et al.; 2006, 2008). The major contribution of this study is the estimation of the FOI for HCV among IDUs using a large cohort study, with more than 25 years of follow up, while assessing the impact of self-reported behavioural risk factors (injection frequency, type of drug injected, sharing of syringes) using an adequate statistical model. Moreover, the inclusion of date of first injection as a factor relaxes the assumption of time homogeneity which is made when cross-sectional data is analysed.

The chapter is organized as follows. In Section 4.2, we describe nonparametric survival models to estimate the time to HCV infection using interval-censored data and parametric survival models to identify potential risk factors. The models are applied to the Amsterdam Cohort Studies data in Section 4.3 while focusing on the estimation of the FOI for HCV and the identification of risk behaviour factors associated with infection. We end with a discussion in Section 4.4.

4.1 Study population and data

The Amsterdam Cohort Studies is a collaboration of several institutions in the Netherlands. ACS is part of the Netherlands HIV Monitoring Foundation and is financially supported by the Netherlands National Institute for Public Health and the Environment.

The cohort study initiated in 1985 to investigate the prevalence, incidence, and risk factors of HIV infections and other blood-borne and/or sexually transmitted diseases, as well as the effects of intervention. Participation in the ACS is voluntary, and informed consent is obtained for every individual at intake. ACS participants visit the Amsterdam Health Service every 4-6 months, they complete a standardized questionnaire about their health, risk behaviour, and sociodemographic situation. Questions at ACS entry refer to the 6 months preceding the visit; questions at follow-up refer to the interim since the preceding visit. Blood is drawn each visit for laboratory testing and storage. Until 2006, 1,663 DUs have been included in the ACS. The recruitment for the DUs was via methadone programs, a sexually transmitted diseases clinic for drug using sex workers and by word of mouth. A drug user was defined as an IDU if he or she reported ever having injected drugs (World Health Organization (WHO); 2011).

4.2 Methods to estimate force of HCV infection and survival function

Several attempts have been made to estimate the force of infection of HCV in the IDU context based on cross-sectional data (Del Fava et al.; 2011; Mathei et al.; 2006; Namata; 2008; Platt et al.; 2009; Sutton et al.; 2006, 2008). All authors assume a parametric function for the prevalence and the force of infection, either assessing the impact of covariates or taking into account the association with other viruses. The proposed methods were applied to the binary data representing the current status of the disease of each IDU. Diverse and more appropriate techniques can be applied to cohort data and therefore, given the data at hand a survival analysis taking into account censoring and truncation should be considered.

Within survival analysis the main interest is in the estimation of the time to event distribution and factors that affect it. One of those factors is the censoring, where only partial information about the event is known. Denote T, the time until an event occurs also called survival time, and d, the censoring indicator, which takes value one if the event occurs and zero if at the end of the study period the event has not been observed. In that case, the subject is said to be right censored and the time to event T is taken to be equal to the follow up period. If the event of interest, in our case infection with HCV, has occurred before the subject enters the study, the data is left censored.

The ACS is a follow up study in which the exact time to infection with the HCV virus (T) is unknown but the time interval in which the infection occurs is known. Let T denote the time to infection and L and R the left and right limit of the interval in which the subject was infected, $L \leq T \leq R$. For current status data $R = \infty$ for a right censored subject (seronegative) or $L = 0$ for a left censored subject (seropositive).

We use the definitions for survival and hazard function described in Chapter 2 Section 2.2.

Considering the characteristics of the study population, the time at risk is given by the self-reported number of years injecting. That is, the time since an IDU starts to inject drugs until he or she becomes infected with HCV. In what follows we first discuss nonparametric approaches to estimate the survival function in case of interval-censored data. We then introduce accelerated failure time models for interval-censored data and estimate the force of infection in case of interval-censored data while accounting for behavioural risk factors and time heterogeneity.

4.2.1 Nonparametric estimation of the survival function

We consider a nonparametric estimate for the survival curve, using the algorithm proposed by Turnbull (1976) which is called a self-consistency algorithm to obtain a nonparametric maximum likelihood estimator (NPMLE) of the survival function. The interval-censored data is treated as incomplete data and the Expectation-Maximization (EM) algorithm Dempster et al. (1977) is applied to take these incomplete data into account.

4.2.2 Accelerated Failure Time Models

Assessing the influence of risk factors in a survival analysis can be done within the accelerated failure time framework, where the time to HCV infection is assumed to follow a specific distribution. We follow the description of the AFT model presented in Section 2.2.3. The regression coefficients have an interpretation similar to those in standard regression.

For the participants in the ACS, the exact time of HCV infection is unknown. Hence we define the limits for the interval in which an IDU was infected as follows: for the seroconverters the lower limit of the interval is the number of years of injection at the last negative result for HCV whereas the upper limit is the number of years of injection at the first positive result; for the individuals who were negative at the end of the follow up the lower limit is the number of years of injection until at last visit, and the upper limit is infinite, that is:

Last negative test result $\leq T \leq$ first positive test result, seroconverter

Last negative test result $\leq T \leq \infty$ seronegative

4.2.3 Left truncation

Left truncation arises when individuals come under observation only some known time after the natural time origin of the phenomenon under study (Klein and Moeschberger; 2003). For this study, the data are left truncated as a condition for inclusion in the study is that individuals are uninfected at cohort entry. To account for left truncation Pencina et al. (2006) proposed five different methods all yielding similar results. The method employed here accounts for left truncation by including the duration of injection at the first visit as a covariate in the model and the results thus warrant a conditional interpretation.

4.3 Application to the Amsterdam Cohort Studies dataset

The description of the ACS dataset was presented in Section 1.2.1. For the analysis presented in this chapter we only included those who entered negative for HCV, totalling 165 individuals: 58 who became seroconverters during the follow up period and 107 who remained negative.

From this group of IDUs 66.1% were males. The average age of first injection was 25.4 years (se 7.8 years), whereas the mean age at first visit was 30 years (se 7.4 years), the mean of the follow up time was 7.9 years (se 5.4 years). With respect to sharing needles, 33.5% stated sharing syringes at least once during the follow up period; concerning the frequency of injection, 41.4% did not recently inject at first visit, 15.4% reported using drugs more than once a day and 16.7% used drugs between 2-6 days per week. The most common drug was a combination of cocaine and heroin: 21.8%; followed by heroin and cocaine use alone with 19.4% and 8.5%, respectively. Individuals started to inject drugs between 1962 and 1980: 12.7%; 1981 and 1990: 43.6%; and 1991 and 2002: 43.6% (table 4.1).

Clearly frequency of injection and type of drug are subject to change during the injecting career. Table 4.1 shows both values at entry and at the last follow up visit. In order to simplify the model, we consider the responses provided in the first follow up visit.

Figure 4.1 shows the Nonparametric Maximum Likelihood estimate for the survival curve (Sun; 2006). Clearly the longer the duration of injecting at first visit the longer the time to HCV infection during follow up. The figure illustrates it is important to not ignore the issue of left truncation as the NPMLE changes according to the level of duration of injecting at first visit.

Given the relatively small sample size the model selection was as follows: our initial model only includes duration of injection at first visit to select the distribution. Then, we performed simple analyses with one of the behavioural risk factors. Finally we consider a multiple model where only the risk factors with at least one significant covariate was included.

Table 4.2 shows the different parametric models with their AIC-values, favouring the generalized gamma model. Therefore, we retain this model as the best model amongst the set of candidate models.

Considering the parametric distributions as introduced above, we performed simple (single covariate) analyses with each of the behavioural risk factors (sharing syringes, frequency of injection and main drug injected) and year of first injection. For each of the models we compared the different distributions in terms of Akaike's

Table 4.1: Amsterdam Cohort Studies dataset. Descriptive statistics for IDUs who enter negative for HCV to the cohort study (n=165)

Individuals (n=165)	n	(%)	n	(%)
HCV serostatus				
Negative	107	64.85		
Positive	58	35.15		
Sharing syringes				
No	109	66.46		
Yes	55	33.54		
Year first injection				
1962- 1980	21	12.73		
1981- 1990	72	43.64		
1991- 2002	72	43.64		
Gender				
Male	109	66.06		
Female	56	33.94		
	First follow up visit		Last follow up visit	
Frequency of injection				
No recent injections	67	41.36	114	71.7
More once per day	25	15.43	10	6.29
Once daily	1	0.62	2	1.26
2-6 days per week	27	16.67	9	5.66
Once a week	3	1.85	3	1.89
2-3 days per month	10	6.17	8	5.03
One day a month	5	3.09	1	0.63
Less than one day a month	24	14.81	12	7.55
Drug of injection				
No recent injections	67	40.61	114	69.09
Heroin	32	19.39	14	8.48
Cocaine	14	8.48	5	3.03
Cocaine and heroin	36	21.82	21	12.73
Amphetamine	6	3.64	4	2.42
Methadone	5	3.03	1	0.61
Unknown drug of injection	5	3.03	6	3.64
	Mean	Std. Dev.		
Duration of injection at first visit	4.59	5.14		
Duration of injection at last visit	12.44	7.43		

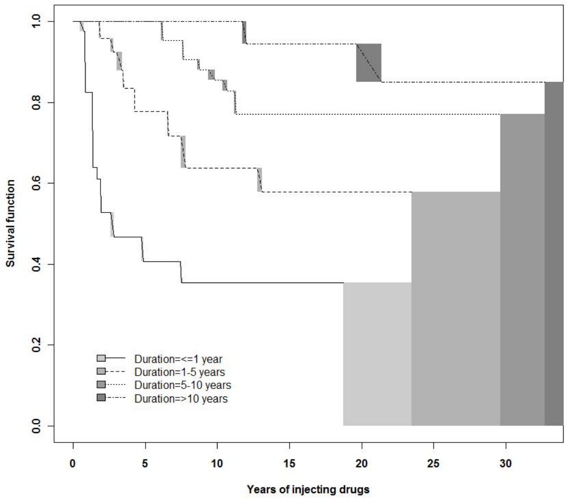


Figure 4.1: Amsterdam Cohort Studies dataset. Non-parametric maximum likelihood estimator of the survival function for time to HCV infection for different levels of duration of injection at first visit.

Information Criterion (AIC) and the likelihood ratio test (results not shown) and found the generalized gamma to be the best distribution for most of the models.

4.3.1 The effect of sharing syringes

To assess the impact of sharing syringes we take into account all responses of the individual during the follow up period, which include information on receptive sharing. Modelling the effect of sharing syringes using the accelerated failure time model was done by extending the model 2.5 including whether the IDU shared syringes ($X = 1$) or not ($X = 0$) where γ is the regression coefficient quantifying the effect of sharing syringes on HCV infection time and W is the error term. Hence, the FOI is given by:

$$\lambda_{HCV}(t, X) = \begin{cases} \exp(-\gamma)\lambda_{0HCV}(t\exp(-\gamma)) \text{ for those who share syringes,} \\ \lambda_{0HCV}(t) \text{ for those who do not share syringes.} \end{cases} \quad (4.1)$$

Under the accelerated failure model, the relation between the survival functions is as follows:

Table 4.2: Amsterdam Cohort Studies dataset. Parametric models for time to HCV infection, including only duration of injection at first visit.

Model	Log likelihood	AIC
Weibull	-235.904	477.807
Log Logistic	-229.367	464.733
Log Normal	-226.057	458.113
Generalized Gamma	-212.282	432.564

$$S_{HCV}(t|\text{sharing}) = S_{HCV}(t\exp(-\gamma)|\text{ no sharing}), \text{ for all } t, \qquad (4.2)$$

implying that the median infection time of those IDUs who share syringes ($X = 1$) is $\exp(\gamma)$ times the median infection time of those IDUs who do not share. Or equivalently, the median survival time of those IDUs not sharing syringes ($X = 0$) is $\exp(-\gamma)$ times the median survival time of those who share.

The acceleration factor, for those who share syringes compared to those who do not, equals $\exp(-\gamma) = \exp(0.22) = 1.25$ and thus the median time to HCV infection for an IDU who does not share is estimated to be 1.2 times longer than that of an IDU sharing syringes. The acceleration factor is adjusted by the duration of injection at the first visit in order to account for left truncation.

4.3.2 Frequency of injection

The frequency of injection at first follow up visit has eight categories, no recent injections, less than one day per month, one day per month, 2-3 days per month, once weekly, 2-6 days per week, once daily and more times daily. We consider a categorization based on four groups: no recent injections (0); less than one day per month, one day per month and 2-3 days per month (1); once weekly and 2-6 days per week (2); and once daily and more times daily (3). The results are shown in Table 4.3.

In this model, significant differences were found between the baseline category (no recent injections) and the remaining three categories. Moreover, we observe a trend in the estimates: when the frequency of injection increases, the acceleration factor increases. For instance the acceleration factor for an IDU who injects once a day or more is $\exp(1.16) = 3.2$, resulting in a threefold increase in median time to HCV infection for an IDU who did not inject recently compared with one who injects once a day or more. Similarly, the acceleration factor for the first and the second group are $\exp(0.88) = 2.4$ and $\exp(0.96) = 2.6$, respectively, leading to similar conclusions.

4.3.3 Drug of injection

There are 7 categories for drug of injection: no recent injections, heroin, cocaine, the combination of heroin and cocaine, amphetamine, methadone and recent IDU with unknown drug of injection. Due to the small number of individuals in the last three categories we recombined them. The results of the models are shown in Table 4.4. The baseline class is no recent injections. Clearly the acceleration factor for injecting any drug as compared to not injecting is very high. For instance, the acceleration factor for heroin (alone) is $\exp(0.78) = 2.2$ that is the median time to HCV infection for the IDUs with no recent injections is two times the median time to HCV infection of those who inject heroin. The remaining three acceleration factors are quite large too, with 3.4 for those who inject cocaine, 2.6 for the combination of heroin and cocaine and 2.9 for those who inject amphetamine, methadone or are recent IDU.

4.3.4 Time dependent force of infection

The models discussed above assume that the baseline hazard depends on the length of the injecting career of the IDU. In this section we include calendar time of the first injection as a covariate in order to investigate if the risk of IDU to be infected changed with time. We consider a categorical variable with three time categories.

$$X_i = \begin{cases} 1 \text{ if first injection between 1962-1980} \\ 2 \text{ if first injection between 1981-1990} \\ 3 \text{ if first injection between 1991-2002 and} \end{cases}$$

The hazard for this model is given by:

$$\lambda_{HCV}(t, X_j) = \exp(\gamma_j)\lambda_{0HCV}(t\exp(-\gamma_j)) \tag{4.3}$$

where γ_j is the effect of time group $j = 1$ and 2 on the hazard rate. Since the model includes a time effect it does not assume time homogeneity (i.e., the assumption that the disease is in a steady state). This is in contrast with models for current status data for which one of the model assumptions is time homogeneity.

Table 4.3: Amsterdam Cohort Studies dataset. Single covariate gamma parametric models for the different risk factors accounting for left truncation. Part I

Parameter	n	Estimate (SE)	P-value	AF	95% CI for AF	
Model with sharing syringes (AIC: 432.682)						
Intercept		0.310 (0.407)	0.4468			
Sharing syringes: yes	55	-0.220 (0.174)	0.2065	1.246	0.886	1.752
Sharing syringes: no	109	0.000		1.000		
Duration of injection at first visit		0.252 (0.019)	<0.0001	0.777	0.749	0.806
Scale		0.707 (0.339)				
Shape		-4.016 (2.493)				
Model with frequency of injection at first visit (AIC: 416.824)						
Intercept		1.573 (0.317)	<0.0001			
<1 day per month, 1 day per month, 2-3 days per month	39	-0.881 (0.271)	0.0011	2.412	1.419	4.101
Once weekly, 2-6 days per week	30	-0.957 (0.274)	0.0005	2.604	1.523	4.452
Once daily, >1 time per day	26	-1.163 (0.317)	0.0002	3.200	1.719	5.956
No recent injections	67	0.000		1.000		
Duration of injection at first visit		0.220 (0.026)	<0.0001	0.803	0.763	0.845
Scale		0.970 (0.123)				
Shape		-1.902 (0.491)				

Table 4.4: Amsterdam Cohort Studies dataset. Single covariate gamma parametric models for the different risk factors accounting for left truncation. Part II.

Parameter	n	Estimate (SE)	P-value	AF	95% CI for AF	
Model with drug injected at first visit (AIC: 426.143)						
Intercept		1.550 (0.357)	<0.0001			
Heroin	32	-0.780 (0.270)	0.0039	2.181	1.284	3.705
Cocaine	14	-1.227 (0.373)	0.0010	3.412	1.643	7.088
Heroin and cocaine	36	-0.943 (0.291)	0.0012	2.569	1.451	4.546
Other	16	-1.067 (0.338)	0.0016	2.907	1.500	5.633
No recent injections	67	0.000		1.000		
Duration of injection at first visit		0.219 (0.026)	<0.0001	0.803	0.763	0.845
Scale		0.954 (0.127)				
Shape		-1.931 (0.653)				
Model including year of first injection (AIC: 495.676)						
Intercept		-0.075 (0.203)	0.7107			
First injection 62-80	21	2.216 (0.142)	<0.0001	0.109	0.083	0.144
First injection 81-90	72	-0.089 (0.144)	0.5378	1.093	0.824	1.449
First injection 91-02	72	0.000		1.000		
Scale		0.477 (0.302)				
Shape		-11.280 (7.063)				
Model including duration of injection at first visit (AIC: 432.564)						
Intercept		0.153 (0.320)	0.6326			
Duration of injection at first visit	165	0.262 (0.016)	<0.0001	0.770	0.746	0.795
Scale		0.689 (0.293)				
Shape		-4.212 (2.241)				

SE, standard error; AF, acceleration factor; CI, confidence interval ; AIC, Akaike's Information Criterion

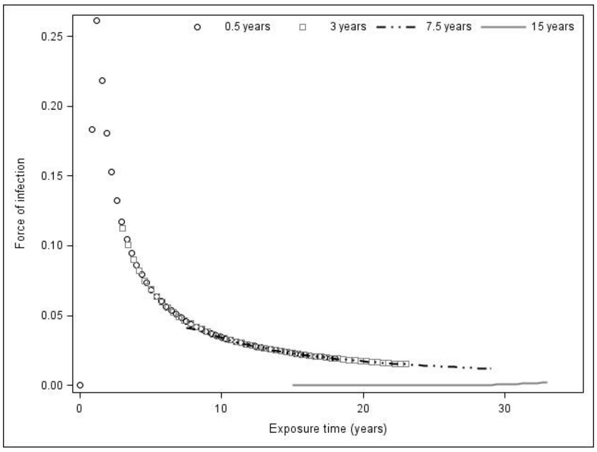


Figure 4.2: Amsterdam Cohort Studies dataset. Force of infection according to duration of injection at first visit.

The parameter estimates for the generalized gamma model are shown in Table 4.4, the reference group is 1991-2002. The acceleration factor for the IDUs with first injection before 1980 compared with IDUs who first inject in 1991-2002 equals $\exp(-2.2) = 0.11$. Hence, the median HCV infection time for the IDUs starting to inject in 1991-2002 is approximately one tenth of the median HCV infection time of the IDUs who start to inject between 1962 and 1980 and were still HCV negative at cohort entry after 1985. This variable is negatively correlated with the duration of injection at first visit and was therefore not considered in the multiple risk factor model. Note that caution should be taken when interpreting the results of this particular analysis because of the omission of the adjustment by left truncation is not explicitly taken into account; besides the calendar time is likely strongly influenced by the recruitment procedure.

Figure 4.2 shows the behaviour of the FOI according to the duration of injection at first visit. The acceleration factor is equal to 0.77, reflecting the fact that those with long exposure time before entering negative to the cohort are low risk IDUs.

4.3.5 Models including several risk factors

Finally we consider a multiple risk factor model, all the risk factors from the single covariate models were included when at least one of their categories was significant (Table 4.5). Comparing the results of the multiple risk factor model with the results of the single risk factor models, the covariates which turns out to be non-significant are sharing syringes and frequency of injection. The acceleration factor for heroin is 2.2, for cocaine 4.8, for the combination of those two 3.2 and for other drugs 3.1. Clearly, current IDUs have a higher risk than non-recent IDUs for HCV infection.

Table 4.5: Amsterdam Cohort Studies dataset. Multiple covariates generalized gamma model for all risk factors accounting for left truncation (AIC: 416.938).

Risk factor	Parameter	Estimate (SE)	P-value	AF	95% CI for AF
	Intercept	1.815 (0.329)	<0.0001		
Drug of injection	Heroin	-0.776 (0.287)	0.0069	2.172	1.237 - 3.814
at first visit	Cocaine	-1.573 (0.412)	0.0001	4.820	2.151 - 10.801
	Heroin and cocaine	-1.160 (0.306)	0.0002	3.190	1.750 - 5.813
	Other	-1.138 (0.348)	0.0011	3.120	1.577 - 6.174
	No recent injections	0.000		1.000	
Frequency of injection at first visit		0.010 (0.007)	0.1290	0.990	0.977 - 1.003
Duration of injection at first visit		0.204 (0.027)	<0.0001	0.815	0.774 - 0.859
Sharing syringes	Sharing syringes: yes	-0.331 (0.203)	0.1024	1.393	0.936 - 2.073
	Sharing syringes: no	0.000		1.000	
	Scale	0.957 (0.118)			
	Shape	-1.701 (0.471)			

SE, standard error; AF, acceleration factor; CI, confidence interval ; AIC, Akaike's Information Criterion

4.4 Concluding remarks

In our study we found a higher risk of HCV infection in the first three years of an IDU career, this is consistent with other studies (Platt et al.; 2009; Sutton et al.; 2006; Van den Berg et al.; 2007a,b). Drug of injection was associated with HCV seroconversion but sharing syringes was not. Our findings provide important additional evidence that it is crucial to target HCV prevention to new injectors as soon as they start to inject and that any efforts to reduce incidence need to take recent injectors into account. However, since it might be hard to find these recent injectors additional efforts are needed to prevent the transition to injecting drug use in non-injecting drug users.

Previous work focused on the estimation of the FOI for HCV in the IDU context based on cross-sectional data thereby relying on time homogeneity. Our study focuses on estimating the FOI based on cohort data, taking into account risk factors as well as the complexities inherent to this type of data while relaxing the time homogeneity assumption. This approach is innovative in the field and it is reassuring to conclude that previous findings can be confirmed. The Amsterdam Cohort Study is a valuable and unique source of information because it includes a follow-up of IDUs of more than 20 years, in this sense allows us to test one crucial assumption that is frequently made and untested when we analyse current status data and is the time heterogeneity. In fact, some studies have confirmed the decrease in the risk behaviour and in the prevalence and incidence of HCV (Van den Berg et al.; 2007a,b; Van de Laar et al.; 2005). Furthermore, a declining trend of injection among groups of drug users, with low or declining rates of injection have been described among opioid users in several European countries although differences between countries are large (Wiessing et al.; 2010); specifically in Amsterdam (Van den Berg et al.; 2007a,b; Van de Laar et al.; 2005; Welp et al.; 2002; Van Ameijden and Coutinho; 2001), notably the decrease in HCV seroprevalence due to impopularity of injecting among drug users and the success of prevention campaigns.

For this study we use interval-censored data methodology, which takes into account the uncertainty about the exact time to event. The nonparametric estimates show the highest risk of HCV infection in the first 3 years of injection; based on the parametric models there is an effect of frequency of injection and drug of injection.

The fact that frequency of injection and the drug of injection were significant risk factors is consistent with previous studies (Van den Berg et al.; 2007a,b; Hahn et al.; 2001; Thorpe et al.; 2000; Miller et al.; 2003a,b). It reflects the cumulative exposure to infected needles and injection paraphernalia. On the other hand, sharing syringes was not identified as a risk factor. A similar result to that observed in Van de Laar et al. (2005).

Analyses include the combined analysis of both HIV and HCV infections considering the time at risk for each of them are shown in Chapter 5; this is done using frailty models considering the bivariate type of data. The general idea is to specify latent variables which act multiplicatively over the baseline hazard, and reflect how frail an individual is to acquire the infections. The frailty could be shared when one latent variable is considered per individual or correlated when a joint latent distribution for both infections is assumed. An illustration of the use of shared frailty models on current status data for Hepatitis B and C has been reported (Sutton et al.; 2006); also for Hepatitis C and HIV infection in (Sutton et al.; 2008) and for Hepatitis A and B with correlated frailties in (Hens et al.; 2009).

In terms of study population, further research could include all IDUs participants in the ACS, like the results presented on Chapter 5. In terms of modelling, we did not take into account all the values of the time dependent covariates during the follow up, therefore more complex models can be developed; like the ones presented on Chapter 3 also a more flexible approach could use splines to incorporate the duration of injection at first visit.

Chapter 5

Correlated frailty model: an application to HIV/HCV co-infection

In survival analysis, frailty models account for correlation between (survival) times and deal with the problem of heterogeneity due to unobserved covariates. For instance, event times from individuals who have common characteristics (siblings, married couples, and so on), or times to ocurrence of different diseases within the same subject.

The frailty concept was introduced by Beard in 1959 as a longevity factor, in order to improve the modelling of the mortality. Then, 20 years later Vaupel (who introduced the frailty term) and Lancaster independently suggested random effects models for durations. Vaupel's goal was to demonstrate that population mortality hazard rates do not reproduce the mortality hazard rates of individuals from that population. Duchateau and Janssen (2008) provide a full description of Beard and Vaupel contributions. In a frailty model a random effect describes the excess of risk, or frailty of an individual or a cluster; those who are more frail experience the event of interest earlier.

Recently, a large amount of literature on the frailty model has been published. The books of Duchateau and Janssen (2008) and Wienke (2011) are valuable sources of information describing concepts and techniques to fit those models. In what follows, we describe the frailty model and mention some applications of such models in infectious disease area. Lastly, we focus on a specific type of frailty model applied to data for which the event times are known to lay in between to observation/inspection times, i.e. type II interval-censored data.

Assuming a proportional hazards model with frailties, the conditional hazard function of the event is given by:

$$\lambda(t|\boldsymbol{\omega})_{ij} \quad = \quad \lambda_0(t)\exp\left(\boldsymbol{\gamma' X_{ij}} + \boldsymbol{\omega_j}\right)$$

where $\boldsymbol{\gamma}$ is the vector of regression coefficients, $\boldsymbol{X}_{ij}$ denotes the vector of covariates for the ith subject in the jth subgroup, and $\boldsymbol{\omega_j}$ denotes the random effects, which follows a certain distribution f_Ω. This model can be rewritten as follows:

$$\lambda(t|\boldsymbol{\omega})_{ij} \quad = \quad \lambda_0(t)\exp\left(\boldsymbol{\gamma' X_{ij}}\right)\boldsymbol{Z_j} \tag{5.1}$$

where $Z_j = \exp(\omega_j)$ is called the frailty which follows a distribution f_Z. In this model all individuals from the same cluster j share the same frailty term, therefore is known as shared frailty model.

For bivariate data, Yashin et al. (1995) proposed an extension of the shared frailty considering correlated frailty, i.e., where instead of a common random effect for individuals within the cluster, two random effects are included (one for each individual or event of interest). A description of the correlated frailty model assuming a gamma frailty distribution is provided in Section 5.3.

This chapter builds on the work of Cattaert (2008) and Hens et al. (2009) considering exact event times, right censored and case II interval-censored data. First we extend the bivariate correlated gamma frailty model describing the individual contribution to the likelihood for case II interval-censored data. Then, we apply several frailty models to the Amsterdam Cohort Studies dataset, previously used in Chapter 4. Subsequently, we describe the procedure to generate the data according to different baseline hazard functions for the simulation studies.

The objectives of the simulation studies are: i) to assess model behaviour of a correlated frailty model in presence of type II interval-censored data and ii) to assess the impact of different frailty variances on a correlated gamma frailty model. In the latter case we know that the estimation procedure becomes challenging due to the restriction imposed on the correlation between the frailties. Finally, we present the results of two simulation studies.

In this chapter we use gamma frailty models for its appealing mathematical properties, although results can be generalized to other frailty distributions. First, we present the univariate gamma frailty model, the restrictions to make the model identifiable, as well as the conditional and the unconditional survival functions.

5.1 Frailty models for current status data

In the infectious disease context, the book by Hens et al. (2012) provides an overview about individual heterogeneity and how to account for that in modeling bivariate serological data. In this section we mention some of the advances in the area, however the list is far from being exhaustive.

Farrington et al. (2001) proposed the use of a shared frailty to model heterogeneity in the acquisition of measles, mumps and rubella. The main objective of the paper was to describe the force of infection for the diseases at different ages. This was a starting point for further applications in the infectious disease area. This model can also be used to estimate the basic reproduction number R_0, defined as the average number of secondary infections generated by an infective individual in a completely susceptible population. The authors note that the individual heterogeneity inflates R_0 with a factor $1 + var(\sigma^2)$ where σ^2 is the variance of the shared frailty. An extension of this model, accounting for time dependency has been proposed by Unkel and Farrington (2012). Recently Farrington et al. (2013) proposes new methods for investigating the extend of heterogeneity in effective contact rates. The authors apply a Dirichlet-multinomial model with an additional overdispersion parameter.

In the context of injecting drug users, Sutton et al. (2006) modelled the force of infection for hepatitis B (HBV) and C (HCV) on injecting drug users, applying a shared frailty model. Later, Sutton et al. (2008) applies a similar model accounting for infection with HIV, HBV and HCV. Del Fava et al. (2011) applied a shared gamma frailty model to HCV and HIV coinfection.

Cattaert (2008) applied several frailty models to seroprevalence data for mumps, rubella, parvo b19 and varicella infection data. Cattaert (2008) and Hens et al. (2009) studied the behaviour of the bivariate-correlated gamma frailty model for case I interval-censored data (current status data) and compared the correlated with the shared frailty model using cross-sectional data on hepatitis A and B. Abrams and Hens (2014) extend some frailty models to account for waning immunity.

Some applications are unrelated to infectious diseases. However, it is worth to mention them since they were applied to current status data. Zhang et al. (2005) presented an additive hazards frailty model accounting for informative observation time, that is the observation time may be related with the underlying survival time. Dunson and Dinse (2002) proposed Bayesian models to analyse current status data with informative censoring. Later on, Chen et al. (2009) described the use of a marginal frailty model on multivariate current status data and applied the model to data coming from a tumorigenicity experiment. A cure model, a special case of a discrete frailty model, where a proportion of individuals are not susceptible to the event of interest, was applied to current status data in Ma (2009).

5.2 Univariate gamma frailty model

In survival analysis one may be interested to assess the impact of several covariates on the event times. In this case, the observed covariates can be included in a proportional hazards model. However, it is impossible to account for all important risk factors that may influence the event times. For instance, some of the factors can be difficult to collect due to financial or time constrains. Then, it is useful to consider two sources of variability: one attributed to the observed risk factors in the model and another caused by unknown covariates. These non-observable risk factors may be described by the frailty in the survival analysis context. Assuming a proportional hazards model conditional upon the random effects (frailty) and a multiplicative effect of the frailties, then the frailty acts multiplicatively on the baseline hazard function as it has been indicated in equation (5.1).

Here, the Laplace transform of the frailty distribution, denoted by $L(u)$ plays a key role to characterize the unconditional density function as well as the frailty distribution. That is, knowing the form of the Laplace transform, we can derive the unconditional survival and the density functions, as well as the mean and the variance of the frailty distribution (Wienke; 2011). These expressions hold for all the models regardless of the frailty distribution.

$$\begin{aligned} S(t) &= E\left[S(t|z)\right] = E\left[\exp(-z\Lambda(t))\right] = L(\Lambda(t)), \\ f(t) &= -\lambda_0(t)L'(\Lambda(t)), \end{aligned} \tag{5.2}$$

where $\Lambda(t)$ is the cumulative baseline hazard function and Z is the frailty term.

Thanks to its mathematical tractablity and its flexibility, the gamma distribution has been widely applied in this context. If the random effects are assumed to follow a gamma distribution, identifiability can be ensured by restricting the parameters of the gamma to be equal, i.e. $Z \sim \Gamma(1/\sigma^2, \sigma^2)$, implying $E(Z) = 1$. $k = 1/\sigma^2$ denotes the shape parameter. Then, the conditional survival function is given by:

$$S(t|z) = e^{-z\Lambda(t)}. \tag{5.3}$$

The unconditional survival function is given by:

$$S(t) = \frac{1}{\left(1+\sigma^2\Lambda(t)\right)^{1/\sigma^2}}. \tag{5.4}$$

Estimation of standard errors

We estimate the standard errors for the baseline hazard parameters and parameter of the shape parameter of the gamma (k) using maximum likelihood. This was possible because Matlab codes used in this chapter also provide the hessian matrix. The estimation of the standard error for σ is based on the delta method as described below.

In the univariate gamma frailty model $\sigma^2 = 1/k$ then $\sigma = k^{-1/2}$. Taking the derivative of σ respect to k: $\frac{d\sigma}{dk} = -\frac{1}{2}k^{-3/2}$. Then, the variance of σ using the delta method is given by the following expression:

$$Var(\sigma) \quad = \quad 1/4k^{-3}Var(k)$$

A natural extension of the univariate gamma frailty model would be a multivariate model where individuals (or events for the same individual) share the same frailty term. This has been described at the begining of this chapter.

5.3 Correlated gamma frailty model

This model is an extension of the univariate gamma frailty model and it has been introduced by Yashin et al. (1995). In what follows, we describe the model as presented in Cattaert (2008) and we include the likelihood contributions when interval-censored data are considered.

We start from three independent gamma distributed random variables $Y_l \sim \Gamma(k_l, 1)$ for $l = 0, 1, 2$, with shape k_l and scale 1. Here Y_0 represents shared risk factors while Y_1 and Y_2 represent non-shared risk factors. The frailty variables Z_i, with $i = 1, 2$, are then composed as:

$$Z_i = \sigma_i^2(Y_0 + Y_i). \tag{5.5}$$

From the properties of the gamma distribution it then follows that $Z_i \sim \Gamma(k_0 + k_i, \sigma_i^2)$. In order to make the frailties identifiable, it is assumed that $E(Z_i) = 1$, implying that $\theta_i \equiv 1/\sigma_i^2 = k_0 + k_i$ and hence $Z_i \sim \Gamma(1/\sigma_i^2, \sigma_i^2)$. This then leads to variances $\text{Var}(Z_i) = \sigma_i^2$, and correlation

$$\rho \equiv \text{Corr}(Z_1, Z_2) = \sigma_1\sigma_2 k_0. \tag{5.6}$$

The restrictions $k_i > 0$ imply a constraint on the correlation, i.e.

$$0 < \rho < \min\left(\frac{\sigma_1}{\sigma_2}, \frac{\sigma_2}{\sigma_1}\right) \leq 1. \tag{5.7}$$

Note that also $\rho = 0$ is possible, but then the corresponding gamma distribution is degenerate, i.e. for $k_0 = 0$ all the probability mass is concentrated at $Y_0 = 0$.

Assuming proportional hazards (i.e. the frailties act multiplicatively on the baseline hazards) and furthermore conditional independence (i.e. conditional on the frailties Z_i the event times T_i are independent), the conditional survival function is given by:

$$S(t_1, t_2|z_1, z_2) = e^{-z_1\Lambda_1(t_1)}e^{-z_2\Lambda_2(t_2)}, \tag{5.8}$$

where $\Lambda_i(t)$ are the cumulative baseline hazard functions, to be calculated from the baseline hazard rates $\lambda_i(t)$ by integration, i.e. $\Lambda_i(t) = \int_0^t \lambda_i(s)ds$.

The unconditional bivariate survival function is found as a Laplace transform. From the Laplace transform $L(u) \equiv E[\exp(-uY_i)] = (1+u)^{-k_i}$ of Y_i, the Laplace transform of (Z_1, Z_2) is seen to be

$$\begin{aligned} L(u_1, u_2) &\equiv E\left[e^{-(u_1Z_1+u_2Z_2)}\right] \\ &= (1+(k_0+k_1)^{-1}u_1 + (k_0+k_2)^{-1}u_2)^{-k_0}(1+(k_0+k_1)^{-1}u_1)^{-k_1} \\ &\quad (1+(k_0+k_2)^{-1}u_2)^{-k_2}. \end{aligned}$$

This is then used to calculate the unconditional bivariate survival function

$$\begin{aligned} S(t_1, t_2) &= E\left(e^{-Z_1\Lambda_1(t_1)-Z_2\Lambda_2(t_2)}\right) \\ &= \left[1+(k_0+k_1)^{-1}\Lambda_1(t_1)\right]^{-k_1}\left[1+(k_0+k_2)^{-1}\Lambda_2(t_2)\right]^{-k_2} \\ &\quad \left[1+(k_0+k_1)^{-1}\Lambda_1(t_1)+(k_0+k_2)^{-1}\Lambda_2(t_2)\right]^{-k_0} \\ &= [S_1(t_1)]^{1-\frac{k_0}{k_0+k_1}}[S_2(t_2)]^{1-\frac{k_0}{k_0+k_2}}\left[[S_1(t_1)]^{k_0+k_1}+[S_2(t_2)]^{k_0+k_2}-1\right]^{-k_0}, \end{aligned}$$

where $S_i(t)$ are the marginal survival functions given by:

$$S_i(t) = \left[1+(k_0+k_i)^{-1}\Lambda_i(t)\right]^{-(k_0+k_i)}.$$

To estimate the model parameters of the gamma frailty model we use maximum likelihood. As it was mentioned in Chapter 2, the contribution to the likelihood depends on the information provided by each individual. Since we are considering two outcomes the contributions for both should be taken into account. If an individual experiences an event, he/she provides information on the probability that the event occurs (density function). The right censored observations provide information regarding the survival function evaluated at the censoring time. Instead for the interval-censored observation we consider the difference between the lower and the upper limit of the survival function. In the following section we describe the contributions considering bivariate outcomes.

5.3.1 Log-likelihood contributions

For uncensored time to event data in each of the two outcomes the log-likelihood is given by the bivariate density function:

$$\ell = \sum_{j=1}^{n} \ln f(t_{1j}, t_{2j}), \tag{5.9}$$

where

$$\begin{aligned} f(t_1, t_2) &= \frac{\partial^2}{\partial t_1 \partial t_2} S(t_1, t_2) \\ &= S(t_1, t_2)[C_1(t_1, t_2) + C_2(t_1, t_2) + C_3(t_1, t_2)](k_0 + k_1)^{-1}(k_0 + k_2)^{-1}\lambda_1(t_1)\lambda_2(t_2), \end{aligned}$$

with

$$\begin{aligned} C_1(t_1, t_2) &= \frac{k_1 k_2}{[1 + (k_0 + k_1)^{-1}\Lambda_1(t_1)]\,[1 + (k_0 + k_2)^{-1}\Lambda_2(t_2)]}, \\ C_2(t_1, t_2) &= \frac{k_0}{1 + (k_0 + k_1)^{-1}\Lambda_1(t_1) + (k_0 + k_2)^{-1}\Lambda_2(t_2)} \\ &\quad \left[\frac{k_1}{1 + (k_0 + k_1)^{-1}\Lambda_1(t_1)} + \frac{k_2}{1 + (k_0 + k_2)^{-1}\Lambda_2(t_2)}\right], \\ C_3(t_1, t_2) &= \frac{k_0(k_0 + 1)}{[1 + (k_0 + k_1)^{-2}\Lambda_1(t_1) + (k_0 + k_2)^{-1}\Lambda_2(t_2)]^2}. \end{aligned}$$

In case at least one of the outcomes is right censored the log-likelihood contribution for the censored observation corresponds to the survival function and for the uncensored time to events to the density function:

$$\ell = \sum_{j=1}^{n} \sum_{i,k=0}^{1} \delta_{\delta_{1j},i}\delta_{\delta_{2j},k} \ln A_{ik}(x_{1j}, x_{2j}), \tag{5.10}$$

where A_{00} corresponds to the case when both event times are right censored; A_{10} and A_{01} represent the cases when one of the outcomes is right censored; $A_{11} = f(t_1, t_2)$ represents the case when both outcomes are uncensored.

$$\begin{aligned}
A_{00}(x_1,x_2) &= S(t_1,t_2), \\
A_{10}(x_1,x_2) &= -\frac{\partial}{\partial t_1}S(t_1,t) = S(t_1,t)[C_4(t_1)+C_6(t_1,t)](k_0+k_1)^{-1}\lambda_1(t_1), \\
A_{01}(x_1,x_2) &= -\frac{\partial}{\partial t_2}S(t,t_2) = S(t,t_2)[C_5(t_2)+C_6(t,t_2)](k_0+k_2)^{-1}\lambda_2(t_2), \\
A_{11}(x_1,x_2) &= \frac{\partial^2}{\partial t_1 \partial t_2}S(t_1,t_2), \qquad (5.11)
\end{aligned}$$

with

$$\begin{aligned}
C_4(t_1) &= \frac{k_1}{1+(k_0+k_1)^{-1}\Lambda_1(t_1)}, \\
C_5(t_2) &= \frac{k_2}{1+(k_0+k_2)^{-1}\Lambda_2(t_2)}, \\
C_6(t_1,t_2) &= \frac{k_0}{1+(k_0+k_1)^{-1}\Lambda_1(t_1)+(k_0+k_2)^{-1}\Lambda_2(t_2)} \qquad (5.12)
\end{aligned}$$

For case II: interval-censored data the log-likelihood is given by:

$$\ell = \sum_{j=1}^{n}\sum_{i,k=0}^{2}\delta_{\delta_{1j},i}\delta_{\delta_{2j},k}\ln A_{ik}(x_{1j},x_{2j}), \qquad (5.13)$$

where A_{00}, A_{01}, A_{10} and A_{11} are defined as in (5.11) and with A_{20} and A_{02} corresponding to the case when one of the events is right censored and the other is interval-censored; A_{21} and A_{12} are the contributions when one observation is interval-censored and the other is an uncensored observation; and A_{22} indicates both times are interval-censored.

$$\begin{aligned}
A_{20}(x_1,x_2) &= S(t_{1L},t_2) - S(t_{1R},t_2), \\
A_{02}(x_1,x_2) &= S(t_1,t_{2L}) - S(t_1,t_{2R}), \\
A_{12}(x_1,x_2) &= -\frac{\partial}{\partial t_1}[S(t_1,t_{2L}) - S(t_1,t_{2R})] \\
&= \{S(t_1,t_{2L})[C_4(t_1)+C_6(t_1,t_{2L})] - S(t_1,t_{2R})[C_4(t_1)+C_6(t_1,t_{2R})]\} \\
&\quad (k_0+k_1)^{-1}\lambda_1(t_1),
\end{aligned}$$

$$\begin{aligned}
A_{21}(x_1, x_2) &= -\frac{\partial}{\partial t_2}[S(t_{1L}, t_2) - S(t_{1R}, t_2)] \\
&= \{S(t_{1L}, t_2)[C_5(t_2) + C_6(t_{1L}, t_2)] - S(t_{1R}, t_2)[C_5(t_2) + C_6(t_{1R}, t_2)]\} \\
&\quad (k_0 + k_2)^{-1}\lambda_2(t_2), \\
A_{22}(x_1, x_2) &= S(t_{1R}, t_{2R}) - S(t_{1R}, t_{2L}) - S(t_{1L}, t_{2R}) + S(t_{1L}, t_{2L}) \qquad (5.14)
\end{aligned}$$

with $C_4(t_1)$ and $C_5(t_2)$ as in (5.12) and:

$$\begin{aligned}
C_6(t_{1L}, t_2) &= \frac{k_0}{1 + (k_0 + k_1)^{-1}\Lambda_1(t_{1L}) + (k_0 + k_2)^{-1}\Lambda_2(t_2)}, \\
C_6(t_{1R}, t_2) &= \frac{k_0}{1 + (k_0 + k_1)^{-1}\Lambda_1(t_{1R}) + (k_0 + k_2)^{-1}\Lambda_2(t_2)}, \\
C_6(t_1, t_{2L}) &= \frac{k_0}{1 + (k_0 + k_1)^{-1}\Lambda_1(t_1) + (k_0 + k_2)^{-1}\Lambda_2(t_{2L})}, \\
C_6(t_1, t_{2R}) &= \frac{k_0}{1 + (k_0 + k_1)^{-1}\Lambda_1(t_{1R}) + (k_0 + k_2)^{-1}\Lambda_2(t_{2R})}. \qquad (5.15)
\end{aligned}$$

We work with the parameters k_i rather than σ_i and ρ because the restriction (5.6) is more naturally stated in terms of these. Positivity of the shape parameters k_i is ensured by logarithmic transformation. In our application we use two baseline hazard distributions, namely, a Weibull and a Gompertz, using parameterization described in Table 5.1. The parameters of both distributions are denoted by α_i and β_i and we use logarithmic transformation to ensure positivity of the baseline hazard parameters.

5.3.2 Estimation of standard errors

It is possible to estimate the standard errors for the baseline hazard parameters, and for k_i using maximum likelihood thanks to the availability of the hessian matrix.

The estimation of the standard errors for σ_1, σ_2 and ρ are based on the delta method as described below.

In the correlated frailty model $\sigma_1^2 = 1/(k_0 + k_1)$, then $\sigma_1 = 1/\sqrt{k_0 + k_1}$. The partial derivatives of σ_1 respect to k_0, k_1 and k_2 are given by the following expressions:

$$\begin{aligned}
\frac{\partial \sigma_1}{\partial k_0} &= -\frac{1}{2(k_0+k_1)^{3/2}} \\
\frac{\partial \sigma_1}{\partial k_1} &= -\frac{1}{2(k_0+k_1)^{3/2}} \\
\frac{\partial \sigma_1}{\partial k_2} &= 0
\end{aligned}$$

On the other hand, $\sigma_2^2 = 1/(k_0+k_2)$, then $\sigma_2 = 1/\sqrt{k_0+k_2}$. The partial derivatives of σ_1 respect to k_0, k_1 and k_2 are given by the following expressions:

$$\begin{aligned}
\frac{\partial \sigma_2}{\partial k_0} &= -\frac{1}{2(k_0+k_2)^{3/2}} \\
\frac{\partial \sigma_2}{\partial k_1} &= 0 \\
\frac{\partial \sigma_2}{\partial k_2} &= -\frac{1}{2(k_0+k_2)^{3/2}}
\end{aligned}$$

Finally, $\rho = \sigma_1\sigma_2 k_0 = \frac{k_0}{[(k_0+k_1)(k_0+k_2)]^{1/2}}$. The partial derivatives of ρ respect to k_0, k_1 and k_2 are given by the following expressions:

$$\begin{aligned}
\frac{\partial \rho}{\partial k_0} &= \frac{1}{[(k_0+k_1)(k_0+k_2)]^{1/2}} - \frac{k_0\,[2k_0+k_1+k_2]}{2\,[(k_0+k_1)(k_0+k_2)]^{3/2}} \\
\frac{\partial \rho}{\partial k_1} &= -\frac{k_0}{2(k_0+k_1)^{3/2}(k_0+k_2)^{1/2}} \\
\frac{\partial \rho}{\partial k_2} &= -\frac{k_0}{2(k_0+k_1)^{1/2}(k_0+k_2)^{3/2}}
\end{aligned}$$

Following the definition of the delta method, the variance of σ_1, σ_2 and ρ are given by:

$$\begin{aligned}
Var(\sigma_1) &= \nabla\sigma_1' Cov(k_0,k_1,k_2)\nabla\sigma_1 \\
Var(\sigma_2) &= \nabla\sigma_2' Cov(k_0,k_1,k_2)\nabla\sigma_2 \\
Var(\rho) &= \nabla\rho' Cov(k_0,k_1,k_2)\nabla\rho \qquad (5.16)
\end{aligned}$$

where $\nabla\sigma_i$ is the gradient vector of σ_i respect to k_0, k_1 and k_2, $\nabla\rho$ is the gradient vector of ρ respect to k_0, k_1 and k_2, while $Cov(k_0, k_1, k_2)$ is the variance-covariance matrix for k_0, k_1, and k_2.

5.4 Application to the Amsterdam Cohort Studies dataset

As in Chapter 4, we use the Amsterdam Cohort Studies dataset as an illustrative example. In this chapter, we consider both time to HCV and HIV infection as outcomes and all 935 individuals are included, whereas in Chapter 4 we focussed only on time to HCV infection for those individuals who entered negative for HCV to the cohort study. In this section, we fit univariate, shared, and correlated frailty models to the data using code in Matlab (an initial version of the code was provided by Cattaert 2008). For each of the models we consider two baseline hazards distribution, namely, a Weibull and a Gompertz, using parameterization described in Table 5.1. The t_i values are the ones we used for the simulation studies as described in Section 5.5.1.

Table 5.1: Baseline hazard functions and their corresponding expressions for time to event data

Distribution	Hazard function $\lambda_i(t)$	Cumulative baseline hazard $\Lambda_i(t)$	t_i values
Weibull $\alpha_i, \beta_i > 0, t \geq 0$	$\alpha_i \beta_i t^{\beta_i - 1}$	$\alpha_i t^{\beta_i}$	$(\Lambda_i / \alpha_i)^{1/\beta_i}$
Gompertz $\alpha_i, \beta_i > 0, t \geq 0$	$\alpha_i \exp(\beta_i t)$	$\frac{\alpha_i}{\beta_i} [\exp(\beta_i t) - 1]$	$\frac{\ln(1+\beta_i \Lambda_i / \alpha_i)}{\beta_i}$

The standard errors were calculated using a maximum likelihood or delta method as described in previous sections; the model fitting requires to provide initial values to the parameter estimates. We tried several initial values to gain confidence that the global maximum has been found. The estimates and the standard errors are shown in Table 5.2.

Univariate frailty models for time to HCV and HIV infection are fitted separately, so we obtain one estimate of the frailty variance per outcome. The shared frailty model assumes a common variance component for both outcomes, while in the correlated frailty model the frailties of outcomes in a cluster are correlated but not necessarily shared. This aspect allows to consider questions about the association between time to events.

The first two columns show the results of applying the univariate frailty model, in this case a model for each outcome is fitted separately allowing one variance of the frailty term per outcome. The individual likelihood values for each model are added up and included in the ninth row in columns 2 and 3.

The log-likelihood value of the univariate frailty model for time to HCV assuming a Weibull baseline hazard is equal to -703.1853 (-703.1870 when we assume Gompertz baseline hazard), whereas the log-likelihood value of the univariate frailty model for time to HIV assuming a Weibull baseline hazard is equal to -1011.6769 (-1011.6322 when we assume Gompertz baseline hazard).

The parameter estimates for the frailty variances indicate a pretty large difference between both outcomes (one is approximately twice the other one). This large difference between the variances is ignored in the shared frailty models, where a common variance is assumed. The results of the shared frailty models are shown in columns 4 and 5.

Finally, with the correlated model (results are shown in columns 6 and 7) we obtain the highest log-likelihood values. The estimates are close to the ones of the univariate frailty model. Assuming a Weibull baseline hazard shows the best fit to the data according to the AIC values. Additionally, one of the frailty variances is 2.5 times the other one and the estimated frailty correlation is 0.42. Note that the correlation is on the boundary of the parameter space, because under the correlated frailty model the correlation is restricted to the smallest ratio of the frailty standard deviations.

Here, the frailty variances measure the population heterogeneity in the susceptibility to HCV and HIV. High values of the parameter ρ indicate similar transmission routes: for IDUs these are sharing syringes and sexual intercourse.

Table 5.2: Amsterdam Cohort Studies dataset. Estimates and standard errors for the baseline hazard and the frailty parameters, assuming univariate frailty, shared frailty, and correlated frailty with Weibull and Gompertz baseline hazard

model	Independent frailty		Shared frailty		Correlated frailty	
baseline	Weibull	Gompertz	Weibull	Gompertz	Weibull	Gompertz
parameter	estimate (SE)	estimate (SE)	estimate (SE)	estimate (SE)	estimate (SE)	estimate (SE)
α_1	1.1116 (0.2813)	1.8107 (0.5350)	1.0007 (0.1124)	1.0149 (0.1535)	1.2153 (0.2159)	1.4326 (0.3250)
β_1	0.6988 (0.2219)	0.0469 (0.0456)	0.5944 (0.0605)	2.8E-07 (7.0E-06)	0.8058 (0.1353)	0.0154 (0.0354)
σ_1	0.9328 (0.2951)	1.3753 (0.1318)			1.0647 (0.1425)	1.2725 (0.1067)
α_2	0.1178 (0.0252)	0.1262 (0.0413)	0.1245 (0.0172)	0.0394 (0.0031)	0.1157 (0.0275)	0.2006 (0.0685)
β_2	0.9459 (0.3592)	0.0142 (0.0435)	0.5010 (0.0472)	2.6E-05 (2.0E-04)	1.1045 (0.2889)	0.0788 (0.0436)
σ_2	2.227 (0.7209)	2.4686 (0.4470)			2.5224 (0.4866)	3.0984 (0.3557)
σ			0.7825 (0.0776)	1.1582 (0.0553)		
ρ					0.4221 (0.0447)	0.4107 (0.0294)
log-likelihood	-1714.8622	-1714.8192	-1697.2228	-1739.5845	-1692.9543	-1693.6627
AIC	3441.7243	3441.6384	3404.4455	3489.1689	3399.9086	3401.3254

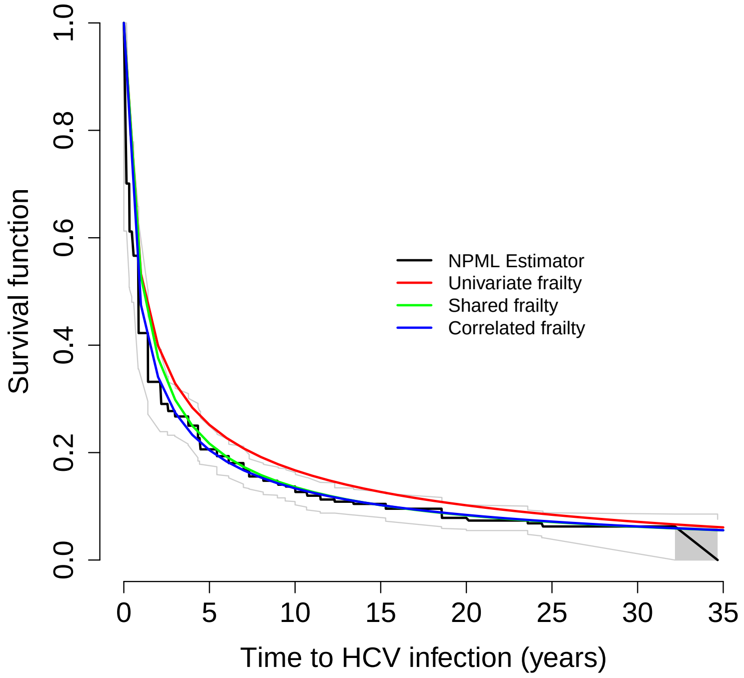


Figure 5.1: Comparison between the NPMLE of the survival function and the fitted frailty models based on a Gompertz baseline hazard for time to HCV infection

We also compare the unconditional survival functions under the different frailty models with the Non-Parametric Maximum Likelihood Estimates for both outcomes. The comparison is shown in Figures 5.1 to 5.4. The figures also include the confidence bands from the Nonparametric Maximum Likelihood Estimations.

Using the Gompertz baseline hazard we observe that the results for the univariate frailty models are better than the results for the shared frailty. Additionally, there is a big improvement under the correlated frailty model compared with the other two models. In fact, Figure 5.1 shows that the unconditional survival function for time to HCV infection under the univariate frailty model does not properly reflect the NPMLE.

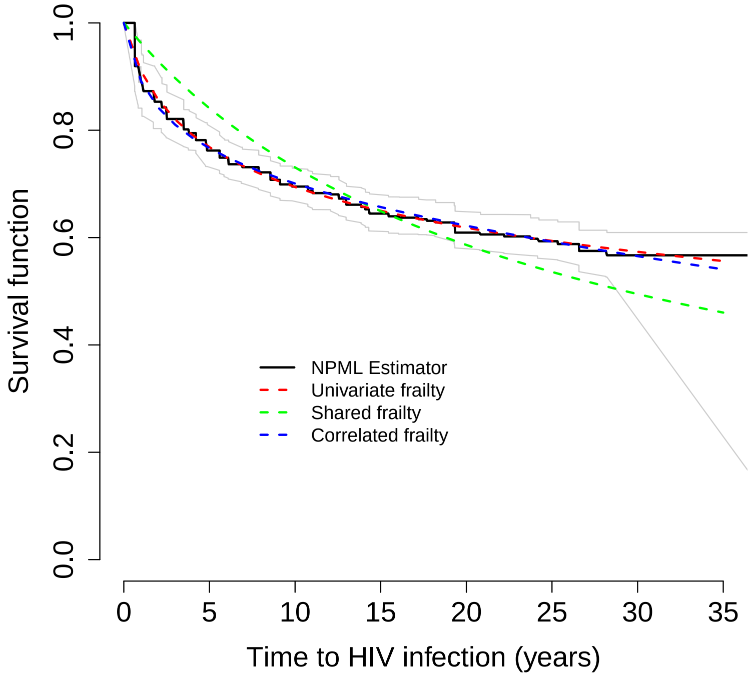


Figure 5.2: Comparison between the NPMLE of the survival function and the fitted frailty models based on Gompertz baseline hazard for time to HIV infection

In Figure 5.2 it becomes clear that the poor goodness of fit results for the shared frailty model are attributed to the time to HIV infection, where the shared frailty model deviates from the NPMLE of the unconditional survival function. Since we may have detected a local maxima based on the starting values we choose for the optimisation. We try to improve the fitted model selecting several sets of starting values but we did not succeed in finding a better solution for this model.

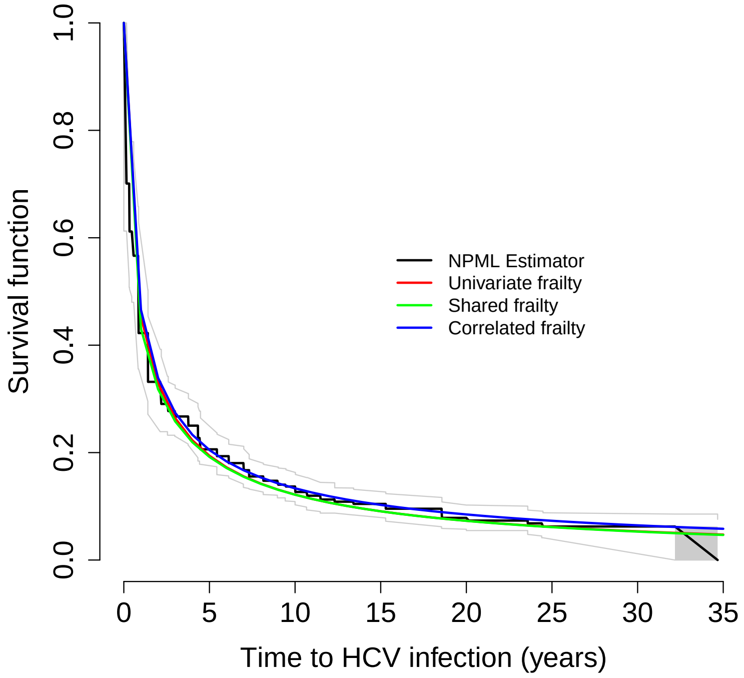


Figure 5.3: Comparison between the NPMLE of the survival function and the fitted frailty models based on Weibull baseline hazard for time to HCV infection

Under the Weibull baseline hazard, for time to HCV infection the unconditional survival functions under the univariate and the correlated frailty models are basically overlapping (Figure 5.3). Instead, for time to HIV infection we observe a small difference between both models (Figure 5.4. For both times to HCV and HIV infection the shared frailty model shows a slightly different fit than the other two models.

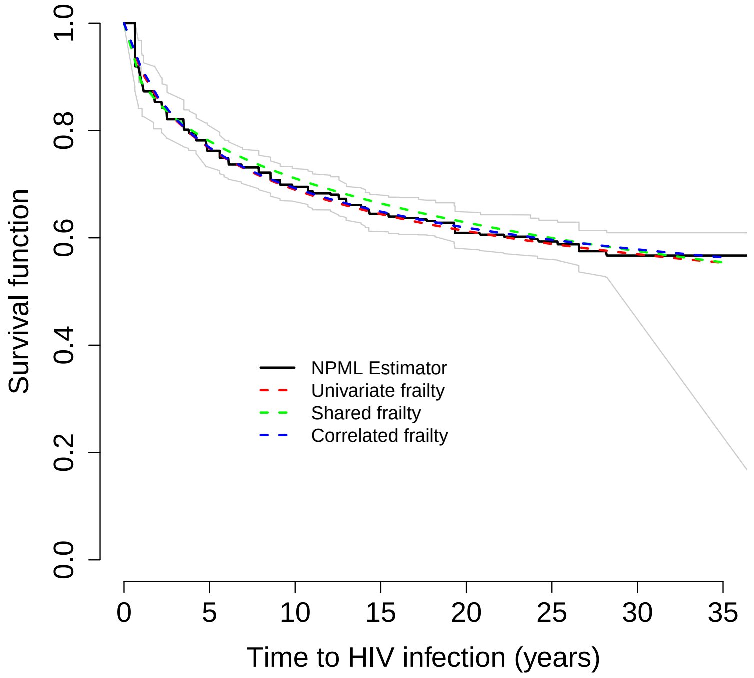


Figure 5.4: Comparison between the NPMLE of the survival function and the fitted frailty models based on a Weibull baseline hazard for time to HIV infection

The large difference between the frailty variances for both outcomes inspires the simulation study presented in the next section. The simulation study has two objectives: first, we want to study the behaviour of the correlated frailty model in presence of interval-censored data; secondly we investigate the effect of the large difference between the frailty variances on the estimation procedure.

5.5 Simulation study based on a correlated frailty model

In the previous section we applied several frailty models to the Amsterdam Cohort Studies dataset. We compared the parameter estimates and assessed the goodness of fit using the log-likehood values, the AIC and a graphical comparison of the unconditional survival function with the NPMLE for both time to events. The results favoured the correlated frailty model with a Weibull baseline hazard.

To get further insights about the correlated frailty model and its properties in presence of interval-censored data, we perform a simulation study with a twofold objective: to assess model behaviour of a correlated frailty model in presence of type II interval-censored data and to assess the impact of different frailty variances on a correlated gamma frailty model. In the latter case we know that the estimation procedure becomes challenging due to the restriction imposed on the correlation between the frailties.

In this section, we first describe how to generate data from a correlated gamma frailty model, then we present the results of the simulations.

5.5.1 Generating data from a correlated gamma frailty model

First z_1 and z_2 are generated from a correlated gamma frailty as follows: for fixed values of σ_1, σ_2 and ρ that satisfy the restriction (5.17), we calculate the corresponding values of k_0, k_1 and k_2.

$$0 < \rho < \min\left(\frac{\sigma_1}{\sigma_2}, \frac{\sigma_2}{\sigma_1}\right) \leq 1. \tag{5.17}$$

Once those values have been calculated we generate the values y_0, y_1 and y_2 from gamma distribution functions $\Gamma(k_0, 1)$, $\Gamma(k_1, 1)$ and $\Gamma(k_2, 1)$ respectively. Next, we calculate $z_1 = \sigma_1^2(y_0 + y_1)$ and $z_2 = \sigma_2^2(y_0 + y_2)$ using the values of y_i and k_i. Conditional on the frailties (z_1 and z_2) we generate the exact time to event using the baseline hazard function for each particular distribution. Table 5.1 includes the distributions considered in this chapter.

The procedure to generate the exact time to event is as follows: we generate uniform random numbers u_i in the unit interval, then we calculate the cumulative distribution function $\Lambda_i = -\ln(u_i/z_i)$, finally using the expressions given in the last column of Table 5.1 we obtain the time to event. The time to event data are also reduced to censored data, i.e. data that are either right censored, uncensored or interval-censored data (case II).

We assume that the censoring times are uniformly distributed on $[0, 40]$, independent of the event times. We thus generate $t \sim \text{Un}[0, 40]$, and define the indicator variable $\delta_i = I(t_i < t)$. Hence $\delta_i = 0$ if $t_i > t$, indicating right censoring, and $\delta_i = 1$ if $t_i < t$, indicating either no censoring or left censoring, depending on the context. Now letting j indicate observations, for uncensored time to event data we observe the event times T_{1j} and T_{2j}. For right censored data we observe $X_{1j} = \min(T_{1j}, T_j)$ and $X_{2j} = \min(T_{2j}, T_j)$, where T_j are the measurement times, and the indicators $\Delta_{1j} = I(t_{1j} < t_j)$ and $\Delta_{2i} = I(t_{2j} < t_j)$.

For the interval-censored data, let T_{ij} denote the failure time random variable following the distribution $F(t_{ij})$. We follow one of the approaches described in Gómez et al. (2009), generating censoring intervals (L_{ij}, R_{ij}) from $F_{L_{ij},R_{ij}}(l_{ij}, r_{ij})$ such that the censoring occurs non informatively.

A proportion of the time to event data are assumed to be interval-censored. For each time to event T_{ij} we generate values from two independent uniform distributions in the interval $(0, c)$: U_{ij}^1 and U_{ij}^2. Then the limits of the interval-censored observations for the ith variable are given by: $L_{ij} = \max(T_{ij} - U_{ij}^1, T_{ij} + U_{ij}^2 - c)$ and $R_{ij} = \min(T_{ij} + U_{ij}^2, T_{ij} - U_{ij}^1 + c)$. We use $c = 1$.

In the first simulation study, around 20% of the observations were exact times whereas the percentage of interval-censored observations ranges from 30% to 50%. In the second simulation study, the proportion of interval-censored data was based on the ACS data. For the first outcome around 65% of the observations are uncensored (exact time to event is observed), and the remaining portion is right censored. However, by construction of the simulation the proportion of interval-censored data decreases when the frailty variance decreases. Then, for the second outcome about half of the observations are right censored and this proportion remains almost constant since there is no change on the frailty variance.

From each dataset we estimate the parameters as well as the standard errors (SE) as was described in subsection 5.3.2. The average of the estimations over the 500 generated samples (mean), its estimated standard deviation or Empirical Standard Error (ESE) , the average of the standard errors, the bias, the variance (based on the Empirical Standard Errors) and the Mean Squared Error (MSE) are computed as:

$$\begin{aligned}
\text{mean} &= \bar{\hat{\theta}} = \sum_{i=1}^{500} \frac{\hat{\theta}_i}{500} \\
\text{ESE} &= SD\left(\bar{\hat{\theta}}\right) = \sqrt{\sum_{i=1}^{500} \frac{\left(\hat{\theta}_i - \bar{\hat{\theta}}\right)^2}{499}}, \\
\text{bias} &= \bar{\hat{\theta}} - \theta, \\
\text{MSE} &= MSE\left(\hat{\theta}\right) = \text{bias}^2\left(\bar{\hat{\theta}}\right) + SD^2\left(\bar{\hat{\theta}}\right), \text{and} \\
\text{mean(SE)} &= \sum_{i=1}^{500} \frac{SE\left(\hat{\theta}_i\right)}{500}
\end{aligned}$$

The ESE provides information about the variability of the results between samples, whereas the SE provides information about the variability of the results within a sample. The average of the estimations (mean) are used to calculated the bias, whereas the variance corresponds to the square of the Empirical Standard Error.

For each of the settings 500 samples were generated. In some cases a solution could not be reached therefore we did not have estimates for the parameters or the hessian matrix was not positive definite. These cases were not included in the summary statistics and we defined the success rate in the estimation procedure as the percentage of samples that produce a solution and the hessian matrix is positive definite over the total number of samples.

5.5.2 Parameter estimation of a correlated frailty model with type II interval-censored

The correlated gamma model has been applied to cross-sectional data (type I interval-censored data). Our simulation study compliments previous work reported on Cattaert (2008) and Hens et al. (2009) to type II interval-censored. Our objective is to investigate the performance of the parameter estimates when the amount of interval-censored observations increases implying that the number of right censored values decreases with the number of exact time to events kept constant.

We consider three different sample sizes n=1,000, 3,787 and 10,000 and assume both outcomes follow a Gompertz baseline hazard. The proportion of exact time to event observations is about 20%, whereas the proportion of right censored observations ranges from 30% (when 50% of the observations are interval-censored) to 50% (when 30% of the observations are interval-censored). The time to event data is generated as described in Section 5.5.1.

Here, to be able to compare results, the parameters of the Gompertz baseline hazard were chosen in accordance with Hens et al. (2009): $\alpha_1 = 0.006$, $\alpha_2 = 0.008$, $\beta_1 = 0.02$ and $\beta_2 = 0.03$, the frailty parameters are $\sigma_1 = 1.6$, $\sigma_2 = 1$ and $\rho = 0.5$.

Table 5.3 shows the true values, the mean of the parameter, its empirical standard error (calculated as described in the previous section) as well as the mean of the standard errors. We assume a correlated frailty model based on 1,000 individuals and with 30%, 40% and 50% of interval-censored observations. The mean of the standard errors and the empirical standard errors decrease for all the parameters when the percentage of interval-censored observations increases. This result is expected given the extra information provided by the increase in the percentage of interval-censored observations.

Table 5.4 presents bias, variance, MSE and in the last column the success rates of the estimation procedure for the first simulation setting. We do not observe any bias on the estimates of the baseline hazard and on the frailty parameters. The estimate for σ_1 with 30% of interval-censored observation has the highest variance, however its variance reduces considerably when the percentage of interval-censored observations increases and by the decrease in the percentage of right censored data.

In general, the MSEs are mainly driven by the variances of the estimates. In addition, the MSEs and the variances decrease when the percentage of interval-censored observations increases as a result of the decrease in the empirical standard error.

The success rate of the estimation procedure is 84% when 30% of the observations are interval-censored and 90% in the other two cases.

When we increase the sample size to 3,787 (Tables 5.5 and 5.6), the empirical standard errors and the mean of the standard errors decrease with respect to the ones based on n=1,000. As a results the estimated MSEs are also smaller.

Results on Tables 5.5 and 5.6 can be compared with the ones of Table II in Hens et al. (2009). The mean estimates in Table 5.5 almost coincides with the ones reported in table II assuming uncensored time to event. Moreover, the ESE are slightly larger than the ones assuming uncensored time to event and smaller than the ones assuming right censored data reported in Table II.

Assuming 30% interval-censored observations, the success rate of the estimation procedure is 97% as indicated in Table 5.6; which is higher than the 84% for the sample size of 1,000. The success rate when 40% of the observations are interval-censored is 97%; whereas the success rate when 50% observations are interval-censored is 93%.

Assuming 40% interval-censored observations, we observe some indication of bias and larger values for the MSE and variance of σ_1 and σ_2 compared to the ones observed when 30% of the observations are interval-censored. An inspection of the results reveals that three samples have estimates for σ_1 larger than 8 and estimates for σ_2 larger than 3. When the results of those three samples are removed we observe the consistent pattern previously described: when we increase the percentage of interval-censored observations bias, variance and MSE decrease.

When we assume 50% of interval-censored observations we observe a similar situation. The variance and the MSE for σ_1 are larger than the ones observed when 30% of the observations are interval-censored. In this case we notice one of the samples have an estimate of σ_1 larger than 8. Results obtained for sample size n=10,000 are presented in Tables 5.7 and 5.8.

For the baseline hazard parameters, as well as k's estimates show consistency when sample size increases. Furthermore, we have a more precise estimates in a larger sample size setting. That is, bias, averaged standard errors and empirical standard error remains constant or decrease when sample size increases as it can be seen in Table 5.7. This decline is also reflected in the variance and the MSE presented in Table 5.8.

However, results for σ_1 and σ_2 assuming 50% of interval-censored observations are a bit counterintuitive (Table 5.8). The bias and empirical standard error are the largest ones comparing n=10,000 with the other two sample sizes (n=1,000 and n=3,787) assumming 50% interval-censored observations. As a result, the variance and the MSE are also the largest of the observed ones. A close inspection at the results reveals that in six out of 500 samples the estimates for σ_1 were larger than 7 and the estimates of k_0, k_1 and k_2 are very small. We perform some exploratory analyses

Table 5.3: Averaged parameter estimates, empirical standard errors and averaged standard errors for the simulation study of the correlated gamma frailty model using a Gompertz baseline hazard ($\lambda_{i0}(t) = \alpha_i \exp(\beta_i t)$; i = 1, 2). n=1,000

Percentage of		α_1	α_2	β_1	β_2	k_0	k_1	k_2	σ_1	σ_2	ρ
interval-censored		0.0060	0.0080	0.0200	0.0300	0.3125	0.0781	0.6875	1.6000	1.0000	0.5000
30%	mean	0.0059	0.0078	0.0319	0.0327	0.3195	0.0909	0.6475	1.8134	1.0501	0.4995
	ESE	0.0007	0.0009	0.0708	0.0074	0.1716	0.1783	0.2216	0.9695	0.1608	0.1144
	mean(SE)	0.0007	0.0008	*0.0149*	0.0069	0.1487	*0.1072*	0.2216	*0.3829*	0.3477	*3.2723*
40%	mean	0.0059	0.0078	0.0242	0.0321	0.3300	0.0849	0.6533	*1.6841*	1.0394	0.5057
	ESE.	0.0006	0.0008	0.0285	0.0068	0.1751	0.1036	0.2097	0.5316	0.1497	0.0954
	mean(SE)	0.0006	0.0008	*0.0099*	0.0061	*0.1337*	*0.0635*	0.1918	*0.2995*	0.1003	*2.7565*
50%	mean	0.0060	0.0079	0.0222	0.0318	0.3150	0.0766	0.6693	*1.6628*	1.0332	0.5009
	ESE	0.0006	0.0007	0.0075	0.0056	0.1169	0.0690	0.1855	0.2781	0.1328	0.0748
	mean(SE)	0.0006	0.0008	0.0063	0.0055	0.1121	0.0539	0.1735	0.2451	*0.0649*	*2.2222*

Table 5.4: Estimated bias, variance and mean square error for 500 datasets assuming a correlated gamma frailty model with time to event using a Gompertz baseline hazard ($\lambda_{i0}(t) = \alpha_i \exp(\beta_i t)$; i = 1, 2). n=1,000

Percentage of		α_1	α_2	β_1	β_2	k_0	k_1	k_2	σ_1	σ_2	ρ	Success
interval-censored		0.0060	0.0080	0.0200	0.0300	0.3125	0.0781	0.6875	1.6000	1.0000	0.5000	Rate
30%	bias^2	4.18E-09	2.44E-08	0.0001	7.36E-06	4.84E-05	0.0002	0.0016	0.0456	0.0025	2.82E-07	83.8%
	variance	4.76E-07	7.46E-07	0.0050	5.45E-05	0.0295	0.0318	0.0491	0.9399	0.0259	0.0131	
	MSE	4.80E-07	7.70E-07	0.0052	6.18E-05	0.0295	0.0319	0.0507	0.9854	0.0284	0.0131	
40%	bias^2	3.14E-09	4.05E-08	1.80E-05	4.55E-06	0.0003	4.60E-05	0.0012	0.0071	0.0016	3.27E-05	90.4%
	variance	4.14E-07	6.26E-07	0.0008	4.58E-05	0.0307	0.0107	0.0440	0.2826	0.0224	0.0091	
	MSE	4.17E-07	6.66E-07	0.0008	5.03E-05	0.0310	0.0108	0.0451	0.2897	0.0240	0.0091	
50%	bias^2	8.68E-10	9.54E-09	5.01E-06	3.37E-06	6.44E-06	2.29E-06	0.0003	0.0039	0.0011	8.50E-07	90.2%
	variance	3.53E-07	5.56E-07	5.61E-05	3.16E-05	0.0137	0.0048	0.0344	0.0773	0.0176	0.0056	
	MSE	3.54E-07	5.65E-07	6.11E-05	3.49E-05	0.0137	0.0048	0.0347	0.0813	0.0187	0.0056	

Table 5.5: Averaged parameter estimates, empirical standard errors and averaged standard errors for the simulation study of the correlated gamma frailty model with time to event using a Gompertz baseline hazard ($\lambda_{i0}(t) = \alpha_i \exp(\beta_i t)$; $i = 1, 2$). n=3,787

Percentage of		α_1	α_2	β_1	β_2	k_0	k_1	k_2	σ_1	σ_2	ρ
interval-censored		0.0060	0.0080	0.0200	0.0300	0.3125	0.0781	0.6875	1.6000	1.0000	0.5000
30%	mean	0.0060	0.0080	0.0205	0.0303	0.3224	0.0773	0.6899	1.6101	1.0031	0.5040
	ESE	0.0003	0.0004	0.0044	0.0033	0.0797	0.0484	0.1182	0.1762	0.0788	0.0551
	mean(SE)	0.0003	0.0004	0.0042	0.0033	0.0807	0.0434	0.1196	0.1734	0.0427	0.8960
40%	mean	0.0060	0.0080	0.0235	0.0313	0.3167	0.0733	0.6881	1.6601	1.0174	0.5009
	ESE	0.0004	0.0005	0.0372	0.0144	0.0760	0.0381	0.1131	0.5521	0.2043	0.0563
	mean(SE)	0.0003	0.0004	0.0035	0.0031	0.0667	0.0350	0.1043	0.1415	0.0360	0.3973
50%	mean	0.0060	0.0080	0.0210	0.0305	0.3235	0.0748	0.6939	1.6119	1.0038	0.5049
	ESE	0.0004	0.0004	0.0179	0.0104	0.0673	0.0314	0.1015	0.3109	0.1451	0.0439
	mean(SE)	0.0003	0.0004	0.0029	0.0027	0.0599	0.0307	0.0944	0.1208	0.0321	0.2646

Table 5.6: Estimated bias, variance and mean square error for 500 datasets assuming a correlated gamma frailty model with time to event using a Gompertz baseline hazard ($\lambda_{i0}(t) = \alpha_i \exp(\beta_i t)$; i = 1, 2). n=3,787

Percentage of		α_1	α_2	β_1	β_2	k_0	k_1	k_2	σ_1	σ_2	ρ	Success
interval-censored		0.0060	0.0080	0.0200	0.0300	0.3125	0.0781	0.6875	1.6000	1.0000	0.5000	Rate
30%	bias^2	4.98E-10	7.48E-10	2.68E-07	7.23E-08	9.84E-05	5.76E-07	5.62E-06	0.0001	9.73E-06	1.60E-05	96.6%
	variance	1.10E-07	1.85E-07	1.96E-05	1.06E-05	0.0063	0.0023	0.0140	0.0310	0.0062	0.0030	
	MSE	1.10E-07	1.86E-07	1.99E-05	1.07E-05	0.0064	0.0023	0.0140	0.0312	0.0062	0.0031	
40%	bias^2	5.05E-10	1.90E-09	1.23E-05	1.82E-06	1.76E-05	2.33E-05	3.42E-07	0.0036	0.0003	8.44E-07	96.8%
	variance	1.47E-07	2.49E-07	0.0014	0.0002	0.0058	0.0015	0.0128	0.3048	0.0418	0.0032	
	MSE	1.48E-07	2.50E-07	0.0014	0.0002	0.0058	0.0015	0.0128	0.3084	0.0421	0.0032	
50%	bias^2	2.33E-10	8.71E-11	9.18E-07	2.51E-07	0.0001	1.08E-05	4.08E-05	0.0001	1.44E-05	2.38E-05	92.8%
	variance	1.33E-07	1.65E-07	0.0003	0.0001	0.0045	0.0010	0.0103	0.0967	0.0210	0.0019	
	MSE	1.34E-07	1.65E-07	0.0003	0.0001	0.0046	0.0010	0.0103	0.0968	0.0211	0.0019	

Table 5.7: Averaged parameter estimates, empirical standard errors and averaged standard errors for the simulation study of the correlated gamma frailty model with time to event using a Gompertz baseline hazard ($\lambda_{i0}(t) = \alpha_i \exp(\beta_i t)$; $i = 1, 2$). n=10,000

Percentage of interval-censored		α_1 0.0060	α_2 0.0080	β_1 0.0200	β_2 0.0300	k_0 0.3125	k_1 0.0781	k_2 0.6875	σ_1 1.6000	σ_2 1.0000	ρ 0.5000
30%	mean	0.0060	0.0080	0.0201	0.0301	0.3186	0.0771	0.6865	1.6002	1.0012	0.5038
	ESE	0.0002	0.0003	0.0026	0.0020	0.0494	0.0279	0.0704	0.1066	0.0501	0.0341
	mean(SE)	0.0002	0.0003	0.0025	0.0021	0.0489	0.0266	0.0734	0.1036	0.0262	0.1062
40%	mean	0.0061	0.0080	0.0237	0.0318	0.3111	0.0786	0.6803	1.6556	1.0251	0.4969
	ESE	0.0005	0.0004	0.0393	0.0179	0.0500	0.0243	0.0845	0.5833	0.2397	0.0419
	mean(SE)	0.0002	0.0003	0.0021	0.0019	0.0405	0.0215	0.0629	0.0857	0.0224	0.0787
50%	mean	0.0061	0.0080	0.0263	0.0330	0.3123	0.0766	0.6841	1.6892	1.0356	0.4965
	ESE	0.0005	0.0004	0.0534	0.0255	0.0526	0.0204	0.0944	0.7692	0.3186	0.0464
	mean(SE)	0.0002	0.0002	0.0018	0.0018	0.0356	0.0184	0.0568	0.0734	0.0196	0.0590

Table 5.8: Estimated bias, variance and mean square error for 500 datasets assuming a correlated gamma frailty model with time to event using a Gompertz baseline hazard ($\lambda_{i0}(t) = \alpha_i \exp(\beta_i t)$; i = 1, 2). n=10,000

Percentage of		α_1	α_2	β_1	β_2	k_0	k_1	k_2	σ_1	σ_2	ρ	Success
interval-censored		0.0060	0.0080	0.0200	0.0300	0.3125	0.0781	0.6875	1.6000	1.0000	0.5000	Rate
30%	bias^2	3.62E-13	1.22E-13	1.08E-08	6.64E-09	3.66E-05	9.42E-07	9.42E-07	5.05E-08	1.47E-06	1.42E-05	97.6%
	variance	3.93E-08	6.30E-08	6.79E-06	4.06E-06	0.0024	0.0008	0.0050	0.0114	0.0025	0.0012	
	MSE	3.93E-08	6.30E-08	6.80E-06	4.07E-06	0.0025	0.0008	0.0050	0.0114	0.0025	0.0012	
40%	bias^2	2.90E-09	1.51E-09	1.38E-05	3.34E-06	2.03E-06	2.09E-07	5.22E-05	0.0031	0.0006	9.46E-06	95.4%
	variance	2.47E-07	1.31E-07	0.0015	0.0003	0.0025	0.0006	0.0071	0.3403	0.0575	0.0018	
	MSE	2.50E-07	1.32E-07	0.0016	0.0003	0.0025	0.0006	0.0072	0.3434	0.0581	0.0018	
50%	bias^2	4.11E-09	6.04E-10	3.97E-05	8.79E-06	5.87E-08	2.16E-06	1.17E-05	0.0080	0.0013	1.26E-05	86.4%
	variance	2.95E-07	1.71E-07	0.0029	0.0006	0.0028	0.0004	0.0089	0.5917	0.1015	0.0022	
	MSE	2.99E-07	1.71E-07	0.0029	0.0007	0.0028	0.0004	0.0089	0.5997	0.1028	0.0022	

(results are not shown) and the same six cases are very distant from other observations for all the parameters. Therefore, we classified those six samples as outliers.

We also notice for σ_2 the bias and the empirical standard error are the largest one compared to previous scenarios. After removing the outliers the results are more inline to what we had expected for this scenario (Table 5.9). Mean estimates, empirical standard error, averaged standard errors, bias, variance and mean square errors when the six samples with high estimates for σ_1 are removed are presented in Table 5.9.

Additionally, the success rate in this scenario (Table 5.8) is the lowest one (86.4%) compared to the ones for n=1,000 and n= 3,787. The results were so different to what we had expected that we re-run this scenario (n=10,000 and 50% interval-censored observation) once more. However, we got similar results to the ones presented here: success rate lower than 90% and few large estimates for σ_1 leading to large variance and MSE. Perhaps the uncertainty attributed to the large proportion of interval-censored observations combined with the large sample size lead to results that can be considered as outliers in some of the generated samples. This topic clearly deserves further investigation.

The success rate assuming 30% of interval-censored observation increases slightly, from 97% when n=3,787 to 98% when n=10,000. However, when we consider 40% interval-censored observations the success rates decrease from 97% to 95%. Although the decrease seems small it may be related to the issue described in the previous paragraph.

The variance and MSE for σ_1 and σ_2 with 40% interval-censored and 10,000 observations are slightly larger to the ones observed when n=3,787. To a lesser extend, we observe a similar situation that has been described.

Our results are in agreement with the results presented by Hens et al. (2009) and Cattaert (2008). The empirical standard errors reported in Table 5.7 are larger than the ones for uncensored observations and smaller than the ones for right censored observations reported by Hens et al. (2009) in Table AI. Clearly, type II interval-censored data provide more information than right censored data but less than exact time to event.

In all the scenarios presented in this section, we observed a correlation smaller than -0.6 between the estimates of σ_1 and ρ, and smaller than -0.4 between σ_2 and ρ. In fact, a negative correlation between the variance and the correlation estimates has been recognized by Wienke (2011), we provide further insights about this issue in the second simulation study.

Table 5.9: Adjusted values after removing six outliers assuming 50% interval-censored observations. Averaged parameter estimates, empirical standard errors, averaged standard errors, bias, variance and mean squared errors for the scenario assuming a correlated gamma frailty model using a Gompertz baseline hazard ($\lambda_{i0}(t) = \alpha_i \exp(\beta_i t)$; i = 1, 2). n=10,000

	α_1	α_2	β_1	β_2	k_0	k_1	k_2	σ_1	σ_2	ρ
	0.0060	0.0080	0.0200	0.0300	0.3125	0.0781	0.6875	1.6000	1.0000	0.5000
mean	0.0060	0.0080	0.0200	0.0300	0.3166	0.0776	0.6928	1.5985	0.9981	0.5011
ESE	0.0002	0.0003	0.0018	0.0018	0.0381	0.0189	0.0598	0.0780	0.0430	0.0254
mean(SE)	0.0002	0.0002	0.0018	0.0017	0.0361	0.0187	0.0575	0.0744	0.0197	0.0598
bias^2	1.35E-10	6.66E-11	1.70E-12	2.06E-09	1.68E-05	2.69E-07	2.82E-05	2.35E-06	3.68E-06	1.10E-06
variance	3.59E-08	6.44E-08	3.41E-06	3.18E-06	0.0015	0.0004	0.0036	0.0061	0.0018	0.0006
MSE	3.61E-08	6.45E-08	3.41E-06	3.18E-06	0.0015	0.0004	0.0036	0.0061	0.0019	0.0006

5.5.3 Impact of restrictive frailty correlation in the parameter estimation of a correlated frailty models

As pointed out before, the correlation should not exceed the smallest ratio of the frailty standard deviations. This property can restrict the model when the variances are very different. In fact in the Amsterdam Cohort example the frailty correlation is located on the boundary of the parameter space. Here, we assess the impact of the restricted correlation. Initially, we assume equal frailty variances and then we gradually decrease one of the variances to assess how the estimation procedure is affected by the restriction.

The parameters for the baseline hazard rates (α_i, β_i) were based on the estimated values for the Amsterdam Cohort Studies data. For each scenario we simulate 500 datasets with 1,000 individuals. The time to event data is generated as described in Section 5.5.1. In the first scenario, the percentage of exact time to event observations for the first outcome is 65%, whereas for the second outcome is 50% reflecting the conditions observed in the ACS data.

The percentage of censored observations for the second outcome remains unchanged throughout all the scenarios. On the other hand, for the first outcome the percentage of censored observations decreases when the frailty variance decreases from 35% in the first six scenarios to 7% in the last three scenarios. The reason is that the cumulative baseline hazard (and therefore the time to event) gets smaller when the frailty variance decreases.

We separate the scenarios in three groups: in the first six scenarios the frailty variances are equal and the correlation ranges from 0.1 to 0.98; in the second group there is a moderate difference between the frailty variances; and in the third group we consider large difference between the frailty variances.

For each group we first present a table describing the frailty parameters and the sucess rates of each of the scenarios in the group (Tables 5.10, 5.13 and 5.16). As in the previous simulation study the rate is defined as the percentage of samples for which a solution is obtained and the hessian matrix is positive definite.

Then, we present another table including mean estimates, ESE and the mean of the standard errors (Tables 5.11, 5.14 and 5.17). Finally we include the bias, the variance and the MSE in Tables 5.12, 5.15 and 5.18.

In the first six scenarios the sucess rate ranges from 90% to 100%. The lowest success rate is observed when ρ is equal to 0.98 (Table 5.10). When the correlation is between 0.3 and 0.7 the success rate was 100%. Only two of the generated samples did not produce standard errors in the real line when the correlation was equal to 0.1

and 0.9.

As can be seen in 5.11, the mean of the standard errors and the empirical standard errors are pretty similar in terms of magnitude for all the estimates, indicating that the variability within sample (standard error) resembles the variability across samples (empirical standar error).

Additionally, for most of the parameters, except k_0 and α_2, we observed a decreasing trend in the empirical standard error and the mean of the standard errors when the frailty correlation increases. For α_2 the mean of the standard errors remains practically without change in all the six scenarios. Instead for k_0 increases in the first five scenarios and then decreases in the last one (Table 5.11).

We do not observe any bias in most of the parameters, except in scenarios 1 and 6 for the frailty variances σ_1, σ_2. We also notice some small biases for α_1. The MSEs are mainly driven by the variance of the estimates. There is a decreasing trend in the variances of the estimates and MSEs as a result of the decreasing trend observed in the ESE (Table 5.12).

Although under equal frailty variances the correlation may range from 0 to 1, we found some indications to be cautious with the results when ρ is on the boundary of the parameter space ($\rho < 0.1$ or $\rho > 0.9$).

It has been recognized by Wienke (2011), that there is a negative correlation between the variance and the correlation estimates. Our results suggest that the negative correlation is not fixed, for instance, in Scenario 1 the Pearson's correlation between the estimates of σ_2 and ρ is -0.05, then in Scenario 2 is -0.15, and in Scenario 5 the correlation is -0.53. We also notice a similar pattern between the correlation σ_1 and ρ, where the Pearson's correlation ranges from -0.08 in Scenario 1 to -0.37 in Scenario 5.

Scenarios 7-10 consider a variance ratio equals to 0.75, here the frailty correlation ranges between 0.2 and 0.71. The success rates of the estimation procedure are above 99%, the lowest values are for those where the correlation is higher (Table 5.13).

As previously described, for most of the parameters we notice a decreasing trend in the empirical standard errors and the means of the standard errors when the correlation increases. The empirical standard errors are of the same magnitude than the means of the standard errors for all the parameters (Table 5.14). The Pearson's correlation coefficient between σ_2 and ρ decreases from -0.13, in Scenario 7 to -0.62 in Scenario 10. The Pearson's correlation coefficient between σ_1 and ρ ranges between -0.09 and -0.21, however in this case we do not observe any clear trend.

In Scenarios 11-13, the variance ratio is 2 and the frailty correlation ranges between 0.3 and 0.49. The success rate ranges from 94% (when the frailty correlation is

Table 5.10: Success rate and frailty parameters assuming equal frailty variances and Weibull baseline hazard. The following parameters remain constant for all the simulation settings: $\alpha_1 = 1.21$, $\alpha_2 = 0.12$, $\beta_1 = 0.81$, $\beta_2 = 1.10$ and $\sigma_1 = \sigma_2 = 2$. $\rho < 1$ n=1,000. Part I

Index	σ_1	σ_2	ρ	k_0	k_1	k_2	Success rate
1	2	2	0.10	0.0250	0.2250	0.2250	99.6%
2	2	2	0.30	0.0750	0.1750	0.1750	100.0%
3	2	2	0.50	0.1250	0.1250	0.1250	100.0%
4	2	2	0.70	0.1750	0.0750	0.0750	100.0%
5	2	2	0.90	0.2250	0.0250	0.0250	99.6%
6	2	2	0.98	0.2450	0.0050	0.0050	89.6%

equal to 0.49) to 99% (Table 5.13). As described before we observe a decreasing trend in the empirical standard errors and the mean of the standard errors when the frailty correlation increases (Table 5.14).

We do not observe any bias in the estimates of the parameters; therefore the MSEs are mainly attributed to the variances of the estimates. The decreasing trends in the ESEs and the means of the standard errors are also reflected here in the variances and the MSEs (Table 5.15).

Regarding the correlation between the variances and the correlation estimates described before. We observe a similar trend in Scenarios 11-13. In Scenario 11, the Pearson correlation coefficient between σ_2 and ρ is -0.12, in Scenario 12 is -0.32 and in Scenario 13 reaches -0.5.

Table 5.11: Mean estimates and empirical standard errors for the baseline hazard and the frailty parameters, assuming equal frailty variances and Weibull baseline hazard. The following parameters remain constant for all the simulation settings: $\alpha_1 = 1.21$, $\alpha_2 = 0.12$, $\beta_1 = 0.81$, $\beta_2 = 1.10$ and $\sigma_1 = \sigma_2 = 2$. $\rho < 1$ n=1,000. Part I

						$\sigma_1 = 2$	$\sigma_2 = 2$	$\alpha_1 = 1.21$	$\alpha_2 = 0.12$	$\beta_1 = 0.81$	$\beta_2 = 1.1$
Index		$\hat{k_0}$	$\hat{k_1}$	$\hat{k_2}$	$\hat{\rho}$	$\hat{\sigma_1}$	$\hat{\sigma_2}$	$\hat{\alpha_1}$	$\hat{\alpha_2}$	$\hat{\beta_1}$	$\hat{\beta_2}$
1	mean	0.0246	0.2249	0.2277	0.0982	2.0165	2.0150	1.2781	0.1207	0.8223	1.1187
	ESE	0.0128	0.0343	0.0455	0.0486	0.1402	0.1790	0.3036	0.0134	0.0681	0.0966
	mean(SE)	0.0125	0.0342	0.0425	0.0472	0.1361	0.1727	0.2670	0.0137	0.0647	0.0956
2	mean	0.0763	0.1760	0.1741	0.3041	2.0050	2.0172	1.2564	0.1213	0.8140	1.1125
	ESE	0.0155	0.0307	0.0335	0.0460	0.1366	0.1588	0.2656	0.0138	0.0655	0.0915
	mean(SE)	0.0156	0.0278	0.0317	0.0455	0.1230	0.1539	0.2329	0.0135	0.0580	0.0856
3	mean	0.1261	0.1247	0.1249	0.5021	2.0061	2.0092	1.2504	0.1210	0.8151	1.1063
	ESE	0.0197	0.0213	0.0216	0.0418	0.1113	0.1334	0.2093	0.0126	0.0517	0.0748
	mean(SE)	0.0192	0.0208	0.0215	0.0416	0.1094	0.1317	0.2051	0.0132	0.0517	0.0744
4	mean	0.1753	0.0746	0.0737	0.7019	2.0081	2.0136	1.2415	0.1208	0.8152	1.1070
	ESE	0.0227	0.0149	0.0141	0.0353	0.1026	0.1158	0.1872	0.0131	0.0485	0.0672
	mean(SE)	0.0222	0.0152	0.0137	0.0360	0.0971	0.1143	0.1805	0.0131	0.0457	0.0658
5	mean	0.2272	0.0246	0.0246	0.9015	1.9982	1.9994	1.2181	0.1200	0.8111	1.1002
	ESE	0.0231	0.0107	0.0082	0.0265	0.0847	0.0937	0.1570	0.0132	0.0386	0.0533
	mean(SE)	0.0241	0.0111	0.0079	0.0276	0.0856	0.0976	0.1551	0.0129	0.0395	0.0572
6	mean	0.2411	0.0068	0.0056	0.9745	2.0131	2.0184	1.2355	0.1205	0.8163	1.1103
	ESE	0.0207	0.0070	0.0045	0.0165	0.0771	0.0830	0.1401	0.0135	0.0363	0.0504
	mean(SE)	0.0217	0.0068	0.0046	0.0186	0.0784	0.0852	0.1457	0.0129	0.0357	0.0514

Table 5.12: Estimated bias, variance and mean squared error for 500 datasets assuming equal frailty variances and Weibull baseline hazards for both outcomes. The following parameters remain constant for all the simulation settings: $\alpha_1 = 1.21$, $\alpha_2 = 0.12$, $\beta_1 = 0.81$, $\beta_2 = 1.10$ and $\sigma_1 = \sigma_2 = 2$. $\rho < 1$ Part I

Index		$\hat{k_0}$	$\hat{k_1}$	$\hat{k_2}$	$\hat{\rho}$	$\hat{\sigma_1}$	$\hat{\sigma_2}$	$\hat{\alpha_1}$	$\hat{\alpha_2}$	$\hat{\beta_1}$	$\hat{\beta_2}$
1	bias^2	1.59E-07	1.12E-08	7.46E-06	3.28E-06	0.0003	0.0002	0.0046	4.66E-07	0.0002	0.0004
	variance	0.0002	0.0012	0.0021	0.0024	0.0197	0.0320	0.0921	0.0002	0.0046	0.0093
	MSE	0.0002	0.0012	0.0021	0.0024	0.0199	0.0323	0.0968	0.0002	0.0048	0.0097
2	bias^2	1.60E-06	9.13E-07	7.92E-07	1.68E-05	2.53E-05	0.0003	0.0021	1.61E-06	1.58E-05	0.0002
	variance	0.0002	0.0009	0.0011	0.0021	0.0187	0.0252	0.0706	0.0002	0.0043	0.0084
	MSE	0.0002	0.0009	0.0011	0.0021	0.0187	0.0255	0.0727	0.0002	0.0043	0.0085
3	bias^2	1.18E-06	9.28E-08	3.73E-09	4.32E-06	3.69E-05	8.49E-05	0.0016	9.71E-07	2.64E-05	3.93E-05
	variance	0.0004	0.0005	0.0005	0.0017	0.0124	0.0178	0.0438	0.0002	0.0027	0.0056
	MSE	0.0004	0.0005	0.0005	0.0018	0.0124	0.0179	0.0455	0.0002	0.0027	0.0056
4	bias^2	1.17E-07	1.62E-07	1.60E-06	3.79E-06	6.51E-05	0.0002	0.0010	6.85E-07	2.72E-05	4.84E-05
	variance	0.0005	0.0002	0.0002	0.0012	0.0105	0.0134	0.0351	0.0002	0.0024	0.0045
	MSE	0.0005	0.0002	0.0002	0.0012	0.0106	0.0136	0.0361	0.0002	0.0024	0.0046
5	bias^2	4.69E-06	1.37E-07	1.28E-07	2.27E-06	3.25E-06	3.64E-07	6.56E-05	2.27E-09	1.32E-06	5.40E-08
	variance	0.0005	0.0001	0.0001	0.0007	0.0072	0.0088	0.0247	0.0002	0.0015	0.0028
	MSE	0.0005	0.0001	0.0001	0.0007	0.0072	0.0088	0.0247	0.0002	0.0015	0.0028
6	bias^2	1.56E-05	3.21E-06	3.96E-07	3.07E-05	0.0002	0.0003	0.0006	2.97E-07	3.97E-05	0.0001
	variance	0.0004	4.85E-05	2.04E-05	0.0003	0.0060	0.0069	0.0196	0.0002	0.0013	0.0025
	MSE	0.0004	5.17E-05	2.08E-05	0.0003	0.0061	0.0072	0.0203	0.0002	0.0014	0.0027

Table 5.13: Success rate and frailty parameters assuming a moderate difference between the frailty variances and Weibull baseline hazard. The following parameters remain constant for all the simulation settings: $\alpha_1 = 1.21$, $\alpha_2 = 0.12$, $\beta_1 = 0.81$ and $\beta_2 = 1.10$. n=1,000.

Index	σ_1	σ_2	ρ	k_0	k_1	k_2	Success rate
			$\sigma_1 = 1.5, \sigma_2 = 2, \rho \leq 0.75$				
7	1.5	2.0	0.20	0.0667	0.3778	0.1833	100.0%
8	1.5	2.0	0.40	0.1333	0.3111	0.1167	100.0%
9	1.5	2.0	0.70	0.2333	0.2111	0.0167	99.2%
10	1.5	2.0	0.71	0.2367	0.2078	0.0133	99.2%
			$\sigma_1 = 1, \sigma_2 = 2, \rho \leq 0.5$				
11	1.0	2.0	0.30	0.1500	0.8500	0.1000	99.4%
12	1.0	2.0	0.40	0.2000	0.8000	0.0500	99.2%
13	1.0	2.0	0.49	0.2450	0.7550	0.0050	94.0%

Last six scenarios show the results for large differences between the frailty variances: the frailty variance ratio ranges between 4 and 20. In these cases, the frailty correlation is heavily restricted as described in Table 5.16.

In Scenarios 14 and 15 the ratio of the frailty variances is 4 and the frailty correlation is 0.2 and 0.24 respectively. In both scenarios the success rate of the estimation procedure is lower than 95%. In most of the samples classified as failures, the standard error for k_2 was given in complex number. In these two scenarios the empirical standard errors and the means of the standard errors have the same magnitude (Table 5.17).

In Scenario 14 the highest bias is observed in the estimates of k_1, whereas in Scenario 15 there is some indication of bias in the estimates of k_1, k_2 and ρ (Table 5.18). From Scenario 15 onwards, the variance and the MSE for σ_1 and σ_2 are also large compared to previous scenarios.

In Scenarios 16 and 17 the ratio between the frailty variances is 6.67 and the frailty correlation is 0.1 in Scenario 16 and 0.145 in Scenario 17. The success rates of the estimation procedure are 89% and 87% respectively (Table 5.16). From Scenario 16 onwards the means of the standard errors are much larger than the empirical standard errors, indicating much larger within sample variability than between sample variability. However, the empirical standard errors are also very large and show an increasing trend when the difference in the frailty variance is larger.

In Scenario 16, in 71 samples (14%) the standard error for k_1 is more than twice the value of the estimate for the parameter (Table 5.17). At this point is difficult to

Table 5.14: Mean estimates and empirical standard errors for the baseline hazard and the frailty parameters, assuming a moderate difference between the frailty variances and Weibull baseline hazard. The following parameters remain constant for all the simulation settings: $\alpha_1 = 1.21$, $\alpha_2 = 0.12$, $\beta_1 = 0.81$ and $\beta_2 = 1.10$. n=1,000

Index		$\hat{k_0}$	$\hat{k_1}$	$\hat{k_2}$	$\hat{\rho}$	$\hat{\sigma_1}$	$\hat{\sigma_2}$	$\hat{\alpha_1}$	$\hat{\alpha_2}$	$\hat{\beta_1}$	$\hat{\beta_2}$
					$\sigma_1 = 1.5, \sigma_2 = 2, \rho \leq 0.75$						
7	mean	0.0691	0.3771	0.1866	0.2046	1.5056	1.9987	1.2352	0.1200	0.8165	1.1043
	ESE	0.0194	0.0524	0.0396	0.0497	0.0914	0.1673	0.1790	0.0138	0.0518	0.0916
	mean(SE)	0.0196	0.0536	0.0390	0.0499	0.0941	0.1668	0.1744	0.0134	0.0515	0.0913
8	mean	0.1328	0.3071	0.1160	0.4010	1.5155	2.0233	1.2434	0.1207	0.8197	1.1156
	ESE	0.0243	0.0426	0.0271	0.0439	0.0884	0.1552	0.1682	0.0138	0.0504	0.0848
	mean(SE)	0.0228	0.0430	0.0259	0.0445	0.0865	0.1448	0.1625	0.0133	0.0478	0.0815
9	mean	0.2340	0.2078	0.0168	0.7026	1.5095	2.0064	1.2282	0.1201	0.8153	1.1061
	ESE	0.0285	0.0290	0.0104	0.0331	0.0711	0.1130	0.1262	0.0124	0.0387	0.0653
	mean(SE)	0.0288	0.0286	0.0106	0.0336	0.0712	0.1136	0.1341	0.0130	0.0393	0.0652
10	mean	0.2393	0.2074	0.0140	0.7108	1.5008	1.9953	1.2219	0.1202	0.8135	1.1032
	ESE	0.0282	0.0274	0.0103	0.0325	0.0683	0.1071	0.1247	0.0128	0.0378	0.0617
	mean(SE)	0.0291	0.0286	0.0099	0.0332	0.0700	0.1124	0.1313	0.0129	0.0387	0.0645
					$\sigma_1 = 1, \sigma_2 = 2, \rho \leq 0.5$						
11	mean	0.1530	0.8515	0.1009	0.3029	1.0043	2.0044	1.2229	0.1195	0.8153	1.1112
	ESE	0.0363	0.1220	0.0391	0.0546	0.0656	0.1630	0.1106	0.0130	0.0402	0.0916
	mean(SE)	0.0347	0.1259	0.0379	0.0531	0.0676	0.1613	0.1140	0.0133	0.0410	0.0893
12	mean	0.2013	0.8068	0.0529	0.3978	1.0031	2.0006	1.2261	0.1205	0.8157	1.1049
	ESE	0.0384	0.1258	0.0305	0.0472	0.0688	0.1503	0.1103	0.0134	0.0427	0.0833
	mean(SE)	0.0368	0.1170	0.0302	0.0473	0.0646	0.1506	0.1103	0.0133	0.0395	0.0836
13	mean	0.2419	0.7541	0.0109	0.4825	1.0068	2.0021	1.2242	0.1205	0.8168	1.1043
	ESE	0.0332	0.1016	0.0151	0.0345	0.0574	0.1329	0.1003	0.0135	0.0343	0.0759
	mean(SE)	0.0362	0.1009	0.0144	0.0376	0.0576	0.1383	0.1008	0.0131	0.0357	0.0778

Table 5.15: Estimated bias, variance and mean squared error for 500 datasets assuming a moderate difference between the frailty variances and Weibull baseline hazards for both outcomes. The following parameters remain constant for all the simulation settings: $\alpha_1 = 1.21, \alpha_2 = 0.12, \beta_1 = 0.81$ and $\beta_2 = 1.10$.

Index		$\hat{k_0}$	$\hat{k_1}$	$\hat{k_2}$	$\hat{\rho}$	$\hat{\sigma_1}$	$\hat{\sigma_2}$	$\hat{\alpha_1}$	$\hat{\alpha_2}$	$\hat{\beta_1}$	$\hat{\beta_2}$
					$\sigma_1 = 1.5, \sigma_2 = 2, \rho \leq 0.75$						
7	bias^2	5.71E-06	4.94E-07	1.09E-05	2.16E-05	3.09E-05	1.60E-06	0.0006	2.30E-09	4.24E-05	1.84E-05
	variance	0.0004	0.0027	0.0016	0.0025	0.0084	0.0280	0.0320	0.0002	0.0027	0.0084
	MSE	0.0004	0.0027	0.0016	0.0025	0.0084	0.0280	0.0327	0.0002	0.0027	0.0084
8	bias^2	3.03E-07	1.64E-05	5.01E-07	1.07E-06	0.0002	0.0005	0.0011	4.36E-07	9.48E-05	0.0002
	variance	0.0006	0.0018	0.0007	0.0019	0.0078	0.0241	0.0283	0.0002	0.0025	0.0072
	MSE	0.0006	0.0018	0.0007	0.0019	0.0081	0.0246	0.0294	0.0002	0.0026	0.0074
9	bias^2	4.81E-07	1.12E-05	9.10E-09	6.50E-06	9.08E-05	4.08E-05	0.0003	2.06E-08	2.80E-05	3.72E-05
	variance	0.0008	0.0008	0.0001	0.0011	0.0051	0.0128	0.0159	0.0002	0.0015	0.0043
	MSE	0.0008	0.0009	0.0001	0.0011	0.0051	0.0128	0.0163	0.0002	0.0015	0.0043
10	bias^2	7.15E-06	1.64E-07	4.55E-07	5.64E-07	6.51E-07	2.25E-05	0.0001	3.48E-08	1.24E-05	1.02E-05
	variance	0.0008	0.0007	0.0001	0.0011	0.0047	0.0115	0.0156	0.0002	0.0014	0.0038
	MSE	0.0008	0.0007	0.0001	0.0011	0.0047	0.0115	0.0157	0.0002	0.0014	0.0038
					$\sigma_1 = 1, \sigma_2 = 2, \rho \leq 0.5$						
11	bias^2	8.96E-06	2.25E-06	8.83E-07	8.37E-06	1.83E-05	1.93E-05	0.0002	2.76E-07	2.83E-05	0.0001
	variance	0.0013	0.0149	0.0015	0.0030	0.0043	0.0266	0.0122	0.0002	0.0016	0.0084
	MSE	0.0013	0.0149	0.0015	0.0030	0.0043	0.0266	0.0124	0.0002	0.0016	0.0085
12	bias^2	1.81E-06	4.57E-05	8.24E-06	4.87E-06	9.89E-06	4.13E-07	0.0003	2.67E-07	3.25E-05	2.45E-05
	variance	0.0015	0.0158	0.0009	0.0022	0.0047	0.0226	0.0122	0.0002	0.0018	0.0069
	MSE	0.0015	0.0159	0.0009	0.0022	0.0047	0.0226	0.0124	0.0002	0.0019	0.0070
13	bias^2	9.33E-06	7.36E-07	3.45E-05	5.56E-05	4.66E-05	4.30E-06	0.0002	2.94E-07	4.66E-05	1.85E-05
	variance	0.0011	0.0103	0.0002	0.0012	0.0033	0.0177	0.0101	0.0002	0.0012	0.0058
	MSE	0.0011	0.0103	0.0003	0.0012	0.0033	0.0177	0.0103	0.0002	0.0012	0.0058

define a common criterion for all scenarios to discard any of the samples with large standard errors. In fact our results illustrate potential problems that arise in presence of large difference in the frailty variances.

When there is large difference between the frailty variances the estimation of the parameters also become challenging. In Scenario 16, in 60 samples (12%) the estimates of k_1 double the value of the estimate of the parameter. Certainly there is overlap, in each sample that the estimate is large the standard error is also large.

In general, in Scenario 17 mean estimates, empirical standard error, mean standard error, variance, MSE and bias for k_1 are larger compared to the ones observed in the previous scenario. On the other hand, Scenarios 16 and 17 have similar values for the bias and the MSE for k_0, k_2 and ρ (Tables 5.17 and 5.18).

The impact of the large difference between the frailty variances is clearly reflected on the estimation of k_1; however, the impact in the estimation of σ_1 is limited. There are two reasons: i) the magnitude of σ_1 does not allow to observe high values for the mean estimates and consequently the bias. Since σ_1 is smaller than one, and the bias^2 can be at most 0.09 (a relatively small value compare to the ones we may observe for other parameters). The second reason is large estimates of k_1 lead to smaller estimates of σ_1 pulling down the mean estimate and bias but not as much as high values can do.

The Pearson correlation coefficient between σ_2 and ρ in Scenarios 16 and 17 is -0.10. Whereas the Pearson correlation coefficient between σ_1 and ρ is 0.24 in Scenario 16 and 0.28 in Scenario 17. The positive correlation between σ_1 and ρ may be influenced by the large estimates of k_1, however this requires more investigation if possible.

In Scenarios 18 and 19 we assume a ratio of the frailty variances of 10 and a frailty correlation smaller than 0.1. The success rate of the estimation procedure for scenario 18 is 72% and 69% for Scenario 19 as can be seen in Table 5.16. In both scenarios means, empirical standard errors and averaged standard errors for k_1 are even larger than the ones observed in scenarios 16 and 17.

In Scenario 18, 76 samples have a very large standard errors for k_1 (twice larger than the correspondent estimate). Moreover, 40 of those 76 samples have also large estimates for k_1 and we see some indication of bias in the baseline hazard parameters, particularly in α_1. We also notice slight bias in the frailty parameters k_0, k_2 and σ_1.

In this scenario, the pearson correlation coefficient between σ_2 and ρ is -0.16 and the pearson correlation between σ_1 and ρ is 0.32.

In Scenario 20 only 52% of the samples were classified as successful. In two of the samples considered as failure the program did not produce standard errors, whereas

Table 5.16: Success rate and frailty parameters assuming a large difference between the frailty variances and Weibull baseline hazard. The following parameters remain constant for all the simulation settings: $\alpha_1 = 1.21, \alpha_2 = 0.12, \beta_1 = 0.81$ and $\beta_2 = 1.10$. n=1,000.

Index	σ_1	σ_2	ρ	k_0	k_1	k_2	Success rate
			$\sigma_1 = 0.5, \sigma_2 = 2, \rho \leq 0.25$				
14	0.50	2.0	0.200	0.2000	3.8000	0.0500	94.0%
15	0.50	2.0	0.240	0.2400	3.7600	0.0100	94.8%
			$\sigma_1 = 0.3, \sigma_2 = 2, \rho \leq 0.15$				
16	0.30	2.0	0.100	0.1667	10.9444	0.0833	88.8%
17	0.30	2.0	0.145	0.2417	10.8694	0.0083	86.8%
			$\sigma_1 = 0.2, \sigma_2 = 2, \rho \leq 0.1$				
18	0.20	2.0	0.080	0.2000	24.8000	0.0500	72.0%
19	0.20	2.0	0.099	0.2475	24.7525	0.0025	68.8%
			$\sigma_1 = 0.1, \sigma_2 = 2, \rho \leq 0.05$				
20	0.10	2.0	0.049	0.2450	99.7550	0.0050	52.4%

in the other cases the hessian matrix was not positive definite. After discarding almost half of the samples, we observe very large values for the estimates and the standard errors for k_1.

The Pearson correlation coefficients between σ_1 and ρ in Scenarios 16 to 20 are positive. Which may be artificial due to the large estimates for k_1 in some of the samples.

Frailty variance differences we consider in Scenarios 16 to 20 are rather extreme and point out how difficult the estimation process can be. The issue is not just the parameter estimation but also the estimation of the standard errors, which hampers statistical inference and model selection.

Table 5.17: Mean estimates and empirical standard errors for the baseline hazard and the frailty parameters, assuming a large difference between the frailty variances and Weibull baseline hazard. The following parameters remain constant for all the simulation settings: $\alpha_1 = 1.21$, $\alpha_2 = 0.12$, $\beta_1 = 0.81$ and $\beta_2 = 1.10$. n=1,000.

Index		$\hat{k_0}$	$\hat{k_1}$	$\hat{k_2}$	$\hat{\rho}$	$\hat{\sigma_1}$	$\hat{\sigma_2}$	$\hat{\alpha_1}$	$\hat{\alpha_2}$	$\hat{\beta_1}$	$\hat{\beta_2}$
$\sigma_1 = 0.5, \sigma_2 = 2, \rho \leq 0.25$											
14	mean	0.1897	4.0047	0.0656	0.1857	0.5024	1.9994	1.2255	0.1210	0.8133	1.1047
	ESE	0.0761	1.2887	0.0717	0.0693	0.0647	0.1660	0.0745	0.0131	0.0323	0.0957
	mean(SE)	0.0887	1.2229	0.0783	0.0786	0.0681	0.1737	0.0781	0.0136	0.0338	0.0948
15	mean	0.2102	3.8200	0.0469	0.2087	0.5102	1.9964	1.2265	0.1203	0.8174	1.1067
	ESE	0.0714	1.0722	0.0630	0.0612	0.0625	0.1785	0.0759	0.0133	0.0323	0.0984
	mean(SE)	0.0783	1.0979	0.0606	0.0695	0.0660	0.1721	0.0773	0.0135	0.0335	0.0944
$\sigma_1 = 0.3, \sigma_2 = 2, \rho \leq 0.15$											
16	mean	0.1498	700.7593	0.1067	0.0878	0.3084	2.0003	1.2296	0.1199	0.8184	1.1100
	ESE	0.1156	6’430.3788	0.1136	0.0712	0.0925	0.1872	0.0683	0.0149	0.0322	0.1007
	mean(SE)	0.1328	50’029.9587	0.1273	0.0810	0.1049	0.1760	0.0715	0.0136	0.0327	0.0964
17	mean	0.1770	768.7461	0.0788	0.1071	0.3176	2.0017	1.2257	0.1203	0.8206	1.1084
	ESE	0.1067	7’959.4158	0.1021	0.0678	0.0847	0.1827	0.0626	0.0139	0.0303	0.0997
	mean(SE)	0.1354	55’015.2091	0.1222	0.0864	0.0996	0.1752	0.0707	0.0136	0.0325	0.0959
$\sigma_1 = 0.2, \sigma_2 = 2, \rho \leq 0.1$											
18	mean	0.1595	5’014.3083	0.0981	0.0708	0.2411	1.9939	1.2438	0.1199	0.8237	1.1033
	ESE	0.1199	53’775.2798	0.1176	0.0634	0.0956	0.1800	0.0646	0.0130	0.0285	0.1006
	mean(SE)	0.1383	309’515.5347	0.1301	0.0838	0.1401	0.1756	0.0697	0.0135	0.0320	0.0955
19	mean	0.1785	14’110.3612	0.0788	0.0707	0.2305	1.9974	1.2419	0.1203	0.8238	1.1074
	ESE	0.1141	95’138.8290	0.1138	0.0612	0.1138	0.1862	0.0603	0.0129	0.0305	0.1027
	mean(SE)	0.1475	1’012’795.7685	0.1386	0.0870	0.1327	0.1765	0.0677	0.0136	0.0312	0.0963
$\sigma_1 = 0.1, \sigma_2 = 2, \rho \leq 0.05$											
20	mean	0.1930	33’231.2394	0.0651	0.0556	0.1744	1.9925	1.2486	0.1215	0.8275	1.1085
	ESE	0.1071	107’413.9819	0.1070	0.0549	0.1183	0.1811	0.0612	0.0143	0.0271	0.1015
	mean(SE)	0.1801	3’084’211.5965	0.1708	0.0920	0.1538	0.1759	0.0640	0.0137	0.0297	0.0965

Table 5.18: Estimated bias, variance and mean squared error for 500 datasets assuming a large difference between the frailty variances and Weibull baseline hazards for both outcomes. The following parameters remain constant for all the simulation settings: $\alpha_1 = 1.21$, $\alpha_2 = 0.12$, $\beta_1 = 0.81$ and $\beta_2 = 1.10$.

Index		$\hat{k_0}$	$\hat{k_1}$	$\hat{k_2}$	$\hat{\rho}$	$\hat{\sigma_1}$	$\hat{\sigma_2}$	$\hat{\alpha_1}$	$\hat{\alpha_2}$	$\hat{\beta_1}$	$\hat{\beta_2}$
					$\sigma_1 = 0.5, \sigma_2 = 2, \rho \leq 0.25$						
	bias^2	0.0001	0.0419	0.0002	0.0002	5.97E-06	3.07E-07	0.0002	9.03E-07	1.08E-05	2.22E-05
14	variance	0.0058	1.6608	0.0051	0.0048	0.0042	0.0276	0.0056	0.0002	0.0010	0.0092
	MSE	0.0059	1.7027	0.0054	0.0050	0.0042	0.0276	0.0058	0.0002	0.0011	0.0092
	bias^2	0.0009	0.0036	0.0014	0.0010	0.0001	1.28E-05	0.0003	6.33E-08	5.47E-05	4.49E-05
15	variance	0.0061	1.2054	0.0037	0.0048	0.0044	0.0296	0.0058	0.0002	0.0011	0.0089
	MSE	0.0070	1.2090	0.0050	0.0058	0.0045	0.0296	0.0060	0.0002	0.0012	0.0090
					$\sigma_1 = 0.3, \sigma_2 = 2, \rho \leq 0.15$						
	bias^2	0.0003	4.76E+05	0.0005	0.0001	7.13E-05	9.63E-08	0.0004	4.72E-09	7.09E-05	0.0001
16	variance	0.0134	4.13E+07	0.0129	0.0051	0.0086	0.0351	0.0047	0.0002	0.0010	0.0101
	MSE	0.0136	4.18E+07	0.0134	0.0052	0.0086	0.0351	0.0051	0.0002	0.0011	0.0102
	bias^2	0.0042	5.74E+05	0.0050	0.0014	0.0003	2.81E-06	0.0002	6.50E-08	0.0001	7.09E-05
17	variance	0.0114	6.34E+07	0.0104	0.0046	0.0072	0.0334	0.0039	0.0002	0.0009	0.0099
	MSE	0.0156	6.39E+07	0.0154	0.0060	0.0075	0.0334	0.0042	0.0002	0.0010	0.0100
					$\sigma_1 = 0.2, \sigma_2 = 2, \rho \leq 0.1$						
	bias^2	0.0016	2.49E+07	0.0023	8.50E-05	0.0017	3.70E-05	0.0011	1.68E-08	0.0002	1.09E-05
18	variance	0.0144	2.89E+09	0.0138	0.0040	0.0091	0.0324	0.0042	0.0002	0.0008	0.0101
	MSE	0.0160	2.92E+09	0.0161	0.0041	0.0108	0.0324	0.0053	0.0002	0.0010	0.0101
	bias^2	0.0048	1.98E+08	0.0058	0.0008	0.0009	6.75E-06	0.0010	7.75E-08	0.0002	5.45E-05
19	variance	0.0218	1.03E+12	0.0192	0.0076	0.0176	0.0311	0.0036	0.0002	0.0010	0.0093
	MSE	0.0265	1.03E+12	0.0250	0.0084	0.0185	0.0312	0.0047	0.0002	0.0012	0.0093
					$\sigma_1 = 0.1, \sigma_2 = 2, \rho \leq 0.05$						
	bias^2	0.0027	1.10E+09	0.0036	4.31E-05	0.0055	5.70E-05	0.0015	2.24E-06	0.0003	7.29E-05
20	variance	0.0115	1.15E+10	0.0114	0.0030	0.0140	0.0328	0.0037	0.0002	0.0007	0.0103
	MSE	0.0142	1.26E+10	0.0151	0.0031	0.0195	0.0329	0.0052	0.0002	0.0010	0.0104

5.6 Concluding remarks

The frailty model offers a powerful tool for the analysis of multivariate survival data, handle clustered survival data and account for heterogeneity due to unobserved covariates.

In case of clustered data, the shared frailty model is frequently used. In this case, all the subjects belonging to the same cluster share the same frailty term and therefore a common variance for all the time to event outcomes is assumed. However, this model may be too restrictive and therefore the correlated frailty model becomes an appealing alternative.

Under the correlated frailty model a natural division within a cluster can be taken into account and a set of random effects for each subcluster is included. In this model, the variances are not necessarily the same and the frailties allowed to be correlated. This model has been described by Wienke (2011) and applied in infectious disease context for current status data by Abrams and Hens (2014), Hens et al. (2009) and Cattaert (2008).

In this chapter, we derive the log-likelihood contributions for the correlated frailty model considering type II interval-censored data. We then apply several frailty models to the ACS data and compare the results. From these results we notice substantial differences in the frailty variances for both outcomes. Additionally, under the correlated frailty model, the correlation is on the boundary of the parameter space (given the constraints of the correlation for the bivariate gamma frailty model). These findings inspired the simulation study with a twofold objective. To assess model behaviour of a correlated frailty model in presence of type II interval-censored data and to assess the impact of the constrained correlation when the frailty variances are different.

The results of the first simulation study indicated that the performance of the model depends on the proportion of type II interval-censored observations. In this sense, our estimates are consistent and in agreement with the ones presented by Hens et al. (2009) and Cattaert (2008). Furthermore, there is information gain (reflected by smaller empirical standard errors), mainly attributed to the interval-censored observations.

We implement successfully the correlated frailty model considering interval-censored data. In general, estimates were more consistent and more precise when the sample size and percentage of interval-censored observation increase.

The success rates also tend to increase when the sample size and/or percentage of interval-censored increase.

Most of the samples we classified as failures a solution could not be reached (the model did not converge) or the parameter estimates are complex numbers. In some other cases, the model converge but the hessian matrix was not positive definite or it was not produced by the software.

A high correlation between the parameter estimates may be an indication of identifiability issues. Several authors have pointed to the conditions under which the correlated frailty model is identifiable. Besides, Wienke (2011) suggests to include observed covariates in order to improve identifiability. Based on the results from Yashin and Iachine (Yashin et al.; 1995; Iachine; 2004) we know that the correlated frailty model is identifiable thanks to the additive decomposition, even without covariates and without parametric shape for the baseline hazard rate. Except in the case of current status data.

It has been recognized by Wienke (2011), there is a negative correlation between the variance and the correlation estimates. In our simulation studies we notice that the negative correlation it is not fixed and in fact decreases when the frailty correlation increases. Then, as it is expected, the correlated gamma frailty model can have serious identifiability issues when the correlation is on the border of the parameter space. Caution should be taken when the frailty correlation is smaller than 0.1 or closer to the smaller ratio between the frailty variances (if that ratio is larger than 0.1).

The results of our second simulation study are limited to the values we consider for σ_1 and σ_2. The model parameters and sample size were chosen to reflect the Amsterdam Cohort Studies example. We assume equal frailty variances, moderate and large difference between frailty variances. In total 20 scenarios were considered assuming different frailty correlation.

When frailty variances are equal some cautious interpretations should be made if the estimated correlation is lower 0.1 or larger than 0.9. In those cases we suggest to perform further sensitivity analyses to assess the reliability of the results.

If the variance ratio is equal to 0.75, we recommend to perform senstivity analyses when the frailty correlation is lower 0.1 or larger than 0.5. If the variance ratio is equal to 0.5 the recommendation is to perform sensitivity analyses regardless of the frailty correlation.

Our conclusions are restricted to the frailty parameters we choose as well as the baseline hazard function. It is possible that other options for baseline hazard functions and different values of frailty parameters lead to different results. Based on the information available for this study it is hard to perform extrapolations.

More research is needed in the area to implement a more complex baseline hazard such as the generalized gamma, the generalized F (Cox; 2008) or the one proposed by Sparling et al. (2006). Another option could be consider a semi-parametric approach where the univariate marginal survival is left unspecified. For the ACS, we implemented the baseline hazard proposed by Sparling et al. (2006), however we face major difficulties with the convergence of the model which to date remain unresolved.

Another aspect that deserves further attention is the impact of misspecification of model components. Further research should examine the misspecification of the baseline hazard, the frailty distribution and the type of frailty (for instance shared frailty versus correlated gamma frailty). Cattaert (2008) reports biases on the frailty and the baseline hazard parameters when a shared gamma frailty is fitted to a dataset coming from correlated gamma frailty model. Hens et al. (2009) presents similar results in terms of the mean estimates, also pointing at differences in the estimated variances.

The models implemented here do not include any covariate. An extension of the analyses in this chapter could include that aspect. An extensive simulation study that includes covariates is presented in Chapter 5 of Wienke (2011). The author adresses the impact of estimation strategies on the correlated frailty models.

The bivariate correlated frailty model can be extended when more than two observations per cluster are considered. In this case, Goethals (2011) presents an overview of a four-dimensional correlated gamma frailty models assuming different correlation structures. The first model is a shared frailty, and the other three are correlated gamma fralty models. She applied the four models to a dataset of mastitis, the inflammation of the udder of a dairy cow. The models also take into account the interval-censored nature of the data.

The four-dimensional frailty distribution imposes constrains on the correlation between the frailty variances. Even though the restrictions are reasonable for practical situations, we may face convergence problems when we are in the border of the parameter space. The impact can be different depending on the correlated gamma frailty model that is applied.

An extension of the correlated frailty models presented here may include the mathematical models presented in Chapter 6, where the survival function is based on the solution of mathematical models.

Chapter 6

Basic mathematical models for HCV and HIV

This chapter introduces two basic transmission model for HIV and HCV respectively, and a joint transmission model accounting for HIV and HCV co-infection; in the three models the transmission is exclusively attributed to sharing syringes. For HCV our models extends the model by Kretzschmar and Wiessing (2004) to account for multiple HCV infections and distinguishes between acute, chronic infected and susceptible individuals who cleared the virus. For HIV we consider two phases: infected with HIV and AIDS.

The joint model combines the two basic transmission models adding extra compartments for those individuals who got infected whit HCV virus in acute stage but afterwards clear the virus spontaneously.

6.1 Mathematical model for HIV

In the context of injecting drug users, to simplify the model we ignore transmission associated with sexual contacts and the time at risk is the duration of injection. The susceptible individuals (S_{HIV}) are infected with HIV at a per capita rate λ_{HIV}, which may depend on: the proportion of people in each disease phase, the syringe sharing rates per unit of time, the sharing partners, the transmission rates, and the proportion of syringes shared. Once an individual acquires HIV it enters into an infected stage (I_{HIV}) that last during ($1/\omega_{HIV}$), after which he/she develops

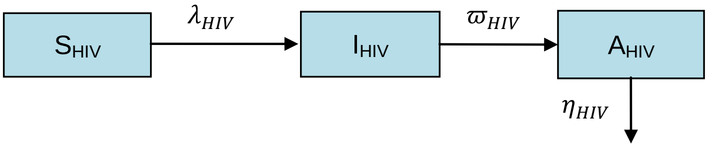


Figure 6.1: Flow diagram of the mathematical model for HIV. S_{HIV}: susceptible HIV, I_{HIV}: HIV positive, and A_{HIV}: AIDS

AIDS(A_{HIV}). The disease associated mortality is given by the parameter γ_{HIV}. The model can be represented by diagram in figure 6.1.

The HIV model assumming a closed population (no entry and exit rates) can be described by the following set of differential equations:

$$\begin{aligned}
\frac{dS_{HIV}(t)}{dt} &= -S_{HIV}(t)\lambda_{HIV} \\
\frac{dI_{HIV}(t)}{dt} &= \lambda_{HIV}S_{HIV}(t) - \omega_{HIV}I_{HIV}(t) \\
\frac{dA_{HIV}(t)}{dt} &= \omega_{HIV}I_{HIV}(t) - \eta_{HIV}A_{HIV}(t)
\end{aligned} \tag{6.1}$$

Solving the differential equation for $S_{HIV}(t)$, taking into account $S_{HIV}(0) = 1$:

$$\begin{aligned}
\frac{dS_{HIV}(t)}{dt}\frac{1}{S_{HIV}(t)} &= -\lambda_{HIV} \\
\int_0^t \frac{dS_{HIV}(s)}{ds}\frac{1}{S_{HIV}(s)}ds &= \int_0^t -\lambda_{HIV}ds \\
\ln S_{HIV}(t) - \ln S_{HIV}(0) &= -\Lambda_{HIV} \\
S_{HIV}(t) &= \exp(-\Lambda_{HIV})
\end{aligned} \tag{6.2}$$

$S_{HIV}(t)$ is the conditional survival function for HIV given the random effects. Assuming proportional hazards, the unconditional survival function $S^*_{HIV}(t)$ can be derived by taking the expectation with respect to the random effects Z. Then, the unconditional survival function coincides with the Laplace transform of the cumulative baseline hazard for HIV.

$S^*_{HIV}(t) = E\{S_{HIV}(t|Z)\} = E\{\exp(-t\lambda_{HIV}(Z))\} = L(\Lambda_{HIV})$

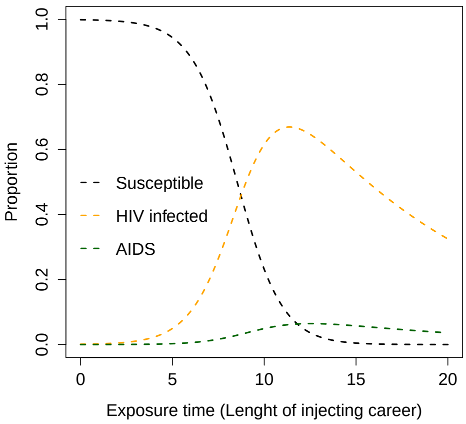


Figure 6.2: Model-based proportions for a HIV transmission model - one risk group. Disease stages: Susceptible, HIV infected, and AIDS

The figure 6.2 shows one fit of the HIV model. We include the proportion of susceptible, infected with HIV and AIDS. For this model fit, the proportion of susceptibles individuals starts to reduce slowly after 5 years of exposure. Whereas the number of HIV infected individuals steadily increases. The proportion of individuals in the AIDS stage shows small increase after ten years..

6.2 Mathematical model for HCV

The time at risk is given by the duration of injection (exposure time) denoted by t. The susceptible individuals (S_{HCV}) are infected with HCV at a per capita rate λ_{HCV}, which depends on: the proportion of people in each disease phase, the number of syringes shared, the transmission probabilities and the proportion of syringes shared. Once an individual acquires HCV he enters into an acute phase (I_{HCV}) that last during ($1/\omega_{HCV}$), after which he becomes a chronic carrier (CC) or clears the virus going back to the susceptibe class. The proportion of IDUs who do not clear spontaneously the virus is denoted by ψ. The individual can be re-infected later one assuming the same force of infection λ_{HCV}. The disease associated mortality is given by the parameter η_{HCV}. The model can be represented by the diagram in figure 6.3:

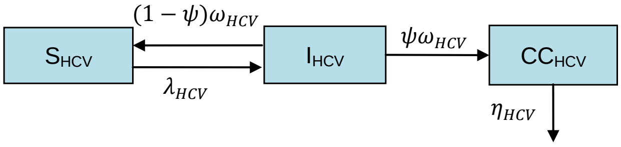


Figure 6.3: Flow diagram of the mathematical model for HCV. S_{HCV}: susceptible HCV, I_{HCV}: acute HCV infected, CC_{HCV}: chronic HCV carrier

The HCV model shown in figure 6.3 can be described by the following set of differential equations:

$$\begin{aligned} \frac{dS_{HCV}(t)}{d(t)} &= (1-\psi)\omega_{HCV}I_{HCV}(t) - \lambda_{HCV}S_{HCV}(t) \\ \frac{dI_{HCV}(t)}{d(t)} &= \lambda_{HCV}S_{HCV}(t) - \omega_{HCV}I_{HCV}(t) \\ \frac{dCC_{HCV}(t)}{d(t)} &= \psi\omega_{HCV}I_{HCV}(t) - \eta_{HCV}CC_{HCV}(t) \qquad (6.3) \end{aligned}$$

The solution of the system is far from trivial. We apply the general solution and the expression we obtain is given by:

$$
\begin{aligned}
S_{HCV}(t) \quad &= \quad \exp\left[t\left(\frac{-\lambda_{HCV}-\omega_{HCV}+\sqrt{\lambda_{HCV}^2+\omega_{HCV}^2+2\lambda_{HCV}\omega_{HCV}(2\psi-1)}}{2}\right)\right] \\
&* \quad \left(\frac{\omega_{HCV}-\lambda_{HCV}+\sqrt{\lambda_{HCV}^2+\omega_{HCV}^2+2\lambda_{HCV}\omega_{HCV}(2\psi-1)}}{2\sqrt{\lambda_{HCV}^2+\omega_{HCV}^2+2\lambda_{HCV}\omega_{HCV}(2\psi-1)}}\right) \\
&- \quad \exp\left[t\left(\frac{-\lambda_{HCV}-\omega_{HCV}-\sqrt{\lambda_{HCV}^2+\omega_{HCV}^2+2\lambda_{HCV}\omega_{HCV}(2\psi-1)}}{2}\right)\right] \\
&* \quad \left(\frac{\omega_{HCV}-\lambda_{HCV}-\sqrt{\lambda_{HCV}^2+\omega_{HCV}^2+2\lambda_{HCV}\omega_{HCV}(2\psi-1)}}{2\sqrt{\lambda_{HCV}^2+\omega_{HCV}^2+2\lambda_{HCV}\omega_{HCV}(2\psi-1)}}\right) \qquad (6.4)
\end{aligned}
$$

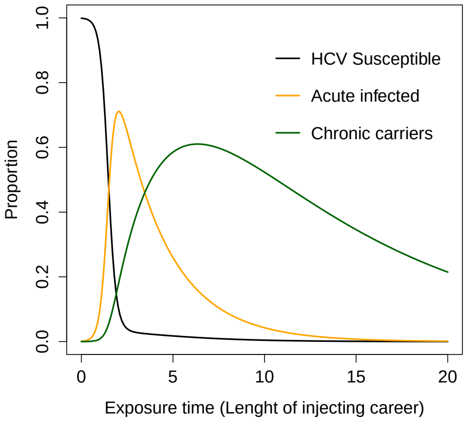


Figure 6.4: Model-based proportions for a HCV transmission model - one risk group. Disease stages: HCV Susceptible, Acute HCV infected, and Chronic HCV carrier

Figure 6.4 shows one fit of the HCV model based on a specific set of parameters. We include the proportion of susceptible, acute infected and chronic carriers. Based on the input parameters in the model, we observe the proportion of susceptibles individuals reduce sharply in the first three years of exposure. Additionally, the peak of acute infected individuals occurs immediately after, whereas the proportion of chronic carriers increase steadily until seven years of exposure and reduce slowly afterwards.

The HCV and HIV models can be combined to generated a joint transmission model for HCV/HIV co-infection as it is shown in the following section. We add an extra stage in the HCV model to differentiate the individuals who are susceptible for HCV but were previously infected by the virus.

6.3 Mathematical model for HCV/HIV co-infection

The joint transmission model we use is shown in Figure 6.5 and consists of the following disease stages for HCV: susceptible (S_{HCV}), acute infected (I_{HCV}), susceptible after spontaneously clearing the virus (S^{+}_{HCV}) (only possible after acute infection) and chronic carrier (CC_{HCV}). For HIV the stages are: susceptible (S_{HIV}), infected with HIV (I_{HIV}) and AIDS (A_{HIV}). An individual can be first infected by HCV and then by HIV or vice versa, then we consider a joint model with 18 compartments as described in Figure 6.5.

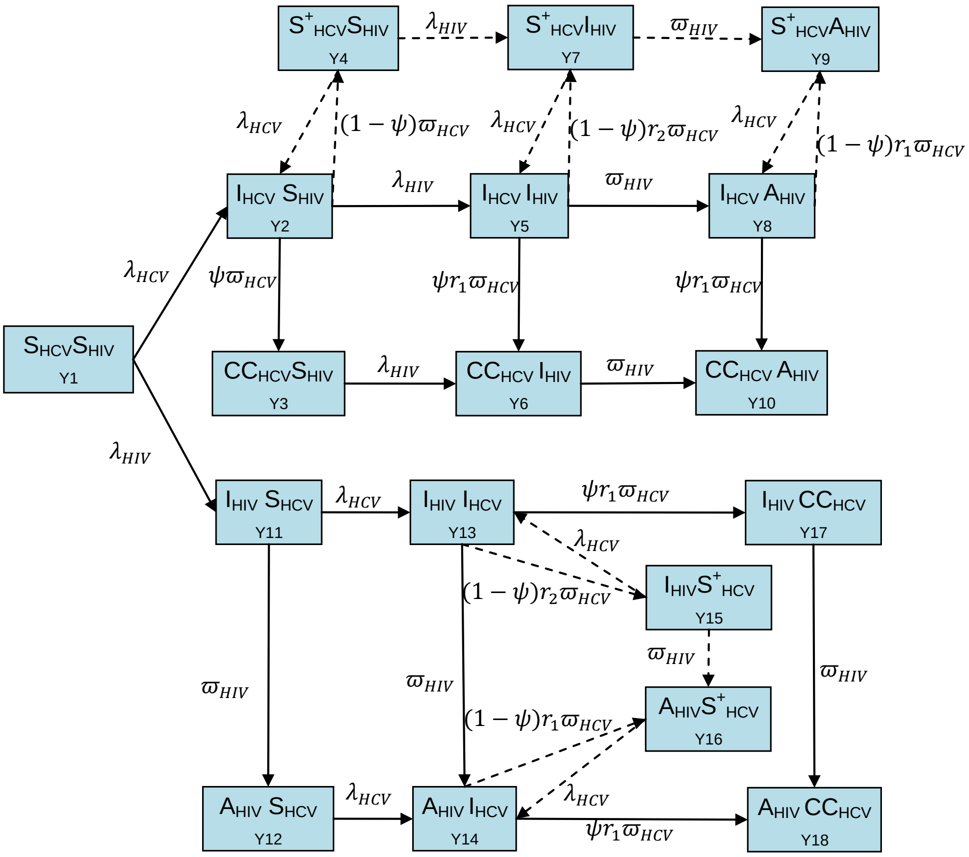


Figure 6.5: Flow diagram of the joint mathematical model for HCV/HIV for one risk group. S_{HCV}: susceptible HCV, S_{HIV} susceptible HIV, S^{+}_{HCV}: susceptible after a previous infection with HCV, I_{HCV}: acute HCV infected, I_{HIV}: HIV infected, CC_{HCV}: chronic HCV carrier and A_{HIV}: AIDS. Entry and exit rates are not shown.

In general the behavior of IDUs is very heterogeneous, some hardly ever share injecting equipment while others do share more frequently. In order to account for that heterogeneity we consider different models extending the single risk group model to a two risk group model (low and high risk) and to a three group risk model (low, moderate and high risk). We assume that once an individual starts injecting drugs, he/she belongs to a specific risk group (denoted by i) without switching groups. The proportion of individuals in a certain risk group is also represented as a model parameter. In figure 6.6 we only consider one risk group, however in the next chapter we consider one, two and three risk groups.

As in the previous models, the time at risk is given by the duration of injection (exposure time). The flow of individuals is as follows: the susceptible individuals are infected with HCV at a per capita rate λ_{HCV_i}, which depends on: the proportion of people in each disease phase, the number of syringes shared, the transmission probabilities and the proportion of syringes shared with members of other risk groups. An HCV positive individual remains acute infected during a period of average length $1/\omega_{HCV}$, whereas the average infectious period for HIV is denoted by $1/\omega_{HIV}$. The diseases specific mortality parameters are denoted by η_{HCV} and η_{HIV}, respectively and cessation rates by ν_i. We did not take into account HIV treatment due to the relatively low use of antiretrovirals (Camoni; 2011). A proportion of the acute infected individuals, ψ, becomes chronic carrier and the remaining proportion, $1 - \psi$, clears the virus spontaneously. Those who clear the virus, can also be reinfected at the same rate λ_{HCV_i}.

The probability of spontaneous clearance of the HCV virus is reduced in case of co-infection (Wasmuth; 2010; Micallef et al.; 2006). The high viral load for HCV could also imply rapid liver disease progression (Wasmuth; 2010). To account for the extra viral load in the presence of co-infection we include the term r_1 to accelerate the disease progression. Additionally, the term r_2 impacts the spontaneous clearance due to co-infection.

The dynamic model for one particular risk group is represented by the following set of ordinary differential equations:

$$
\begin{aligned}
\frac{dY_{1i}}{dt} &= B - Y_{1i}(\lambda_{HCV_i} + \lambda_{HIV_i} + \nu_i) \\
\frac{dY_{2i}}{dt} &= \lambda_{HCV_i} Y_{1i} + \lambda_{HCV_i} Y_{4i} - Y_{2i}(\omega_{HCV} + \lambda_{HIV_i} + \nu_i) \\
\frac{dY_{3i}}{dt} &= \psi\omega_{HCV} Y_{2i} - Y_{3i}(\lambda_{HIV_i} + \eta_{HCV} + \nu_i) \\
\frac{dY_{4i}}{dt} &= (1-\psi)\omega_{HCV} Y_{2i} - Y_{4i}(\lambda_{HIV_i} + \lambda_{HCV_i} + \nu_i) \\
\frac{dY_{5i}}{dt} &= \lambda_{HIV_i} Y_{2i} + \lambda_{HCV_i} Y_{7i} - Y_{5i}(\omega_{HIV} + \psi r_1 \omega_{HCV} + (1-\psi) r_2 \omega_{HCV} + \nu_i) \\
\frac{dY_{6i}}{dt} &= \lambda_{HIV_i} Y_{3i} + \psi r_1 \omega_{HCV} Y_{5i} - Y_{6i}(r_1 \eta_{HCV} + \omega_{HIV} + \nu_i) \\
\frac{dY_{7i}}{dt} &= (1-\psi) r_2 \omega_{HCV} Y_{5i} + \lambda_{HIV_i} Y_{4i} - Y_{7i}(\lambda_{HCV_i} + \omega_{HIV} + \nu_i) \\
\frac{dY_{8i}}{dt} &= \omega_{HIV} Y_{5i} + \lambda_{HCV_i} Y_{9i} - Y_{8i}(\eta_{HIV} + \psi r_1 \omega_{HCV} + (1-\psi) r_2 \omega_{HCV} + \nu_i) \\
\frac{dY_{9i}}{dt} &= (1-\psi) r_2 \omega_{HCV} Y_{8i} + \omega_{HCV} Y_{7i} - Y_{9i}(\eta_{HIV} + \lambda_{HCV_i} + \nu_i) \\
\frac{dY_{10i}}{dt} &= \psi r_1 \omega_{HCV} Y_{8i} + \omega_{HIV} Y_{6i} - Y_{10i}(\eta_{HIV} + r_1 \eta_{HCV_i} + \nu_i) \\
\frac{dY_{11i}}{dt} &= \lambda_{HIV_i} Y_{1i} - Y_{11i}(\omega_{HIV} + \lambda_{HCV_i} + \nu_i) \\
\frac{dY_{12i}}{dt} &= Y_{11i}\omega_{HCV} - Y_{12i}(\eta_{HIV} + \lambda_{HCV_i} + \nu_i) \\
\frac{dY_{13i}}{dt} &= \lambda_{HCV_i}(Y_{11i} + Y_{15i}) - Y_{13i}(\omega_{HIV} + \psi r_1 \omega_{HCV} + (1-\psi) r_2 \omega_{HCV} + \nu_i) \\
\frac{dY_{14i}}{dt} &= \lambda_{HCV_i}(Y_{12i} + Y_{16i}) + \omega_{HIV} Y_{13i} - Y_{14i}(\psi r_1 \omega_{HCV} + (1-\psi) r_2 \omega_{HCV} + \eta_{HIV} + \nu_i) \\
\frac{dY_{15i}}{dt} &= (1-\psi) r_2 \omega_{HCV} Y_{13i} - Y_{15i}(\lambda_{HCV_i} + \omega_{HIV} + \nu_i) \\
\frac{dY_{16i}}{dt} &= \omega_{HIV} Y_{15i} + (1-\psi) r_2 \omega_{HCV} Y_{14i} - Y_{16i}(\lambda_{HCV_i} + \eta_{HIV} + \nu_i) \\
\frac{dY_{17i}}{dt} &= \psi r_1 \omega_{HCV} Y_{13i} - Y_{17i}(\omega_{HIV} + r_1 \eta_{HCV} + \nu_i) \\
\frac{dY_{18i}}{dt} &= \psi r_1 \omega_{HCV} Y_{14i} + \omega_{HIV} Y_{17i} - Y_{18i}(\eta_{HIV} + r_1 \eta_{HCV} + \nu_i)
\end{aligned}
$$

It is far more complex to find a general solution for this model, therefore we only focus on the solution of the model based on integration routines. To illustrate the complexity of the model we refer to De Vos et al. (2012), where a detailed analysis of a similar model is provided assuming HCV equilibrium.

For the joint model, we consider two definitions for the force of infection. The first corresponds to the model proposed by Kretzschmar and Wiessing (2004) and assumes the same transmission rate for every syringe-sharing; while the second definition, following Garnett and Anderson (1994) takes the number of partners into account.

6.3.1 A first definition of the force of infection

The force of infection in each of the risk groups i (λ_{HCV_i}), is a function of the sharing rates (κ_i for the i-th risk group), the transmission rate per syringe-sharing event at each infection stage ($b_{HCV_I}, b_{HCV_{CC}}$), the proportion of syringes shared with members of other risk groups (m_{ij}: mixing proportions among the risk groups), and the proportion of infected individuals in each risk group for every disease stage (endemic prevalences: $\text{Prev}_{HCV_{Ij}}$, $\text{Prev}_{HCV_{CCj}}$). For this parametrization the force of infection is given by:

$$\lambda_{HCV_i} = \kappa_i \sum_{j=1}^{R} m_{ij} \left(b_{HCV_I} \text{Prev}_{HCV_{Ij}} + b_{HCV_{CC}} \text{Prev}_{HCV_{CCj}} \right), \tag{6.5}$$

with R the number of risk groups. Note that $\sum_j m_{ij} = \sum_i m_{ij} = 1$. Similarly, we define the force of infection for HIV by:

$$\lambda_{HIV_i} = \kappa_i \sum_{j=1}^{R} m_{ij} \left(b_{HIV_I} \text{Prev}_{HIV_{Ij}} + b_{HIV_A} \text{Prev}_{HIV_{Aj}} \right). \tag{6.6}$$

Two issues arise when using this definition of the force of infection. First, it assumes that every syringe-sharing event has the same transmission rate (although, it does depend on the stage of infection); when in fact the transmission depends not only on the syringe-sharing event but also on the number of sharing partners. Secondly, it is difficult to quantify the degree of mixing, because this depends on the parameters m_{ij} (probability that a member of risk group i shares syringes with a member of group j). Although the extremes are fully assortative (for $i = j$, $m_{ij} = 1$

whereas for $i \neq j, m_{ij} = 0$) or fully disassortative (for $i = j$, $m_{ij} = 0$), it is not straightforward to interpret the degree of mixing for different values of m_{ij}.

6.3.2 A second definition of the force of infection

Following Garnett and Anderson (1994), the force of infection may account for the number of sharing partners per unit of time (n_i). An individual may be infected by a syringe partner in a certain disease stage in one of τ contacts. The probability of infection per unit of time is then given by $B_{HCV_k} = 1 - (1 - b_{HCV_k})^{\tau}$, where $k = I$ for infected and $k = CC$ for chronic carriers. The second definition for the force of infection is

$$\lambda_{HCV_i} = \sum_{j=1}^{R} n_i m_{ij} \left(B_{HCV_I} \mathrm{Prev}_{HCV_{Ij}} + B_{HCV_{CC}} \mathrm{Prev}_{HCV_{CCj}} \right). \tag{6.7}$$

The mixing patterns are given by $m_{ij} = \frac{T_j n_j}{\sum T_i n_i}(1 - \nu) + \nu \delta_{ij}$ where δ_{ij} is a dirac-delta function equal to one if $i = j$ and zero otherwise. The parameter ν denotes the degree of assortative mixing: $\nu = 0$ corresponds to random mixing and $\nu = 1$ to fully assortative mixing. T_i denotes the number of IDUs in the i-th risk group. In an equivalent way the force of infection for HIV is defined.

6.3.3 Illustration of the mathematical model for HCV/HIV co-infection

An example of a joint model is presented on figure 6.6. Here we assumed only one risk group and the first definition of the force of infection. Since the information about the serostatus for HCV and HIV is available, in figure 6.6 we include the proportion of susceptible for both infection, the proportion of positive for at least one of the viruses and the proportion of individuals positive for both viruses.

The proportion of individuals susceptible for both diseases decreases steadily over the exposure time until about 15 years of exposure, then remains almost constant. On the other hand, the proportion of individuals positive for HIV and negative for HCV increases reaching a peak around 12 years, after dimishes due to the increas of the individuals who become positive for HCV.

The fitted values presented in this chapter for the different models are examples, they have not been calibrated to any real data example and therefore do not reflect any real life trend. In the following chapter we apply a seven step procedure to callibrated a model to the data.

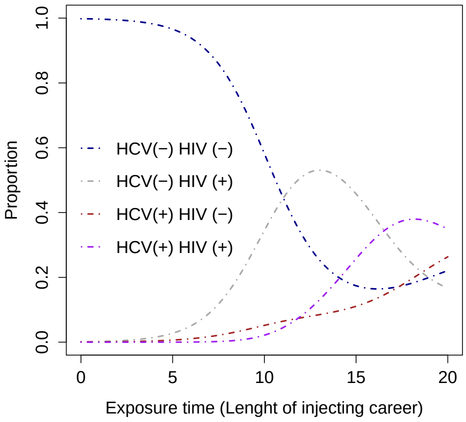


Figure 6.6: Model-based proportions for a joint HCV/HIV transmission model - one risk group with the first definition of the force of infection

This chapter presented two basic transmission models for HCV and HIV and a joint model that accounts simultaneously for the transmission of both viruses. The last model will be considered in the next chapter, where we calibrate the joint model to two different datasets (the Vedette and the Itinere) and assess the model from a statistical perspective.

Chapter 7

A mathematical model for HIV/HCV co-infection and its assessment from a statistical perspective

In the IDU setting serological data constitutes one valuable source of information for understanding HIV and HCV epidemiology. Serological data comprises the serostatus for each individual and self-reported data on the duration of injection can be considered as a measurement of the time at risk. Several studies have addressed modelling the force of infection and co-infection for HCV and HIV in IDU populations based on serological data (Del Fava et al.; 2011; Sutton et al.; 2006, 2008). In these models, the force of infection has been estimated as a function of the exposure time and a term reflecting the individual heterogeneity in the acquisition of the virus; a frailty term. Those models are very flexible and are useful for identifying risk factors for the infection at hand; however, they do not focus on the transmission process of the viruses (Garnett et al.; 2011).

In infectious disease epidemiology, mathematical models are used to model mechanistically the spread of certain diseases and to assess the impact of intervention policies. In mathematical models, the force of infection reflects the transmission process explicitly using the effective contact rate, transmission probabilities and the number of infected individuals. One of the first examples of a joint dynamic model for HCV/HIV aimed at assessing the cost-effectiveness of needle and syringe programmes (Vickerman et al.; 2008). A similar model was used to explore the hypothesis of a low prevalence of HCV despite the high rates of sharing needles/syringes, and to project future HIV/HCV co-infection while assessing the impact of interventions (Vickerman et al.; 2009). Recently, Vickerman et al. (2011) used a mathematical model to understand the trends in HIV and HCV prevalences, determining epidemiological profiles and De Vos et al. (2012) investigated the relationship between the prevalences and the heterogeneity of injecting risk behavior.

This chapter build up on the joint transmission model described in Chapter 6 and the methods and results based on the Vedette dataset have been published on Castro Sanchez et al. (2013).

Here, we present a procedure to assess the model from a statistical perspective, using bivariate serological data from both infections. For HCV our model extends the model by Kretzschmar and Wiessing (2004) to account for multiple HCV infections and distinguishes between acute, chronic infected and susceptible individuals who cleared the virus. For HIV we consider two phases: infected with HIV and AIDS.

In model development, the estimation or 'calibration' of the model to data and the assessment of the sensitivity to capital parameters and uncertainty in projections are key aspects of any analysis. Several authors have pointed to the importance of rigorous sensitivity analysis to model the dynamics of a given infectious disease (Bilcke et al.; 2011; Garnett et al.; 2011; Jit and Brisson; 2011; Okais et al.; 2010; Vanni et al.; 2011). For example, Vanni et al. (2011) and Bilcke et al. (2011) provide a methodological framework to account for different sources of uncertainty and to calibrate the model to observational data.

The model assessment of the proposed joint transmission model includes the estimation of the parameters or the calibration of the model to data, the quantification of model uncertainty and model selection, the assessment of the statistical variability, and analyses of the model parameters in the high-dimensional parameter space. Although maximum likelihood (ML) theory is not entirely applicable in our overparameterized setting, ML concepts are used together with modern statistical techniques developed for high-dimensional problems. In what follows, model uncertainty refers to the joint mathematical model itself and the option to consider different alternatives for certain components of the model. We will consider two components of special interest: i) the number of risk groups (one, two or three), and ii) the formulation of the force of infection (two options). As it is uncertain which option is best, the models corresponding to the different options are examined and compared with statistical measures. Six different models will be compared and empirical evidence will be used to select the best model.

7.1 Overview of the datasets

The deterministic model was calibrated separately to two different datasets. Here we briefly described them, some exploratory data analyses are provided on Chapter 1.

7.1.1 The Vedette IDU data

The dataset comes from a longitudinal study of heroin users in Italy. The main goal was to evaluate the effectiveness of treatments provided by the National Health Services. Antibody levels were determined for HIV, HCV and Hepatitis B. The subset used for the analyses presented in this manuscript corresponds to the Piedmont region totalling 2,628 IDUs.

7.1.2 The Itinere IDU data

The dataset corresponds to the baseline measurement of a cohort study of heroin drug users in Spain. The main goals were to monitor the health impact of drug use and to identify risk factors. Determination of antibodies was made for HIV, HCV, Hepatitis B, and Human T-lymphotropic virus, using dried blood samples.

7.2 Description of the statistical methods for model assessment

The deterministic joint model for HCV/HIV co-infection has been introduced in Chapter 6. Here, we present the techniques to assess the model from a statistical perspective before turning to the results section.

7.2.1 Identification and estimation of the model parameters

An overview of all model parameters is shown in Table 7.1 and Table 7.2: 14 parameters for the single risk group model, 19 for the two risk group and 26 for the three risk group model, regardless of which definition is used for the force of infection. The right column of Table 7.1 and Table 7.2 gives the range of plausible values for each parameter, based on literature as far as available (only for Table 7.1). Denote θ the vector of all parameters

$$\theta = (b_{HCV_I}, b_{HCV_{CC}}, b_{HIV_I}, b_{HIV_A}, \dots),$$

and Θ the Cartesian product space of all intervals of plausible ranges. We assume that the estimate $\hat{\theta}$ that maximizes the likelihood given the data belongs to Θ.

However, standard application of ML-estimation (with Newton-Raphson-like numeric approaches based on derivatives) is not applicable in this case: the model is defined through an extensive set of differential equations with 13 or more parameters and the serological data provide no direct information (through e.g. summary or sufficient statistics) about any of the individual parameters. For such an overparameterized model, one can expect several combinations of parameter values (solutions) to lead to similar (log)likelihood values (flat likelihood surface).

The approach we take is as follows. For every parameter we draw values from a uniform distribution (using the ranges from Table 7.1 and 7.2) and select 500,000 parameter sets using Latin Hypercube Sampling (LHS). LHS is an efficient technique to generate parameter values from the high dimensional space Θ (Stein; 1987; Blower and Dowlatabadi; 1994; Hoare et al.; 2008). In order to "estimate" the parameter vector θ (or to calibrate the model, i.e. to find the value of θ most supported by the observed data), we use ML concepts and measures. In our case, the estimate for θ corresponds to a mathematical model that reflects as closely as possible the trends in the observed joint prevalences for HCV and HIV as a function of exposure time.

In cross-sectional serological studies the serostatus for HCV and HIV as well as

Table 7.1: List of all parameters in the joint mathematical model for HCV/HIV: part I.

Model parameter definition	Range	References
Transmission rate per syringe-sharing event of HCV		
In acute stage (b_{HCV_I})	0.005 - 0.2	Yazdanpanah et al. (2005); De Carli et al. (2003)
In chronic stage ($b_{HCV_{CC}}$)	0.005 - 0.2	
Transmission rate per syringe-sharing event of HIV		
In infection stage (b_{HIV_I})	0.001 - 0.1	Baggaley et al. (2006); White et al. (2007)
In AIDS stage (b_{HIV_A})	0.000 - 0.1	
Duration in every stage HCV		
Duration in acute infection stage HCV ($1/\omega_{HCV_I}$)	4-6 months	Wasmuth (2010); Hutchinson et al. (2006a,b); Deuffic-Burban et al. (2004)
Duration as chronic carrier of HCV ($1/\eta_{HCV_{CC}}$)	$\geq$ 10 years	
Duration in every stage HIV		
Duration in the infection stage HIV ($1/\omega_{HIV_I}$)	3-30 years	UNAIDS (2010); Vickerman et al. (2009); Garcia de la Hera et al. (2004); Jarrin et al. (2008); Serraino et al. (2009); Todd et al. (2007)
Duration as AIDS ($1/\eta_{HIV_A}$)	1 - 10 years	
Spontaneous clearance and co-infection		
Proportion of people who do not spontaneously clear the HCV in acute stage ψ	0.4 - 1.0	Wasmuth (2010); Micallef et al. (2006); Hutchinson et al. (2006a,b); Deuffic-Burban et al. (2004)
Acceleration factor for disease progression of HCV in presence of co-infection r_1	1.0 - 4.0	
Factor to modify spontaneous clearance of HCV in presence of co-infection r_2	0.0 - 4.0	
Entry and exit rate parameters		
Entry rate (E)	0.0 - 0.2	Bouhnik et al. (2004); Galai et al. (2003); Steensma et al. (2005)
Exit rates, including cessation (ν_i), assuming a injecting career length at least 5 years	0.0 - 0.2	

Table 7.2: List of all parameters in the joint mathematical model for HCV/HIV: part II.

Model parameter definition	Range
Behavioral parameters	
Mixing proportions m_{ij}	0-1
Degree of assortativeness ν	0-1
Sharing syringe rate per trimester κ_1 (one RG)	1-300
Sharing syringe rate per trimester low risk group κ_1 (2 RGs)	1-100
Sharing syringe rate parameter $f_1 = \kappa_2/\kappa_1$ (2 RGs)	1-50
Sharing syringe rate low risk group κ_1 (3 RGs)	1-100
Sharing syringe rate parameter $f_2 = \kappa_2/\kappa_1$ (3 RGs)	1-50
Sharing syringe rate parameter $f_3 = \kappa_3/\kappa_2$ (3 RGs)	1-20
Mixing proportions m_{ij}	0-1
Degree of assortativeness ν	0-1
Number of sharing partners partners n_1 (1 RG)	1-50
Sharing syringe events per partner τ (1 RG)	1-20
Number of sharing partners low RG n_1 (2 RGs)	1-50
Number of sharing partners parameter $g_1 = n_2/n_1$ (2 RGs)	1-20
Sharing syringe events per partner τ (2 RGs)	1-20
Number of sharing partners low RG n_1 (3 RGs)	1-50
Number of sharing partners parameter $g_2 = n_2/n_1$ (3 RGs)	1-20
Number of sharing partners parameter $g_3 = n_3/n_2$ (3 RGs)	1-20
Sharing syringe events per partner τ (3 RGs)	1-10

the self-reported duration of injection (exposure time) are available for each participant. An individual with a certain exposure time is either positive for both viruses, negative for both or positive for only one of them, leading to a multinomial distribution. Let p_{00t}, p_{01t}, p_{10t} and p_{11t} denote the proportion of IDUs with an injecting career length t that are uninfected, infected by HIV and not by HCV, infected by HCV and not by HIV, and infected by both HIV and HCV, respectively. Given a specific parameter vector θ, we use integration routines to solve the differential equations, from which the corresponding values for $p_{00t}(\theta)$, $p_{01t}(\theta)$, $p_{10t}(\theta)$ and $p_{11t}(\theta)$ are derived. Denote T the set of distinct observed exposure times. The multinomial likelihood function is given by

$$L(\theta|\{y_{rst}\}_{t\in T}) = \sum_{t\in T}(p_{00t}(\theta))^{y_{00t}}(p_{01t}(\theta))^{y_{01t}}(p_{10t}(\theta))^{y_{10t}}(p_{11t}(\theta))^{y_{11t}}, \tag{7.1}$$

where y_{rst} is the observed number of individuals for the corresponding combination $r = 0,1; s = 0,1$ with exposure time d. According to the ML principle, the higher the likelihood $L(\theta)$ the better the data supports the parameter vector θ. In the next two subsections we describe how the bootstrap can be used the get additional guidance in selecting a final model and in assessing variability. In a last subsection we discuss different ways to further evaluate the model parameters from a statistical point of view.

7.2.2 Model selection

In general, there exist several statistical criteria for model selection. Here we focus on Akaike's information criterion (AIC, see e.g. Akaike (1974)), which rewards goodness of fit (measured by the likelihood) but also penalizes for complexity. The AIC value for each candidate model with likelihood L is defined as

$$\mathrm{AIC}_L = -2\log\{\max_\theta L(\theta|\{y_{rst}\}_{t\in T})\} + 2 \times (\text{number of model parameters}),$$

and the candidate model with the smallest AIC-value is considered to be the best model. To assess which model would perform best across several samples (of the same size), we extend the model selection exercise with the application of the non-parametric bootstrap (i.e. resampling the data, see e.g. Davison and Hinkley (1997)). Using this nonparametric bootstrap we address model selection uncertainty. The individuals in the sample are resampled (with replacement) in order to get a bootstrap sample (of the same size). For each bootstrap sample b ($b = 1, ..., B$), the likelihood function $L^{(b)}(\theta|\{y^*_{rst}(t)\}_{t\in T})$ (where $y^*_{rst}(b)$ are the observed counts in the bootstrap sample) is maximized (again over the 500,000 LHS-generated parameter vectors) for each candidate model, leading to B AIC values $\mathrm{AIC}^*_L(b)$ for each candidate model. The model that is most often selected as best model over all bootstrap samples, is retained as the final model.

7.2.3 Assessing variability

As an alternative to ML-standard errors, we propose to apply again the nonparametric bootstrap as a method to quantify sampling variability. We follow the same bootstrap approach as in the previous subsection but limited to the final model, and leading to B estimates $\hat{\theta}^*(b)$. The number of times each different LHS-generated parameter vector is selected as maximizer of the likelihood is calculated. This frequency distribution characterizes the variability.

7.2.4 Statistical evaluation of model parameters

As mentioned in previous sections, standard ML estimation and inference cannot be applied. In this section we do not focus on the value of θ that maximizes $L(\theta|\{y_{rst}\}_{t\in T})$ but to the subset $\Theta_{1\%} \subset \Theta$ that corresponds to the top 1% of highest values of $L(\theta|\{y_{rst}\}_{t\in T})$. This subset $\Theta_{1\%}$ is then examined in various ways to get further insights in the parameters, their association and their impact on the model. We used several standard multivariate techniques including principal component analysis, cluster analyses, etc as well as some recently developed methods shown to be useful in the context of data mining and analysis of genetic data. We limit ourselves to describing the results of a specific version of classification trees, generalized additive models and a very recently developed association measure, the maximum information coefficient; but we start with exploring the parameter space univariately.

7.2.4.1 Univariate explorative analysis of $\Theta_{1\%}$

A parameterwise graphical comparison of the density f_U of the initial uniform distribution (range as in Table 7.1) with the density $f_{\Theta_{1\%}}$ corresponding to the values within the subspace $\Theta_{1\%}$, allows us to indicate which parameters are (highly) influenced by the data (with a peaked unimodal density $f_{\Theta_{1\%}}$). It resembles the comparison of prior and posterior densities in a Bayesian approach.

7.2.4.2 Activity region finder

Amaratunga and Cabrera (2004) developed a recursive partitioning classification tree (ARF = Activity Region Finder) to characterize a subset of cases that respond positively or have high response values (such as highly expressed genes in microarray experiments). Consider each of the 500,000 LHS-generated parameter vectors θ as input (explanatory variables) and the corresponding binary indicators that equal 1 if $\theta \in \Theta_{1\%}$ and 0 otherwise as output (response variable). Applying ARF on these 500,000 'observations' gives insights in which patterns and structures are characterizing the subspace $\Theta_{1\%}$. The procedure is available as an R package (downloadable from one of the author's website Amaratunga and Cabrera (2004)).

7.2.4.3 Generalized additive models

Generalized additive models (GAM, see e.g. Wood (2006)) are well-established flexible regression models. In essence, it is a generalized linear model with a linear predictor involving a sum of smooth functions of the covariates. Here we apply

GAM on the same input-output setting as used with the ARF. More precisely it relates the linear combination

$$\alpha + f_1(b_{HCV_I}) + f_2(b_{HCV_{CC}}) + f_1(b_{HIV_I}) + f_2(b_{HIV_A}) + \dots, \tag{7.2}$$

where f_j's are cubic splines functions to the probability that the corresponding likelihood belongs to the top 1%. Finally, as in logistic regression, a logit link is used. This is again an alternative and flexible way to gain further insights in which patterns of parameter vectors θ characterize the subspace $\Theta_{1\%}$.

7.2.4.4 Maximal information coefficient

Based on our sampling algorithm we do not expect to see any association between the parameters in the full space Θ. We however expect some association structure between the parameters in the subset $\Theta_{1\%}$. To quantify the association one can use the Maximal Information Coefficient (MIC) proposed by Reshef et al. (2011). The MIC captures a wide range of associations not limited to functional relationships (generality property). In addition it gives the same score to equally noisy relationships (equitability property). The MIC ranges between 0 and 1, the higher the value the stronger the relationship between the variables. There is no closed form expression to calculate the MIC, but software provided by the authors of Reshef et al. (2011) can be used.

7.3 Application to the Vedette dataset

The AIC values in the second column of Table 7.3 suggest that the model using the first definition of the force of infection with two risk groups fits best to the data, with only a very small difference in favor of the second definition of the force of infection. Of course, for a similar sample of the same size the result could be different. Therefore, we motivate our final decision on the results of a nonparametric bootstrap exercise (with $B = 500$ runs). The last column of Table 7.3 indicates that the two-risk group model with the first definition for the force of infection is selected more often (41%) than any other model; and therefore we select that one as our final model. The parameter values of this final model at the maximal multinomial likelihood are listed in Table 7.4. As mention before standard ML inference cannot be applied, therefore, we cannot report standard errors in Table 7.4.

Panels (a) and (b) of Figure 7.1 show the model-based fits for p_{00t}, p_{01t}, p_{10t}, p_{11t} considering one, two and three risk groups with the first definition of the force of infection on top of the observed probabilities. The fits follow the data pattern reasonably well for short and medium exposure times, whereas for longer exposure times all three models deviate from the observed data, specially when only one risk group is considered in the model. It is worth noting that the limited number of individuals at longer exposure times may explain the poor fit in this region. A closer inspection confirms that the two-risk group model fits better to the data. Panel (c) depicts the marginal probabilities $p_{\cdot 1t}$ and $p_{1\cdot t}$. The fitted curves p_{11t} and p_{01t} for the two-risk group do however exhibit a peculiar bump at an exposure time of 5 years.

7.3.1 Assessing variability

It turns out that only 29 out of 500,000 (that is only 0.01%) parameter vectors θ were associated with the highest likelihood values $L^{(b)}(\theta|\{y^*_{rst}(b)\}_{t\in T})$ in at least one of the $B = 500$ bootstrap samples.

The model-based prevalence curves for HCV and HIV for these 29 parameter vectors are shown in Figure 7.2 (gray lines). Furthermore, only six of them were selected in 74.4% of the bootstrap samples. The colored lines and symbols represent the model-based prevalence curves corresponding to the three most often selected parameter sets (29.4%, 17.4% and 9.6%, in total 56.4% of all bootstrap samples). Not unexpectedly, the model-based curves show more variability for exposure times more than 15 years (due to less data in that region) Additionally, the model-based curves for HCV are less variable compared with those for HIV.

Table 7.3: Vedette IDU dataset. Results of all models combining one to three risk groups with two definitions for the force of infection: value of the multinomial likelihood, AIC value, and the number of times that the respective combination was selected as best model, out of 500 bootstrap resamples.

# Risk groups	Multinomial likelihood	AIC based on Multinomial	Boostrap Frequency
	First definition for the force of infection		
One	-1,071.723	2,171.447	0
Two	-1,050.930	2,139.860	205
Three	-1,047.111	2,146.222	31
	Second definition for the force of infection		
One	-1,056.612	2,143.224	57
Two	-1,056.952	2,151.904	65
Three	-1,048.735	2,141.470	142

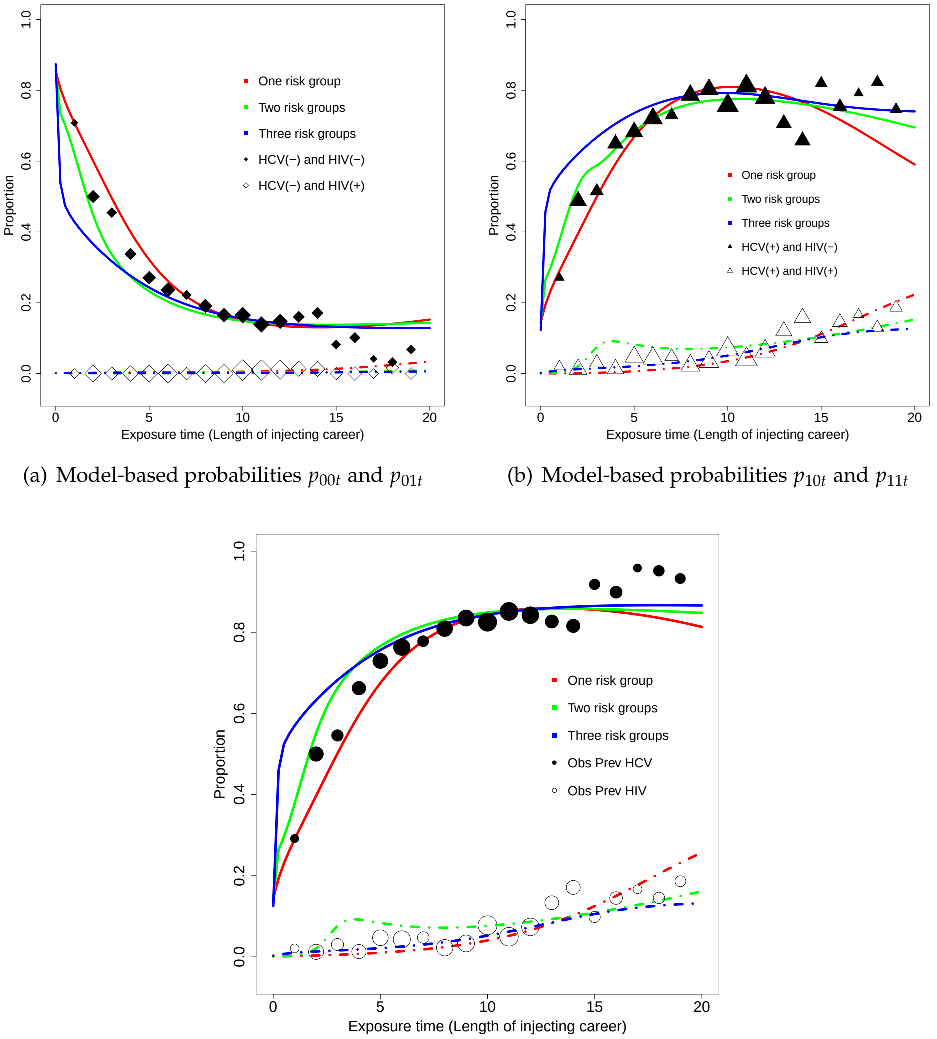
For the six parameter vectors θ selected by 74.4% of the bootstrap samples, the rate of sharing for the low risk groups is between 5.7 and 23.7 syringes per three months period and the ratio between the rates of sharing for both groups (κ_2/κ_1) is between 31 and 48.3. Thus, the highest risk group shares between about thirty and fifty times more syringes than the low risk group. The ratio between the transmission rates per syringe-sharing event in the acute infected stage of HCV against infected stage of HIV is between 1.23 and 29.6, indicating HCV is more transmissible than HIV. The percentage of individuals in the low risk group varies between 29% and 95%.

7.3.2 Univariate explorative analysis of $\Theta_{1\%}$

We limit ourselves to the sharing rates parameters ($\kappa_1, \kappa_2/\kappa_1$), the transmission rate (b_{HIV_I}), the proportion of individuals in the low risk group ($prop$) the mixing parameters (m_{11} and m_{22}), as the results were most pronounced for these parameters and because they will repeatedly play a more prominent role in the further analyses. Figure 7.3 demonstrates more peaked densities $f_{\Theta_{1\%}}$ (as compared to f_U) for the sharing rate κ_1 and the transmission rate b_{HIV_I} and less pronounced peaks of mass concentration for $\kappa_2/\kappa_1, prop, m_{11}$, and m_{22}. Note also that the location of the mode of each marginal density plot $f_{\Theta_{1\%}}$ deviates substantially from the multidimensional location of the mode of the multinomial likelihood (Table 7.4), indicating that the parameters are not independent in the full dimensional parameter space. The dependence structure of the parameters is further investigated in the next sections.

Table 7.4: Vedette IDU dataset. Parameter values that maximize the multinomial likelihood for the model with two risk groups and the first definition for the force of infection.

Model parameter	Value
Transmission rate of HCV in acute stage (b_{HCV_I})	0.133
Transmission rate of HCV in chronic stage ($b_{HCV_{CC}}$)	0.032
Transmission rate of HIV in infection stage (b_{HIV_I})	0.005
Transmission rate of HIV in AIDS stage (b_{HIV_A})	0.046
Duration in the acute infection stage HCV ($1/\omega_{HCV_I}$)	5.8 months
Duration as chronic carrier of HCV ($1/\eta_{HCV_{CC}}$)	28.6 years
Duration in the infection stage HIV ($1/\omega_{HIV_I}$)	3.7 years
Duration in AIDS ($1/\eta_{HIV_A}$)	1.9 years
Proportion of not spontaneously clearance of the HCV (ψ)	0.43
Acceleration factor for HCV disease progression due to co-infection (r_1)	2.23
Factor affecting the spontaneous clearance probability due to co-infection (r_2)	1.15
Percentage of IDUs in the low risk group (two RGs)	63.4%
Sharing syringe rate low risk group per trimester κ_1 (two RGs)	23.69
Factor increasing sharing syringe rate high risk group f_1 (two RGs)	46.653
Mixing proportion m_{11}	0.993
Mixing proportion m_{22}	0.288
Entry rate (E)	0.059
Injecting career length low risk group ($1/\nu_1$)	24.4 years
Injecting career length high risk group ($1/\nu_2$)	5.7 years



(a) Model-based probabilities p_{00t} and p_{01t}

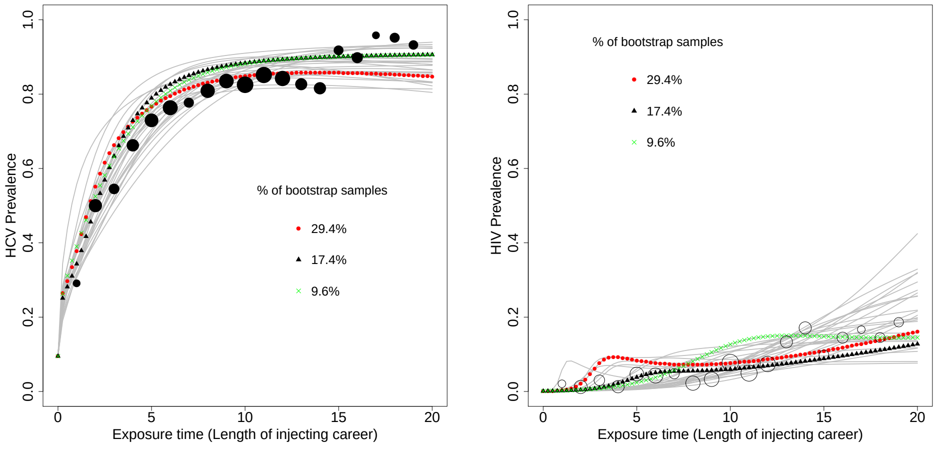
(b) Model-based probabilities p_{10t} and p_{11t}

(c) Model-based probabilities $p_{r.t}$ and $p_{.st}$

Figure 7.1: Vedette IDU dataset. Observed proportions together with model-based joint and marginal prevalence fits for the model with one, two and three risk groups, combined with the first definition for the force of infection. The size of the symbols is proportional to the observed number of individuals at each exposure time.

7.3.2.1 Activity region finder

Although all parameters were included, only the sharing rates parameters κ_1 and $f_1 = \kappa_2/\kappa_1$, the transmission rates per syringe-sharing event b_{HIV_I} and $b_{HCV_{CC}}$, the



(a) HCV prevalence

(b) HIV prevalence

Figure 7.2: Vedette IDU dataset. Bootstrap results - observed prevalences (circles) and model-based prevalences. The gray lines correspond to parameter sets that lead to the highest likelihood value for at least one bootstrap sample. The other colored lines indicate the parameter sets that produce the highest likelihood values more frequently.

entry rate (E), the proportion of people in the low risk group $prop$ and the proportion of individuals who become chronic carriers after being acute infected with HCV (ψ) take part in the recursive partitioning classification process.

The sharing rate in the low risk group κ_1 is responsible for the first split, followed by splits using the transmission rate per syringe-sharing event $b_{HCV_{CC}}$. In total 37 regions were classified as high activity regions. The most relevant regions can be described as follows. A sharing rate in the low risk group (κ_1) between 1.24 and 3.88 combined with a transmission rate per syringe-sharing event of HCV at chronic stage ($b_{HCV_{CC}}$) between 0.087 and 0.2, with proportion of individuals who become chronic carriers (after being HCV acute infected) ψ between 0.73 and 0.995, with a transmission rate per syringe-sharing event of HIV at infectious stage (b_{HIV_I}) produce on average log-likelihood values of -1472.9. Another significant region is obtained combining κ_1 between 0.132 and 6.176 with $b_{HCV_{CC}}$ between 0.087 and 0.2, and with $f_1 = \kappa_2/\kappa_1$ between 1.35 and 10.53; here the average of the log-likelihood values is -1557.78.

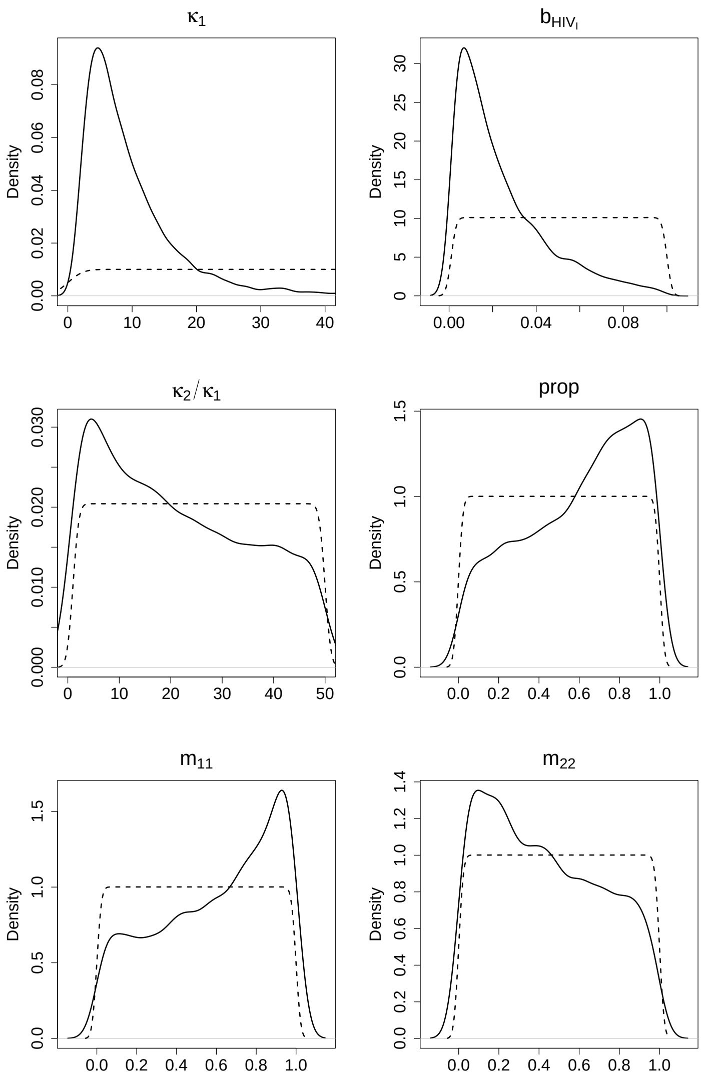


Figure 7.3: Analysis of the Vedette IDU dataset. Overlaid smoothed density plots $f_{\Theta_{1\%}}$ and f_U for the sharing rates $\kappa_1, \kappa_2/\kappa_1$ and the transmission rate b_{HIV_I}, the proportion of individuals in the low risk group $prop$, and the mixing parameters m_{11} and m_{22}.

7.3.2.2 Generalized additive models

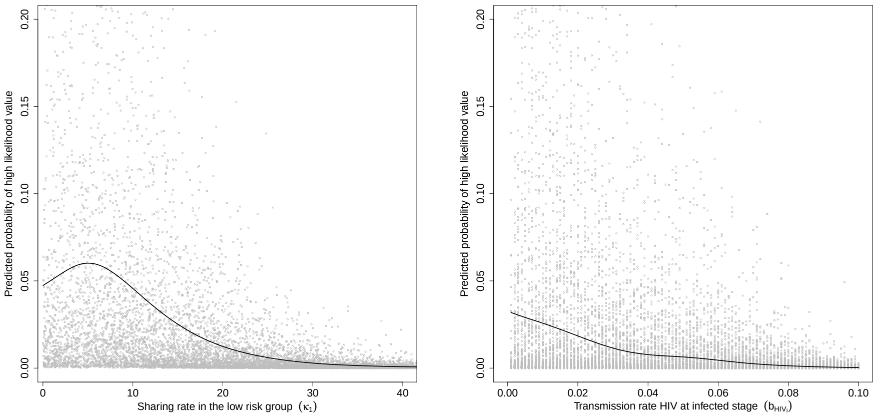
The parameters that show a significant p-value in this analysis are the sharing rates parameters (κ_1 and $f_1 = \kappa_2/\kappa_1$), the transmission rate per syringe-sharing event of HIV at infection stage (b_{HIV_I}), the transmission rate of HCV per syringe-sharing event at chronic carrier stage ($b_{HCV_{CC}}$), the proportion of syringes shared within the low risk group (m_{11}) and the proportion of individuals in the high risk group. Table 7.5 shows the χ^2 and corresponding p-values for all the parameters in the model significant at the 5% level. In Figure 7.4 the fitted probabilities of a high likelihood are shown for the two most significant parameters in the additive model. Values smaller than 20 for the syringes rate in the low risk group (κ_1) are associated with a high likelihood; whereas transmission rates of HIV at the infection stage smaller than 0.04 increase the probability of obtaining a high likelihood value.

Table 7.5: Vedette IDU dataset. Results of the generalized additive model: χ^2 and p-values for the selection of parameters that have a significant effect on the probability of a high (top 1%) multinomial likelihood.

Parameter	χ^2	p-value
Sharing rate in the low risk group (κ_1)	750.05	<0.0001
Factor increasing sharing syringe rate high risk group (f_1)	65.39	<0.0001
Proportion of syringes shared within the low risk group (m_{11})	116.01	<0.0001
Proportion of individuals in the low risk group ($prop$)	10.17	0.0225
HCV transmission rate at chronic stage ($b_{HCV_{CC}}$)	102.03	<0.0001
HIV transmission rate at infected stage (b_{HIV_I})	13.42	0.0062

7.3.2.3 Maximal information coefficient

The highest MIC is 0.25 and measures the association between the sharing rate parameter in the low risk group κ_1 and the transmission rate per syringe-sharing event in the chronic carrier of HCV $b_{HCV_{CC}}$. The second highest MIC is 0.20 and quantifies the strength of association between the transmission rate b_{HIV_I} per syringe-sharing event in the infected stage of HIV and the sharing rate parameter in the low risk group κ_1. The MIC between the transmission rate b_{HIV_I} and the sharing rate parameter $f_1 = \kappa_2/\kappa_1$ also equals 0.14. All other MIC values are lower than 0.12. With respect to the directionality of the association, the sharing syringe rate in the low risk group κ_1 is negatively correlated with the transmission rates $b_{HCV_{CC}}$ and b_{HIV_i}; the same occurs between the transmission rate b_{HIV_I} and the sharing rate parameter $f_1 = \kappa_2/\kappa_1$.



(a) Sharing rate low risk group κ_1

(b) Transmission rate HIV per syringe-sharing event at infected stage b_{HIV_I}

Figure 7.4: Vedette IDU dataset. Results of the generalized additive model: Predicted probabilities of high likelihood values considering estimated components κ_1 and b_{HIV_I}.

7.4 Application to the Itinere dataset

The AIC values in the second column of Table 7.6 suggest that the model using the first definition of the force of infection with two risk groups fits best to the data, with only a very small difference in favor of the second definition of the force of infection. Of course, for a similar sample of the same size the result could be different. Therefore, we motivate our final decision on the results of a nonparametric bootstrap exercise (with $B = 500$ runs). The last column of Table 7.6 indicates that the two-risk group model with the first definition for the force of infection is selected more often (81%) than any other model; and therefore we select that one as our final model. The parameter values of this final model at the maximal multinomial likelihood are listed in Table 7.7.

Panels (a) and (b) of Figure 7.5 show the model-based fits for p_{00t}, p_{01t}, p_{10t}, p_{11t} considering one, two and three risk groups with the first definition of the force of infection on top of the observed probabilities. The fit for one risk group does not reflect the data pattern observed in the joint and the marginal prevalences. At longer exposure times there is small number of individuals and none of the models follow observed trends. As for the Vedette dataset the two-risk group model fits better to the data. Panel (c) depicts the marginal probabilities $p_{\cdot 1t}$ and $p_{1\cdot t}$.

Table 7.6: Itinere IDU dataset. Results of all models combining one to three risk groups with two definitions for the force of infection: value of the multinomial likelihood, AIC value, and the number of times that the respective combination was selected as best model, out of 500 bootstrap resamples.

# Risk groups	Multinomial likelihood	AIC based on Multinomial	Boostrap Frequency
	First definition for the force of infection		
One	-724.114	1,476.227	0
Two	-674.921	1,387.841	404
Three	-682.827	1,417.654	1
	Second definition for the force of infection		
One	-764.484	1,558.968	0
Two	-681.636	1,401.273	50
Three	-678.364	1,400.728	45

7.4.1 Assessing variability

It turns out that only 28 out of 500,000 (that is only 0.01%) parameter vectors θ were associated with the highest likelihood values $L^{(b)}(\theta|\{y^*_{rst}(b)\}_{t\in T})$ in at least one of the $B = 500$ bootstrap samples.

The model-based prevalence curves for HCV and HIV for these 28 parameter vectors are shown in Figure 7.6 (gray lines). Four of the 28 parameter vector were selected in 88% of the bootstrap samples. The colored lines and symbols represent the model-based prevalence curves corresponding to the four most often selected parameter sets (49.4%, 21.8%, 9.4% and 7.8%).

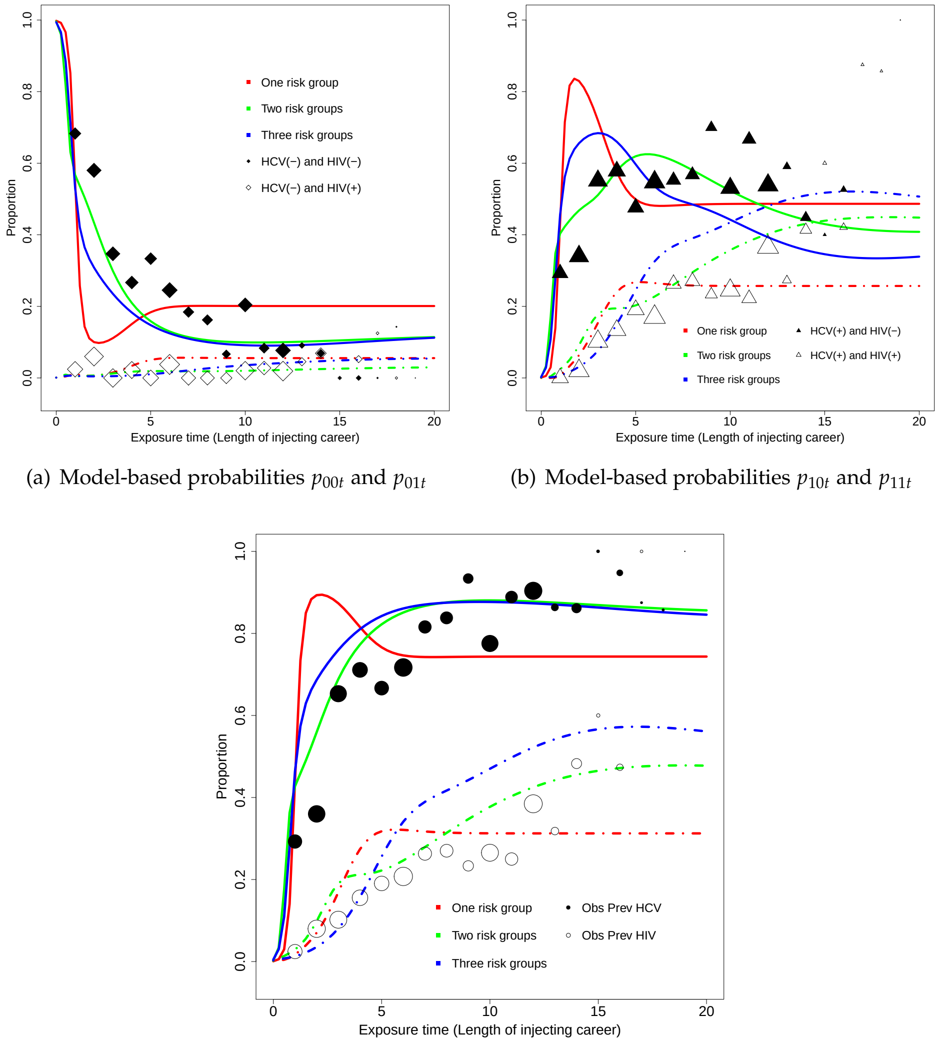
The model-based curves show more variability for the Itinere study than the ones for the Vedette data. Additionally, the model-based curves for HCV are less variable compared with those for HIV.

For the five parameter vectors θ selected by 88% of the bootstrap samples, the rate of sharing for the low risk group is between 10.0 and 20.1 syringes per three months period (larger than the values for the Vedette data) and the ratio between the rates of sharing for both groups (κ_2/κ_1) is between 24.6 and 41.2. Thus, the highest risk group shares between twenty to forty times more syringes than the low risk group. The ratio between the transmission rates per syringe-sharing event in the acute infected stage of HCV against infected stage of HIV is between 4.4 and 15.4, indicating HCV is more transmissible than HIV. The percentage of individuals in the low risk group varies between 70% and 98%.

Comparing the fitted models for both datasets, we notice that the percentage of

Table 7.7: Itinere IDU dataset. Parameter values that maximize the multinomial likelihood for the model with two risk groups and the first definition for the force of infection.

Model parameter	Value
Transmission rate of HCV in acute stage (b_{HCV_I})	0.184
Transmission rate of HCV in chronic stage ($b_{HCV_{CC}}$)	0.072
Transmission rate of HIV in infection stage (b_{HIV_I})	0.019
Transmission rate of HIV in AIDS stage (b_{HIV_A})	0.090
Duration in the acute infection stage HCV ($1/\omega_{HCV_I}$)	4.1 months
Duration as chronic carrier of HCV ($1/\eta_{HCV_{CC}}$)	18.9 years
Duration in the infection stage HIV ($1/\omega_{HIV_I}$)	15.6 years
Duration in AIDS ($1/\eta_{HIV_A}$)	8.4 years
Proportion of not spontaneously clearance of the HCV (ψ)	0.475
Acceleration factor for HCV disease progression due to co-infection (r_1)	1.40
Factor affecting the spontaneous clearance probability due to co-infection (r_2)	0.91
Percentage of IDUs in the low risk group (two RGs)	92.8%
Sharing syringe rate low risk group per trimester κ_1 (two RGs)	20.12
Factor increasing sharing syringe rate high risk group f_1 (two RGs)	24.59
Mixing proportion m_{11}	0.982
Mixing proportion m_{22}	0.075
Entry rate (E)	0.180
Injecting career length low risk group ($1/\nu_1$)	33.3 years
Injecting career length high risk group ($1/\nu_2$)	10.1 years



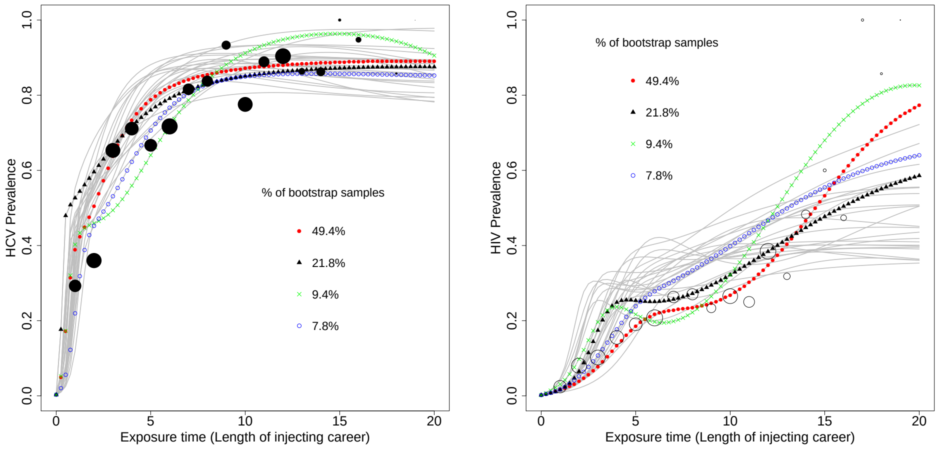
(a) Model-based probabilities p_{00t} and p_{01t}

(b) Model-based probabilities p_{10t} and p_{11t}

(c) Model-based probabilities $p_{r.t}$ and $p_{.st}$

Figure 7.5: Itinere IDU dataset. Observed proportions together with model-based joint and marginal prevalence fits for the model with one, two and three risk groups, combined with the first definition for the force of infection. The size of the symbols is proportional to the observed number of individuals at each exposure time.

individuals in the low risk group is larger in the Itinere study, however the individuals in this group exhibit a more risky behaviour compared to those in the Vedette study. The results reflect the differences in the study populations, whereas the Itinere



(a) HCV prevalence

(b) HIV prevalence

Figure 7.6: Itinere IDU dataset. Bootstrap results - observed prevalences (circles) and model-based prevalences. The gray lines correspond to parameter sets that lead to the highest likelihood value for at least one bootstrap sample. The other colored lines indicate the parameter sets that produce the highest likelihood values more frequently.

focuses on street young drug users, the Vedette data focusses on IDUs attending treatment centers.

7.4.2 Univariate explorative analysis of $\Theta_{1\%}$

We limit ourselves to the sharing rate parameter in the low risk group (κ_1), the transmission rate of HIV at infectious stage (b_{HIV_I}), the mixing parameter (m_{11}), and the proportion of individuals in the low risk group ($prop$), as the results were most pronounced for these parameters and because they will repeatedly play a more prominent role in the further analyses. Figure 7.7 demonstrates more peaked densities $f_{\Theta_{1\%}}$ (as compared to f_U) for the sharing rate κ_1 and the transmission rate b_{HIV_I} and less pronounced peaks of mass concentration for m_{11} and $prop$.

7.4.2.1 Activity region finder

For Itinere dataset ten parameters take part in the recursive partioning classificatio process: the sharing rate parameters κ_1 and $f_1 = \kappa_2/\kappa_1$, the transmission rates per syringe-sharing event b_{HIV_I}, b_{HCV_I} and $b_{HCV_{CC}}$, the entry rate (E), the proportion of

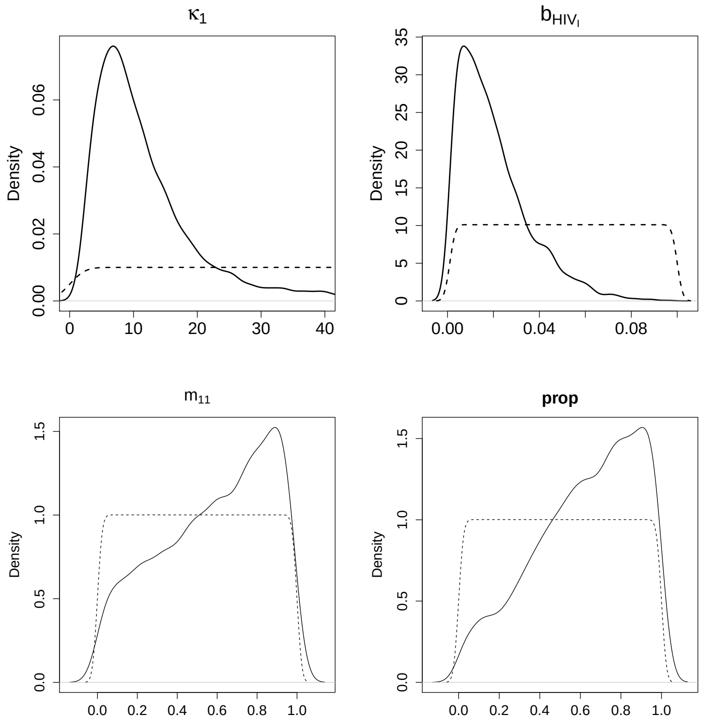


Figure 7.7: Analysis of the Itinere IDU dataset. Overlaid smoothed density plots $f_{\Theta_{1\%}}$ and f_U for the sharing rates κ_1, the transmission rate b_{HIV_I}, the mixing parameter m_{11}, and the proportion of individuals in the low risk group *prop*.

people in the low risk group *prop*, the proportion of individuals who become chronic carriers after being acute infected with HCV (ψ), the mixing proportion m_{22} and the acceleration factor r_1.

The sharing rate in the low risk group κ_1 is responsible for the first split, followed by splits using the transmission rate per syringe-sharing event $b_{HCV_{CC}}$. In total 40 regions were classified as high activity regions. The most relevant regions can be described as follows. A sharing rate in the low risk group (κ_1) between 3.5 and 8 combined with a transmission rate per syringe-sharing event of HIV at infected stage (b_{HIV_I}) between 0.008 and 0.029, with a factor of increasing sharing syringe rate ($f_1 = \kappa_2/\kappa_1$) between 8.8 and 49.7, and with a transmission rate per syringe-sharing event of HCV at chronic stage ($b_{HCV_{CC}}$) between 0.082 and 0.199 produce on average log-likelihood values of -963.7.

Table 7.8: Itinere IDU dataset. Results of the generalized additive model: χ^2 and p-values for the selection of parameters that have a significant effect on the probability of a high (top 1%) multinomial likelihood.

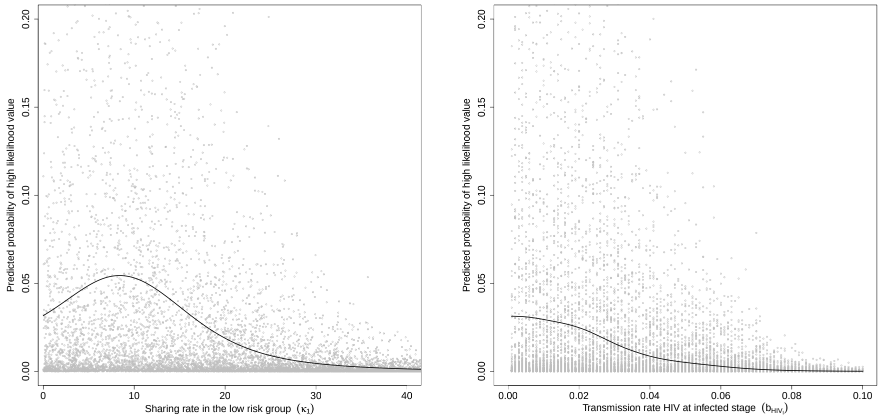
Parameter	χ^2	p-value
Sharing rate in the low risk group (κ_1)	662.42	<0.0001
Factor increasing sharing syringe rate high risk group (f_1)	188.40	<0.0001
Proportion of syringes shared within the low risk group (m_{11})	17.96	0.0028
Proportion of syringes shared within the low risk group (m_{22})	25.76	0.0001
Proportion of individuals in the low risk group ($prop$)	48.28	<0.0001
HCV transmission rate at infectious stage (b_{HCV_I})	18.22	0.0033
HCV transmission rate at chronic stage ($b_{HCV_{CC}}$)	82.71	<0.0001
HIV transmission rate at infected stage (b_{HIV_I})	79.76	<0.0001

7.4.2.2 Generalized additive models

The parameters that show a significant p-value in this analysis are the sharing rates parameters (κ_1 and $f_1 = \kappa_2/\kappa_1$), the transmission rate per syringe-sharing event of HIV at infection stage (b_{HIV_I}), the transmission rates of HCV per syringe-sharing event (b_{HCV_I} and $b_{HCV_{CC}}$), the mixing parameters (m_{11} and m_{m22}) and the proportion of individuals in the high risk group. Table 7.8 shows the χ^2 and corresponding p-values for all the parameters in the model significant at the 5% level. In Figure 7.8 the fitted probabilities of a high likelihood are shown for the two most significant parameters in the additive model. Values smaller than 20 for the syringes rate in the low risk group (κ_1) are associated with a high likelihood; whereas transmission rates of HIV at the infection stage smaller than 0.04 increase the probability of obtaining a high likelihood value.

7.4.2.3 Maximal information coefficient

The highest MIC is 0.22 and measures the association between the sharing rate parameter in the low risk group κ_1 and the transmission rate per syringe-sharing event in the chronic carrier of HCV $b_{HCV_{CC}}$. The second highest MIC is 0.2 and quantifies the association between sharing rate parameter in the low risk group κ_1 and the transmission rate b_{HIV_I} per syringe-sharing event in the infected stage of HIV. The MIC between κ_1 and the mixing parameter m_{11} equals to 0.17. The MIC between b_{HIV_I} and $b_{HCV_{CC}}$ equals to 0.16. All other MIC values are lower than 0.14. With respect to the directionality of the association, the sharing syringe rate in the low risk group (κ_1) is negatively correlated with the transmission rates $b_{HCV_{CC}}$ and b_{HIV_i}.



(a) Sharing rate low risk group κ_1

(b) Transmission rate HIV per syringe-sharing event at infected stage b_{HIV_I}

Figure 7.8: Itinere IDU dataset. Results of the generalized additive model: Predicted probabilities of high likelihood values considering estimated components κ_1 and b_{HIV_I}.

7.5 Concluding remarks

This study proposes and studies a joint mathematical model for co-infection of HCV and HIV in the context of injecting drug users. We propose the use of statistical concepts and measures to calibrate the model to bivariate data and illustrate the procedure with data from two studies of heroin IDUs in Italy and Spain.

The proposed joint mathematical model takes into account some of the biological complexities observed in the dynamics of the transmission of HCV and HIV in the IDU context. It accounts for parenteral transmission of the viruses, different transmission rates per syringe-sharing event for both viruses in each disease stage, the impact of co-infection on the transmissibility, the duration of the individual in every disease stage and the length of the injecting career. In contrast, the sexual transmission, which would increase HIV prevalence, is not explicitly represented in the model because of the relatively safe sexual behavior reported in Barrio et al. (2007) and Sabbatini et al. (2001). In fact our model cannot be used to evaluate the amount of sexual HIV transmission.

For both data examples, there are some similarities: the results support the assumption of two different risk groups (high and low risk group) in combination with our first definition for the force of infection. Based on the statistical assessment, we observe a better fit in early exposure times, mainly due to the limited amount of individuals with a longer duration of injection. The statistical analyses identified the sharing rates parameters κ_1 and $f_1 = \kappa_2/\kappa_1$, the HIV transmission rate at infected stage b_{HIV_I} and the HCV transmission rate at chronic stage $b_{HCV_{CC}}$ to play a major role whereas the HCV transmission rate at infected stage b_{HCV_I}, the mixing parameter m_{11} and the proportion of individuals in the low risk group $prop$ have a moderate impact on the multinomial likelihood. Additionally, parameters such as the cessation rates (μ_1 and μ_2), the duration in every disease stage $1/omega_{HCV_I}$, $1/\gamma_{HCV_{CC}}$, $1/\omega_{HIV_I}$ and $1/\gamma_{HIV_A}$, the term to reflect the extra transmissibility of HCV (and faster liver disease progression) in presence of co-infection r_1 seem to be less relevant based on our analyses and given the data at hand. The main difference between the parameter estimates for both datasets is the HIV transmission rate in AIDS stage b_{HIV_A} is larger for the Itinere dataset than for the Vedette dataset; the result is expected considering the prevalence of HIV is larger for this Spanish study. The difference in the HIV prevalence may be attributed to the study populations: the individuals in the Itinere project were mainly street users whereas on the Vedette data they were attending treatment centers.

We compare the parameter estimates of our model obtained in the variability assessment section with those reported in other mathematical models for HCV infection and HCV/HIV co-infection. For this comparison we discard the model proposed by Kretzschmar and Wiessing (2004) because it ignores spontaneous clearance of the HCV virus and secondary infections.

We notice similarities between some of the parameter estimates presented here for both datasets and those described by Vickerman et al. (2008, 2009) and De Vos et al. (2012). For instance, the HCV transmission rates (b_{HCV_I} and $b_{HCV_{CC}}$), HIV transmission rates (b_{HIV_I} and b_{HIV_A}) and proportion of individuals that resolve their infection ($1 - \rho$) reported here are in agreement with those found by Vickerman et al. (2008, 2009) and De Vos et al. (2012). Additionally, the factor difference between HIV and HCV transmission probability of our model supports results reported by Vickerman et al. (2008). In our model, the estimates for the HCV transmission rate at chronic stage $b_{HCV_{CC}}$ seem slightly larger than the estimates for the HCV transmission rate at acute stage b_{HCV_I} (median ratio 1.2), while De Vos et al. (2012) assumed both transmission rates are equal.

On the other hand, the estimates of the exit rates (μ_1 and μ_2) in our model for the Vedette dataset are larger than the ones reported in De Vos et al. (2012) and Vickerman et al. (2007). A large exit rate implies short injecting career, which may be attributed to a large cessation of injection. This discrepancy may be due to differences in the study populations, that is, the individuals in the Vedette study were admitted to treatment centers run by the National Health Service, whereas the models by Vickerman et al. (2007) and De Vos et al. (2012) were parameterized based on prospective cohort studies and surveillance data.

Our model has several limitations such as the assumption of invariant force of infection with respect to calendar time. Some studies have reported temporal differences in HIV prevalence and HIV risk related behavior among IDUs in Europe and antiretroviral therapy use. For instance, one of the findings of Davoli et al. (1995) was a decrease in sharing syringes among the self-reported HIV positive drug users in Rome between 1990 and 1992; Suligoi et al. (2004) reported a decrease in the HIV prevalence in the first half of the 1990s partly due to the implementation of prevention programs aiming to modify risk behaviour among IDUs. In Spain, De La Fuente et al. (2006) evaluated changes in the prevalence of HIV infection among young heroin users, pointing at the decrease in proportion of individuals who ever inject. In fact, the declining trend of injection in drug users has been also described in several European countries (Wiessing et al.; 2010; Castro-Sanchez et al.; 2012). Additionally, the disease-free survival time has been extended with the use antiretroviral therapy that started in 1996 (Poundstone et al.; 2001), leading to a reduction in the HIV prevalence. Both aspects may have an impact on the time homogeneity assumption.

Even though several previous studies have found that injectors with recently initiated IDUs can have higher risk behaviour than more experienced IDUs (Doherty et al.; 2000); our model did not allow for changes in the risk over an injecting career. However, earlier models allowed for this additonal risk but it did not improve the goodness of fit of the model, so it was not included in the final version of the model.

Additionally, our model omits some injecting practices that prevent transmission, such as syringe cleaning or disinfection. Even though some of these practices do not substitute the use of sterile needles or cessation of injection, they may help to prevent blood-borne infections such as HCV (Kapadia et al.; 2002), although their impact at the population level is inconclusive (Hagan et al.; 2011). Unfortunately, information regarding the frequency or method of syringe cleaning was scarce for the Vedette and the Itinere data and so could not be included in our model.

Chapter 8

Discussion

This thesis presents several statistical and mathematical models applied to HIV and HCV co-infection and to nosocomial infections. The models were applied to four different studies, taking into account the objectives and characteristics of each of the studies.

Regarding the statistical methods we focus on those applied to type II interval-censored data. We include an overview of existing methods and software availability to analyse this type of data.

In the literature review presented in Chapter 2, the first part describes statistical models applied to survival analysis ranging from completely non-parametric to fully parametric methods. We present the options that deal with interval-censored data. Here, we did not attempt to be exhaustive, excluding topics such as bayesian framework. We also did not mention flexible modelling, such as smoothing splines for continuous covariates or mixture distributions for the error terms, which can be applied.

Although, interval-censored data is very common when the event of interest can only be monitored at specific time points, many of the proposed methods have not been implemented in any statistical software. The lack of software is notorious in the extensions of the Cox proportional hazard models, due to the lack of a unified approach in this setting in combination with the large computational efforts that are needed.

In the analysis of interval-censored data, there are a lot of open questions such as model assessment techniques and joint modelling with a longitudinal outcome. Sun (2006) Chapter 10 describes goodness of fit tests when the baseline hazard is fully specified with a parametric form. In case of proportional hazard model the author describes some procedures that can be applied as well as several graphical model-checking techniques. He also presents regression diagnostics based on residual-based procedures when an additive hazards model is assumed.

All the statistical models, regarding survival analyses methodology, discussed in this thesis assumed that the censoring mechanism is independent of the time to event (independent interval censoring). To deal with dependent or informative interval censoring we refer to Dunson and Dinse (2002), Zhang et al. (2005) and Sun (2006).

Chapters 3, 4 and 5 present our contribution applied to two different studies: one clinical trial from a hospital and one observational study with long follow up (ACS).

In Chapter 3 the objective is to quantify the effect of the use probiotics and antibiotics on time to colonization with ampicillin-resistant *Enterococcus faecium*. Here the risk factors are time dependent covariates, adding some complexity to the model.

The study was performed in a hospital with documented high prevalence of intestinal ARE carriage, in this setting we did not find significant impact of daily probiotics intake on the reduction of the time to ARE acquisition. In the same sense, a recent meta-analysis by Hempel et al. (2012) mentions that most of the trials did not show a statistically significant advantage of probiotics use and a review made by Oudhuis et al. (2011) shows conflicting results regarding the effects of probiotics on infection rates.

When we compare distribution of the age of admission for the two groups (with and without probiotics) we notice the patients who receive probiotics tend to be older than those who did not receive them. This may indicate a selection bias which can have a negative impact on the results of probiotics use.

The type of probiotics, the type of antibiotics the patient had received and the clinical condition of the patients are aspects that become relevant to assess the impact of probiotics intake on the reduction of the time to ARE acquisition according to Hempel et al. (2012) and Oudhuis et al. (2011).

In Chapter 4 the goal is to estimate the force of infection and assess the impact of risk factors on the time to HCV infection, this is possible the first time that the HCV force of infection has been estimated using time to event data in the context of injecting drug users. We use the ACS a study with more than 20 years follow up, focusing on patients who enter negative to the cohort.

We found a higher risk of HCV infection in the first three years of an IDU career. This is consistent with other studies like Platt et al. (2009); Sutton et al. (2006); Van den Berg et al. (2007a,b).

Drug of choice to inject was associated with HCV seroconversion but sharing syringes was not. Similar results have been reported by Van den Berg et al. (2007a,b); Hahn et al. (2001); Thorpe et al. (2000); Miller et al. (2003a,b); Van de Laar et al. (2005). It reflects the cumulative exposure to infected needles and injection paraphernalia.

Our findings provide important additional evidence that it is crucial to target HCV prevention to new injectors as soon as they start to inject and that any efforts to reduce incidence need to take recent injectors into account. However, since it might be hard to find these recent injectors additional efforts are needed to prevent the transition to injecting drug use in non-injecting drug users.

Chapter 5 is dedicated to bivariate clustered data. Here, we present an overview of frailty models and the theoretical framework of correlated gamma frailty model when consider interval-censored data. Then we apply several frailty models to the ACS data. The results inspire a simulation study with a twofold objective: to assess model behaviour to assess model behaviour of a correlated frailty model in presence of interval-censored data and to assess the impact of different frailty variances on a correlated gamma frailty model.

In the first simulation study we show that it is possible to apply correlated frailty model when part of the data is interval censored. Our estimates are consistent with the ones presented by Hens et al. (2009) and Cattaert (2008).

From the second simulation study, we provide some suggestions on when it would be suitable to perform sensitivity analyses based on the difference between frailty variances and the frailty correlation.

A high correlation between the parameter estimates may be an indication of identifiability issues. Several authors have pointed to the conditions under which the correlated frailty model is identifiable. Besides, Wienke (2011) suggests to include observed covariates in order to improve identifiability. Based on the results from Yashin and Iachine (Yashin et al.; 1995; Iachine; 2004) we know that the correlated frailty model is identifiable thanks to the additive decomposition, even without covariates and without parametric shape for the baseline hazard rate. Except in the case of current status data.

It has been recognized by Wienke (2011), there is a negative correlation between the variance and the correlation estimates. In our simulations we notice that the negative correlation it is not fixed and in fact decreases when the frailty correlation increases. Then, as it is expected, the correlated gamma frailty model can have serious identifiability issues when the correlation is on the border of the parameter space. Caution should be taken when the frailty correlation is smaller than 0.1 or closer to the smaller ratio between the frailty variances (if that ratio is larger than 0.1).

The results of our second simulation study are limited to the values we consider for frailty variances. The model parameters and sample size were chosen to reflect the Amsterdam Cohort Studies example. We assume equal frailty variances, moderate and large difference between frailty variances. In total 20 scenarios were considered assuming different frailty correlation.

When frailty variances are equal some cautious interpretations should be made if the estimated correlation is lower 0.1 or larger than 0.9. In those cases we suggest to perform further sensitivity analyses to assess the reliability of the results.

If the variance ratio is equal to 0.75, we recommend to perform senstivity analyses when the frailty correlation is lower 0.1 or larger than 0.5. If the variance ratio is equal to 0.5 the recommendation is to perform sensitivity analyses regardless of the frailty correlation.

Our conclusions are restricted to the frailty parameters we choose as well as the baseline hazard function. It is possible that other options for baseline hazard functions and different values of frailty parameters lead to different results. Based on the information available for this study is hard to perform extrapolations.

More research is needed in the area to implement a more complex baseline hazard such as the generalized gamma, the generalized F (Cox; 2008) or the one proposed by Sparling et al. (2006). Another option could be to consider a semi-parametric approach where the univariate marginal survival is left unspecified. For the ACS, we implemented the baseline hazard proposed by Sparling et al. (2006), however we face major difficulties with the convergence of the model.

Chapters 6 and 7 are dedicated to the mathematical models. Here we are interested in the transmission process itself rather than on the risk factors or the shape of the force of infection.

In Chapter 6 we introduce two basic transmission models for HCV and HIV and a joint model that accounts simulataneously for the transmission of both viruses.

The joint transmission model is considered in Chapter 7, where we calibrate the joint model to two different datasets and assess the model from a statistical perspective. The proposed joint mathematical model takes into account some of the biological complexities observed in the dynamics of the transmission of HCV and HIV in the IDU context.

For both data examples, there are some similarities: the results support the assumption of two different risk groups (high and low risk group) in combination with our first definition for the force of infection. Based on the statistical assessment, we observe a better fit in early exposure times, mainly due to the limited amount of individuals with a longer duration of injection.

The sharing rates parameters, the HIV transmission rate at infected stage and the HCV transmission rate at chronic stage play a major role according to the statistical analyses performed. Whereas the HCV transmission rate at infected stage, one of the mixing parameter and the proportion of individuals in the low risk group have a moderate impact.

Additionally, parameters such as the cessation rates, the duration in every disease stage, the term to reflect the extra transmissibility of HCV (and faster liver disease progression) in presence of co-infection seem to be less relevant based on our analyses and given the data at hand.

The main difference between the parameter estimates for both datasets is the HIV transmission rate in AIDS stage is larger for the Itinere dataset than for the Vedette dataset; the result is expected considering the prevalence of HIV is larger for this Spanish study. The difference in the HIV prevalence may be attributed to the study populations: the individuals in the Itinere project were mainly street users whereas on the Vedette data they were attending treatment centers.

Our results regarding transmission rates and proportion of individuals that resolve their infection are consistent with the ones reported by Vickerman et al. (2008, 2009) and De Vos et al. (2012).

On the other hand, the estimates of the exit rates in our model for the Vedette dataset are larger than the ones reported in De Vos et al. (2012) and Vickerman et al. (2007). A large exit rate implies short injecting career, which may be attributed to a large cessation of injection. This discrepancy may be due to differences in the study populations, that is, the individuals in the Vedette study were admitted to treatment centers run by the National Health Service, whereas the models by Vickerman et al. (2007) and De Vos et al. (2012) were parameterized based on prospective cohort studies and surveillance data.

Our model has several limitations such as the assumption of invariant force of infection with respect to calendar time. Some studies have reported temporal differences in HIV prevalence and HIV risk related behavior among IDUs in Europe and antiretroviral therapy use. For instance, in Italy Davoli et al. (1995) reports a decrease in sharing syringes whereas Suligoi et al. (2004) describes a decrease in the HIV prevalence in the first half of the 1990s. In Spain, De La Fuente et al. (2006) points at the decrease in proportion of ever injection. A similar trend has been described several European countries by Wiessing et al. (2010) and Castro-Sanchez et al. (2012). This may have an impact on the time homogeneity assumption.

Even though several previous studies have found that injectors with recently initiated IDUs can have higher risk behaviour than more experienced IDUs (Doherty et al.; 2000); our model did not allow for changes in the risk over an injecting career. However, earlier models allowed for this additonal risk but it did not improve the goodness of fit of the model, so this was not included in the final version of the model.

Additionally, our model omits some injecting practices that prevent transmission, such as syringe cleaning or disinfection. Even though some of these practices do not substitute the use of sterile needles or cessation of injection, they may help to prevent blood-borne infections such as HCV (Kapadia et al.; 2002), although their impact at the population level is inconclusive (Hagan et al.; 2011). Unfortunately, information regarding the frequency or method of syringe cleaning was scarce for the Vedette and the Itinere data and so could not be included in our model.

Bibliography

Aalen, O. (1978). Nonparametric inference for a family of counting processes, *The Annals of Statistics* **6**(4): 701–726.

Abrams, S. and Hens, N. (2014). Modeling individual heterogeneity in the acquisition of recurrent infections: an application to parvovirus b19, *Biostatistics* p. kxu031.

Aceijas, C., Stimson, G., Hickman, M., Rhodes, T. et al. (2004). Global overview of injecting drug use and hiv infection among injecting drug users, *Aids* **18**(17): 2295–2303.

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control* **19**: 716–723.

Amaratunga, D. and Cabrera, J. (2004). Mining data to find subsets of high activity, *Journal of Statistical Planning and Inference* **122**(1): 23–41.

Anderson, R. and May, R. (1991). *Infectious diseases of humans: dynamics and control*, Vol. 28, Wiley Online Library.

Baggaley, R., Boily, M., White, R. and Alary, M. (2006). Risk of HIV-1 transmission for parenteral exposure and blood transfusion: a systematic review and meta-analysis, *AIDS* **20**(6): 805.

Bargagli, A., Faggiano, F., Amato, L., Salamina, G., Davoli, M., Mathis, F., Cuomo, L., Schifano, P., Burroni, P. and Perucci, C. (2006). VEdeTTE, a longitudinal study on effectiveness of treatments for heroin addiction in Italy: study protocol and characteristics of study population, *Subs. Use Misuse* **41**(14): 1861–1879.

Barrio, G., De La Fuente, L., Toro, C., Brugal, T., Soriano, V., Gonzalez, F., Bravo, M., Vallejo, F., Silva, T. et al. (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**(4): 592–603.

Betensky, R. A., Lindsey, J. C., Ryan, L. M. and Wand, M. P. (1999). Local EM estimation of the hazard function for interval-censored data, *Biometrics* **55**: 238–245.

Betensky, R. A., Lindsey, J. C., Ryan, L. M. and Wand, M. P. (2002). A local likelihood proportional hazards model for interval censored data, *Statistics in Medicine* **21**: 263–275.

Bilcke, J., Beutels, P., Brisson, M. and Jit, M. (2011). Accounting for methodological, structural, and parameter uncertainty in decision-analytic models, *Medical Decision Making* **31**(4): 675–692.

Blower, S. and Dowlatabadi, H. (1994). Sensitivity and uncertainty analysis of complex models of disease transmission: an HIV model, as an example, *International Statistical Review/Revue Internationale de Statistique* pp. 229–243.

Bouhnik, A., Carrieri, M., Rey, D., Spire, B., Gastaut, J., Gallais, H., Obadia, Y. et al. (2004). Drug injection cessation among HIV-infected injecting drug users, *Addictive Behaviors* **29**(6): 1189–1197.

Brown, D., Brown, N., Cookson, B., Duckworth, G., Farrington, M., French, G., King, L., Lewis, D., Livermore, D., Macrae, B., Scott, G., Williams, D. and Woodford, N. (2006). National glycopeptide resistant enterococcal bacteraemia surveillance Working Group report to the Department of Health August 2004, *Journal of Hospital Infection* **62**(Suppl1): 1–S27.

Cai, T. and Betensky, R. A. (2003). Hazard regression for interval-censored data with penalized spline, *Biometrics* **59**: 570–579.

Camoni, L. (2011). Linfezione da HIV e le epatiti B e C nella popolazione tossicodipendente: i dati del Centro Operativo AIDS, `http://www.politicheantidroga.it/media/449962/dati_coa_camoni.pdf`. Accesed September 18, 2012.

Castro-Sanchez, A., Shkedy, Z., Hens, N., Aerts, M., Geskus, R., Prins, M., Wiessing, L. and Kretzschmar, M. (2012). Estimating the force of infection for HCV in injecting drug users using interval-censored data, *Epidemiology and Infection* pp. 1064–1074.

Castro Sanchez, A. Y., Aerts, M., Shkedy, Z., Vickerman, P., Faggiano, F., Salamina, G. and Hens, N. (2013). A mathematical model for hiv and hepatitis c co-infection and its assessment from a statistical perspective, *Epidemics* **5**(1): 56–66.

Cattaert, T. (2008). Impact of heterogeneity on estimation of infectious disease parameters [master thesis]. Hasselt University, Diepenbeek, Belgium.

Chang, C.-K. (2006). Sample size calculation and timeline estimate for Progression-Free Survival, *PharmaSUG proceedings: Statistics & Pharmacokinetics* **sp07**.

Chen, M.-H., Tong, X. and Sun, J. (2009). A frailty model approach for regression analysis of multivariate current status data, *Statistics in Medicine* **28**(27): 3424–3436.

Cox, C. (2008). The generalized f distribution: an umbrella for parametric survival analysis, *Statistics in Medicine* **27**(21): 4301–4312.

Cox, D. (1972). Regression models and life-tables, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 187–220.

Davey, P., Brown, E., Fenelon, L., Finch, R., Gould, I., Hartman, G., Holmes, A., Ramsay, C., Taylor, E., Wilcox, M. et al. (2005). Interventions to improve antibiotic prescribing practices for hospital inpatients, *Cochrane Database System Review* **4**(4).

Davison, A. and Hinkley, D. (1997). *Bootstrap methods and their application*, Cambridge University Press.

Davoli, M., Bargagli, A., Perucci, C., Schifano, P., Belleudi, V., Hickman, M., Salamina, G., Diecidue, R., Vigna-Taglianti, F. and Faggiano, F. (2007). Risk of fatal overdose during and after specialist drug treatment: the vedette study, a national multi-site prospective cohort study, *Addiction* **102**(12): 1954–1959.

Davoli, M., Perucci, C., Abeni, D., Arcà, M., Brancato, G., Forastiere, F., Montiroli, P. and Zampieri, F. (1995). HIV risk-related behaviors among injection drug users in Rome: differences between 1990 and 1992, *American Journal of Public Health* **85**(6): 829–832.

De Carli, G., Puro, V., Ippolito, G., Studio, I. et al. (2003). Risk of hepatitis C virus transmission following percutaneous exposure in healthcare workers., *Infection* **31**: 22.

De La Fuente, L., Bravo, M., Toro, C., Brugal, M., Barrio, G., Soriano, V., Vallejo, F. and Ballesta, R. (2006). Injecting and HIV prevalence among young heroin users in three Spanish cities and their association with the delayed implementation of harm reduction programmes, *Journal of Epidemiology and Community Health* **60**(6): 537–542.

de Regt, M., van der Wagen, L. E., Top, J., Blok, H. E., Hopmans, T. E., Dekker, A. W., Hené, R. J., Siersema, P. D., Willems, R. J. and Bonten, M. J. (2008). High acquisition and environmental contamination rates of cc17 ampicillin-resistant enterococcus faecium in a dutch hospital, *Journal of Antimicrobial Chemotherapy* **62**: 1401–1406.

de Regt, M., Willems, R. J., Hené, R. J., Siersema, P. D., Verhaar, H. J., Hopmans, T. E. and Bonten, M. J. (2010). Effects of probiotics on acquisition and spread of multiresistant enterococci, *Antimicrobial Agents and Chemotherapy* **54**(7): 2801–2805.

De Vos, A., Van der Helm, J., Prins, M. and Kretzschmar, M. (2012). Determinants of persistent spread of HIV in HCV-infected populations of injecting drug users, *Epidemics* **4**(2): 57–67.

Del Fava, E., Shkedy, Z., Hens, N., Aerts, M., Suligoi, B., Camoni, L., Vallejo, F., Wiessing, L. and Kretzschmar, M. (2011). Joint modeling of HCV and HIV co-infection among injecting drug users in Italy and Spain using individual cross-sectional data, *Statistical Communications in Infectious Diseases* **3**(1): 3.

Dempster, A., Laird, N. and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society. Series B (Methodological)* **39**(1): 1–38.

Deuffic-Burban, S., Wong, J., Valleron, A., Costagliola, D., Delfraissy, J. and Poynard, T. (2004). Comparing the public health burden of chronic hepatitis C and HIV infection in France, *Journal of Hepatology* **40**(2): 319–326.

Dietz, K. and Heesterbeek, J. (2002). Daniel Bernoulli's epidemiological model revisited, *Mathematical Biosciences* **180**(1): 1–21.

Doherty, M., Garfein, R., Monterroso, E., Brown, D. and Vlahov, D. (2000). Correlates of HIV infection among young adult short-term injection drug users, *AIDS* **14**(6): 717–726.

Dorey, F., Little, R. and Schenker, N. (1993). Multiple imputation for threshold-crossing data with interval censoring, *Statistics in Medicine* **12**(17): 1589–1603.

DSouza, A., Rajkumar, C., Cooke, J. and Bulpitt, C. (2002). Probiotics in prevention of antibiotic associated diarrhoea: meta-analysis, *BMJ* **324**(7350): 1361.

Duchateau, L. and Janssen, P. (2008). *The frailty model*, Springer Verlag.

Dunson, D. B. and Dinse, G. E. (2002). Bayesian models for multivariate current status data with informative censoring, *Biometrics* **58**(1): 79–88.

Farrington, C. (1996). Interval censored survival data: a generalized linear modelling approach, *Statistics in Medicine* **15**(3): 283–292.

Farrington, C., Kanaan, M. and Gay, N. (2001). Estimation of the basic reproduction number for infectious diseases from age-stratified serological survey data, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **50**(3): 251–292.

Farrington, C. P., Whitaker, H. J., Unkel, S. and Pebody, R. (2013). Correlated infections: quantifying individual heterogeneity in the spread of infectious diseases, *American journal of epidemiology* **177**(5): 474–486.

Fay, M. P. (1999). Comparing several score tests for interval censored data, *Statistics in Medicine* **18**(3): 273–285.

Fay, M. and Shaw, P. (2010). Exact and asymptotic weighted logrank tests for interval censored data: the interval R package, *Journal of Statistical Software* **36**(2): 1–34.

Finkelstein, D. (1986). A proportional hazards model for interval-censored failure time data, *Biometrics* pp. 845–854.

Fisher, K. and Phillips, C. (2009). The ecology, epidemiology and virulence of enterococcus, *Microbiology* **155**: 1749–1757.

Galai, N., Safaeian, M., Vlahov, D., Bolotin, A. and Celentano, D. (2003). Longitudinal patterns of drug injection behavior in the ALIVE study cohort, 1988–2000: description and determinants, *American Journal of Epidemiology* **158**(7): 695.

Gani, J. (2010). Modelling epidemic diseases, *Australian & New Zealand Journal of Statistics* **52**(3): 321–329.

Garcia de la Hera, M., Ferreros, I., del Amo, J., de Olalla, P., Hoyos, S., Muga, R., del Romero, J., Guerrero, R. and Hernández-Aguado, I. (2004). Gender differences in progression to AIDS and death from HIV seroconversion in a cohort of injecting dug users from 1986 to 2001, *Journal of Epidemiology and Community Health* **58**(11): 944–950.

Garnett, G. and Anderson, R. (1994). Balancing sexual partnership in an age and activity stratified model of HIV transmission in heterosexual populations, *Mathematical Medicine and Biology* **11**(3): 161–192.

Garnett, G., Cousens, S., Hallett, T., Steketee, R. and Walker, N. (2011). Mathematical models in the evaluation of health programmes, *Lancet* **378**: 515–525.

Goetghebeur, E. and Ryan, L. (2000). Semiparametric regression analysis of interval-censored data, *Biometrics* **56**: 1139–1144.

Goethals, K. (2011). Multivariate survival models for interval-censored udder quarter infection times [dissertation]. Ghent University, Ghent, Belgium.

Goggins, W., Finkelstein, D., Schoenfeld, D. and Zaslavsky, A. (1998). A markov chain monte carlo em algorithm for analyzing interval-censored data under the cox proportional hazards model, *Biometrics* pp. 1498–1507.

Gómez, G., Calle, M., Oller, R. and Langohr, K. (2009). Tutorial on methods for interval-censored data and their implementation in r, *Statistical Modelling* **9**(4): 259–297.

Gupta, V. and Garg, R. (2009). Probiotics, *Indian Journal of Medical Microbiology* **27**: 202–209.

Hagan, H., Pouget, E. and Des Jarlais, D. (2011). A systematic review and meta-analysis of interventions to prevent hepatitis C virus infection in people who inject drugs, *J Infect Dis* **204**(1): 74–83.

Hahn, J., Page-Shafer, K., Lum, P., Ochoa, K. and Moss, A. (2001). Hepatitis C virus infection and needle exchange use among young injection drug users in San Francisco, *Hepatology* **34**(1): 180–187.

Hempel, S., Newberry, S., Maher, A., Wang, Z., Miles, J., Shanman, R., Johnsen, B. and Shekelle, P. (2012). Probiotics for the prevention and treatment of antibiotic-associated diarrhea a systematic review and meta-analysis, *JAMA: The Journal of the American Medical Association* **307**(18): 1959–1969.

Hens, N., Aerts, M., Faes, C., Shkedy, Z., Lejeune, O., Van Damme, P. and Beutels, P. (2010). Seventy-five years of estimating the force of infection from current status data, *Epidemiology and Infection* **138**(6): 802.

Hens, N., Shkedy, Z., Aerts, M., Faes, C., Damme, P. V. and Beutels, P. (2012). *Modeling infectious disease parameters based on serological and social contact data A modern statistical perspective*, Springer, New York.

Hens, N., Wienke, A., Aerts, M. and Molenberghs, G. (2009). The correlated and shared gamma frailty model for bivariate current status data: An illustration for cross-sectional serological data, *Statistics in Medicine* **28**(22): 2785–2800.

Hickson, M., D'Souza, A., Muthu, N., Rogers, T., Want, S., Rajkumar, C. and Bulpitt, C. (2007). Use of probiotic lactobacillus preparation to prevent diarrhoea associated with antibiotics: randomised double blind placebo controlled trial, *bmj* **335**(7610): 80.

Hoare, A., DG, R. and DP, W. (2008). Sampling and sensitivity analyses tools (SaSAT) for computational modelling, *Theoretical Biology and Medical Modelling* **5**.

Hosmer, D. W., Lemeshow, S. and May, S. (2008). *Applied survival analysis: regression modeling of time-to-event data*, Wiley-Interscience, Hoboken, NJ.

Huang, J., Lee, C. and Yu, Q. (2008). A generalized log-rank test for interval-censored failure time data via multiple imputation, *Statistics in Medicine* **27**(17): 3217–3226.

Hutchinson, S., Bird, S. and Goldberg, D. (2005). Modeling the current and future disease burden of hepatitis c among injection drug users in scotland, *Hepatology* **42**(3): 711–723.

Hutchinson, S., Bird, S., Taylor, A. and Goldberg, D. (2006a). Modelling the spread of hepatitis C virus infection among injecting drug users in Glasgow: implications for prevention, *Int. J. Drug Policy* **17**(3): 211–221.

Hutchinson, S., Roy, K., Wadd, S., Bird, S., Taylor, A., Anderson, E., Shaw, L., Codere, G. and Goldberg, D. (2006b). Hepatitis C virus infection in Scotland: epidemiological review and public health challenges, *Scottish Medical Journal* **51**(2): 8–15.

Iachine, I. (2004). Identifiability of bivariate frailty models. Preprint 5. Department of Statistics, University of Southern Denmark, Odense.

Jarrin, I., Geskus, R., Bhaskaran, K., Prins, M., Perez-Hoyos, S., Muga, R., Hernández-Aguado, I., Meyer, L., Porter, K. and Amo, J. (2008). Gender differences in HIV progression to AIDS and death in industrialized countries: slower disease progression following HIV seroconversion in women, *American Journal of Epidemiology* **168**(5): 532.

Jit, M. and Brisson, M. (2011). Modelling the epidemiology of infectious diseases for decision analysis: a primer, *Pharmacoeconomics* **29**(5): 371–386.

Kapadia, F., Vlahov, D., Des Jarlais, D., Strathdee, S., Ouellet, L., Kerndt, P., Williams, I., Garfein, R. et al. (2002). Does bleach disinfection of syringes protect against hepatitis C infection among young adult injection drug users?, *Epidemiology* **13**(6): 738–741.

Kaplan, E. and Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association* **53**(282): 457–481.

Klein, J. and Moeschberger, M. (2003). *Survival analysis techniques for censored and truncated data*, Springer, New York.

Kooperberg, C. and Clarkson, D. (1997). Hazard regression with interval-censored data, *Biometrics* pp. 1485–1494.

Kretzschmar, M. and Wiessing, L. (2004). Modelling the transmission of hepatitis C in injecting drug users, *in* J. Jager, W. Limburg and M. Kretzschmar (eds), *Hepatitis C and injecting drug use: impact, costs and policy options*, OOPEC, pp. 143–159.

Law, M., Dore, G., Bath, N., Thompson, S., Crofts, N., Dolan, K., Giles, W., Gow, P., Kaldor, J., Loveday, S. et al. (2003). Modelling hepatitis C virus incidence, prevalence and long-term sequelae in Australia, 2001, *International Journal of Epidemiology* **32**(5): 717–724.

Lesaffre, E., Komárek, A. and Declerck, D. (2005). An overview of methods for interval-censored data with an emphasis on applications in dentistry, *Statistical Methods in Medical Research* **14**(6): 539–552.

Lindsey, J. (2007). *Applying generalized linear models*, Diepenbeek.

Lindsey, J. and Ryan, L. (1998). Methods for interval-censored data, *Statistics in Medicine* **17**(2): 219–238.

Ma, S. (2009). Cure model with current status data, *Statistica Sinica* **19**(1): 233.

Martin, N., Vickerman, P., Foster, G., Hutchinson, S., Goldberg, D. and Hickman, M. (2011a). Can antiviral therapy for hepatitis C reduce the prevalence of HCV among injecting drug user populations? A modeling analysis of its prevention utility, *Journal of hepatology* **54**(6): 1137–1144.

Martin, N., Vickerman, P. and Hickman, M. (2011b). Mathematical modelling of hepatitis C treatment for injecting drug users, *Journal of Theoretical Biology* **274**(1): 58–66.

Mathei, C., Shkedy, Z., Denis, B., Kabali, C., Aerts, M., Molenberghs, G., Van Damme, P. and Buntinx, F. (2006). Evidence for a substantial role of sharing of injecting paraphernalia other than syringes/needles to the spread of hepatitis C among injecting drug users, *Journal of Viral Hepatitis* **13**(8): 560–570.

Micallef, J. M., Kaldor, J. M. and Dore, G. J. (2006). Spontaneous viral clearance following acute hepatitis C infection: a systematic review of longitudinal studies, *Journal of Viral Hepatitis* **13**(1): 34–41.

Miller, C., Johnston, C., Spittal, P., Li, K., LaLiberté, N., Montaner, J. and Schechter, M. (2003a). Opportunities for prevention: hepatitis C prevalence and incidence in a cohort of young injection drug users, *Hepatology* **36**(3): 737–742.

Miller, M., Mella, I., Moi, H. and Eskild, A. (2003b). HIV and hepatitis C virus risk in new and longer-term injecting drug users in Oslo, Norway, *JAIDS Journal of Acquired Immune Deficiency Syndromes* **33**(3): 373–379.

Namata, H. (2008). Flexible statistical models for microbial risk assessment and infectious diseases [dissertation]. Hasselt University, Diepenbeek, Belgium.

Namata, H., Shkedy, Z., Faes, C., Aerts, M., Molenberghs, G. and Theeten, H. (2007). Estimation of the force of infection from currents status data using generalized linear mixed models, *Journal of Applied Statistics* (34): 923–939.

NCHECR (2010). Epidemiological and economic impact of potential increased hepatitis C treatment uptake in Australia, *Technical report*, National Centre in HIV Epidemiology and Clinical Research, The University of New South Wales. Sidney, Australia.

Nelson, P., Mathers, B., Cowie, B., Hagan, H., Des Jarlais, D., Horyniak, D. and Degenhardt, L. (2011). Global epidemiology of hepatitis B and hepatitis C in people who inject drugs: results of systematic reviews, *Lancet* **378**(9791): 571–583.

Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data, *Technometrics* **14**(4): 945–966.

Okais, C., Roche, S., Kurzinger, M., Riche, B., Bricout, H., Derrough, T., Simondon, F. and Ecochard, R. (2010). Methodology of the sensitivity analysis used for modeling an infectious disease, *Vaccine* **28**(51): 8132–8140.

Oudhuis, G., Bergmans, D. and Verbon, A. (2011). Probiotics for prevention of nosocomial infections: efficacy and adverse effects, *Current Opinion in Critical Care* **17**(5): 487–492.

Pan, W. (2000). A multiple imputation approach to cox regression with interval-censored data, *Biometrics* **56**: 199–203.

Pencina, M., Larson, M. and D'Agostino, R. (2006). Choice of time scale and its effect on significance of predictors in longitudinal studies, *Statistics in Medicine* **26**(6): 1343–1359.

Peto, R. and Lee, P. (1973). Weibull distributions for continuous-carcinogenesis experiments, *Biometrics* pp. 457–470.

Platt, L., Sutton, A., Vickerman, P., Koshkina, E., Maximova, S., Latishevskaya, N., Hickman, M., Bonell, C., Parry, J. and Rhodes, T. (2009). Measuring risk of HIV and HCV among injecting drug users in the Russian Federation, *The European Journal of Public Health* **19**(4): 428–433.

Poundstone, K. E., Chaisson, R. E. and Moore, R. D. (2001). Differences in HIV disease progression by injection drug use and by sex in the era of highly active antiretroviral therapy, *AIDS* **15**(9): 1115–1123.

Reshef, D., Reshef, Y., Finucane, H., Grossman, S., McVean, G., Turnbaugh, P., Lander, E., Mitzenmacher, M. and Sabeti, P. (2011). Detecting novel associations in large data sets, *Science* **334**(6062): 1518–1524.

Rockstroh, J. K. and Spengler, U. (2004). HIV and hepatitis C virus co-infection, *Lancet Infectious Diseases* **4**(7): 437–444.

Sabbatini, A., Carulli, B., Villa, M., Corrêa Leite, M., Nicolosi, A. et al. (2001). Recent trends in the HIV epidemic among injecting drug users in Northern Italy, 1993-1999, *AIDS* **15**(16): 2181.

Salamina, G., Diecidue, R., Vigna-Taglianti, F., Jarre, P., Schifano, P., Bargagli, A., Davoli, M., Amato, L., Perucci, C. and Faggiano, F. (2010). Effectiveness of therapies for heroin addiction in retaining patients in treatment: results from the VEdeTTE study, *Substance Use & Misuse* **45**(12): 2076–2092.

Satten, G. (1996). Rank-based inference in the proportional hazards model for interval censored data, *Biometrika* **83**(2): 355–370.

Satten, G., Datta, S. and Williamson, J. (1998). Inference based on imputed failure times for the proportional hazards model with interval-censored data, *Journal of the American Statistical Association* **93**(441): 318–327.

Serraino, D., Zucchetto, A., Suligoi, B., Bruzzone, S., Camoni, L., Boros, S., Paoli, A., Maso, L., Franceschi, S. and Rezza, G. (2009). Survival after AIDS diagnosis in Italy, 1999-2006: a population-based study, *JAIDS* **52**(1): 99.

Shkedy, Z., Aerts, M., Molenberghs, G., Beutels, P. and Van Damme, P. (2003). Modelling forces of infection by using monotone local polynomials., *Applied Statistics* (52): 469–485.

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.

Sparling, Y. H., Younes, N. and Lachin, J. M. (2006). Parametric survival models for interval-censored data with time-dependent covariates, *Biostatistics* **7**: 599–617.

Steensma, C., Boivin, J., Blais, L. and Roy, É. (2005). Cessation of injecting drug use among street-based youth, *Journal of Urban Health* **82**(4): 622–637.

Stein, M. (1987). Large sample properties of simulations using latin hypercube sampling, *Technometrics* **29**(2): 143–151.

Suligoi, B., Magliochetti, N., Nicoletti, G., Pezzotti, P. and Rezza, G. (2004). Trends in HIV prevalence among drug-users attending public drug-treatment centres in Italy: 1990–2000, *Journal of Medical Virology* **73**(1): 1–6.

Sun, J. (2006). *The Statistical Analysis of Interval-Censored Failure Time Data*, Springer, New York.

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**(1): 93.

Sutton, A., Hope, V., Mathei, C., Mravcik, V., Sebakova, H., Vallejo, F., Suligoi, B., Brugal, M., Ncube, F., Wiessing, L. et al. (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**(11): 809–816.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling survival data: extending the Cox model*, Springer Verlag.

Thorpe, L., Ouellet, L., Levy, J., Williams, I. and Monterroso, E. (2000). Hepatitis C virus infection: prevalence, risk factors, and prevention opportunities among young injection drug users in Chicago, 1997–1999, *Journal of Infectious Diseases* **182**(6): 1588–1594.

Todd, J., Glynn, J., Marston, M., Lutalo, T., Biraro, S., Mwita, W., Suriyanon, V., Rangsin, R., Nelson, K., Sonnenberg, P. et al. (2007). Time from HIV seroconversion to death: a collaborative analysis of eight studies in six low and middle-income countries before highly active antiretroviral therapy, *AIDS* **21**: S55.

Turnbull, B. W. (1976). The empirical distribution with arbitrarily grouped censored and truncated data, *Journal of the Royal Statistical Society, Series B* **38**: 290–295.

UNAIDS (2010). Global HIV/AIDS Response: Progress report 2011, *Technical report*, Joint United Nations Programme on HIV/AIDS (UNAIDS).

Unkel, S. and Farrington, C. P. (2012). A new measure of time-varying association for shared frailty models with bivariate current status data, *Biostatistics* **13**(4): 665–679.

Van Ameijden, E. and Coutinho, R. (2001). Large decline in injecting drug use in Amsterdam, 1986–1998: explanatory mechanisms and determinants of injecting transitions, *Journal of Epidemiology and Community Health* **55**(5): 356–363.

Van de Laar, T., Langendam, M., Bruisten, S., Welp, E., Verhaest, I., van Ameijden, E., Coutinho, R. and Prins, M. (2005). Changes in risk behavior and dynamics of hepatitis C virus infections among young drug users in Amsterdam, the Netherlands, *Journal of Medical Virology* **77**(4): 509–518.

Van den Berg, C., Smit, C., Bakker, M., Geskus, R., Berkhout, B., Jurriaans, S., Coutinho, R., Wolthers, K. and Prins, M. (2007a). Major decline of hepatitis C virus incidence rate over two decades in a cohort of drug users, *European Journal of Epidemiology* **22**(3): 183–193.

Van den Berg, C., Smit, C., Van Brussel, G., Coutinho, R. and Prins, M. (2007b). Full participation in harm reduction programmes is associated with decreased risk for human immunodeficiency virus and hepatitis C virus: evidence from the Amsterdam Cohort Studies among drug users, *Addiction* **102**(9): 1454–1462.

Vanni, T., Karnon, J., Madan, J., White, R., Edmunds, W., Foss, A. and Legood, R. (2011). Calibrating models in economic evaluation: a seven-step approach, *Pharmacoeconomics* **29**(1): 35–49.

Vaupel, J., Manton, K. and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality, *Demography* **16**(3): 439–454.

Vickerman, P., Hickman, M. and Judd, A. (2007). Modelling the impact on hepatitis c transmission of reducing syringe sharing: London case study, *International Journal of Epidemiology* **36**(2): 396–405.

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

Vickerman, P., Martin, N. and Hickman, M. (2011). Understanding the trends in HIV and hepatitis C prevalence amongst injecting drug users in different settings Implications for intervention impact, *Drug and Alcohol Depence* **123**(1-3): 122–131.

Vickerman, P., Miners, A. and Williams, J. (2008). Assessing the cost-effectiveness of interventions linked to needle and syringe programmes for injecting drug users: An economic modelling report, *Technical report*, National Institute for Health and Clinical Excellence.

Vickerman, P., Platt, L. and Hawkes, S. (2009). Modelling the transmission of HIV and HCV among injecting drug users in Rawalpindi, a low HCV prevalence setting in Pakistan, *Sexually Transmitted Infections* **85**(Suppl 2): ii23–ii30.

Vynnycky, E. and White, R. (2010). *An introduction to infectious disease modelling*, OUP Oxford.

Wasmuth, J.-C. (2010). Hepatitis C - Epidemiology, transmission and natural history, *in* S. Mauss, T. Berg, J. Rockstroh, C. Sarrazin and H. Wedemeyer (eds), *Hepatology - A Clinical text book - 2nd Edition*, Flying Publisher, Dusseldorf, Germany.
URL: *http://www.hepatologytextbook.com/index.htm*

Welp, E., Lodder, A., Langendam, M., Coutinho, R. and van Ameijden, E. (2002). HIV prevalence and risk behaviour in young drug users in Amsterdam, *AIDS* **16**(9): 1279–1284.

White, R., Ben, S., Kedhar, A., Orroth, K., Biraro, S., Baggaley, R., Whitworth, J., Korenromp, E., Ghani, A., Boily, M. et al. (2007). Quantifying HIV-1 transmission due to contaminated injections, *Proceedings of the National Academy of Sciences USA* **104**(23): 9794–9799.

Wienke, A. (2011). *Frailty models in survival analysis*, Chapman & Hall/CRC.

Wienke, A., Arbeev, K., Locatelli, I. and Yashin, A. (2005). A comparison of different bivariate correlated frailty models and estimation strategies, *Mathematical biosciences* **198**(1): 1–13.

Wiessing, L., Klempova, D., Hedrich, D., Montanari, L. and Gyarmathy, V. (2010). Injecting drug use in Europe: stable or declining, *Euro Surveillance: European Communicable Disease Bulletin* **15**.

Wood, S. (2006). *Generalized additive models: an introduction with R*, CRC Press, Boca Ratón, Florida.

World Health Organization (WHO) (2011). Hepatitis C, `http://www.who.int/mediacentre/factsheets/fs164/en/`. Accessed 12 July, 2012.

Yashin, A., Vaupel, J. and Iachine, I. (1995). Correlated individual frailty: an advantageous approach to survival analysis of bivariate data, *Mathematical Population Studies* **5**(2): 145–159.

Yazdanpanah, Y., De Carli, G., Migueres, B., Lot, F., Campins, M., Colombo, C., Thomas, T., Deuffic-Burban, S., Prevot, M., Domart, M. et al. (2005). Risk factors for hepatitis C virus transmission to health care workers after occupational exposure: a European case-control study, *Clinical Infectious Diseases* **41**(10): 1423.

Zhang, M. and Davidian, M. (2008). Smooth semiparametric regression analysis for arbitrarily censored time-to-event data, *Biometrics* **64**: 567–576.

Zhang, Z., Sun, J. and Sun, L. (2005). Statistical analysis of current status data with informative observation times, *Statistics in Medicine* **24**(9): 1399–1407.

Zhao, Q. and Sun, J. (2004). Generalized log-rank test for mixed interval-censored failure time data, *Statistics in Medicine* **23**(10): 1621–1629.

Zhao, X., Zhao, Q., Sun, J. and Kim, J. S. (2008). Generalized log-rank tests for partly interval-censored failure time data, *Biometrical Journal* **50**(3): 375–385.

Appendix A

Predicted probabilities survival models Chapter 4

Table A.1: Amsterdam Cohort Studies dataset. Predicted probabilities for an IDU to remain HCV negative after *t*-years of injecting drugs based on the model that includes sharing syringes accounting for left truncation.

Exposure time (*t*-years)	Sharing syringes							
	No				Yes			
	Duration of injection at first visit				Duration of injection at first visit			
	0.5 years	3 years	7.5 years	15 years	0.5 years	3 years	7.5 years	15 years
1	0.976	1.000	1.000	1.000	0.927	1.000	1.000	1.000
2	0.791	0.968	1.000	1.000	0.735	0.910	1.000	1.000
3	0.686	0.861	1.000	1.000	0.637	0.795	1.000	1.000
4	0.620	0.780	1.000	1.000	0.576	0.719	0.996	1.000
5	0.573	0.721	0.996	1.000	0.532	0.665	0.964	1.000
6	0.538	0.676	0.975	1.000	0.499	0.623	0.919	1.000
7	0.509	0.641	0.941	1.000	0.473	0.590	0.876	1.000
8	0.486	0.611	0.905	1.000	0.451	0.563	0.837	1.000
9	0.466	0.586	0.871	1.000	0.433	0.540	0.804	1.000
10	0.449	0.565	0.841	1.000	0.417	0.521	0.775	1.000

Table A.2: Amsterdam Cohort Studies dataset. Predicted probabilities of an IDU will be HCV negative after *t*-years of injecting drugs based on the model that year of first injection.

Exposure time	Year of first injection		
(*t*-years)	1962-1980	1981-1990	1991-2002
1	1.000	0.937	0.953
2	1.000	0.823	0.837
3	1.000	0.763	0.776
4	1.000	0.723	0.735
5	1.000	0.693	0.705
6	1.000	0.670	0.681
7	0.997	0.651	0.662
8	0.977	0.635	0.645
9	0.956	0.621	0.631
10	0.938	0.609	0.619

Table A.3: Amsterdam Cohort Studies dataset. Predicted probabilities of an IDU will be HCV negative after *t*-years of injecting drugs based on the model with duration of injection at first visit.

Exposure time	Duration of injection at first injection			
(*t*-years)	0.5 years	3 years	7.5 years	15 years
1	0.950	1.000	1.000	1.000
2	0.759	0.945	1.000	1.000
3	0.660	0.832	1.000	1.000
4	0.598	0.754	0.999	1.000
5	0.554	0.698	0.981	1.000
6	0.520	0.656	0.943	1.000
7	0.493	0.622	0.902	1.000
8	0.471	0.594	0.864	1.000
9	0.452	0.570	0.830	1.000
10	0.436	0.550	0.801	1.000

Samenvatting

In dit proefschrift worden diverse statistische en wiskundige modellen voor HIV en HCV co-infectie en voor nosocomiale infecties gepresenteerd. De modellen werden toegepast op vier verschillende studies, rekening houdend met de doelstellingen en kenmerken van elk onderzoek. In Hoofdstuk 2 presenteren we een literatuuronderzoek van statistische modellen die worden toegepast op survival analyse, variërend van niet-parametrische tot volledig parametrische methoden. Daarnaast presenteren we een kort overzicht van wiskundige modellen voor de overdracht van infectieziekten. Het hoofddoel van Hoofdstuk 3 is om het effect te kwantificeren van het probiotica- en antibioticagebruik op de kolonisatietijd van de *Enterococcus faecium* bacteria die resistentie vertoond tegen ampicilline (ARE). In deze studie zijn de risicofactoren tijdsafhankelijk, wat een extra complexiteit aan het model toevoegt. Het onderzoek werd uitgevoerd in een ziekenhuis met gedocumenteerde hoge prevalentie van intestinaal ARE-transport. In deze studie was er geen significante invloed aangetoond van de dagelijkse probiotica-inname op de vermindering van de tijd tot het oplopen van ARE. Soortgelijke conclusies waren getrokken uit een recente meta-analyse van Hempel et al. (2012) waar werd vermeld dat in de meeste studies geen statistisch significant voordeel van probioticagebruik werd aangetoond. Bovendien heeft een overzicht in Oudhuis et al. (2011) aangetoond dat er tegensprekende resultaten bestaan over probiotica-effecten op de infectie ratios. Als we de verdeling van opnameleeftijd voor de twee groepen (met en zonder probiotica) vergelijken, merken we dat de patiÃ«nten die probiotica krijgen vaak ouder zijn dan diegenen die geen probiotica krijgen. Dit kan een selectievertekening aanwijzen die een negatieve impact zou kunnen hebben op de resultaten van probioticagebruik. Het type probiotica, het type antibiotica dat de patiënt ontvangen heeft en de klinische toestand van de patiënten zijn relevante aspecten om de invloed van probiotica-inname op de vermindering van de tijd tot het oplopen van ARE te beoordelen, aldus Hempel et al. (2012) en Oudhuis et al. (2011).

In Hoofdstuk 4 schatten we de infectiedruk van HCV en beoordelen we de invloed van risicofactoren op de tijd tot HCV-infectie. Waarschijnlijk is dit de eerste keer dat HCV-infectiedruk geschat is op basis van time-to-event-gegevens in het kader van injecterende drugsgebruikers (IDG). We maken gebruik van de ACS-studie met meer dan 20 jaar opvolging, gericht op patiënten die negatief zijn op het moment van opname in de cohort. We vonden een hoger risico op HCV-infectie in de eerste drie jaar van een IDG-carrière. Dit wordt ook bewezen in andere studies, zoals Platt et al. (2009); Sutton et al. (2006); Van den Berg et al. (2007a,b). De keuze van het geïnjecteerde drug werd geassocieerd met HCV-seroconversie, het delen van spuiten was er niet mee geassocieerd. Soortgelijke resultaten zijn gerapporteerd door Van den Berg et al. (2007a,b); Hahn et al. (2001); Thorpe et al. (2000); Miller et al. (2003a,b); Van de Laar et al. (2005). Dit geeft de cumulatieve blootstelling aan besmette naalden en injectie uitrusting weer.

Onze resultaten geven aanvullende aanwijzingen dat het cruciaal is om de HCV-preventie op nieuwe injecterende druggebruikers te richten zodra ze met injecties beginnen en dat alle inspanningen om de HCV-incidentie te verminderen rekening moeten houden met recente injecterende gebruikers. Desalniettemin, aangezien het moeilijk is om die recente injecterende gebruikers op te sporen, zijn er bijkomende inspanningen nodig om de overgang van niet-injecterende drugsgebruikers naar injecterend drugsgebruik te voorkomen.

Hoofdstuk 5 is gewijd aan geclusterde bivariate gegevens. Hierin presenteren we een overzicht van frailty modellen en het theoretische kader van gecorreleerde gamma frailty model voor interval gecensureerde data. Wij passen verschillende frailty modellen toe op de ACS gegevens. Gebaseerd op deze resultaten, wordt er een simulatiestudie met een tweeledig doel uitgevoerd: i) om het gedrag van een gecorreleerd frailty model te evalueren in aanwezigheid van interval gecensureerde data en ii) om de impact te beoordelen van verschillende frailty varianties op een gecorreleerde gamma frailty model. In de eerste simulatiestudie tonen we aan dat het mogelijk is om gecorreleerde frailty model toe te passen wanneer een deel van de data interval gecensureerd is. Onze schattingen zijn consistent met resultaten van Hens et al. (2009) en Cattaert (2008). Gebaseerd op de tweede simulatiestudie geven we een aantal voorstellen wanneer het geschikt is om sensitiviteitsanalyses te uitvoeren. Deze voorstellen zijn gebaseerd op het verschil tussen de frailty varianties en de frailty correlatie. Wienke (2011) heeft aangetoond dat er een negatieve correlatie bestaat tussen de variantie en de correlatie schattingen. In onze simulaties zien we dat deze negatieve correlatie niet constant is en, in feite, afneemt wanneer de frailty correlatie toeneemt. Verder, volgens de verwachtingen, kan het gecorreleerde gamma frailty model ernstige identificeerbaarheidsproblemen ervaren wanneer de correlatie op de grens ligt van de parameterruimte. Voorzichtigheid moet genomen worden bij frailty correlaties kleiner dan 0,1 of wanneer deze dichter bij de kleinere verhouding tussen frailty varianties is (als die verhouding groter is dan 0,1).

De resultaten van onze tweede simulatiestudie zijn beperkt tot de gekozen waarden van de frailty varianties. Modelparameters en steekproefgrootte werden gekozen om de ACS-voorbeeld na te botsen. Gelijke frailty varianties, matig en groot verschil tussen frailty varianties worden verondersteld. In totaal werden 20 scenario's beschouwd met variërende frailty correlaties. Wij bieden een aantal algemene richtlijnenaan om de betrouwbaarheid van de resultaten te beoordelen wanneer het gecorreleerde frailty model toegepast wordt. Onze conclusies zijn beperkt tot de gekozen frailty parameters en de baseline hazard functie. Het is mogelijk dat andere opties voor de baseline hazard functies en andere waarden van de frailty parameters leiden tot verschillende resultaten. Het is moeilijk om extrapolaties uit te voeren op basis van dit onderzoek. Hoofdstukken 6 en 7 zijn gericht op de wiskundige modellen. Hier zijn we geïnteresseerd in het transmissie proces zelf in plaats van de risicofactoren of de vormen van infectiedruk. In Hoofdstuk 6 introduceren we twee fundamentele transmissiemodellen voor HCV en HIV en een gezamenlijk model dat gelijktijdig de overdracht van beide virussen modelleert.

Het gezamenlijke transmissiemodel wordt beschreven in Hoofdstuk 7, waar we gezamenlijke modellen toepassen op twee verschillende datasets. Bovendien worden de gezamenlijke modellen vanuit een statistisch perspectief beoordeeld. Het voorgestelde gezamenlijke wiskundige model houdt rekening met de biologische complexiteiten, waargenomen in de dynamiek van HCV- en HIV-transmissie in de IDG-context. Voor beide datavoorbeelden zijn er enkele overeenkomsten: de resultaten ondersteunen de modelonderstelling van twee verschillende risicogroepen (hoge en lage risico groep) in combinatie met onze eerste definitie van infectiedruk. Op basis van de statistische beoordeling zien we een betere modelfit voor vroegere blootstellingstijden, voornamelijk vanwege de beperkte hoeveelheid individuen met een langere duur van drugsinjecties. Op basis van de uitgevoerde statistische zijn analyses cruciale modelparameters bepaald. Enkele verschillen in de parameterschattingen voor beide datasets kunnen worden toegeschreven aan de studiepopulaties: individuen in het Itinere-project waren vooral druggebruikers op straat terwijl in het Vedette-project individuen in behandelingscentra werden opgevolgd. Onze resultaten met betrekking tot transmissieratios en percentage van personen die hersteld zijn van hun besmetting in overeenstemming met degenen vermeld door Vickerman et al. (2008, 2009) en De Vos et al. (2012). Ons model heeft een aantal beperkingen, zoals de veronderstelling van een invariante infectiedruk met de kalendertijd. Sommige studies hebben tijdsgebonden verschillen gerapporteerd in de HIV-prevalentie en het HIV-risico gerelateerd gedrag tussen IDG in Europa en het antiretrovirale therapie gebruik. De modellen, die in dit proefschrift gepresenteerd werden, dragen bij aan ons begrip van HCV- en HIV- transmissie in het kader van injecterende drugsgebruikers. De beperkingen van de modellen wijzen op nieuwe vragen die in verder onderzoek aangepakt zouden kunnen worden.