DOCTORAATSPROEFSCHRIFT

2008 | Faculteit Wetenschappen

Flexible Statistical Models for Microbial Risk Assessment and Infectious Diseases

Proefschrift voorgelegd tot het behalen van de graad van
Doctor in de Wetenschappen, richting wiskunde, te verdedigen door:

Harriet NAMATA

Promotor: Prof. dr. Marc Aerts
Copromotor: Prof. dr. Christel Faes

universiteit hasselt
CENTRUM VOOR STATISTIEK

D/2008/2451/44

To David Makonzi, my husband. Thank you for your patient endurance for all the time we have been far from each other.

When the Lord brought back
the captives to Zion,
we were like men who dreamed.
Our mouths were filled with
laughter,
our tongues with songs of joy.
Then it was said among the nations,
"The Lord has done great things for them."
The Lord has done great things for us,
and we are filled with joy.
Those who sow in tears will reap
with songs of joy.
He who goes out weeping,
carrying seed to sow,
will return with songs of joy,
carrying sheaves with him.

(Psalm 126, NIV Bible)

Acknowledgments

A longing fulfilled is sweet to the soul. As a girl in high school I desired to do a Bachelors degree in Statistics but I instead did Bachelor of Science with Education, Mathematics and Physics as my teaching subjects. I am grateful to this mathematical background which has led to masters degrees and now PhD in Statistics. Reflecting back whenceforth I have come I simply marvel. The mere term 'model' was a mystery I thought I would never grasp; even worse when categorical variables were involved. But now behold, what seemed a mystery has become a comprehensible and useful tool to provide answers to real problems. Indeed the road to success is not always straight. I acknowledge support from several people who have enabled me to walk through this journey.

My sincere thanks go to my supervisor Prof. Dr. Marc Aerts, to my co-supervisor Prof. Dr. Christel Faes and to Prof. Dr. Ziv Shkedy for guiding me and mentoring me in all the work that has led to this PhD. Thank you so very much. There were tough moments when I felt I missed it but you always led me back on the road. You have been a great inspiration to me; a student is not above his teacher, but everyone who is fully trained will be like his teacher (Luke 6:40). Every opportunity and challenge has been a building block to a greater learning and I am grateful to be a part of the world of statistical/mathematical modellers.

I would like to extend my thanks to the collaborating teams who provided the data and with whom I have worked to produce publications. Thanks to Dr. Peter Teunis, the biostatistician at the National Institute of Public Health and the Environment (RIVM), Bilthoven, The Netherlands. Thanks to the epidemiologists at the Veterinary and Agrochemical Research Center in Brussels: Dr. Estelle Méroc, Dr. Sarah Welby, Dr. Koen Mintiens and those I did not have a chance to meet in person. Furthermore I would like to thank Dr. Lucas Wiessing of the European Monitoring Centre for Drugs and Drug Abuse (EMCDDA), Lisbon, Portugal and Dr. Mirjam Kretzschmar at the Julius Centre for Health Sciences & Primary Care, University Medical Centre

Utrecht, The Netherlands and at the Center for Infectious Disease Control, RIVM, The Netherlands.

I recall my first day in Belgium, Saturday the 14th September 2002, which turned from joy to grief when I arrived at Brussels airport and there was no one to pick me as it was supposed to be. Luckily the good samaritan, Evans from California (who I lost contact with), led me to the Limburgs Universtair Centrum (LUC) at the time. Then I met Prof. Dr. Noel Veraverbeke who drove me to the student house, carried my heavy suitcase to the student room and took me to Spar supermarket to buy some necessities. My day became bright knowing I was finally at my destination. Thank you very much Noel and I remain grateful for what you did for me that day. With time I got to know many CENSTAT members and I must say it has been a wonderful experience at CENSTAT. Thanks to all CENSTAT members for the free and friendly environment to learn and work.

My special thanks go to the D3 group where I first celebrated my birthday. The nice decorations and gifts for such a special day coupled with the spirit of togetherness made me feel at home. Many thanks to: Prof. Dr. Ziv Shkedy, Veerle Vandersmissen, Dr. Caroline Beunckens, Suzy Van Sanden, and Arthur Gitome. I would also like to thank Annouschka Laenen who drove me to and from the Spanish classes (for the one year I followed the course). Thanks indeed, you saved me walking there especially in the winter season. To Martine Bernaert and all the Spanish class colleagues, thank you very much for your friendship. Haha! you gave me a joyful surprise when you celebrated my 30th birthday in class. Muchas gracias!

It was during the pursuit of this PhD that David and I got married. Our special thanks go to all who supported us and most especially the friends in Uganda who worked tirelessly, in our absence, to make our wedding a success. Our deep appreciation go to all who attended our wedding. Mwebale nnyo, Mukama abawe omukisa ogutaliko buyinike.

I am grateful for the spiritual nourishment I have received over the six years of my stay in Belgium. My heartfelt thanks go to the Evangelische Kerk in Diepenbeek, to The Same Anointing Ministries in Hasselt and to Christ Centered Church in Leuven. To: Luc and Nicole, and Els Gevaert of the Evangelische Kerk, Seth and Cecilia Yeboah, and to the Pastors Grace and Amponsah of The Same Anointing Ministries; thank you, respectively, for your hospitality, kindness, and advice. Pastor & Mrs Robinah Kirunda of Revival Tabernacle Ministries in Uganda, you are a real blessing in my life. I have seen God fulfill the promises He spoke to me through you in the year 2002 and I cannot forget the wonderful moments of prayer, free sharing and encouragement we had at Lungujja. Many thanks indeed and thanks to the members

of the Revival Tabernacle Ministries.

Just as wisdom is a shelter so money is a shelter (Ecclesiastes 7:12). Without the financial support from the Vlaamse Interuniversitaire Raad (VLIR) (for the masters) and the Bijzonder Onderzoeksfonds (BOF) "BOF04G01" (for the PhD), all these experiences in Belgium would not have happened. More to the academic achievement, I had the opportunity to travel to other nations for conferences or visits and this to me was a superb adventure of the globe. Thank you very much for your generosity.

The name of the Lord, Jesus Christ, be praised for his unfailing love and his wonderful deeds for men.

Harriet Namata

September 15, 2008

Contents

1 Introduction **1**
1.1 Research Problem 3
1.2 Modeling Human to Human Infectious Diseases Data 4
1.3 Dose-Response Modeling of Food-Borne Infectious Diseases 6
1.4 Modeling Data on *Salmonella* Infection in Belgian Chicken Flocks . . 9

I Human to Human Infectious Diseases Data Modeling **13**

2 Estimation of the Force of Infection from Current Status Data Using Generalized Linear Mixed Models **15**
2.1 Estimation of Prevalence from Current Status Data 18
2.2 Smoothing Binary Data Using GLMM 19
2.2.1 Generalized Linear Mixed Models 19
2.2.2 Penalized Spline Formulation as GLMM 20
2.2.3 Parameter Estimation 21
2.2.4 Estimation of the Force of Infection 22
2.3 Data Analysis 23
2.3.1 Bootstrap Confidence Intervals 25
2.4 Simulation 26
2.5 Discussion 32

3 Modeling the Force of Infection for Parvovirus B19 in Europe Using Penalized Spline Models **39**
3.1 Data 41
3.2 Estimating the Force of Infection Using Penalized Splines 43
3.2.1 Simple GLMM Spline Model 43
3.2.2 Extension of the Basic Model 44

3.2.3 Model (3.2) . . . 45
3.2.4 Model (3.3) . . . 45
3.2.5 Model (3.4) . . . 45
3.2.6 Proportional Odds and Proportional Hazard Models . . . 46
3.3 Estimation and Model Selection . . . 48
3.3.1 Quasi-Likelihood Estimation . . . 48
3.3.2 Hierarchical Bayesian Modeling . . . 49
3.4 Application to the Data . . . 50
3.4.1 Quasi-Likelihood Estimation . . . 50
3.4.2 Full Bayesian Approach . . . 57
3.5 Piecewise Constant Force of Infection . . . 60
3.6 Discussion and Conclusion . . . 66

4 Estimation of the Prevalence and Force of Infection of Hepatitis C Among Injecting Drug Users in Five European Countries 69
4.1 Data and Methods . . . 71
4.1.1 Study Design . . . 71
4.1.2 The Exposure Time - The Length of the Injection Career . . . 77
4.1.3 Statistical Methodology . . . 79
4.2 Data Analysis . . . 81
4.2.1 Descriptive Analysis . . . 81
4.2.2 Modeling the Prevalence and Force of Infection . . . 84
4.2.3 Second Analysis: IDUs With Recent Injecting Career . . . 91
4.3 Discussion . . . 92

II Dose-Response Modeling of Food-Borne Infectious Diseases 95

5 Model Averaging in Microbial Risk Assessment Using Modified Fractional Polynomials and Generalized Linear Mixed Models 97
5.1 Microbial Dose-Response Models . . . 99
5.1.1 A Generic Mechanistic Dose-Response Model . . . 100
5.1.2 Fractional Polynomials . . . 104
5.2 Model Averaging Approach . . . 106
5.3 Application to Single Strain Data . . . 107
5.3.1 Salmonella Typhi . . . 109
5.3.2 Campylobacter Jejuni . . . 111

5.4 Simulation Study for Single Strain Data . . . 116
5.4.1 First Setting . . . 117
5.4.2 Second Setting . . . 119
5.4.3 To Include Fractional Polynomials or Not . . . 123
5.5 Application to Multi-Strain Data . . . 124
5.6 Simulation Study for Multi-Strain Data . . . 131
5.7 Discussion . . . 134

III Modeling Data on *Salmonella* Infection in Broiler and Layer Chicken Flocks 137

6 Risk Factor Identification for *Salmonella* in Belgian Laying Hens 139
6.1 Material and Methods . . . 141
6.1.1 Data Collection . . . 141
6.1.2 Single-Level Analysis . . . 142
6.1.3 Two-Level Analysis . . . 142
6.1.4 Three-Level Analysis . . . 144
6.2 Results . . . 145
6.2.1 Data Exploration . . . 145
6.2.2 Data Analysis . . . 148
6.3 Discussion . . . 152

7 Prevalence and Persistence of *Salmonella* in Belgian Broiler Chicken Flocks: An Identification of Risk Factors. 155
7.1 Materials and Methods . . . 157
7.1.1 Data Collection . . . 157
7.1.2 Data Description . . . 158
7.1.3 Data Analysis . . . 161
7.2 Results . . . 163
7.2.1 Data Description . . . 163
7.2.2 Conditional Analysis . . . 168
7.2.3 Joint Analysis . . . 173
7.3 Discussion . . . 175

8 Concluding Remarks and Future Research 179
8.1 Modeling Human to Human Infectious Diseases Data . . . 179
8.2 Dose-Response Modeling for Food-borne Infectious Diseases . . . 180

8.2.1 Single Strain 180
8.2.2 Several Strains 181
8.3 Modeling Data on *Salmonella* Infection in Broiler and Layer Chicken Flocks 183

References **185**

Samenvatting **203**

Chapter 1

Introduction

Pathogenic microorganisms, such as bacteria, viruses, parasites or fungi are responsible for several infectious diseases that bother human and animal health worldwide (Haas *et al.* 1999; FAO/WHO, 2003). Infectious diseases can range from the common illnesses, such as the cold, to deadly illnesses, such as HIV/AIDS. They are referred to as infectious due to their potentiality to be transmitted from one person or species to another. Human to human infectious diseases can be spread through the following ways: sexual transmission e.g. hepatitis B, HIV/AIDS; airborne transmission through inhaling airborne droplets of the organism, which may exist in the air as a result of a cough or sneeze from an infected person e.g. influenza; blood-borne transmission through contact with infected blood, such as through blood transfusions or when sharing contaminated needles and syringes e.g. hepatitis B and C, HIV/AIDS; and through direct skin contact with an infected person e.g. measles. Infectious diseases, such as malaria, can also be transmitted to humans through insects, such as mosquitoes, which draw blood from an infected person and then bite a healthy person. Food- and water-borne infectious diseases are transmitted to humans by consumption of contaminated food and water e.g. typhoid fever. Furthermore consumption of foods of animal origin, particularly eggs, meat and milk products, can lead to zoonotic diseases (transferred from animals to humans) like *Salmonella*. Zoonotic diseases can also be airborne like avian influenza.

Although all these infectious diseases are caused by harmful microorganisms, the literature on microbiological risk assessment is devoted to food- and waterborne infectious pathogens (Haas *et al.* 1999; FAO/WHO, 2003). However, it can be seen in general that the steps of microbiological risk assessment apply for infectious diseases

of humans. The first step in microbial risk assessment deals with identifying the infectious agent and the associated adverse effect based on various data sources such as clinical literature, clinical microbiologists, case studies, hospitalization studies, laboratory animal studies and other epidemiological data. This step can be generalized to all infectious diseases as the source of the disease must be identified before control and prevention interventions can be established.

In the second step of microbial risk assessment, which is exposure assessment, scientists seek to determine the size and population exposed to a pathogenic microorganism, the routes of exposure, the quantities exposed to, duration of exposure and whether the exposure was continuous or intermittent. While the quantities of exposure are often termed as dose for food-borne diseases, they are equivalent to the frequency of injecting among injecting drug users for human to human infectious diseases such as hepatitis C (HCV). Increased doses can occur due to exposure to a single large dose of the pathogen, repeated doses of the pathogen or prolonged duration of exposure or a cumulative dose that survives in the body. For HCV and sexually transmitted diseases, the duration of exposure is the exposure time to the contaminated objects or infected persons while for diseases like measles the duration is the age.

The key concept in the third step of microbial risk assessment, which is dose-response assessment, is to evaluate the relationship between the microbial dose and the adverse effect. Figure 1.1 shows the scatter plots of food-borne disease data (panel a) and human to human infectious diseases data (panels b to d). Strong relationships can be seen between *Salmonella Typhi* and log(dose); between HCV seroprevalence and the frequency of injection among injecting drug users in Czech Republic – per month or per week or per day; between HCV seroprevalence and the exposure time in Belgium; and between rubella seroprevalence in the UK versus the age of the individual.

The final step of microbial risk assessment is risk characterization. This step integrates the information from the previous steps in order to estimate the magnitude of the public health problem in an exposed population taking into account variability and uncertainty at each step. For food- and water-borne diseases the quantification of risk is a function of the hazard and exposure dose. For human to human infectious diseases the risk can be expressed in terms of age or number of contacts or exposure time. The curves for the predicted probabilities added to the scatter plots in Figure 1.1 give an easy-to-comprehend view on the relationship of the exposure variable (horizontal-axis) on the occurrence of the diseases (vertical-axis). An important issue in the risk characterization of human to human infectious diseases considers the ratio of the first derivative of the estimated probability function and the complement of the

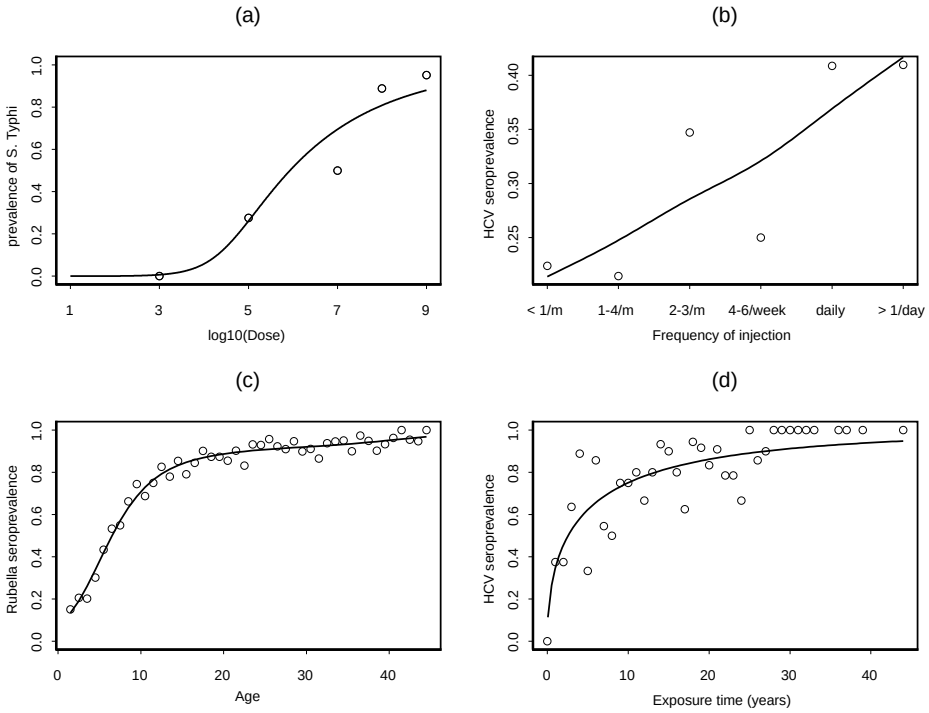

Figure 1.1: *Scatter plot of the observed and estimated probability of (a) Salmonella Typhi as a function of dose, (b) hepatitis C against the frequency of injecting (c) rubella as a function of age and (d) hepatitis C as a function of exposure time.*

probability to give the rate at which people become infected. The rest of the thesis focuses on specific risk characterizations as will be discussed further on.

1.1 Research Problem

The main area of work of this thesis was using data and mathematical and/or statistical models to estimate risks and trends as well as to identify risk factors associated with some bacterial and viral microbial agents in order to enable epidemiologists to improve the understanding of the epidemiology of these infectious diseases and evaluate the impact of intervention programmes against the diseases. It weaves together different research problems and depending on the data at hand different aspects regarding the statistical methods are emphasized. The thesis is divided into three

parts which deal, respectively, with modeling data on infectious diseases of humans, dose response models for foodborne infectious diseases and identifying risk factors for *Salmonella* infection in Belgian chicken flocks. How the various modeling techniques have been integrated into these parts is explained in the following sections.

1.2 Modeling Human to Human Infectious Diseases Data

In the first part we analyze cross-sectional current-status data from serological diagnostics for less to more severe viral infections like rubella, varicella, mumps, parvovirus B19, hepatitis C virus, hepatitis B virus and HIV. For an example, Figure 1.2 depicts the data sets, for the occurrences of rubella and mumps in the UK and varicella in Belgium, with age. There is relevant information represented by the points in the plots but it is very difficult to draw any conclusion from this alone. With mathematical or statistical modeling the data sets can be reduced to summaries that can give insights in the epidemiology of the disease and that can be used for prediction and can be integrated in the mathematical modeling of the disease. For a typical childhood infectious disease, the flow diagram in Figure 1.3 shows how the individuals move from the susceptible class (X) at birth, become infected (Y), recover and acquire life-long immunity (Z). Figure 1.3 represents a basic SIR (Susceptible-Infected-Removed) model but more complex models, assuming maternal antibody and a latent period can be used to describe the flow of individuals between the disease stata. One very important parameter of interest in infectious disease epidemiology is the force of infection $\ell(a)$, which is the rate at which the susceptible individuals become infected. It is assumed to vary across age groups (discussed further in Part I). The modeling of the force of infection can take on parametric models, nonparametric models or semi-parametric models. Muench (1934) modeled the force of infection using a constant model while Griffiths (1974) used the linear model and Grenfell and Anderson (1985) employed a more flexible polynomial model. Shkedy *et al. (2006)* used fractional polynomials to model age dependent force of infection. Hens *et al.* (2007) illustrates joint modeling of the force of infection for varicella-zoster virus and the parvo B19-virus in Belgium using flexible marginal and conditional models. Shkedy *et al* (2003) proposed to use local polynomials as a nonparametric approach. Keiding (1991) proposed modeling the force of infection using isotonic regression models. Keiding *et al* (1996) used an alternative modeling approach based on natural cubic splines. Nagelkerke *et al* (1999) estimated the force of infection using a semi-parametric approach via smoothing cu-

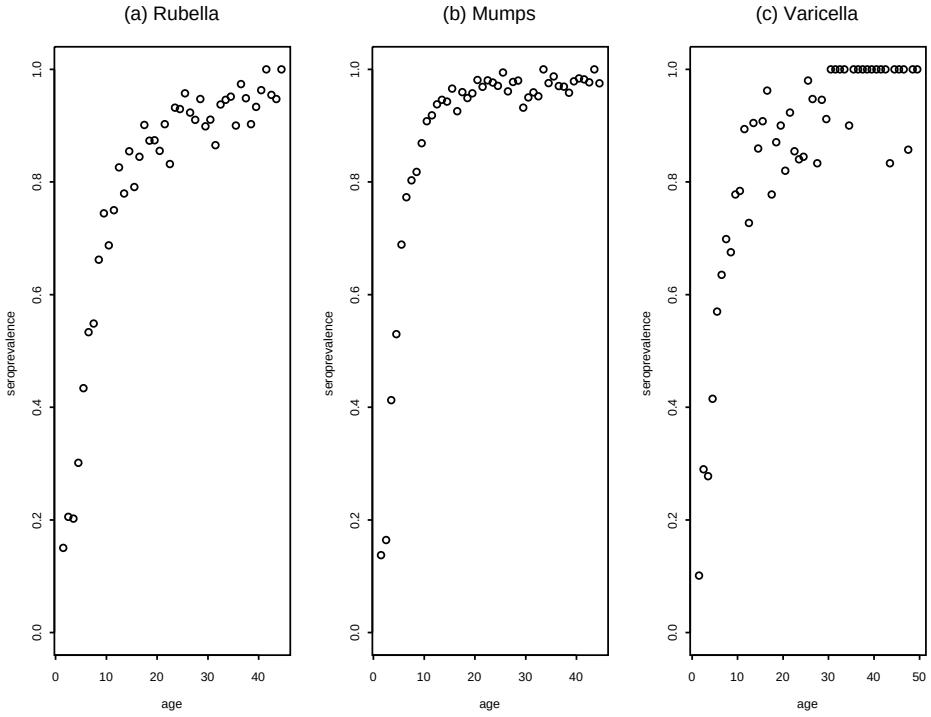

Figure 1.2: *Serological dataset: (a) rubella and (b) mumps in the UK and (c) varicella in Belgium.*

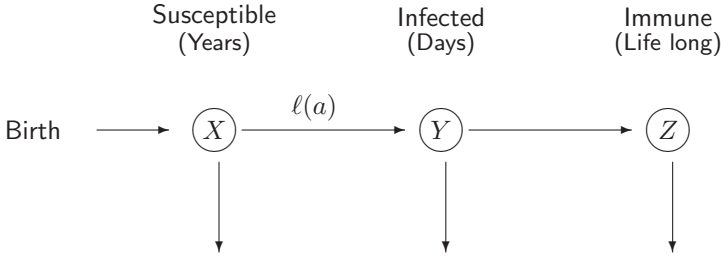

Figure 1.3: *Illustration of the SIR model. The individuals are entered into the susceptible class, then move to the infected class and after recovering they move into the immune class. The force of infection, $\ell(a)$, will be discussed further in Part I.*

bic splines and the proportional hazards model. In Chapter 2 we estimate the force of infection using nonparametric regression based on penalized regression splines and

generalized linear mixed models, with the nonparametric component involving a single continuous predictor (e.g. age). Chapter 3 extends the nonparametric approach in Chapter 2 with a discrete predictor in various ways and shows the flexibility of penalized splines as they relate to proportional odds, proportional hazard models, and constant piecewise force of infection. In Chapter 4, in addition to parametric and nonparametric modeling of the prevalence and force of infection, we investigate risk factor behaviors associated with hepatitis C virus among injecting drug users in five European countries.

1.3 Dose-Response Modeling of Food-Borne Infectious Diseases

Food-borne illness is among the most widespread public health problems and creates social and economic burdens in addition to human suffering. Figure 1.4 (Haas *et al.* 1999) shows some of the ways microbes can be transferred from stool to food or water, which can result in diseases like salmonellosis. Once exposure via ingestion has taken place the major steps involved in the food-borne disease process are shown in Figure 1.5 (FAO/WHO, 2003). Each ingested organism has a probability of surviving all barriers to reach a target site for growth or multiplication. While infections may be asymptomatic, where a host does not develop any adverse reactions to the infection and eliminates the pathogens within a limited period of time (i.e. recovers), infections may also lead to symptomatic illness. In a small fraction of ill cases, chronic infection or sequelae may occur and there may be risk to mortality related to sequelae or acute disease.

The key concept in developing dose response models is the relation between the actual surviving organisms (the effective dose) and the probability of occurrence of an adverse event to the host when considering dichotomous responses. This relation has been described using mathematical and statistical functions based on biological and empirical rationales (Haas *et al.* 1999). The exponential and the Beta-Poisson models have received much attention in dose response modeling owing to their biological derivation (Haas *et al.* 1999). These models fall in the class of hit-theory models, which assume that a single infectious microorganism surviving within the host may result in infection (Teunis *et al.* 1996). Alternative models, though not widely used in microbial risk assessment, which assume the existence of a threshold level of pathogens that must be ingested in order for the microorganism to produce infection or disease are discussed by (Haas *et al.* 1999). In addition, empirical mod-

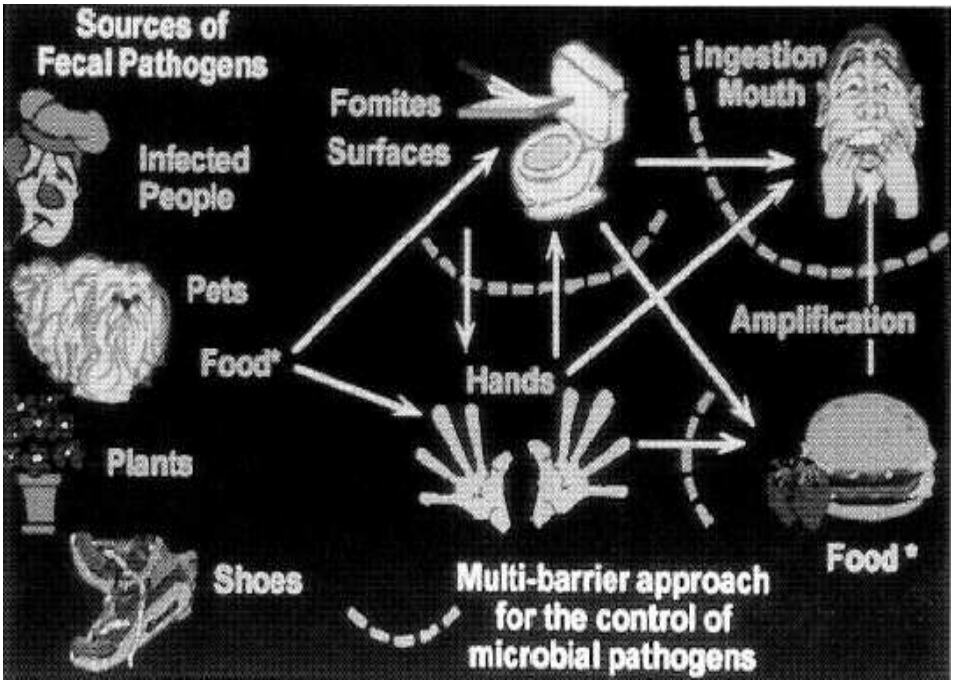

Figure 1.4: *Routes of transmission in the home for fecal-oral microorganisms*

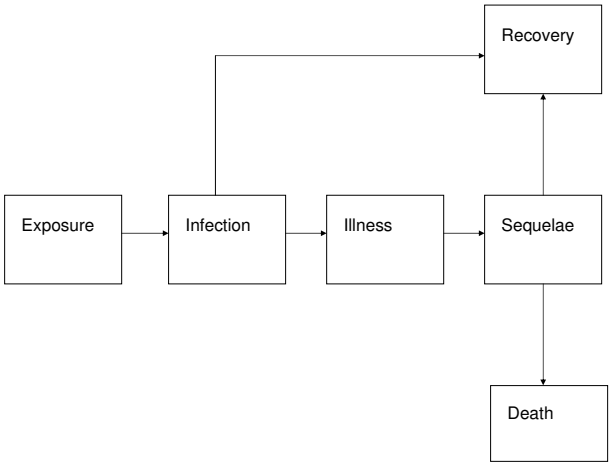

Figure 1.5: *The major steps in the foodborne infectious disease process*

els (primarily for chemical agents), which assume that the population exposed has a tolerance distribution for an adverse effect have been used in microbial dose response assessment (Haas *et al.* 1999). If the population is exposed to the pathogen at a certain level, all individuals of the population who have a tolerance less or equal to the dosed level will exhibit the adverse effect. Essentially any probability density function with support over the positive line can be a tolerance distribution. Despite the use of empirical models they were not regarded as biological plausible (Haas *et al.* 1999). However, starting from the general mechanistic framework to derive the exponential and Beta-Poisson models, Kodell *et al.* (2002) derive the empirical models such as the log logistic, log probit, the extreme value and other models thus rendering them biologically plausible. This shows that in essence many functions can be derived that are flexible enough for dose-response relations. Chapter 5 of part II extends the above mentioned dose response models with modified fractional polynomial models (starting with the fractional polynomials by Royston and Altman, 1994) that are formulated to satisfy biological plausibility. However, when extrapolating outside the region of observed data, all possible models may predict widely differing results (Coleman and Marks, 1998; Holcomb et al. 1999). This necessitates a selection of the best model or a set of models to use.

The traditional approach was based on one best model selected according to some statistical criterion for goodness of fit such as the Akaike Information Criterion (AIC), Kullback Information Criterion (KIC) or Deviance Information Criterion (DIC) in the case of full Bayesian models. This approach, however, ignores the other possible models and one makes statistical inference based on the single selected model. In addition, many different models (as shown in Chapter 5) will usually fit a given data set equally well and therefore goodness of fit is not a sufficient criterion for model selection. Over the past decade research in risk assessment has been directed to study and incorporate uncertainty arising from the alternative dose response models. Buckland *et al.* (1997) proposed a way to incorporate model uncertainty by averaging across a plausible set of candidate models using Akaike weights and this has been employed by other researchers (Burnham and Anderson, 2002). In Chapter 5 we present this model averaging approach to estimate the risks of *Salmonella Typhi* and *Campylobacter jejuni* at low doses using the proposed modified set of fractional polynomials in addition to the Beta-Poisson model and the classical empirical models.

1.4 Modeling Data on *Salmonella* Infection in Belgian Chicken Flocks

Salmonella, named after the American veterinary pathologist Daniel Elmer Salmon, was first isolated in 1885 from pigs (Microsoft® Encarta®, 2008). The bacterium is a genus of rod-shaped infectious bacteria that is transmitted to humans through consumption of contaminated poultry, eggs, pork and certain other foods and can cause diseases of the intestines. *Salmonella Enteritidis* is one of the species that infects chicken flocks without causing visible disease, and can spread from hen to hen rapidly. The left panel of Figure 1.6 (Microsoft® Encarta®, 2008) shows a *Salmonella* bacterium which can move by means of fine threadlike projections called flagella. The arrangement of flagella across the surface of the bacterium differs from species to species; they can be present at the ends of the bacterium or all across the body surface. The right panel of Figure 1.6 shows the *Salmonella Enteritidis* species (Kunkel, 2007).

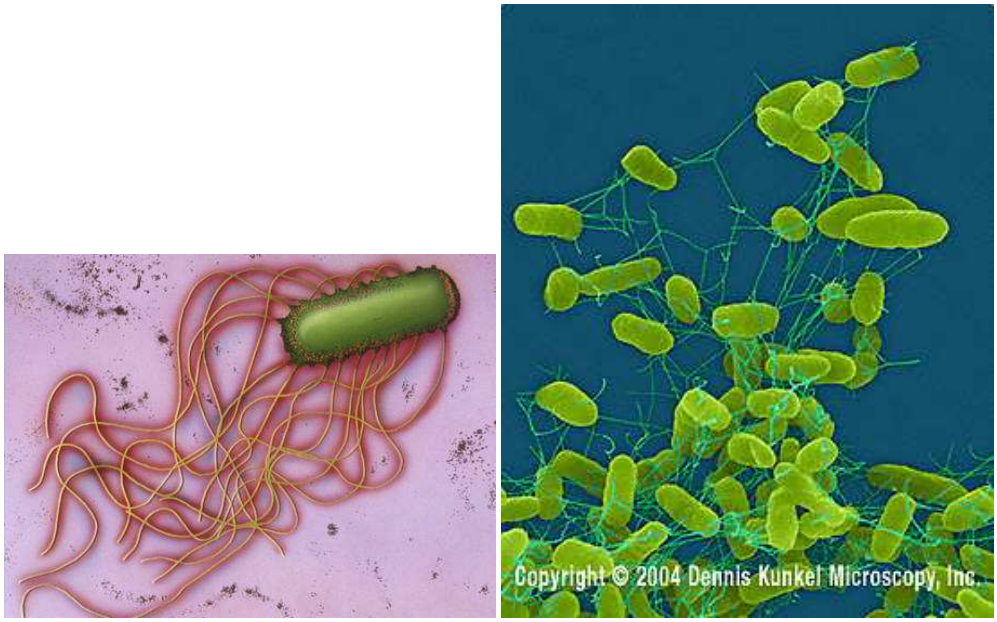

Figure 1.6: *Left Panel: Salmonella bacterium showing flagella. Right panel: Salmonella Enteritidis - rod prokaryote (bacterium). This zoonotic microorganism causes salmonellosis (food poisoning) in humans when infected poultry contaminate eggs, poultry meat (which humans ingest).*

In Belgium, the identification of the flock's or farm's *Salmonella* status is often times based on testing a number of different samples from the same flock or farm, giving rise to correlated data. Such clustered binary responses, disease status in this case, also frequently arise in other epidemiologic applications. The scientific objectives involve: (i) modeling the marginal mean responses, such as the probability of disease, and the within-cluster association of the multivariate responses and (ii) modeling the cluster-specific responses and the heterogeneity of clusters. In this regard, statistical models which incorporate and study the clustered type of data are a useful procedure. They are extensions of the well-known logistic regression that is a particular case of the generalized linear models with binary response data and a logit link function (McCullagh and Nelder, 1989). They are usually classified into marginal (a population averaged) and random-effects models. We will briefly describe two marginal models, generalized estimating equations (GEE) and the alternating logistic regression (ALR) models and the general form of random effects models. For details about these models the reader is referred to Liang and Zeger (1986), Carey *et al.* (1993), Agresti (2002), Molenberghs and Verbeke (2005) and Aerts *et al.* (2002).

The generalized estimating equations method, originally proposed by Liang and Zeger (1986) also outlined by Bobashev and Anthony (1998) is a commonly used marginal model for clustered data which accounts for the correlation of a disease within clusters. Let Y_{ij} denote the jth response at time point $t_{ij}(j = 1, \ldots, n_i)$ for cluster $i(i = 1, \ldots, N)$ with expectation π_i and a working covariance matrix V_i. This covariance matrix V_i is an $n_i \times n_i$ matrix where the jth diagonal elements denote the variance for the jth observation in the ith cluster and the off diagonal elements specify the covariance between two different units (j, k) in the ith cluster. Formally, this amounts to

$$V_i = \text{cov}(Y_{ij}, Y_{ik}) = \begin{cases} \pi_{ij}(1 - \pi_{ij}) & \text{if } j = k \\ \\ \text{corr}(Y_{ij}, Y_{ik}) \times [\pi_{ij}(1 - \pi_{ij})\pi_{ik}(1 - \pi_{ik})]^{1/2} & \text{if } j \neq k \end{cases}$$

where $\pi_{ij} = \text{E}(Y_{ij} = 1)$. The term $\text{corr}(Y_{ij}, Y_{ij})$ must be given a working correlation pattern in the analysis. Several choices are possible for the working form of the covariance matrix, ranging from the most simple assumption of independence ($\text{corr}(Y_{ij}, Y_{ij}) = 0$ if $j \neq k$) within clusters to the most complex unstructured form, where all parameters vary. It must be emphasized that estimation of the mean structure is consistent whatever the true correlation structure is, but efficiency is optimal

when using an appropriate working covariance structure (Liang and Zeger, 1986). The intra-cluster correlation, however, is treated as a nuisance with the GEE approach.

The alternating logistic regression (ALR) proposed by Carey *et al.* (1993) is another marginal model that explicitly models the clustering of a disease within clusters. The model yields a readily interpretable statistical index of a disease clustering in the form of a "pairwise odds ratio" (PWOR). In the literal sense, the PWOR reflects how strongly a disease occurs in clusters. In more technical terms, the PWOR reflects odds of a disease for a unit in a cluster given that another randomly chosen unit from that cluster has a disease, relative to the odds if that randomly chosen unit does not have a disease. The logarithm of the PWOR can be expressed as a function of an indicator variable coded to show whether units j and k in a pair belong to the same or different clusters:

$$log(PWOR_{jk}) = \alpha F_{jk},$$

where F_{jk}, takes values 1 or 0, depending on whether the pair (j, k) belongs to the same cluster. The ALR model, therefore, alternates between estimating the mean structure using the logistic regression and estimating the disease clustering using the pairwise odds ratio. It should be noted that when the association is of interest, the ALR model is preferred to the GEE approach.

The third method incorporates clustering of a disease in clusters through shared random effects. This involves the random components inside the linear predictor of ordinary logistic regression model, i.e random effects logistic regression model

$$\text{logit}(E(Y_{ij}|X_{ij}, Z_{ij}, u_i)) = X'_{ij}\beta + Z'_{ij}u_i$$

where the random effects u_i are assumed to vary independently from one cluster to another according to a common distribution, usually the normal distribution with mean 0 and an unknown variance, σ^2. Z_{ij} is often a subvector of X_{ij}, which means that random effects apply only to a part of the covariates and/or the intercept. The random effect variance is interpreted as the variation in $\text{logit}(\pi_i)$ between clusters after having accounted for fixed effects. With an approximate variance for the binary outcome the intra-class correlation (ICC) (correlation between two units in the same cluster) can be computed as the sum of variance components of common random effects divided by the total variation (fixed effects variation plus random variation).

Part III of this thesis investigates the risk factors for Salmonella in broiler and egg laying chicken flocks in Belgium and since these data are clustered the above statistical models have been employed.

Part I

Human to Human Infectious Diseases Data Modeling

Chapter 2

Estimation of the Force of Infection from Current Status Data Using Generalized Linear Mixed Models

The occurrence of infectious diseases, both in industrialized and economically developing countries, cause substantial health and economic impacts. This has provoked the emergence of various techniques to estimate the disease burden and the impact of interventions aiming to prevent and control the spread of infectious diseases in populations.

The use of quantitative methods based on mathematical models to study the transmission dynamics of infectious diseases has increased in importance for scientists, policy makers, and health professionals. Many of these models are deterministic and based on a set of differential equations which describe the course of individuals from one phase to another. In this chapter, we presume lifelong immunity and negligible mortality caused by the infection, two commonly made assumptions (Keiding *et al.* 1996; Shkedy *et al.* 2003). Let $q(a, t)$ denote the fraction of individuals at risk for the infection at age a and time t. Under the assumptions specified above, the change in

this fraction can be described by the partial differential equation

$$\frac{\partial}{\partial a}q(a,t) + \frac{\partial}{\partial t}q(a,t) = -\ell(a,t)q(a,t), \tag{2.1}$$

where $\ell(a,t)$ represents the rate at which susceptible individuals become infected, the force of infection. The natural death rate is assumed to be zero up to the life expectancy and infinity thereafter. In a steady state, that is the time homogeneous form, $\partial q(a,t)/\partial t = 0$, (2.1) reduces to the differential equation

$$\frac{\partial}{\partial a}q(a) = -\ell(a)q(a), \tag{2.2}$$

which describes the change in the fraction of susceptible individuals with age of the individual.

The force of infection can be estimated from an age-specific cross-sectional prevalence sample, which is a sample taken at a certain time point and for each of the individuals in the sample the observed information consists of whether the individual has been infected or not before his or her age at the test. Assuming that the disease is in a steady state, then the age-dependent force of infection can be modelled according to equation (2.2).

Viewing a cross-sectional serological sample as a special case of current status data allows using terminology from modeling survival data. This type of data consists of information about the individual's age and whether or not a specific event occurred before the individual's age at the time of the test. In our setting an event is infection by a disease. Individuals who experienced the event before age at test are left censored while those who experienced it not before their age at test are right censored. Non-parametric approaches for the estimation of the prevalence and of the force of infection, discussed by Keiding (1991) and Keiding *et al.* (1996) used isotonic regression. This method estimates the prevalence, $\pi(a) = 1 - q(a)$ by a step function $\hat{\pi}(a)$. For the force of infection, Keiding (1991) suggested a kernel smoothed estimate $\int K\{(a-u)/h\}/h\{1-\hat{\pi}(u)\}d\hat{\pi}(u)$ where K is a kernel and h a bandwidth. In order to avoid the two-step procedure based on isotonic regression, Keiding *et al.* (1996), used an alternative modeling approach based on natural cubic splines. Nagelkerke *et al.* (1999) estimated the force of infection using a semiparametric approach via smoothing cubic splines and the proportional hazards model. Shkedy *et al.* (2003) proposed to use local polynomials which simultaneously estimate prevalence and force of infection. For a given link function, Shkedy *et al.* (2003) estimated the local force of infection by $\ell(a) = \eta\prime(a)\delta\{\eta(a)\}$ where the form of δ is determined by the link function and $\eta(a)$ is the functional form of the predictor, age, locally approximated

by a polynomial of order p. By using the local polynomial method, both $\pi(a)$ and $\ell(a)$ are estimated simultaneously as a smooth function of age.

In this chapter, we extend the alternative modeling approach of Keiding *et al.* (1996) by using penalized splines with truncated polynomial basis and generalized linear mixed models (Ruppert *et al.*, 2003) which can be estimated using the SAS GLIMMIX procedure and macro. Eilers and Marx (1996) introduced penalized splines as a regression with B-splines penalizing the $(q+1)$-th order difference in the B-spline coefficients, for a B-spline of degree q. The underlying idea of penalized spline smoothing is to fit a smooth curve by using a high dimensional basis but, instead of simple parametric fitting, a penalized version is pursued to provide a smooth fit. This approach resembles smoothing splines, the major difference being that for smoothing splines the dimension of the corresponding spline basis grows with sample size while with penalized spline smoothing a finite dimensional basis is used. A connection between smoothing splines and mixed models is discussed in Verbyla *et al.* (1999). Not only does penalized spline smoothing permit flexible choices of the spline model, it also has strong links to linear mixed models and to penalized quasi-likelihood (PQL) estimation in generalized linear mixed models (Ruppert *et al.*, 2003). The smoothing parameter is selected based on the generalized linear mixed model (GLMM) framework which is equipped with an automatic smoothing parameter choice which corresponds to PQL and restricted maximum likelihood (REML) estimation of the variance components. A practical advantage for this methodology is that software to implement it is now accessible through statistical packages such as SAS.

The proposed method was applied to serological datasets of rubella and mumps in the UK and varicella in Belgium (October 1999 to April 2001) shown again in Figure 2.1. The first two datasets were discussed in Whitaker and Farrington (2004) and Shkedy *et al.* (2003). They consist of 4230 and 8179 individuals for rubella and mumps, respectively, aged 1.5 to 44.5 years old (Figures 2.1(a) and 2.1(b)). The varicella dataset contains the serological results of 2027 Belgian individuals together with their age in years, ranging from 1.5 to 49.5 years (Figure 2.1(c)). These diseases are common airborne childhood infectious diseases spread by droplet and airborne transmission. This chapter is organized as follows. Section 2.1 sets out on the estimation of the prevalence from current status data. We discuss how binary data can be smoothed using GLMM in Section 2.2. Section 2.3 presents data analyses based on the proposed approach. The results of simulation studies follow in Section 2.4 and

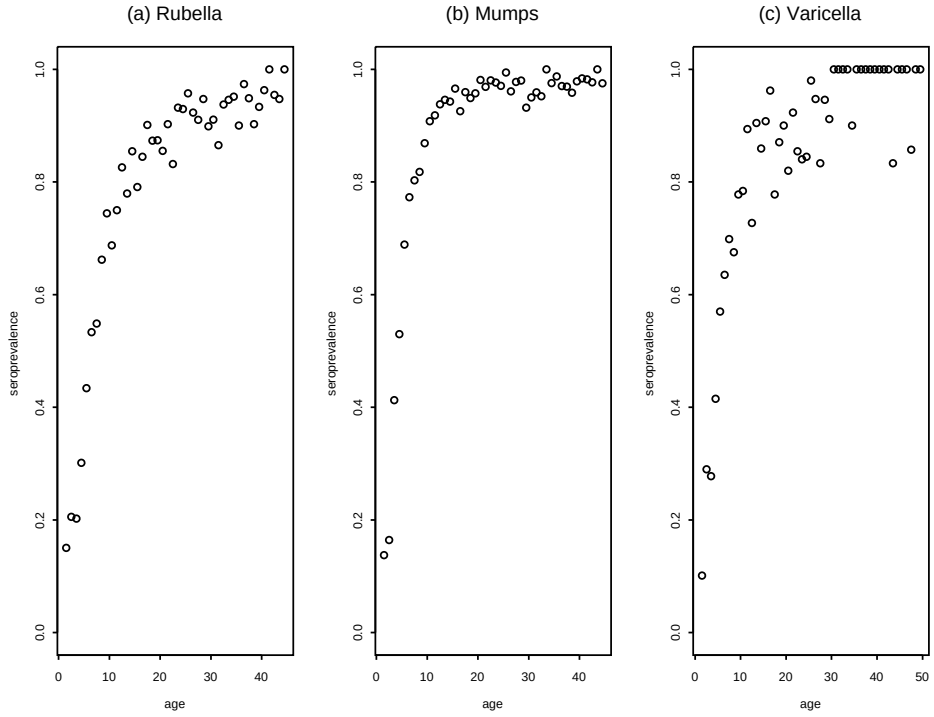

Figure 2.1: *Serological dataset: (a) rubella and (b) mumps in the UK and (c) varicella in Belgium.*

finally a discussion of the results is presented in Section 2.5. The work of this chapter has been published in Namata *et al.* (2007).

2.1 Estimation of Prevalence from Current Status Data

We consider an age-specific cross-sectional prevalence sample of size N and let a_i ($i = 1, 2, \ldots, N$) be the age of the ith individual. A cross-sectional prevalence sample is a current status sample in which all individuals are censored. The individuals who experienced an infection before their age a_i at test are left censored while those who did not experience an infection before their age at test are right censored. Instead of

observing the age at infection, we observe a binary response indicator

$$Y_i = \begin{cases} 1 & \text{if individual } i \text{ experienced an infection before age } a_i \text{ (left-censored),} \\ 0 & \text{otherwise (right-censored).} \end{cases}$$

This gives rise to an independent and identically distributed sample $(a_1, Y_1), \ldots, (a_N, Y_N)$. Keiding *et al.* (1996) estimated the distribution function $\pi(a) = 1 - \exp\left\{\int_0^{a_0} \ell_0(a)da\right\}$, the cumulative probability of being infected by age a as a non-parametric maximum likelihood estimator which is a step function and the force of infection, $\ell_0(a)$ was estimated by natural cubic splines with the smoothing parameter chosen by inspection. We propose to semi-parametrically estimate the prevalence by a smooth curve using penalized splines and generalized linear mixed models, an approach which automatically selects a smoothing parameter. The force of infection easily derives from the calculated derivative of the fitted curve. When there are decreases in the prevalence the derived force of infection is negative, which would be nonsensical from an epidemiological perspective. To ensure positivity of the force of infection, we monotonize the prevalence using a pool-adjacent-violators (PAV) algorithm proposed by Robertson *et al.* (1988).

2.2 Smoothing Binary Data Using GLMM

2.2.1 Generalized Linear Mixed Models

Generalized linear mixed models are commonly used as an extension of the generalized linear models, formulated for univariate data, since they allow for correlated responses through the inclusion of random-effect terms in the linear component (McCulloch and Searle 2001). The framework can also be used to smooth data (Ruppert *et al.*, 2003; Verbyla *et al.* 1999). Let us first review the generalized linear mixed model and associated parameter estimation. Consider the cross-sectional sample (a_i, Y_i) from Section 2.1. Let us assume a canonical link function (McCullagh and Nelder, 1989) and no overdispersion, then the GLMM can be written in one-parameter exponential family notation as

$$f(\mathbf{y}|\mathbf{u}) = \exp\left[\mathbf{y}^T(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) - \mathbf{1}^T b(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) + \mathbf{1}^T c(\mathbf{y})\right], \tag{2.3}$$

where $\mathbf{X}$ and $\mathbf{Z}$ are p-dimensional and q-dimensional vectors of age values a_i and in general possibly other variables, $\mathbf{1}$ is the vector of ones, $\boldsymbol{\beta}$ is the fixed effects vector and $\mathbf{u}$ is the random effects vector which has the normal density,

$$f(\mathbf{u}) \quad = \quad (2\pi)^{-q/2}|\sigma_u^2\mathbf{I}|^{-1/2}\exp\left[-\frac{1}{2}\mathbf{u}^T(\sigma_u^2\mathbf{I})^{-1}\mathbf{u}\right].$$

The marginal likelihood as a function of $(\boldsymbol{\beta}, \sigma_u^2)$ is given by,

$$\begin{aligned} l(\boldsymbol{\beta}, \sigma_u^2) &= \int_{\Re^q} f(\mathbf{y}|\mathbf{u}) f(\mathbf{u}) d\mathbf{u} \\ &= (2\pi)^{-q/2} |\sigma_u^2 \mathbf{I}|^{-1/2} \exp\{1^T c(\mathbf{y})\} J(\boldsymbol{\beta}, \sigma_u^2) \end{aligned}$$

with $J(\boldsymbol{\beta}, \sigma_u^2) = \int_{\Re^q} \exp\{\mathbf{y}^T(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) - 1^T b(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) - \frac{1}{2}\mathbf{u}^T(\sigma_u^2\mathbf{I})^{-1}\mathbf{u})\} d\mathbf{u}$. Based on a Laplace approximation of $J(\boldsymbol{\beta}, \sigma_u^2)$, the likelihood is approximated by a quasi-likelihood function which is penalized, to achieve smoothness and numerical stability, by introducing a penalty term, $-1/2\mathbf{u}^T(\sigma_u^2\mathbf{I})^{-1}\mathbf{u}$, from which the term penalized quasi-likelihood derives. Next, we show how penalized splines can be cast into the GLMM framework.

2.2.2 Penalized Spline Formulation as GLMM

Suppose that observations y_i at values of age a_i satisfy the relationship

$$\text{logit}[P(Y_i = 1|a_i)] = \text{logit}[\pi(a_i)] = \eta(a_i), \qquad i = 1, 2, \cdots, N. \tag{2.4}$$

The linear predictor $\eta(a_i)$ can be estimated non-parametrically using penalized splines in the following way. Consider a pth degree spline model with K knots given by

$$\eta(a_i) = \beta_0 + \beta_1 a_i + \cdots + \beta_p a_i^p + \sum_{k=1}^{K} u_k (a_i - t_k)_+^p, \tag{2.5}$$

with the truncated power basis function defined as

$$(a_i - t_k)_+^p = \begin{cases} 0, & a_i \leq t_k \\ (a_i - t_k)^p, & a_i > t_k, \end{cases} \tag{2.6}$$

where $a_1 \leq a_2 \leq \cdots \leq a_N$ and t_k denotes the kth knot. In vector form, the mean structure model for $\eta(a_i)$ becomes

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}. \tag{2.7}$$

Here, $\boldsymbol{\eta} = [\ \eta(a_1) \quad \cdots \quad \eta(a_N)\ ]^T$, $\boldsymbol{\beta} = [\ \beta_0 \quad \beta_1 \quad \cdots \quad \beta_p\ ]^T$ is the vector of fixed effects and $\mathbf{u} = [\ u_1 \quad u_2 \quad \cdots \quad u_k\ ]^T$ is the vector of random effects and the design matrices are given by

$$\mathbf{X} = \begin{bmatrix} 1 & a_1 & a_1^2 & \cdots & a_1^p \\ 1 & a_2 & a_2^2 & \cdots & a_2^p \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & a_N & a_N^2 & \cdots & a_N^p \end{bmatrix}, \ \mathbf{Z} = \begin{bmatrix} (a_1 - t_1)_+^p & \cdots & \cdots & (a_1 - t_K)_+^p \\ (a_2 - t_1)_+^p & \cdots & \cdots & (a_2 - t_K)_+^p \\ \vdots & \vdots & \vdots & \vdots \\ (a_N - t_1)_+^p & \cdots & \cdots & (a_N - t_K)_+^p \end{bmatrix}.$$

Because a large number of knots t_k leads to too rough a fit, the nonlinear part $\mathbf{Z}$ is penalized by assuming that the coefficients, $\mathbf{u}$, are random effects and are constrained to reduce the influence of the knots and hence ensure stable estimation. It is further assumed that $\mathbf{u} \sim N(\mathbf{0}, \sigma_{\mathbf{u}}^{\mathbf{2}}\mathbf{I})$. These two assumptions provide a strong connection between penalized splines and mixed models. The choice of the knots t_k is made in a way that mimics the distribution of the predictor space. Since we only have one predictor, age, a simple solution proposed by Ruppert *et al.* (2003) is to select equally spaced knots based on the quantiles

$$t_k = \left(\frac{k+1}{K+2}\right) th,$$

which is the sample quantile of the unique age values a_i, where $1 \leq k \leq K$. Throughout this chapter we employ second degree up to fourth degree penalized splines models because any linear combination of these spline basis functions will have a continuous first derivative and hence smooth estimates for both prevalence and force of infection will be obtained. However, for comparison purposes, we include results for linear splines in Section 2.4 where simulation studies about the sensitivity of the estimated prevalence and force of infection on the degree of the spline model and the number of knots taken are performed.

2.2.3 Parameter Estimation

Now, substituting the penalized spline model (2.5) for the linear predictor into (2.3), fully specifies the GLMM. The parameters of (2.3) are estimated by penalized quasi-likelihood (Ruppert *et al.*, 2003) based upon writing it in linear mixed model form:

$$\mathbf{y}^* = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \sigma_{\epsilon}^{-\mathbf{2}}(\mathbf{y} - \boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}^*, \qquad \boldsymbol{\varepsilon}^* \sim N(0, \sigma_{\epsilon}^2) \tag{2.8}$$

using pseudo-data $\mathbf{y}^*$ as response. Briefly, for given values of $(\boldsymbol{\beta}, \sigma_u^2, \sigma_\epsilon^2)$, empirical Bayes estimates of $\mathbf{u}$ are obtained and substituted into (2.8) and this results into a pseudo-variable $\mathbf{y}^*$. The linear mixed model is then fit to the pseudo-data to obtain updated values of $(\boldsymbol{\beta}, \sigma_u^2, \sigma_\epsilon^2)$ which, when re-substituted into the model, yield updated pseudo-data. This fitting process continues and upon convergence produces PQL estimates of these parameters. We refer to Molenberghs and Verbeke (2005) for a review of this formulation.

Estimating the Smoothing Parameter

Smoothing the data using penalized splines requires choosing the value for the smoothing parameter, which controls the trade-off between the smoothness and goodness of fit of the fitted model. Ruppert *et al.* (2003) suggest a smoothing parameter within the framework of generalized linear mixed models via PQL and REML estimation techniques. Maximum likelihood can also be used but REML is advantageous as it produces less biased estimates of the variance components. Thus, smoothing parameter selection trims down to variance component estimation with a small variance of random effects corresponding to more smoothness of the curve. Therefore the penalized spline fitting criterion in (2.8), when divided by the pseudo-error variance, σ_ϵ^2, can be written as

$$\frac{1}{\sigma_\epsilon^2}||\mathbf{y}^* - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{Z}\hat{\mathbf{u}}||^2 + \frac{\lambda^{2p}}{\sigma_\epsilon^2}\mathbf{u}^T\mathbf{I}\mathbf{u},$$

where the ratio $\sigma_u^2 = \sigma_\epsilon^2/\lambda^{2p}$ (for a pth degree P-spline) underscores the connection between the smoothing parameter, λ, and variance components. The power of $2p$ is based on scale arguments that if the covariate undergoes a transformation, the same transformation is applied to the smoothing parameter (Ruppert *et al.* (2003)).

2.2.4 Estimation of the Force of Infection

In this section we discuss the estimation of the force of infection for model with logit link function. Using the general form for the hazard function in the current status data framework, the estimate for the force of infection is given by

$$\hat{\ell}(a) = \frac{\hat{\pi}'(a)}{1 - \hat{\pi}(a)} = \hat{\eta}'(a)\hat{\pi}(a). \tag{2.9}$$

For quadratic penalized spline model we have

$$\hat{\eta}(a) = \hat{\beta}_0 + \hat{\beta}_1 a + \hat{\beta}_2 a^2 + \sum_{k=1}^{K} \hat{u}_k (a - t_k)_+^2,$$

and the force of infection is obtained as

$$\hat{\ell}(a) = \left[\hat{\beta}_1 + 2\hat{\beta}_2 a + \sum_{k=1}^{K} 2\hat{u}_k (a - t_k)_+\right] \hat{\pi}(a).$$

For models with other link functions, one can estimate the force of infection by $\hat{\ell}(a) = \hat{\eta}'(a)\delta(\hat{\eta})$, where $\delta()$ is determined by the link function used in the model as discussed by Shkedy *et al.* (2003).

2.3 Data Analysis

The proposed approach is applied to three serological data sets: rubella, varicella and mumps. Second degree up to fourth degree spline models are fitted to these data sets with 10 and 20 knots. To the rubella dataset, cubic and fourth-degree spline models were fitted but they produced zero estimates for the random effects hence zero random effects variance. The same case applied to the varicella dataset when the fourth degree spline model was fitted. As a result these models were not considered further for rubella and varicella respectively. In this section we are mainly interested in smooth fit of the force of infection, so we do not consider linear penalized splines since they produce piecewise constant estimate for the force of infection compared to high-degree penalized spline fits. A summary of the models considered together with their measure of smoothness are shown in Table 2.1. Models 1(a) and 1(b) represent the 10-knot and 20-knot quadratic penalized spline models for rubella, respectively. To the varicella dataset, quadratic and cubic penalized spline models were fitted and these are shown by Models 2(a) to 2(d). Finally, to the mumps dataset results are given up to fourth-degree penalized splines designated by Models 3(a) to 3(f).

Clearly, from Table 2.1, we see that the random-effects variance gets smaller and smaller with high-degree penalized spline models which consequently results in larger values of the smoothing parameter. However, the difference in the smoothness is not excessive for the different knot locations. Akaike Information Criterion (AIC) in the last column, for all the data sets, is observed to be smallest for quadratic penalized spline models with 10 knots which makes them better models relative to the other models. Figure 2.2 shows the curve fits to these datasets. It can be observed from the plots that there is not much effect of the number of knots on the estimated prevalence and force of infection. However, the degree of the penalized spline has some effect. Although the effect cannot be seen on the logit scale, it is present on the scale of interest, the force of infection. Figures 2.2(a) and 2.2(b) depict the results on rubella. For this dataset, quadratic penalized spline fits were sufficient to obtain smooth estimates. The highest force of infection is estimated at the age of 7.5 afterwhich the force of infection declines, then gradually rise again at older ages. Figures 2.2(c) and 2.2(d) pertain to the varicella dataset. Here, two peaks in the estimated force of infection are exhibited but these slightly differ according to the degree of the penalized splines. With quadratic penalized spline models, the estimated force of infection peaks at the age of 6.5 and 32.5 while for the cubic penalized spline models the peaks occur at the age of 5.5 and 33.5. From age 43.5 onwards the force of infection is constrained to zero to avoid negative forces of infection. The fits to the mumps dataset are shown

Table 2.1: *Model selection based on the smoothing parameter λ and Akaike Information Criterion (AIC).*

Data	Model No.	knots k	Model Parameters Int=intercept	Variance components σ^2_ϵ	σ^2_u	λ	AIC
Rubella	1a	10	Int, $a, a^2, (a-t_k)^2_+$	1.0017	0.000025	14.148	20645.3
	1b	20	Int, $a, a^2, (a-t_k)^2_+$	1.0017	0.000013	16.661	20645.4
Varicella	2a	10	Int, $a, a^2, (a-t_k)^2_+$	0.9893	0.000147	9.057	10041.1
	2b	20	Int, $a, a^2, (a-t_k)^2_+$	0.9892	0.000078	10.612	10041.2
	2c	10	Int, $a, a^2, a^3, (a-t_k)^3_+$	0.9951	2.933E-7	12.258	10073.7
	2d	20	Int, $a, a^2, a^3, (a-t_k)^3_+$	0.9951	1.533E-7	13.658	10073.7
Mumps	3a	10	Int, $a, a^2, (a-t_k)^2_+$	0.9983	0.000221	8.198	45470.6
	3b	20	Int, $a, a^2, (a-t_k)^2_+$	0.9982	0.000126	9.434	45472.3
	3c	10	Int, $a, a^2, a^3, (a-t_k)^3_+$	1.0001	5.092E-7	11.191	45488.3
	3d	20	Int, $a, a^2, a^3, (a-t_k)^3_+$	1.0001	2.646E-7	12.481	45488.1
	3e	10	Int, $a, a^2, a^3, a^4, (a-t_k)^4_+$	1.0015	9.21E-10	13.476	45523.8
	3f	20	Int, $a, a^2, a^3, a^4, (a-t_k)^4_+$	1.0015	4.83E-10	14.608	45523.8

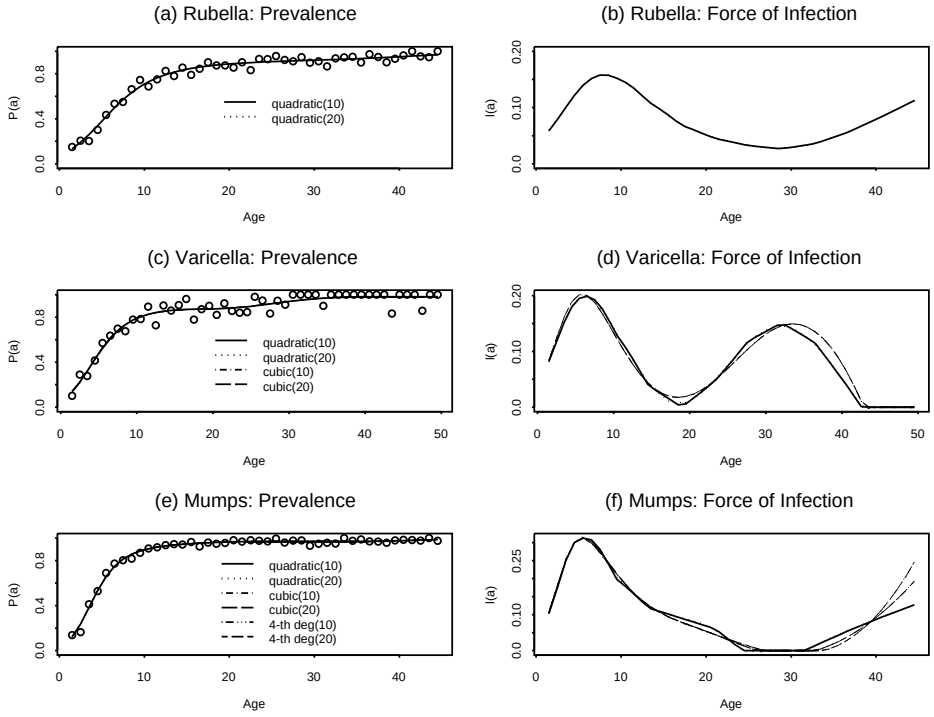

Figure 2.2: *Prevalence data and estimates for (a) rubella, (c) varicella and (e) mumps and the estimated force of infection for (b) rubella, (d) varicella and (f) mumps; using quadratic spline, cubic and/or 4th-degree Penalized splines with 10 and 20 knots.*

in Figures 2.2(e) and 2.2(f). The maximum estimated force of infection is estimated at the age of 5.5 for all models. However, beyond this age, an effect of the degree of the penalized splines on the estimated force of infection is seen between the ages of 15 and 24 and among adults over 32 years old.

2.3.1 Bootstrap Confidence Intervals

To quantify the variability of the estimated prevalences and forces of infection, we applied the percentile method of bootstrap confidence intervals, which take α and $1-\alpha$ percentiles of the bootstrap distribution to define the interval. We sampled, with replacement, B bootstrap samples from the original data, each sample containing N pairs (a_i^*, Y_i^*) and obtained $100(1-2\alpha)\%$ percentile confidence intervals

$(\hat{\ell}^*(a)_{[(B+1)\alpha]}, \hat{\ell}^*(a)_{[(B+1)(1-\alpha)]})$, where $\hat{\ell}^*(a)_{[(B+1)\alpha]}$ is the $(B+1)\alpha th-$order statistic of the bootstrap replicated forces of infection $\hat{\ell}^*_1(a), \ldots, \hat{\ell}^*_B(a)$. Since the bootstrap procedure was not constrained, we defined the lower and upper confidence limits to be $\max\{0, \hat{\ell}^*(a)_{[(B+1)\alpha]}\}$ and $\max\{0, \hat{\ell}^*(a)_{[(B+1)(1-\alpha)]}\}$, respectively, as a counterpart to the PAV algorithm in order to avoid negative estimates of forces of infection at higher ages. Because of the presence of small sample sizes at higher age values, the confidence bounds are very wide. Therefore, to have meaningful graphical results, we restrict the age up to 40.5 years old.

Figure 2.3 shows the estimated probability curves and forces of infection together with their 95% percentile bootstrap confidence intervals for the 10-knot quadratic penalized spline models. Though it appears as if the bootstrap confidence intervals do not differ on the probability scale, this is not so when we consider the derivative scale, the force of infection. We note an increase in the variability around the estimated forces of infection at older age groups which can be explained by smaller sample sizes at these age levels as mentioned earlier.

2.4 Simulation

The smoothness of the estimated force of infection is oftentimes sensitive to the number of knots and the degree of the penalized spline. To determine whether the number of knots and the degree of the spline are set to appropriate sizes, different number of knots and different degrees of penalized splines are used and the results compared to determine the sensitivity of both the estimated prevalence and force of infection with respect to these tuning constants. Two simulation studies were conducted, the first used the sample size at each age value, as the one in the mumps dataset, while the second used sample size 200 per age value. The age values used are according to the mumps dataset. The true prevalence was taken according to the log-logistic fit to the mumps dataset as

$$\pi(a) = \frac{\alpha_1 a^{\alpha_2}}{1 + \alpha_1 a^{\alpha_2}}$$

and the true force of infection as

$$\ell(a) = \frac{\alpha_2 \alpha_1 a^{\alpha_2 - 1}}{1 + \alpha_1 a^{\alpha_2}},$$

where α_1 is the exponent of the intercept and α_2 is the slope obtained from the fitted log-logistic model. The true age at which the maximum force of infection occurs is 4.5. There were $M = 500$ simulated datasets for each simulation study.

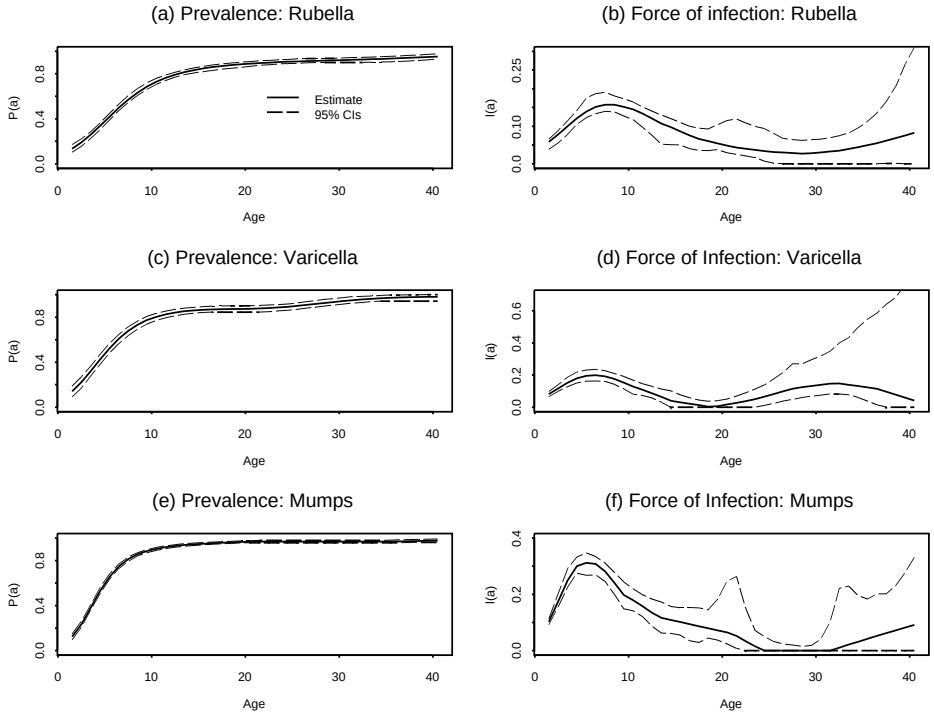

Figure 2.3: *Bootstrap confidence intervals for rubella, varicella and mumps using quadratic penalized splines with 10 knots: Left hand panels denote prevalence* $\hat{\pi}(a)$ *and right hand panels denote force of infection* $\hat{\ell}(a)$.

On each data set, generalized linear mixed models were fitted with penalized splines from the second to the fourth-degree and a set of 5, 10 and 20 knots. Using the estimate of prevalence $\pi_{jK}(a)$, the force of infection $\ell_{jK}(a)$ at age a for the jth simulation and K knots was estimated according to (2.9). For prevalence and force of infection, respectively, the local squared bias is estimated as $\hat{b}^2(a) = \{\bar{\hat{\pi}}(a) - \pi(a)\}^2$ and $\hat{b}^2(a) = \{\bar{\hat{\ell}}(a) - \ell(a)\}^2$ with $\bar{\hat{\pi}}(a) = \sum_{j=1}^{M} \hat{\pi}_j(a)/M$ and $\bar{\hat{\ell}}(a) = \sum_{j=1}^{M} \hat{\ell}_j(a)/M$, while the local variances are estimated as $\hat{v}(a) = \sum_{j=1}^{M}\{\hat{\pi}_j(a) - \bar{\hat{\pi}}(a)\}^2/M$ and $\hat{v}(a) = \sum_{j=1}^{M}\{\hat{\ell}_j(a) - \bar{\hat{\ell}}(a)\}^2/M$. Hence, the estimate of the local mean-squared error (MSE) is given by $\widehat{MSE}(a) = \hat{b}^2(a) + \hat{v}(a)$.

Figures 2.4 and 2.5 show the results from the first simulation study. Figure 2.4 shows the true and estimated average prevalences over all simulations for the second to the fourth degree penalized splines with 5, 10, and 20 knots. The first row shows the

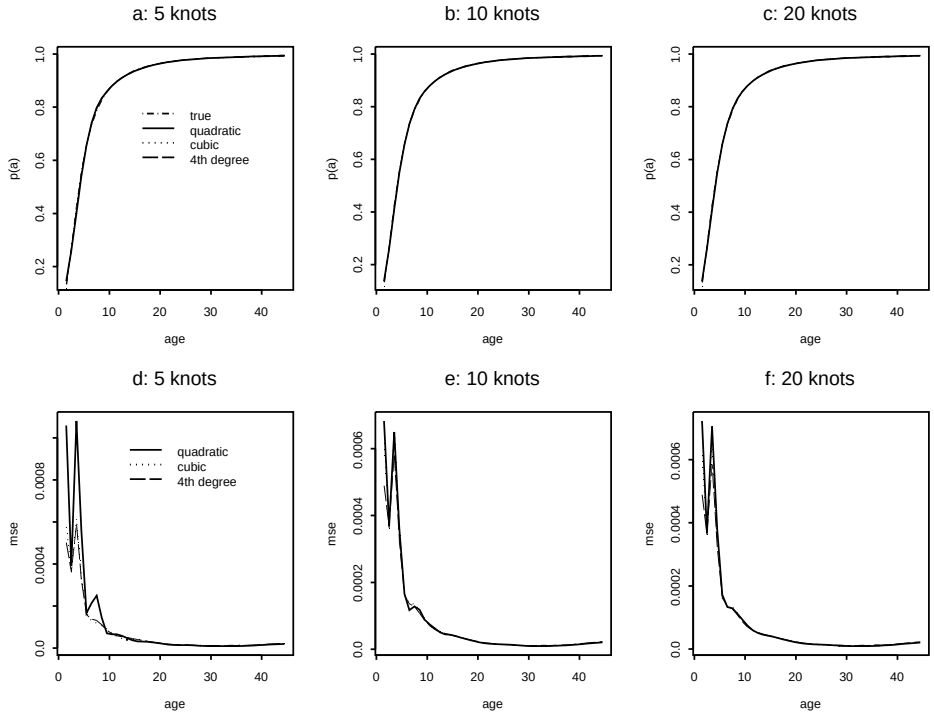

Figure 2.4: *Simulation results for Prevalence. Upper panels: prevalence, lower panels: mean square error. Sample size at each age equals that of the mumps dataset.*

averaged estimate of the prevalences over all simulated data sets versus age. As seen before, no clear distinction among the models can be observed on the logit scale. However, when we turn to the mean squared error plots in the second row, the MSE is high for the 5-knot quadratic penalized spline model at ages below 10 years as compared to other models.

On the other hand, in Figure 2.5, we observe the average estimates of forces of infection and their MSEs. Here, slight differences among the models can be seen. In the first row, all models estimated the maximum force of infection at the age 4.5 except for the 5-knot quadratic penalized spline model which estimates the maximum at the age of 5.5 years. All models estimate an increase in the force of infection after the age of 37.5 but the increase is smaller for quadratic models. Looking at the mean squared error plots in the second row, we see smaller MSEs for quadratic penalized

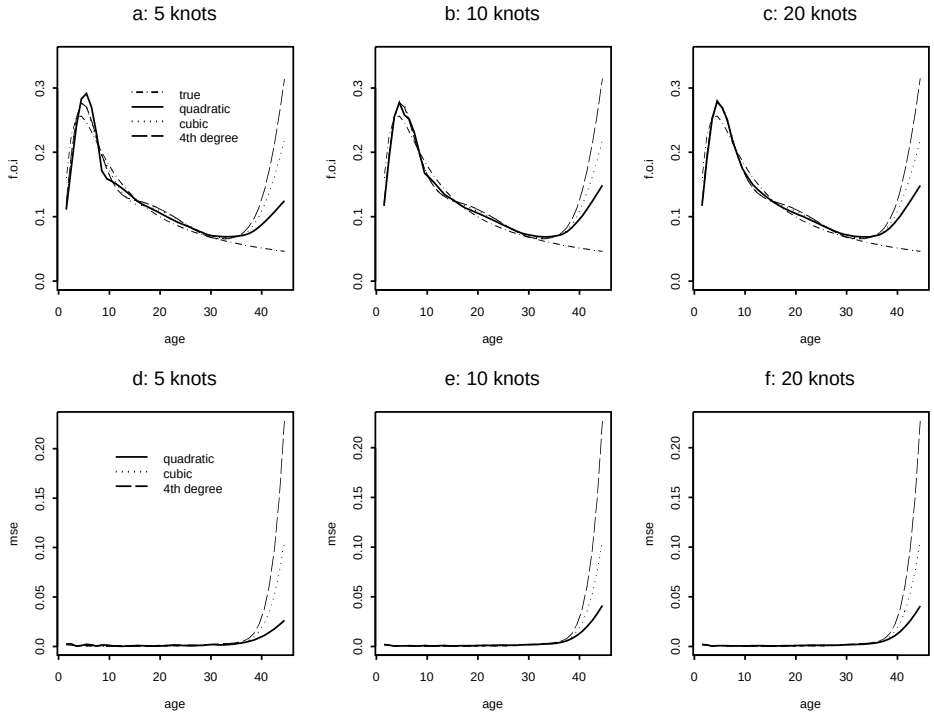

Figure 2.5: *Simulation results for Force of infection. Upper panels: force of infection, lower panels: mean square error. Sample size at each age equals that of the mumps dataset.*

spline models relative to the other models. The results from the second simulation study are presented in Figures 2.6 and 2.7. Figure 2.6 shows the averaged estimates of prevalence and their MSEs. As seen previously, on the logit scale, models cannot be distinguished. Compared to the results from the previous simulation study, the mean squared errors have slightly increased. Nevertheless, the 5-knot quadratic penalized spline model still gives high MSE at young ages relative to other models.

Figure 2.7 shows the averaged estimates of force of infection in the first row and their mean squared errors in the second row. The 10-knot quadratic and the cubic penalized spline models estimate the maximum force of infection at the age of 4.5 while the rest estimate it at the age of 5.5 years. Although the estimate of the force of infection still increases at higher ages for all models, quadratic penalized spline models give smaller estimates. The MSEs dramatically decrease relative to the results of the first simulation study with quadratic penalized spline models still giving the

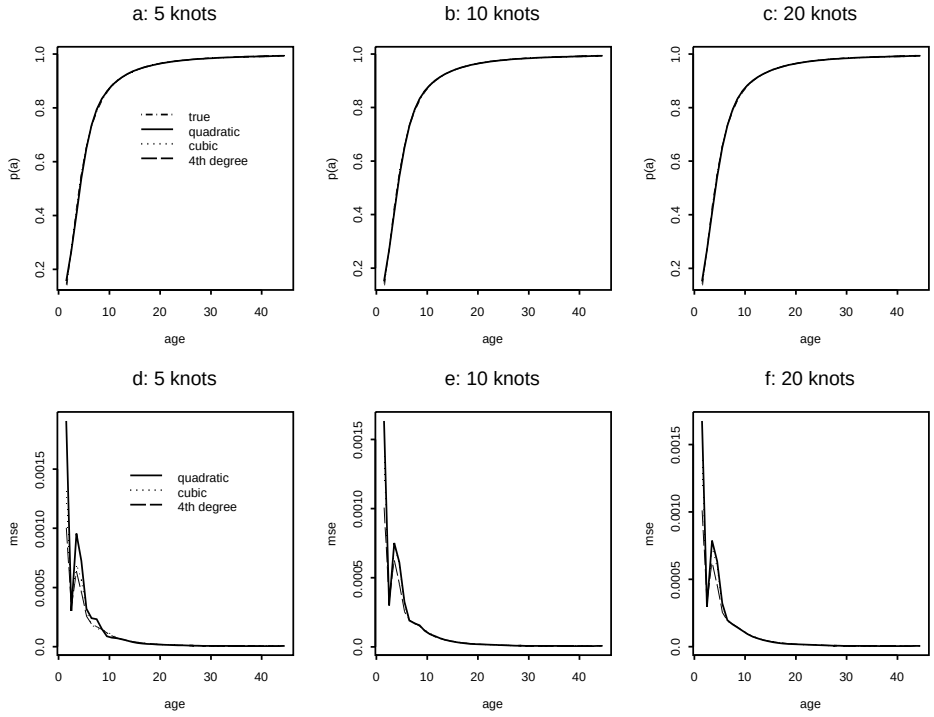

Figure 2.6: *Simulation results for Prevalence. Upper panels: prevalence, lower panels: mean square error. Sample size at each age equals 200.*

smallest MSEs. Tables 2.2 and 2.3 show global squared bias, variance and MSEs for all models including linear penalized spline models for comparison purposes. These global estimates are presented at two age scales: the truncated scale 1.5 to 30.5 and the full scale 1.5 to 44.5. Table 2.2 shows the results for prevalence. On both age scales, the global MSE is highest for linear models and smallest for fourth degree penalized spline models. Table 2.3 shows the results of force of infection. On the truncated age scale the cubic penalized spline models give the smallest global MSE while on the full scale global MSE is smallest for the 10-knot linear penalized spline model. Nevertheless, results on the full scale also account for small sample sizes at higher age values.

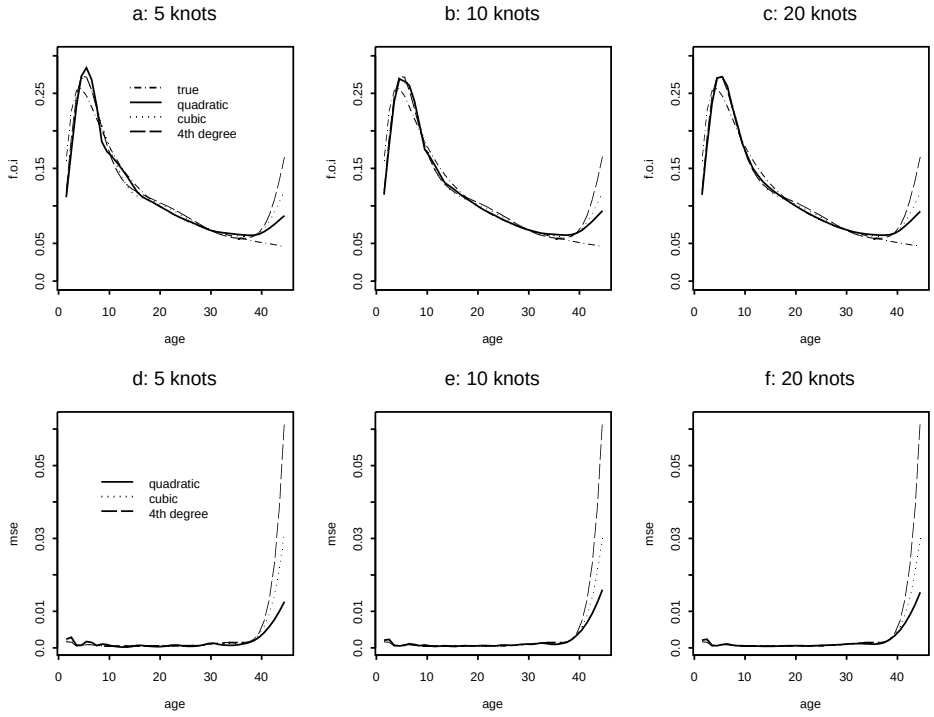

Figure 2.7: *Simulation results for Force of infection. Upper panels: force of infection, lower panels: mean square error. Sample size at each age equals 200.*

We now examine results from the second simulation study which uses a reasonably large sample at each age. Tables 2.4 and 2.5 present the results for prevalence and force of infection, respectively. The results for prevalence in Table 2.4 still follow the same trend as before, highest global estimates for linear penalized splines and lowest estimates for the fourth-degree spline models. Table 2.5 shows results for force of infection. On the truncated age scale the fourth-degreee penalized spline model still has the lowest global MSE. However, on the full age scale, the quadratic spline models, regardless of the number of knots, produce the lowest MSEs relative to other models. This agrees with what we see in Figures 2.5 and 2.7. So our observation in Table 2.3, on the full age scale, can be attributed to small sample size. Thus, from the simulation studies conducted, we observe considerable improvements from linear to quadratic spline models and from 5 to 10-knots models, but beyond this the improvements are minor.

Table 2.2: *Simulation results for prevalence: global simulated squared bias, variance and mean squared error for linear, quadratic, cubic and 4-th degree penalized spline fits with 5, 10 and 20 knots at two age scales. Sample size at each age equals that of mumps dataset.*

		age scale: 1.5 to 30.5 years				age scale: 1.5 to 44.5 years			
		linear	quadratic	cubic	4th-deg	linear	quadratic	cubic	4th-deg
		$\times 10^{-5}$	$\times 10^{-5}$	$\times 10^{-5}$	$\times 10^{-5}$	$\times 10^{-5}$	$\times 10^{-5}$	$\times 10^{-5}$	$\times 10^{-5}$
5 knots	$\bar{b}^2$	59.71	9.20	3.75	3.58	40.71	6.28	2.56	2.45
	$\bar{v}$	5.32	5.77	6.46	6.35	4.00	4.33	4.84	4.79
	$\overline{MSE}$	65.04	14.97	10.21	9.93	44.71	10.61	7.40	7.24
10 knots	$\bar{b}^2$	15.46	4.16	4.37	3.45	10.54	2.84	2.99	2.36
	$\bar{v}$	6.99	6.71	6.32	6.41	5.15	5.00	4.74	4.83
	$\overline{MSE}$	22.45	10.87	10.69	9.86	15.69	7.84	7.73	7.19
20 knots	$\bar{b}^2$	3.90	4.66	4.33	3.45	2.66	3.18	2.96	2.36
	$\bar{v}$	7.95	6.65	6.34	6.41	5.82	4.95	4.75	4.83
	$\overline{MSE}$	11.85	11.31	10.67	9.86	8.48	8.13	7.71	7.19

2.5 Discussion

In this chapter we focussed on estimating the force of infection semi-parametrically by modeling the functional form of the predictor by penalized splines which were then fitted with generalized linear mixed models. The data were modelled on the logit scale and the final model was selected based on AIC. However, estimates on this scale should be interpreted with caution as it might be tempting to conclude that estimates were close yet on the actual scale of interest, the derivative scale, they were less. Indeed, turning to the derivative scale we observed altered estimates for the forces of infection for the different models.

The estimated forces of infection were sensitive to the degree of the spline used and

Table 2.3: *Simulation results for force of infection: global simulated squared bias, variance and mean squared error for linear, quadratic, cubic and 4-th degree penalized spline fits with 5, 10 and 20 knots at two age scales. Sample size at each age equals that of mumps dataset.*

		age scale: 1.5 to 30.5 years				age scale: 1.5 to 44.5 years			
		linear	quadratic	cubic	4th-deg	linear	quadratic	cubic	4th-deg
		$\times 10^{-4}$	$\times 10^{-4}$	$\times 10^{-4}$	$\times 10^{-4}$	$\times 10^{-4}$	$\times 10^{-4}$	$\times 10^{-4}$	$\times 10^{-4}$
5 knots	$\bar{b}^2$	38.19	3.87	1.77	1.95	27.49	7.46	19.41	40.93
	$\bar{v}$	8.66	6.04	5.81	6.20	18.78	27.84	62.76	110.32
	$\overline{MSE}$	46.85	9.91	7.58	8.15	46.27	35.30	82.17	151.25
10 knots	$\bar{b}^2$	7.76	1.76	2.00	1.84	7.25	8.99	19.52	40.80
	$\bar{v}$	13.75	7.31	5.87	6.28	23.24	37.46	63.65	109.12
	$\overline{MSE}$	21.51	9.07	7.87	8.12	30.49	46.44	83.17	149.92
20 knots	$\bar{b}^2$	5.41	1.81	1.99	1.84	6.67	8.96	19.54	40.81
	$\bar{v}$	16.56	7.15	5.86	6.28	28.70	36.78	63.71	109.11
	$\overline{MSE}$	21.97	8.96	7.85	8.12	35.37	45.73	83.24	149.92

the number of knots therein. However, as long as the choice of the knots covers the range of the predictor well, not much difference is seen in the estimates. To this end, we visualized only a marginal impact of the number of knots used on the estimated forces of infection but some effect on the estimates with respect to the degree of the spline. Cubic and fourth degree spline models yielded smoother estimates than quadratic splines but not only do they estimate higher forces of infection at higher ages, they also produce very small random effects variances that tend to zero which apparently suggests that these higher-degree penalized splines might not be necessary. However, the estimates from quadratic penalized spline models were relatively lower and their smoothness was reasonably sufficient.

Further, although the primary aim of this chapter was not to contrast with paramet-

Table 2.4: *Simulation results for prevalence: global simulated squared bias, variance and mean squared error for linear, quadratic, cubic and 4-th degree penalized spline fits with 5,10 and 20 knots at two age scales. Sample size at each age equals 200.*

		age scale: 1.5 to 30.5 years				age scale: 1.5 to 44.5 years			
		linear	quadratic	cubic	4th-deg	linear	quadratic	cubic	4th-deg
		$\times 10^{-5}$	$\times 10^{-5}$	$\times 10^{-5}$	$\times 10^{-5}$	$\times 10^{-5}$	$\times 10^{-5}$	$\times 10^{-5}$	$\times 10^{-5}$
5 knots	$\bar{b}^2$	64.56	10.45	6.22	4.88	44.02	7.13	4.24	3.33
	$\bar{v}$	14.98	7.76	8.34	8.05	10.54	5.46	5.87	5.68
	$\overline{MSE}$	79.54	18.21	14.56	12.93	54.56	12.59	10.11	9.01
10 knots	$\bar{b}^2$	19.18	7.81	6.98	4.94	13.08	5.32	4.76	3.37
	$\bar{v}$	18.57	8.26	8.12	7.94	13.00	5.81	5.72	5.60
	$\overline{MSE}$	37.75	16.07	15.10	12.88	26.08	11.13	10.48	8.97
20 knots	$\bar{b}^2$	12.78	8.19	6.97	4.87	8.72	5.59	4.75	3.33
	$\bar{v}$	19.28	8.19	8.12	8.05	13.49	5.76	5.71	5.69
	$\overline{MSE}$	32.06	16.38	15.09	12.92	22.21	11.35	10.47	9.02

ric modeling, it is comforting to note that the patterns in our estimates identify with those of Whitaker and Farrington (2004) for rubella and mumps. The only difference was seen in the peaks, our semi-parametric approach yielded slightly higher estimates than the parametric counterparts, which may be ascribed to the flexibility of the semi-parametric method.

The variability around the estimated probability curves and forces of infection was studied using the percentile bootstrap confidence intervals. Considering computational time constraints, 500 bootstrap samples were considered reasonable. The intervals were wider at older age groups for reasons of small sample sizes at these age groups. Particularly for varicella, the sample sizes were less than 15 from the age of 30.5 onwards and indeed this dataset exhibited wider intervals at high ages as compared to the rubella and mumps datasets. This variation might also be a consequence

Table 2.5: *Simulation results for force of infection: global simulated squared bias, variance and mean squared error for linear, quadratic, cubic and 4-th degree penalized spline fits with 5,10 and 20 knots at two age scales. Sample size at each age equals 200.*

		age scale: 1.5 to 30.5 years				age scale: 1.5 to 44.5 years			
		linear	quadratic	cubic	4th-deg	linear	quadratic	cubic	4th-deg
		$\times 10^{-4}$	$\times 10^{-4}$	$\times 10^{-4}$	$\times 10^{-4}$	$\times 10^{-4}$	$\times 10^{-4}$	$\times 10^{-4}$	$\times 10^{-4}$
5 knots	$\bar{b}^2$	30.02	3.15	2.03	1.88	21.00	3.18	4.17	7.52
	$\bar{v}$	7.55	4.81	4.70	4.72	13.10	14.02	22.50	33.52
	$\overline{MSE}$	37.72	7.96	6.73	6.60	34.10	17.20	26.67	41.04
10 knots	$\bar{b}^2$	6.82	2.24	2.23	1.80	5.38	2.89	4.15	7.51
	$\bar{v}$	11.51	5.27	4.51	4.77	16.92	16.71	21.55	33.53
	$\overline{MSE}$	18.33	7.51	6.74	6.58	22.30	19.60	25.70	41.04
20 knots	$\bar{b}^2$	4.87	2.27	2.23	1.81	4.31	2.87	4.15	7.50
	$\bar{v}$	13.56	5.19	4.50	4.78	20.75	16.18	21.58	33.52
	$\overline{MSE}$	18.43	7.46	6.73	6.59	25.06	19.05	25.73	41.02

of primary varicella infection being a relatively rarer event in adults versus children, than primary rubella or mumps infection.

The non linear mixed procedure NLMIXED also fits generalized linear mixed models but the class of models it can accommodate is more limited. It attempts to maximize the log likelihood directly by adaptive Gaussian quadrature. If the number of quadrature points is large enough it gives exact maximum likelihood estimates of the parameters but this is only theoretical. Although these are very accurate, the number of random effects that can be handled is limited. Also, residual association cannot be accommodated with the NLMIXED procedure. The GLIMMIX procedure and macro allow for multiple random effects and residual association. A disadvantage

is the inaccessibility of the actual likelihood and severe biases that can result for a number of applications. Nevertheless, we considered PQL approach as implemented in the GLIMMIX macro as our preferred choice.

Alternative approaches to PQL such as the use of standard likelihood to obtain maximum likelihood estimates and a Bayesian Markov Chain Monte Carlo analysis could be employed. However, these methods are substantially more challenging to implement than the GLIMMIX procedure and macro.

The simulation study highlights that the proposed penalized spline-based GLMM method exhibits good performance. Therefore, considering the examples of the data sets employed and simulation results, for estimating a smooth curve estimate of force of infection, we recommend using a quadratic penalized spline with 10 knots which can be fitted as suggested using generalized linear mixed models.

Appendix

We fitted our GLMMs using the SAS macro GLIMMIX. However, compared with the GLIMMIX procedure (SAS Institute Inc. 2004), the results obtained were found to be exactly the same. In this section we show how the GLIMMIX procedure can be used to fit our models. Given the scatterplot vectors $\mathbf{x}$ and $\mathbf{y}$ and a set of knots t_k the GLMMs were fitted using the SAS code below. The dataset constits of the binary response $\mathbf{y}$, the polynomial components are contained in `xlist` that forms the design matrix $\mathbf{X}$ and the `zlist` comprises of the truncated power basis functions $(x_i - t_k)_+^p$ forming the design matrix $\mathbf{Z}$. As a result the kth column of matrix $\mathbf{Z}$ comprises the truncated power function corresponding to the kth knot. The Toeplitz (1) covariance structure is used to group together the constructed columns of $\mathbf{Z}$ with continuous variables to have a common variance component.

```
proc glimmix data=dataset;
model y(event=1)=&xlist / dist=binomial link=logit solution ;
random &zlist/ type=toep(1) solution;
random _residual_/solution;
output out=yhat /allstats;
nloptions tech=nmsimp;
run;
```

For the model of rubella with quadratic splines with 10 knots we have

Cov Parm	Estimate	Standard Error
Variance	0.000025	0.000026
Residual (VC)	1.0017	0.02179

Now, the cube of the smoothing parameter is 1.0017/0.000025.

Chapter 3

Modeling the Force of Infection for Parvovirus B19 in Europe Using Penalized Spline Models

The previous chapter introduced how to model the prevalence and force of infection of a disease as a smooth function of a continuous predictor such as age based on penalized splines, which are fitted using the generalized linear mixed model framework. However, apart from age it is possible to include other variables whether discrete or continuous, in the spline model. In this chapter we focus on inclusion of a discrete variable and also show the flexibility associated with the spline model based on parvovirus B19 data.

Parvovirus B19 is a virus that commonly (and only) infects humans. It was discovered in the 1970's, while healthy blood donors were being screened for hepatitis B (Cossart *et al.* 1975). The virus is spread by contact with infected respiratory secretions (for example, saliva, sputum, or nasal mucus), and from mother to unborn baby also known as perinatal transmission. Parvovirus B19 infection is common and occurs worldwide and affects both sex groups. Seroepidemiologic studies from several countries show that infection is most common in children aged 6-10 years, but can occur at any age. The incubation period of the virus varies from 4-20 days from infection to the development of a characteristic rash or other symptoms. The most

common illness caused by parvovirus B19 infection is called the 'fifth disease', a mild rash that occurs most often in children. The ill child typically has a 'slapped-cheek' rash on the face and a lacy red rash on the trunk and limbs. Once a child recovers from the parvovirus infection he or she develops lasting immunity, and is protected against future infection (Anderson *et al.* 1985). An adult who is infected with parvovirus B19 may have no symptoms at all or, may develop a rash, joint pain or swelling or both. Patients who have a compromised immune system, sickle cell anemia and women who are pregnant are at a greater risk for developing fifth disease (Anderson, 1987; Koch and Adler, 1989). While the disease is generally mild, most studies have focused on risk factors in pregnant women because of the risk to the fetus (Valeur-Jensen *et al.* 1999).

A key epidemiological parameter governing the transmission of infection within a given population is the force of infection. It is defined as the rate at which a susceptible individual is transferred from the susceptible class to the infection class. A major effort has been devoted in the past to model force of infection assuming a constant force of infection. Empirical evidence of age-related changes in the force of infection have been documented for childhood infections (Griffiths, 1974; Anderson and May, 1982; Farrington, 1990; Keiding 1991; Shkedy *et al.* 2003; Shkedy *et al.* 2006).

Serological surveys are a useful source of information about epidemiological parameters for infectious diseases. In particular they may be used to estimate contact rates, forces of infection, the reproduction number and the critical vaccination threshold (Farrington *et al.* 2001; Van Effelterre *et al.* 2008). These methods require the assumption of life-long immunity following initial infection and that the disease is in a steady state (Grenfell and Anderson, 1985). Several approaches were proposed to model the prevalence and force of infection for different infectious diseases such as measles, mumps, rubella, hepatitis A and varicella. Muench (1934; 1959) considered models in which the force of infection is constant and hence independent of age. Grenfell and Anderson (1985) used polynomial functions to model age dependent force of infection. Farrington (1990), Farrington *et al.* (2001) and Edmunds *et al.* (2000) proposed a non-linear model for the prevalence. However, the approach requires prior knowledge about the dependence of the force of infection on age. Shkedy *et al.* (2006) proposed to model age dependent force of infection from seroprevalence data using fractional polynomials, a method which provides a variety of different types of relationships between the force of infection and age. Faes *et al.* (2006a) and Hens *et al.* (2007) estimated the force of infection using monotone fractional polynomials from clustered seroprevalence data. Hens *et al.* (2008) estimated the force of infection using fractional polynomials and cubic regression splines for multi-sera data. Shkedy

et al. (2003) proposed to use local polynomials which simultaneously estimate prevalence and force of infection.
In the previous chapter we estimated the force of infection by using penalized spline basis model fitted by GLMM approach. In the current chapter we extend the basic model of the previous chapter and discuss a modeling approach which allows us to include other covariates in the model in addition to the host age. In particular the covariate of primary interest is country. Serological data from five European countries: Belgium, England and Wales, Finland, Italy and Poland will be used. Recently, Mossong *et al.* (2008) investigated country effect on the force of infection using local polynomials and piecewise constant models. However, Mossong *et al.* (2008) estimate the force of infection for each country separately. In contrast with Mossong *et al.* (2008), in this chapter we model both the prevalence and force of infection using data from all countries. The GLMMs discussed in this chapter allow us to estimate the effect of the country by either including a country specific fixed effect or a country specific smoother. Section 3.1 presents the data to which apply the proposed method discussed in Section 3.2, is applied. The GLMM with logit and cloglog link functions are used in order to take possible proportionality into account (i.e., proportional odds model or proportional force of infection model). Section 3.3 describes the model fitting procedures and the criteria to select among the models and their application to the data in Section 3.4. In Section 3.5 we show how the piecewise constant force of infection can be estimated using penalized splines and finally a discussion of the results and conclusion are given in Section 3.6. The work of this chapter can be found in Namata *et al.* (2008d).

3.1 Data

The data (see Table 3.1) consists of 14070 individuals from five EU countries: Belgium (BE), England and Wales (EW), Finland (FI), Italy (IT) and Poland (PL) collected, respectively, in the years 2001-2003, 1996, 1997-1998, 2003-2004, and 1995-2004. The data set contains the individuals' status of parvovirus B19 infection (outcome=1, if infected and outcome=0, otherwise), their gender and age. Table 3.1 shows the summary of these data. 4.4% of the data were missing while 0.3% individuals had non equitable outcomes. There were 497 individuals older than 65 years. The overall prevalence is 52.1% and it ranges from 41.76% in Finland to 66.02% in Belgium. For Belgium, in particular, there were only 5 individuals above 65 years and they were seropositive. The presence of only one type of outcome causes the problem of

Table 3.1: *Summary of the Parvovirus B19 data set.*

Variable	No. subjects	No. missing	Equivocal results	missing gender	No. >65yrs	Effective N	seropositive % out of N
Response	14070	621	39	2	497	12911	52.10
Gender							
female	7232	278	20		281	6675	52.58
male	6836	343	19		255	6236	51.59
Age	14070	621	39	2	497	12911	52.10
Country							
BE	3374	276	18	0	5	3075	66.02
EW	3179	343	14	1	106	2715	50.90
FI	2500	0	1	0	107	2392	41.76
IT	2517	2	1	1	139	2374	47.22
PL	2500	0	5	0	140	2355	50.74

inestimable parameters. So in order to avoid this problem, we considered the data set up to age 65 years and also excluded all the missing and non equitable observations. This results in an effective sample size of 12911 individuals. Figure 3.1 shows the five seroprevalence samples. Note that the age is grouped in one year intervals and the maximum age is 65 years.

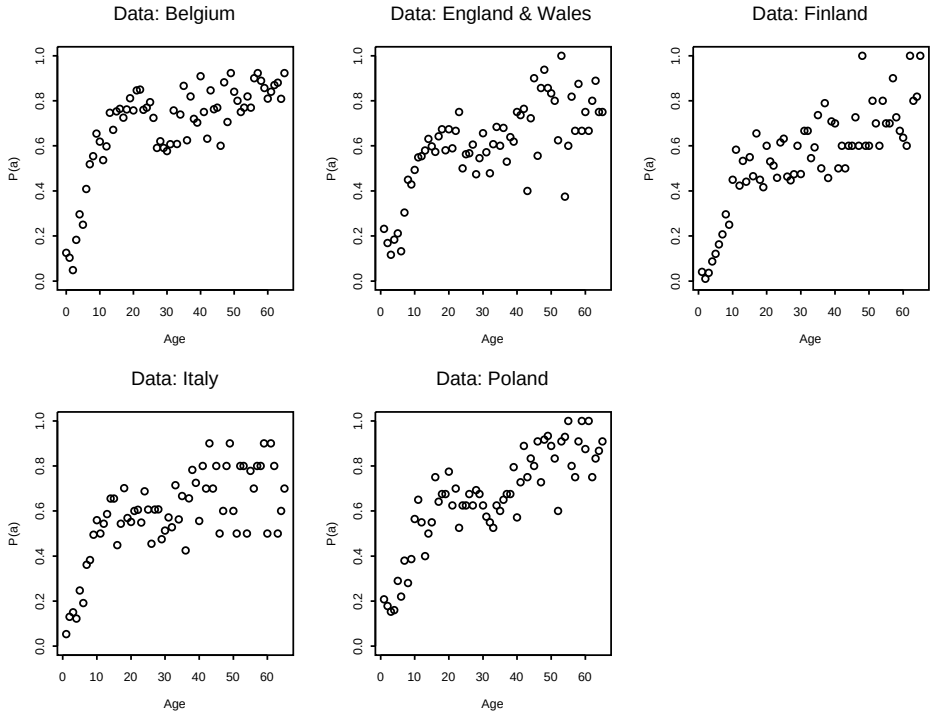

Figure 3.1: *Country-specific raw data for the proportions of parvovirus B19 infection versus age.*

3.2 Estimating the Force of Infection Using Penalized Splines

3.2.1 Simple GLMM Spline Model

In the previous chapter we implemented a spline analysis of rubella, mumps and varicella considering linear, quadratic and cubic spline models with various knots (5, 10 to 20). In the simulation study the penalized quadratic splines emerged as the best performing regardless of the number of knots. Following this simulation finding in the previous chapter, we apply 20-knots penalized quadratic spline models throughout this chapter. For a single continuous predictor (age) and a binary response Y_i $i = 1, \ldots, N$, which takes the value 1 if the ith individual was infected with B19 before age a_i, the disease prevalence, $\pi(a_i) = P(Y_i = 1)$ can be modeled using the

GLMM for which the linear predictor is given by

$$g(\pi(a_i)) = \eta(a) = \beta_0 + \beta_1 a_i + \beta_2 a_i^2 + \sum_{k=1}^{K} u_k (a_i - t_k)_+^2 \tag{3.1}$$

with the truncated power basis function defined as

$$(a_i - t_k)_+^2 = \begin{cases} 0, & a_i \leq t_k \\ (a_i - t_k)^2, & a_i > t_k. \end{cases}$$

Here, g is the logit link function, u_k are random effects for which we assume $u_k \sim N(0, \sigma_u^2)$, i.e a common variance component and zero covariances. The parameter vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ are fixed regression coefficients, and $t_1 < t_2 < \cdots < t_K$ are fixed knots. We considered $K = 20$ and t_k is the sample quantile of age values corresponding to probability $(k+1)/(K+2)$ but other choices of knots can be used (Namata *et al.* 2007; Ruppert *et al.* 2003). In matrix notation the GLMM (3.1) can be written as

$$g(\pi(\mathbf{a})) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \;\; \mathbf{u} \sim N\left[\mathbf{0}, \mathbf{G}\right],$$

where $\mathbf{X} = [1 \; a_i \; a_i^2]_{1 \leq i \leq N}$, $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2]^T$, $\mathbf{u} = [u_1, u_2, \ldots, u_K]^T$, $\mathbf{G} = \text{diag}(\sigma_u^2 \mathbf{1_K})$ and $\mathbf{Z} = \left[(a_i - t_1)_+^2 \; (a_i - t_2)_+^2 \; \ldots \; (a_i - t_K)_+^2\right]_{1 \leq i \leq N}$ for an effective sample size N. Therefore $\mathbf{X}\boldsymbol{\beta}$ is the pure polynomial component of the spline and $\mathbf{Zu}$ is the component with truncated power basis functions. The only covariate included in the GLMM (3.1) is age. In the current study five cross-sectional samples are available and it is of primary interest to include the country, from which the cross-sectional sample was taken, as a covariate in the model.

3.2.2 Extension of the Basic Model

An advantage of spline modeling is its flexibility. In particular model (3.1) can be extended in several ways to include other predictors from binary to discrete to continuous variables. The predictors can be added additively but also interactive terms can be included. In what follows we discuss some of these extensions. Let w_{ij}, $i = 1, \cdots, N, j = 1, 2, 3, 4, 5$ be an indicator variable given by

$$w_{ij} = \begin{cases} 1 & \text{if country=j,} \\ 0 & \text{otherwise.} \end{cases}$$

Here, $j = 1$ to 5 respectively for Belgium, England and Wales, Finland, Italy and Poland respectively.

3.2.3 Model (3.2)

The second model we consider is a semi-parametric model that allows to add the country as a fixed effect into the model. This model assumes that country and age act additively on the prevalence of B19 on the scale of the linear predictor. For a link function g, the model can be expressed as

$$g(\pi(\mathbf{a})) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{W}\boldsymbol{\gamma}, \quad \mathbf{u} \sim N\left[\mathbf{0}, \mathbf{G}\right], \tag{3.2}$$

with $\boldsymbol{\gamma} = [\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5]^T$, and $\mathbf{W} = [w_{i1}\ w_{i2}\ \ldots w_{i5}]^T$ while the rest of the components are given as in (3.1). Note that the country is added to the fixed parametric part of the model while the dependency on age is modeled nonparametrically by the random part of the model. The additional term $\mathbf{W}\boldsymbol{\gamma}$ represents the vertical shift between the five country curves on the scale of the linear predictor.

3.2.4 Model (3.3)

In contrast with model (3.2) which includes country as a fixed effect in the linear predictor the next model allows for the possibility of country and age interacting with one another. The effect of age on prevalence in the smooth term is thus allowed to depend upon country. This nonparametric interaction model is denoted as

$$g(\pi(a_i)) = \beta_0 + \beta_1 a_i + \beta_2 a_i^2 + \sum_{j=1}^{5} w_{ij} \left(\sum_{k=1}^{K} u_k^j (a_i - t_k)_+^2 \right), \tag{3.3}$$

where $u_k^j \sim N(0, \sigma_u^2)$. The model can be written as

$$g(\pi(\mathbf{a})) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}^*\mathbf{u}^* \quad \mathbf{u}^* \sim N\left[\mathbf{0}, \mathbf{G}^*\right],$$

with $\mathbf{u}^* = [u_1^1, \ldots, u_K^1, u_1^2, \ldots, u_K^2, \ldots, u_1^5, \ldots, u_K^5]^T$, $\mathbf{G}^* = \text{diag}(\sigma_u^2\mathbf{1_K}, \sigma_u^2\mathbf{1_K}, \ldots, \sigma_u^2\mathbf{1_K})$ and $\mathbf{Z}^* = \text{diag}(\mathbf{Z_1}, \mathbf{Z_2}, \ldots, \mathbf{Z_5})$. Note that $\mathbf{Z}^*$ is a block diagonal matrix in which each block consists of the design matrix for the random effects of the j'th country. Hence, the model specifies a different smooth function for each subset of observations defined by the levels of country by having the random effects independent from function to function and so there is no implied similarity between the effects. However, the model assumes the same amount of smoothness for the different functions and as such a constant variance component, σ_u^2 across the countries.

3.2.5 Model (3.4)

Model (3.4) combines between models (3.2) and (3.3); that is the country effect is entered in the linear predictor and the smooth term depends upon country. This

gives rise to the semiparametric model

$$g(\pi(a_i)) = \mathbf{W}\boldsymbol{\gamma} + \beta_1 a_i + \beta_2 a_i^2 + \sum_{j=1}^{5} w_{ij} \left(\sum_{k=1}^{K} u_k^j (a_i - t_k)_+^2 \right), \tag{3.4}$$

which in matrix notation becomes

$$g(\pi(\mathbf{a})) = \mathbf{W}\boldsymbol{\gamma} + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}^*\mathbf{u}^* \quad \mathbf{u}^* \sim N\left[\mathbf{0}, \mathbf{G}^*\right].$$

The design matrices and the coefficients are as defined above for models (3.2) and (3.3). A constant variance component, σ_u^2 across the countries is assumed.

Table 3.2 summarises the different models discussed above. The number of parameters in model (3.1) is equal to 5: the fixed effects β_0, β_1 and β_2 and the variance components σ_u^2 and σ_ϵ^2. Note that models (3.3a) and (3.4a) have the same mean structures as models (3.3) and (3.4), respectively. However, these models assume country specific smoothing parameter, i.e., $\mathbf{G}^* = \text{diag}(\sigma_{u1}^2\mathbf{1_K}, \sigma_{u2}^2\mathbf{1_K}, \cdots, \sigma_{u5}^2\mathbf{1_K})$.

Table 3.2: *Terms considered for inclusion in the model*

Model	Terms	Smoothing parameters	# of model parameters
(3.1)	age^1, age^2, (age - t$_k$)$_+^2$	1	5
(3.2)	age^1, age^2, country(5 levels), (age - t$_k$)$_+^2$	1	9
(3.3)	age^1, age^2, country(5 levels)*(age - t$_k$)$_+^2$	1	5
(3.3a)	age^1, age^2, country(5 levels)*(age - t$_k$)$_+^2$	5	9
(3.4)	age^1, age^2, country(5 levels), country*(age - t$_k$)$_+^2$	1	9
(3.4a)	age^1, age^2, country(5 levels), country*(age - t$_k$)$_+^2$	5	13

3.2.6 Proportional Odds and Proportional Hazard Models

In the previous section we have described different models to estimate the prevalence of B19. In this section we derive the force of infection from the estimated prevalences and investigate the properties of the ratios of the estimated prevalences and forces of infection between the country levels. Let us denote the linear predictor as $\eta(a)$. For the logit link function the prevalence can be obtained as

$$\pi(a) = \frac{e^{\eta(a)}}{1 + e^{\eta(a)}},$$

and force of infection as (Shkedy *et al.* 2003)

$$\ell(a) = \eta'(a)\pi(a).$$

For the cloglog link function, the prevalence and the force of infection, respectively, are obtained as

$$\pi(a) = 1 - e^{-e^{\eta(a)}} \text{ and } \ell(a) = \eta'(a)e^{\eta(a)}.$$

Using the above equations, the expressions for prevalence and force of infection can easily be written down for all the models considered. In Table 3.3 we summarise the expressions for the odds ratio and the ratio of the forces of infection between the j'th country and the baseline country for the three models (3.2) to (3.4). These expressions

Table 3.3: *Expressions for the ratio of the odds of the prevalences and the ratio of the forces of infection for country j in comparison to Poland, the baseline category for models (3.2) to (3.4) using logit and cloglog link functions. The expressions consider the spline models at K knots.*

Model	link	$\frac{\text{odds}(\pi_j)}{\text{odds}(\pi_5)}$	$\frac{\ell_j}{\ell_5}$
(3.2)			
	logit	e^{γ_j}	$\frac{\pi_{2j}}{\pi_{25}}$
	cloglog	$\frac{\pi_{2j}}{\pi_{25}} e^{e^{\boldsymbol{X}\boldsymbol{\beta} + \sum_{k=1}^{K} u_k Z_k (e^{\gamma_j} - 1)}}$	e^{γ_j}
(3.3)			
	logit	$e^{\sum_{k=1}^{K} (u_{kj} - u_{k5}) Z_k}$	$\frac{\eta'_{3j}(a)}{\eta'_{35}(a)} \frac{\pi_{3j}}{\pi_{35}}$
	cloglog	$\frac{\pi_{3j}}{\pi_{35}} e^{e^{\boldsymbol{X}\boldsymbol{\beta}} (e^{\sum_{k=1}^{K} u_{kj} Z_k} - e^{\sum_{k=1}^{K} u_{k5} Z_k})}$	$\frac{\eta'_{3j}(a)}{\eta'_{35}(a)} e^{\sum_{k=1}^{K} (u_{kj} - u_{k5}) Z_k}$
(3.4)			
	logit	$e^{\gamma_j} e^{\sum_{k=1}^{K} (u_{kj} - u_{k5}) Z_k}$	$\frac{\eta'_{4j}(a)}{\eta'_{45}(a)} \frac{\pi_{4j}}{\pi_{45}}$
	cloglog	$\frac{\pi_{4j}}{\pi_{45}} e^{e^{\boldsymbol{X}\boldsymbol{\beta}} (e^{\gamma_j} e^{\sum_{k=1}^{K} u_{kj} Z_k} - e^{\sum_{k=1}^{K} u_{k5} Z_k})}$	$\frac{\eta'_{4j}(a)}{\eta'_{45}(a)} e^{\gamma_j} e^{\sum_{k=1}^{K} (u_{kj} - u_{k5}) Z_k}$

display meaningful interpretations of the prevalence and the force of infection between any two countries. Let us consider model (3.2). When the logit link function is used, the odds ratio between the j'th country and the baseline country depends

upon the country effect and it is not dependent on age. The ratio of the forces of infection (π_{2j}/π_{25}) is age-dependent. For a model with cloglog link the force of infection ratio is proportional ($\exp(\gamma_j)$) while the odds ratio is age-dependent. Hence, model (3.2) with logit link function implies a proportional odds model while a model with cloglog link implies a proportional force of infection model (proportional hazard model). The properties of proportional odds or proportional force of infection do not remain when model (3.3) and (3.4) are used. For example the odds ratio for model (3.4) which includes country as a fixed and random effect is $e^{\gamma_j} e^{\sum_{k=1}^{K}(u_{kj}-u_{k5})Z_k}$ and it depends on age. For the cloglog model, the ratio of the force of infection is $\frac{\eta'_{4j}(a)}{\eta'_{45}(a)} e^{\gamma_j} e^{\sum_{k=1}^{K}(u_{kj}-u_{k5})Z_k}$. Thus, different model formulations lead to different interpretations of the odds ratio and the force of infection ratios. This implies that a model selection procedure is needed in order to select the best model among the fitted models. A selection of the best model will allow us to draw conclusions not only about the dependency of the force of infection on age but also on the proportionality.

3.3 Estimation and Model Selection

3.3.1 Quasi-Likelihood Estimation

All the models discussed above were fitted using the SAS procedure GLIMMIX. Unfortunately a model selection procedure based on the information criteria reported in GLIMMIX output is not possible. SAS procedure GLIMMIX does not provide a likelihood value for the estimated models, instead a pseudo-likelihood is calculated and therefore one cannot use the information criteria in GLIMMIX output nor the likelihood ratio tests. Following Ruppert *et al.* (2003) we constructed the deviance conditional on the pseudo-likelihood parameter estimates for both $\boldsymbol{\beta}$ and $\mathbf{u}$,

$$D(\mathbf{y}; \hat{\boldsymbol{\pi}}(\hat{\boldsymbol{\beta}}, \hat{\mathbf{u}})) = 2 \sum_{i=1}^{n} \left\{ y_i \log\left(\frac{y_i}{\hat{\pi}_i}\right) + (1-y_i) \log\left(\frac{1-y_i}{1-\hat{\pi}_i}\right) \right\}. \tag{3.5}$$

According to Ruppert *et al.* (2003) the effective number of parameters associated with the fit were computed as the trace of the hat matrix as

$$p_D = \text{trace}\left\{ \left(\boldsymbol{C}^T \boldsymbol{W} \boldsymbol{C} + \frac{1}{2}\boldsymbol{\Lambda} \right)^{-1} \boldsymbol{C}^T \boldsymbol{W} \boldsymbol{C} \right\}, \tag{3.6}$$

where $\boldsymbol{C} = [\boldsymbol{X}, \boldsymbol{Z}]$, $\boldsymbol{W} = \text{var}(\mathbf{y}|\boldsymbol{X}, \boldsymbol{Z}, \mathbf{u})$ and

$$\boldsymbol{\Lambda} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma_\epsilon^2 \text{cov}(\mathbf{u})^{-1} \end{bmatrix}.$$

The selection between the models based upon the adjusted AIC (Ruppert *et al.*, 2003) is given by

$$AIC = n^{-1}[D(\mathbf{y}; \hat{\boldsymbol{\pi}} : \Lambda) + 2p_D] \tag{3.7}$$

Note that for a given model, significant test for the fixed effects is still valid when GLIMMIX is used.

3.3.2 Hierarchical Bayesian Modeling

Each of the above mentioned models can be formulated as a hierarchical Bayesian model. For example, the prevalence of B19 infection can be estimated for model (3.1) using a hierarchical Bayesian model

$$\begin{aligned} y_i &\sim \text{Bernoulli}(\pi_i) && (3.8)\\ g(\pi_i) &= \boldsymbol{X\beta} + \boldsymbol{Z}\mathbf{u} && (3.9)\\ \boldsymbol{\beta} \sim \text{N}(0, 10^6) \quad &\text{and} \quad u_k \sim \text{N}(0, \sigma_u^2) && (3.10)\\ \sigma_u^{-2} &\sim \text{Gamma}(10^{-3}, 10^{-3}). && (3.11) \end{aligned}$$

In Equation 3.8 we specify the likelihood for the response with π_i estimated according to 3.9. The equations (3.10) and (3.11) represent the prior and hyperprior distribution for the parameters in the model, respectively. Following Crainiceanu *et al.* (2004), a normal prior distribution centered at zero with a standard error of 1000 is considered to be sufficiently noninformative for the $\boldsymbol{\beta}$ parameter vector. The parametrization of the Gamma(a, b) is chosen so that its mean $a/b = 1$ and its variance is $a/b^2 = 1000$. Given data y_i and parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{u})$, where $\mathbf{u}$ in turn depends on the hyperparameter σ_u^2 that is not mentioned in the likelihood, a Bayesian analysis starts with prior probabilities $\text{P}(\boldsymbol{\beta})$ and $\text{P}(\mathbf{u}|\sigma_u^2)$ and the likelihood $\text{L}(\mathbf{y}|\boldsymbol{\theta})$ to compute a posterior probability

$$P(\boldsymbol{\beta}, \mathbf{u}, \sigma_u^2|\mathbf{y}) = \frac{\text{L}(\mathbf{y}|\boldsymbol{\beta}, u)P(\boldsymbol{\beta})P(\mathbf{u}|\sigma_u^2)P(\sigma_u^2)}{\sum \text{L}(\mathbf{y}|\boldsymbol{\beta}, u)P(\boldsymbol{\beta})P(u|\sigma_u^2)P(\sigma_u^2)} \tag{3.12}$$

using Markov chain Monte Carlo (MCMC) simulation (Gilks *et al.*, 1996). The value $\bar{\boldsymbol{\theta}}$ which is the average of the samples of $\boldsymbol{\theta}$ is the Bayesian analogue of the maximum likelihood estimator. For model selection, we use the deviance information criterion proposed by Spiegelhalter *et al.*, (1998, 2002) and also employed by Erkanali *et al.*, (1999), Rahmann *et al.* (1999) and Gelfand *et al.* (2000). Define the deviance as $D(\boldsymbol{\theta}) = -2\text{log}(\text{L}(\mathbf{y}|\boldsymbol{\theta}))$. The average of $D(\boldsymbol{\theta})$ over the samples of $\boldsymbol{\theta}$ (i.e. $\bar{\text{D}} = E_{\boldsymbol{\theta}|\mathbf{y}}[D(\boldsymbol{\theta})]$) is a measure of how well the model fits the data; the larger this is,

the worse the fit. The effective number of parameters of the model is computed as $p_D = \bar{\mathrm{D}} - D(\bar{\boldsymbol{\theta}})$. The larger this is, the easier it is for the model to fit the data. The Deviance Information Criterion (DIC) is given by

$$DIC = \bar{\mathrm{D}} + p_D = \mathrm{D}(\bar{\boldsymbol{\theta}}) + 2p_D. \tag{3.13}$$

The models with smaller DIC indicate better fits than those with larger DIC.

3.4 Application to the Data

3.4.1 Quasi-Likelihood Estimation

The results for selecting between the models fitted with the GLIMMIX procedure are presented in Table 3.4. The table also shows the deviance, the effective number of parameters (pD) and the adjusted AIC according to formulas (3.5), (3.6) and (3.7). model (3.1), which corresponds to fitting one smooth function of age to all data for the five countries simultaneously, compared to all other models, does not appear to fit the data well, yielding a deviance of 15481.88 on 9.99 degrees of freedom (pD) for the logit model and a deviance of 15486.98 on 8.25 degrees of freedom for the cloglog models. Considering the semiparametric model, model (3.2), which specifies linear effects for country, the penalized quasilikelihood fit of the quadratic spline model yields a deviance of 15242.94 on 13.85 df for the logit model and a deviance of 15261.17 on 12.18 degrees of freedom for the cloglog model, or a decrease of about 1.5% relative to the deviances from model (3.1) with both link functions. Among all the fitted models, model (3.4a), which (i) allows countries to differ by their intercepts and (ii) specifies a country-by-smooth term interaction with separate smoothing parameters, had the lowest AIC value for both link functions; 1.18199 and 1.18134 for logit and cloglog link functions respectively. Therefore having identified the best fitting model, significance tests can be used to guide what individual estimates remain in the model.

Tables 3.5 and 3.6, respectively, show the estimated pseudo-likelihood coefficients and corresponding standard errors, for all the fitted models. model (3.4a) shows that the intercepts for England & Wales and Italy are not significantly different from that of Poland (the baseline category) suggesting a common homogeneous effect for England & Wales and Italy and Poland. Since Poland is the baseline category, this is equivalent to testing whether the coefficients γ_2 (for England & Wales) and γ_4 (for Italy) are significantly different from zero i.e., $\mathrm{H}_0 : \gamma_2 = \gamma_4 = 0$. For both link functions, a reduced model, model (3.5), yielded lower AICs than model (3.4a) implying that we

Table 3.4: *Adjusted AIC and equivalent number of parameters.*

	logit link			cloglog link			# of
Model	Deviance	pD	AIC	Deviance	pD	AIC	parameters
(3.1)	15481.88	9.98958	1.20067	15486.98	8.24893	1.20080	5
(3.2)	15242.94	13.8530	1.18276	15261.17	12.1835	1.18392	9
(3.3)	15270.76	30.2643	1.18746	15273.70	26.7318	1.18714	5
(3.3a)	15235.20	33.7608	1.18525	15236.55	30.2063	1.18480	9
(3.4)	15203.41	31.1436	1.18238	15202.53	28.6969	1.18193	9
(3.4a)	15196.22	32.2396	1.18199	15192.16	30.0286	1.18134	13
(3.5)	15197.73	30.3544	1.18182	15193.04	28.1962	1.18112	11

cannot reject the null hypothesis that the intercepts for England & Wales, Italy and Poland are equal.

Table 3.5: *Pseudo-Likelihood estimates of the parameters from non- and semi-parametric models by modeling the probability of B19 infection using the logit link function. The asterisk * indicates significant fixed effects.*

model effect	model (3.1)	model (3.2)	model (3.3)	model (3.3a)	model (3.4)	model (3.4a)
Fixed Effects						
Intercept	-2.444(0.164)*	-2.384(0.167)*	-2.697(0.127)*	-2.635(0.122)*	-2.641(0.137)*	-2.610(0.138)*
Age						
age^1	0.205(0.067)*	0.220(0.066)*	0.331(0.037)*	0.308(0.035)*	0.352(0.032)*	0.3448(0.0330)*
age^2	0.006(0.008)	0.005(0.008)	-0.007(0.003)*	-0.007(0.003)*	-0.009(0.003)*	-0.008(0.003)*
Country						
Belgium		0.377(0.062)*			0.353(0.120)*	0.354(0.120)*
England&Wales		-0.148(0.063)*			-0.082(0.120)	-0.078(0.119)
Finland		-0.532(0.064)*			-0.843(0.133)*	-0.962(0.147)*
Italy		-0.203(0.065)*			-0.100(0.119)	-0.099(0.119)
Poland		baseline				
Variance components						
$\hat{\sigma}^2_u$	4.7E-5(4.8E-5)	4.3E-5(4.8E-5)	2.6E-5(1.0E-5)		1.5E-5(6.6E-6)	
$\hat{\sigma}^2_\epsilon$	1.0017(0.0125)	0.9998(0.0125)	1.0031(0.0125)	1.0015(0.0125)	0.9992(0.0125)	0.9990(0.0125)
$\hat{\sigma}^2_{u1}$				0.00138(0.00109)		1.5E-5(1.5E-5)
$\hat{\sigma}^2_{u2}$				1.7E-5(1.5E-5)		1.2E-5(9.6E-6)
$\hat{\sigma}^2_{u3}$				1.2E-5(9.97E-6)		5.9E-5(8.1E-5)
$\hat{\sigma}^2_{u4}$				2.1E-5(2.1E-5)		1.5E-5(1.2E-5)
$\hat{\sigma}^2_{u5}$				1.5E-5(1.2E-5)		1.5E-5(1.2E-5)

Table 3.6: *Pseudo-Likelihood estimates of the parameters from non- and semi-parametric models by modeling the probability of B19 infection using the complementary log log link function. The asterisk * indicates significant fixed effects.*

model effect	model (3.1)	model (3.2)	model (3.3)	model (3.3a)	model (3.4)	model (3.4a)
Fixed Efects						
Intercept	-2.619(0.132)*	-2.550(0.133)*	-2.670(0.106)*	-2.619(0.104)*	-2.598(0.116)*	-2.559(0.117)*
Aage						
age^1	0.272(0.044)*	0.277(0.044)*	0.307(0.028)*	0.285(0.027)*	0.309(0.026)*	0.301(0.026)*
age^2	-0.005(0.005)	-0.005(0.005)	-0.009(0.002)*	-0.008(0.002)*	-0.009(0.002)*	-0.009(0.002)*
Country						
Belgium		0.218(0.039)*			0.310(0.097)*	0.315(0.095)*
England&Wales		-0.107(0.042)*			-0.056(0.099)	-0.050(0.098)
Finland		-0.353(0.044)*			-0.668(0.112)*	-0.834(0.131)*
Italy		-0.146(0.043)*			-0.057(0.099)	-0.054(0.098)
Poland		baseline				
Variance components						
$\hat{\sigma}^2_u$	1.2E-5(9.81E-6)	1.1E-5(9.9E-6)	1.2E-5(4.23E-6)		8.43E-6(3.31E-6)	
$\hat{\sigma}^2_\epsilon$	1.0031(0.0125)	1.0005(0.0125)	1.0011(0.0125)	0.9998(0.0125)	0.9966(0.0124)	0.9962(0.0124)
$\hat{\sigma}^2_{u1}$				0.00061(0.00048)		6.9E-6(5.6E-6)
$\hat{\sigma}^2_{u2}$				8.1E-6(6.2E-6)		7.2E-6(5.2E-6)
$\hat{\sigma}^2_{u3}$				8.6E-6(6.4E-6)		5.6E-5(6.9E-5)
$\hat{\sigma}^2_{u4}$				9.9E-6(8.8E-6)		8.5E-6(6.5E-6)
$\hat{\sigma}^2_{u5}$				6.4E-6(4.5E-6)		7.1E-6(5.1E-6)

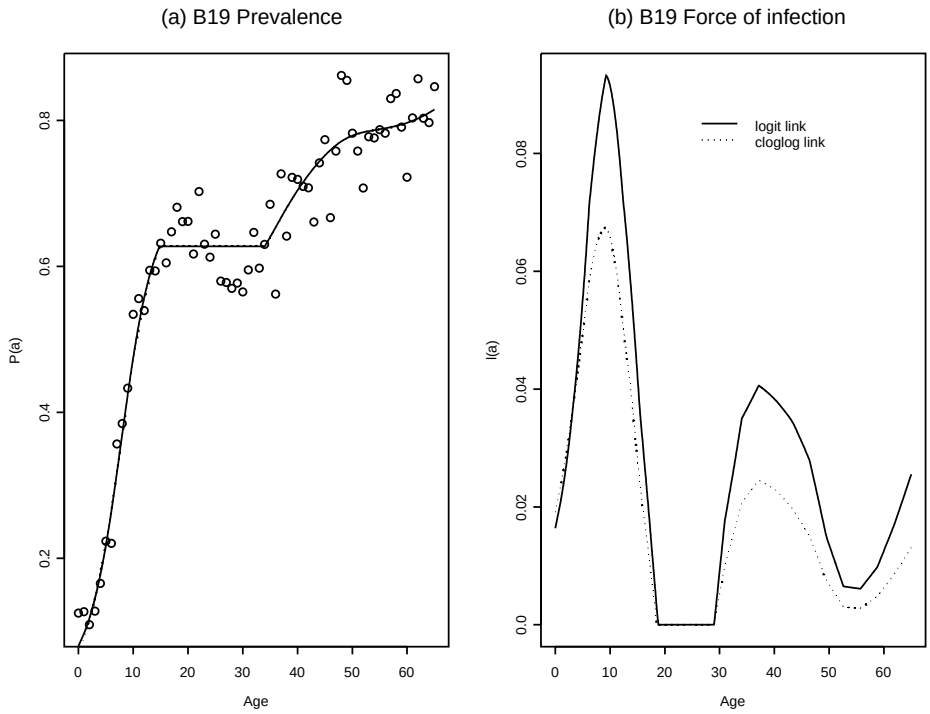

Figure 3.2: *Model (3.1). Fitted curves for prevalence and force of infection for the data to all five countries simultaneously*

Figure 3.2 shows the estimated prevalence and force of infection across all countries according to model (3.1). The force of infection increases until the age of 9 years and then decreases until 18 years. Between 18 and 28 years the force of infection goes to negative, due to a corresponding decrease in the prevalence, but it is set to zero in order to conform with the pool adjacent violaters algorithm (Robertson *et al.* 1988) because a negative force of infection has no meaning in epidemiology. A second peak is observed at the age of 37 whereafter the force of infection decreases but rises again from the mid 50s. Note however that the logit model estimates higher force of infection compared to that of the cloglog model. The fitted curves from model (3.2) are shown in Figure 3.3. Belgium has the highest estimate for prevalence and force of infection while Finland showed the lowest estimates for prevalence and force of infection. The curves for England & Wales, Italy and Poland are estimated to be close to each other. For all countries the force of infection increases until the peak at the age of around

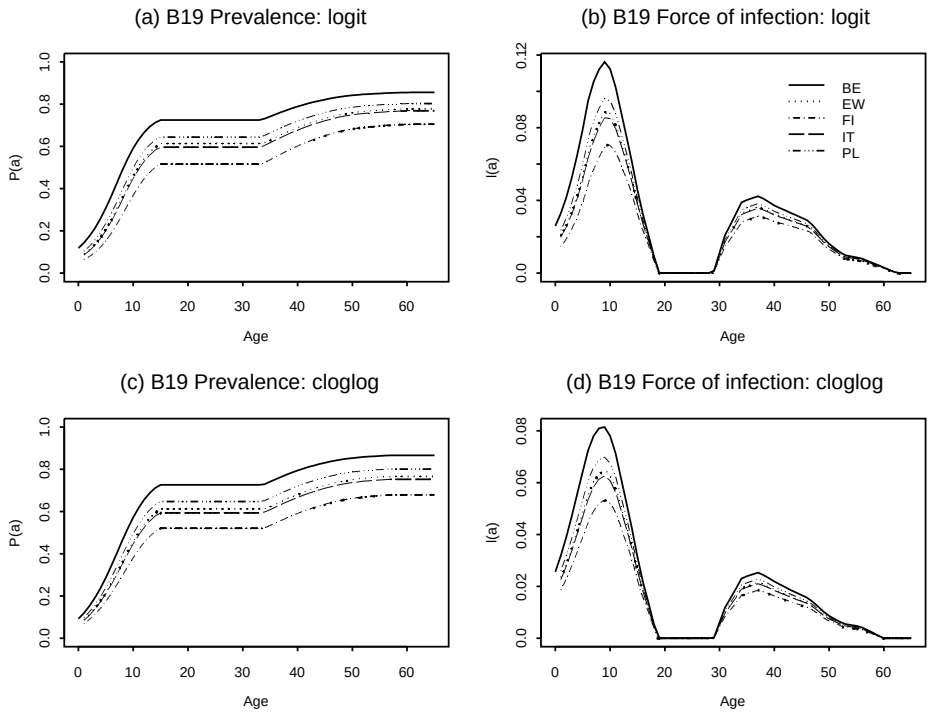

Figure 3.3: *Model (3.2). Estimated prevalence and force of infection for Belgium, England & Wales, Finland, Italy and Poland*

9 years then decreases steadily to the ages between 18 and 28 from where it rises to another peak at age 37 and finally decrease at older age values.

Figure 3.4 shows plots for the estimated curves for prevalence and force of infection for the model which assumes a common intercept and equivalent smoothness for the different smoothing functions for countries. The first peak of the force of infection ranges from age 7 for Finland to age 10 for Belgium. Unlike model (3.2), the second peak from model (3.3) is more clearly pronounced for Poland, England & Wales and Italy than that for Belgium which appears to flatten out. The force of infection for Finland, however, increases from about the age of 45 years onwards. Figure 3.5, which allows separate intercepts for country and separate smoothness for the different smoothing functions for countries suggests that a common intercept for Poland, England & Wales and Italy might be reasonable as their curves barely differ at young age groups. The first peak of the forces of infection now appears from the

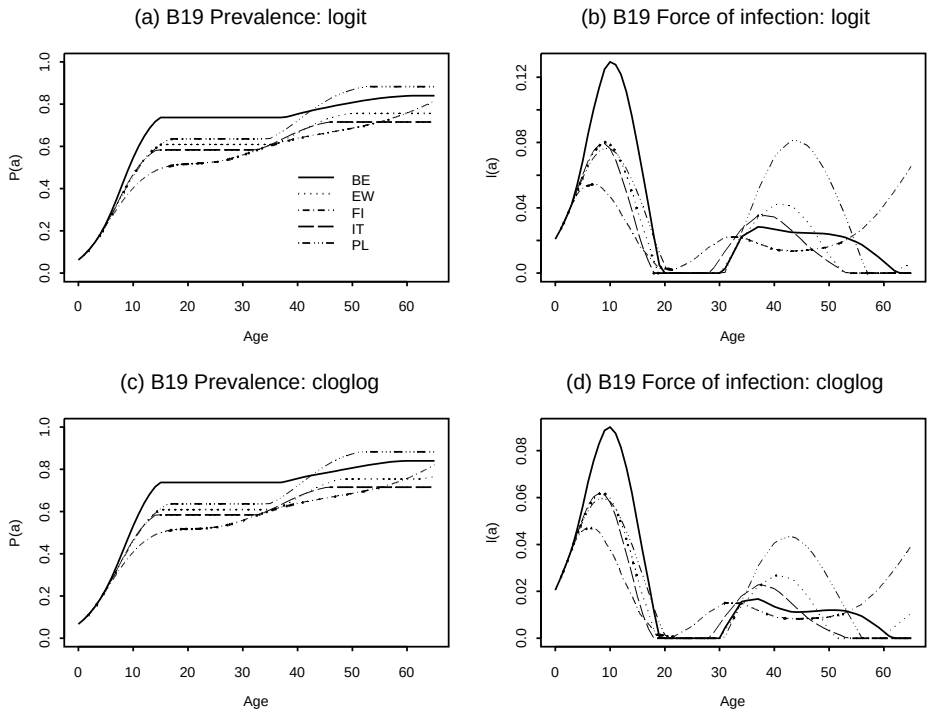

Figure 3.4: *Model (3.3). Estimated prevalence and force of infection for Belgium, England & Wales, Finland, Italy and Poland.*

age of 8 for England & Wales and Poland to the age of 10 for Finland while the second peak appears at the ages of 44, 43 and 40 for Poland, England & Wales and Italy respectively. Figure 3.6 is the fit when a common intercept for Poland, England & Wales and Italy is allowed for in model (3.5).

In summary we have seen that the force of infection for parvovirus-B19 infection increases at young ages until about 10 years. This finding confirms other findings (Gilbert, 2000) that the disease is common in preschool and primary school-aged children. Relating to the current countries, the enrolment age in primary school is: 6 years in Belgium, 5 years in England and Wales, 7 years in Finland, 6 years in Italy and 7 years in Poland. For secondary school enrolment the ages in years are 12, 11, 13, 11 and 13 for the five countries respectively. There is also some suggestion in the data that the force of infection increases among adults in their 30s, meaning that these adults may have acquired the infection from an infected child at home

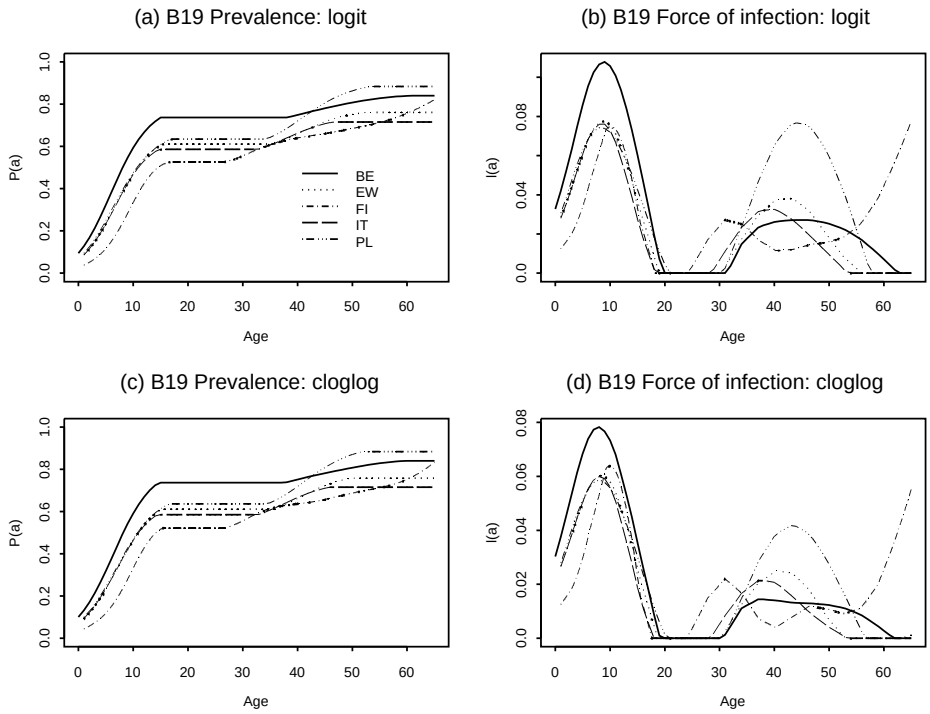

Figure 3.5: *Model (3.4a). Estimated prevalence and force of infection for Belgium, England & Wales, Finland, Italy and Poland*

or at occupational exposure such as, primary school teachers and child-care workers (Gilbert, 2000).

3.4.2 Full Bayesian Approach

The models discussed in Section 3.2 were fitted using WinBUGS software based on one chain of several samples, in order to reduce on the computing time. The convergence of the Markov chain monte Carlo simulations to the posterior distribution were investigated using Geweke diagnostic. The Geweke diagnostic tests the null hypothesis that the Markov chain is in the stationary distribution and produces z-statistics for each estimated parameter. Convergence was checked based on various samples before deciding on the final number of samples needed and how many to remove for burn-in. It was found that the $\boldsymbol{\beta}$ parameters reached their stationary distribution but the variance component was far from stationarity. We further investigated this

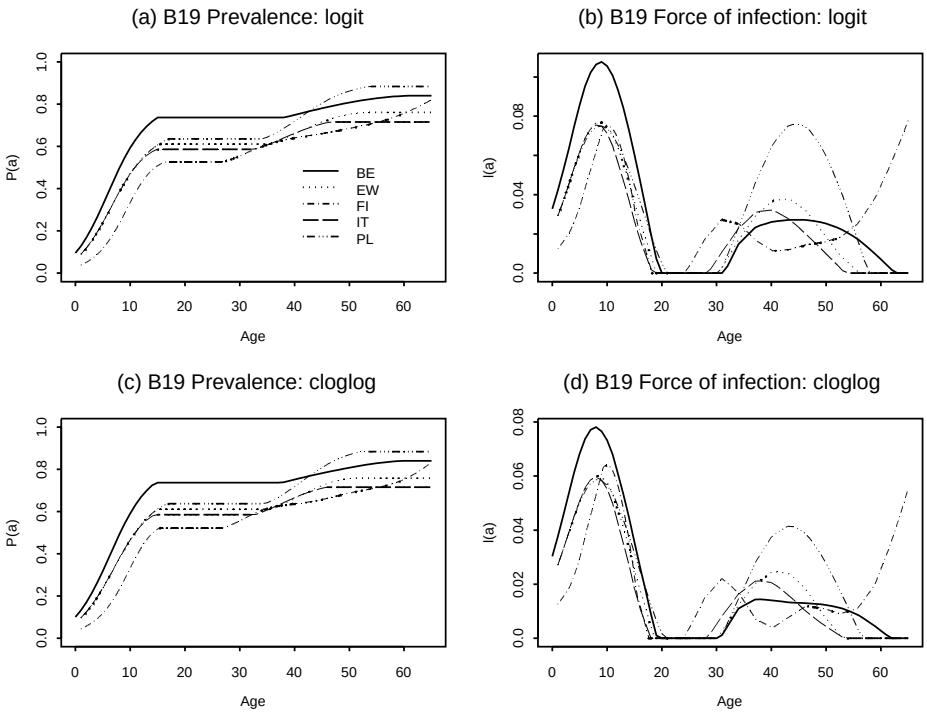

Figure 3.6: *Model (3.5). Estimated prevalence and force of infection for Belgium, Finland, and for England & Wales, Italy and Poland combined.*

using two other prior distributions for the variance components (Crainiceanu *et al.* 2004); the uniform prior on $(0, M]$, and the uniform prior on $[-M, M]$ for the log of σ_u^2, where M is very large. However, convergence for all parameters at one time could not be achieved. The Bayesian fits were based on a chain of 100000 samples, of which the first 10000 were removed as a burn-in period.

Table 3.7 summarizes the DIC differences between the fitted models. Among all the models, model (3.4a) emerged as the best model as it has the lowest DIC. Note that the model is only slightly better than model (3.5) (adjusted AIC equal to 15246.00 and 15246.20 respectively). Comparing the results with those of the frequentist approach, we see that the effective number of parameters in the Bayesian framework is always larger than that of the frequentist (see Figure 3.7). However, this is expected since Bayesian inference takes into account the uncertainty of all parameters. This inherent additional variability results in larger estimated variance components than

those obtained with GLIMMIX.

Table 3.7: *Measures for the goodness-of-fit and complexity of logit and cloglog Bayesian models over 100000 runs with 10000 left out as burn-in.*

	logit link				cloglog link			
Model	$\bar{D}$	$D(\bar{\pi})$	p_D	DIC	$\bar{D}$	$D(\bar{\pi})$	p_D	DIC
(3.1)	15492.60	15481.90	10.65	15503.20	15576.00	15562.50	13.45	15589.40
(3.2)	15255.20	15241.10	14.11	15269.30	15441.50	15424.90	16.59	15458.10
(3.3)	15273.80	15237.20	36.60	15310.40	15455.80	15424.20	31.57	15487.30
(3.3a)	15270.40	15230.80	39.61	15310.00	15447.70	15410.00	37.68	15485.40
(3.4)	15209.80	15172.20	37.55	15247.30	15354.60	15320.50	34.14	15388.80
(3.4a)	15204.50	15163.00	41.49	15246.00	15219.90	15179.00	40.92	15260.80
(3.5)	15205.30	15164.30	40.94	15246.20	15222.30	15183.50	38.72	15261.00

Figure 3.8 and 3.9 compares the estimated force of infection for the five countries using pseudo-likelihood and Bayesian estimation techniques for logit and cloglog link functions respectively for model (3.5), the best model. Also presented in the figures are the 95% credible intervals for the Bayesian fit. It can be seen that the force of infection peaks higher for the Bayesian analysis than the pseudo-likelihood estimate. The Bayesian estimate reveals three peaks of the force of infection for Belgium, Finland and Italy while the pseudo-likelihood fit reveals two peaks. The credible intervals widen at higher age values, revealing limited information available at these age points. In general the two methods show, for all the countries, that the forces of infection are in the same direction. A difference in the fits is observed for England & Wales: while, from the age of 55 onwards, the force of infection increases with the Bayesian fit, the pseudo-likelihood fit becomes negative and thus is constrained to zero.

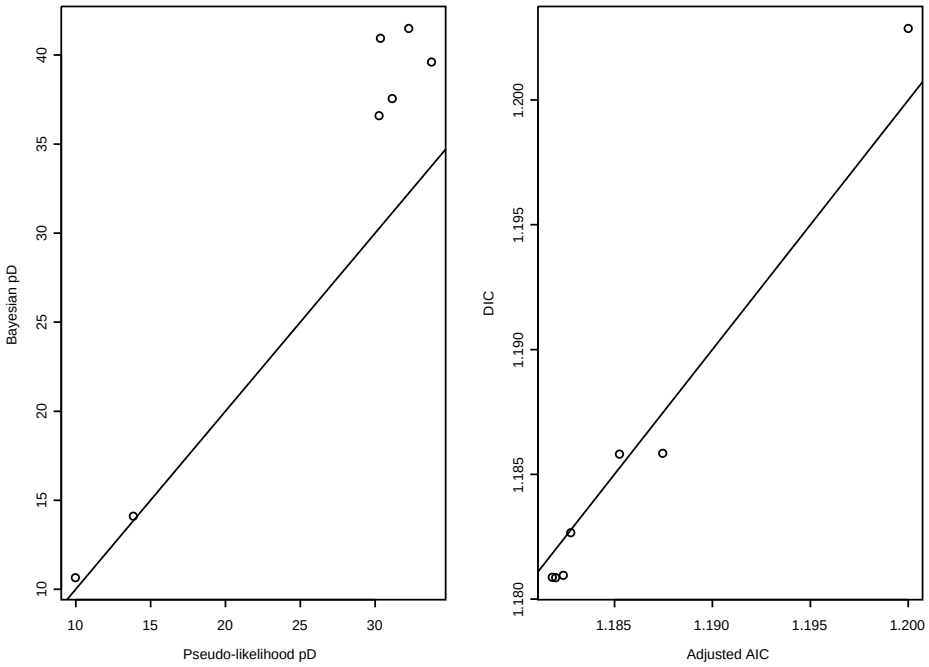

Figure 3.7: *Information criteria and pD for the logit link models:Left panel gives the effective number of parameters obtained for the pseudo-likelihood and full Bayesian approach. Right panel gives the adjusted AIC and DIC*

3.5 Piecewise Constant Force of Infection

In the previous sections the force of infection was assumed to be a smooth function of age with flexible shape. Interestingly, models with constant, linear and piecewise constant force of infection can be easily formulated as GLMM as well. Our starting point is a model which assumes a piecewise constant force of infection. In such a model the force of infection is assumed to be constant within an age group. Consider a finite age-classes population (Anderson and May, 1991) in which the population is sub divided into K age classes. Let $\ell_1, \ell_2, \ldots, \ell_K$ be the force of infection in the age classes. The piecewise constant force of infection model is of primary interest since it is required for the estimation of the WAIFW (who acquires infection from whom) matrix (Anderson and May, 1991). Becker (1989) considered a model with piecewise

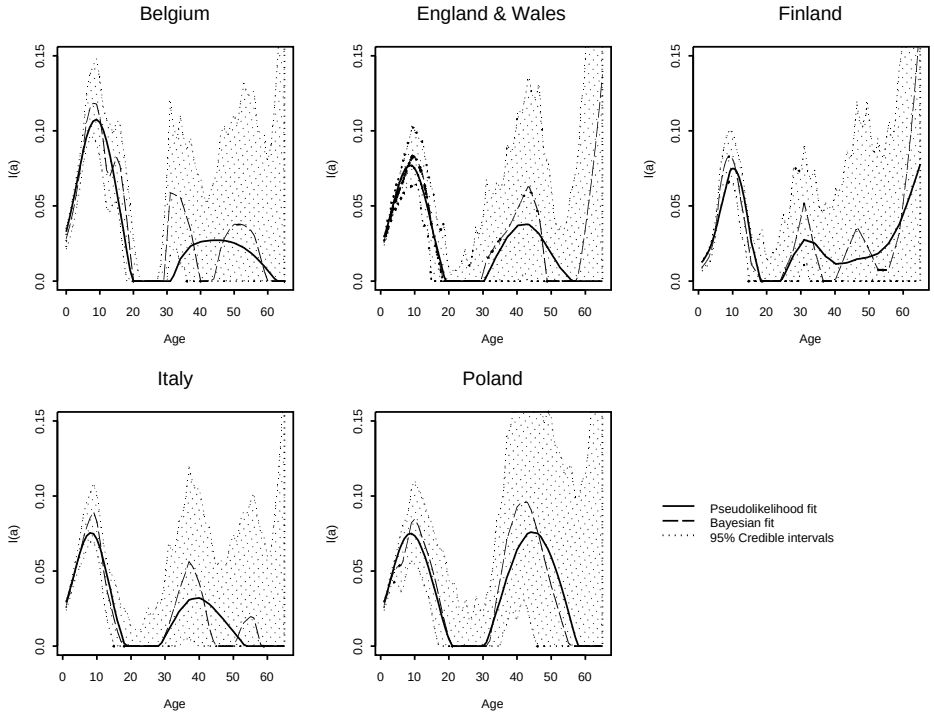

Figure 3.8: *model (3.5) pseudo-likelihood estimate of force of infection (full line), Bayesian estimate of force of infection (broken line) and 95% credible intervals (dotted) using logit link function*

constant force of infection of the form

$$\pi(a) = \exp\left(-\left[\Sigma_{k=1}^{K-1}\ell_k(a_k - a_{k-1}) + \ell_K(a - a_{K-1})\right]\right), \quad \text{for} \quad a_{K-1} \leq a < a_K. \tag{3.14}$$

Note that the piecewise constant model (3.14) was discussed by Becker (1989) as a fixed effects model with the number of parameters equal to the number of age classes which can be fitted to the data as GLM with log link function. In such a model the prevalence is given by

$$\pi(a) = 1 - e^{\eta(a)}.$$

Figure 3.10 shows the linear predictor for the piecewise constant model for a case in which the population is divided into three age classes with force of infection equal to

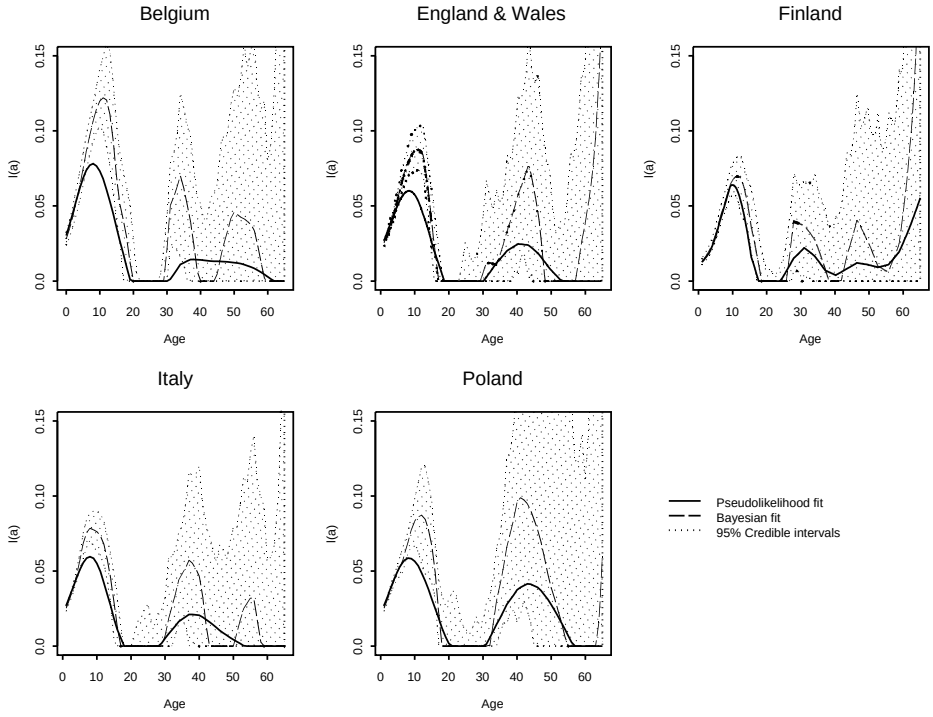

Figure 3.9: *model (3.5) pseudo-likelihood estimate of force of infection (full line), Bayesian estimate of force of infection (broken line) and 95% credible intervals (dotted) using cloglog link function*

ℓ_1, ℓ_2 and ℓ_3, respectively. The linear predictor for this case is given by

$$\begin{aligned}
\eta_1 &= \ell_1 a && \rightarrow \tfrac{d\eta_1}{da} = \ell_1, \quad \text{for} \quad a < a_1, \\
\eta_2 &= \ell_1 a_1 + \ell_2(a - a_1) && \rightarrow \tfrac{d\eta_2}{da} = \ell_2, \quad \text{for} \quad a_1 \leq a < a_2, \\
\eta_3 &= \ell_1 a_1 + \ell_2(a_2 - a_1) + \ell_3(a - a_2) && \rightarrow \tfrac{d\eta_3}{da} = \ell_3, \quad \text{for} \quad a_2 \leq a < a_3.
\end{aligned}$$

As illustrated in Figure 3.10 the force of infection is the slope of the linear predictor at each age class. We turn now to formulate the piecewise constant model as a special case of a GLMM with a linear penalized spline and a log link. For this model, the linear predictor is given according to (3.1), for a polynomial of degree one, as

$$\eta(a) = \beta_0 + \beta_1 a + \sum_{k=1}^{K} u_k (a - t_k)_+, \tag{3.15}$$

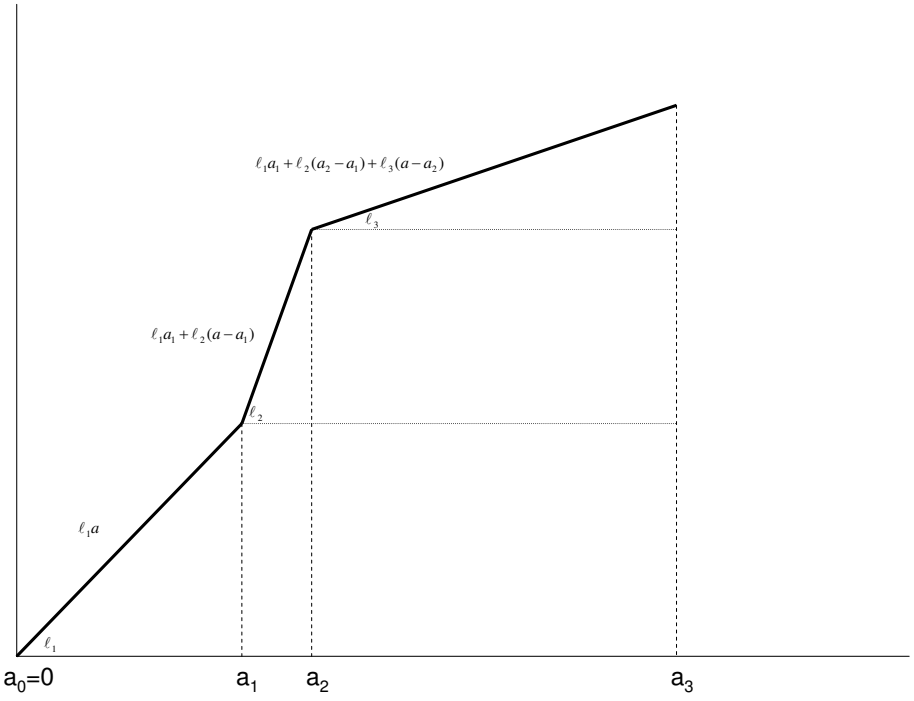

Figure 3.10: *The linear predictor for a model with three age groups.*

where the truncated power basis function can be adapted from (3.1). Hence, for a population with K age classes, $t_k = a_k$. Note the GLMM in (3.15), regardless of the number of age classes, consists of 4 parameters: the two fixed effects and the two variance components. However, the effective number of parameters is not fixed. Using log link function implies $\pi(a) = 1 - \exp(-\eta(a))$. Taking the first derivative of the linear predictor with respect to age, the force of infection is obtained as

$$\begin{aligned}
&\eta_1 = \beta_0 + \beta_1 a && \rightarrow \tfrac{d\eta_1}{da} = \beta_1, && \text{for} \quad a < a_1,\\
&\eta_2 = \beta_0 + \beta_1 a + u_1(a - a_1)_+ && \rightarrow \tfrac{d\eta_2}{da} = \beta_1 + u_1, && \text{for } a_1 \leq a < a_2,\\
&\eta_3 = \beta_0 + \beta_1 a + u_1(a - a_1)_+ + u_2(a - a_2)_+ && \rightarrow \tfrac{d\eta_3}{da} = \beta_1 + u_1 + u_2, && \text{for } a_2 \leq a < a_3.
\end{aligned}$$

Figure 3.11 shows the linear predictor for the GLMM (3.15). The force of infection for the GLMM is the sum of the slopes at each age group. Hence, at the k'th age group, the force of infection is given by

$$\ell_k = \frac{\pi_k^{'}(a)}{1 - \pi_k(a)} = \eta^{'}(a) = \beta_1 + \sum_{k=1}^{i} u_k.$$

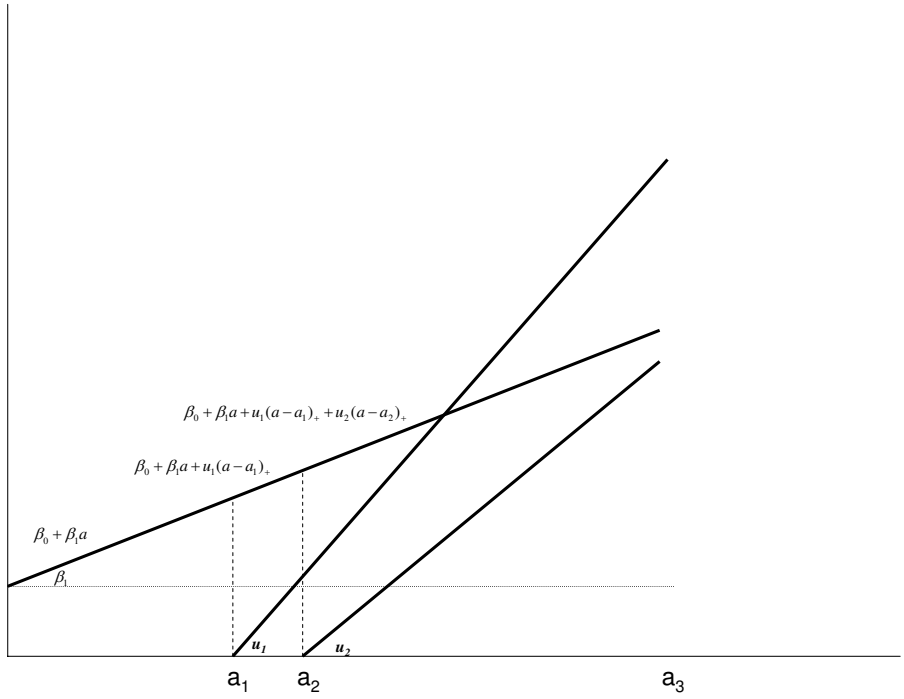

Figure 3.11: *GLMM. The linear predictor for a model with three age groups.*

The parametrization of the constant piecewise model as GLMM is not unique for this specific model but models with constant or linear force of infection can be expressed as GLMM as well. For all models the linear predictor is given by $\eta(a) = X\beta + Zu$.

Table 3.8: *Models for the force of infection using GLMM.*

Model for the force of infection	Linear Predictor	Basis function	σ_u^2	link function
Constant	$\beta_0 + \beta_1 a$	linear	$\sigma_u^2 = 0$	log
Linear	$\beta_0 + \beta_1 a + \beta_2 a^2$	quadratic	$\sigma_u^2 = 0$	log
Piecewise constant	$X\beta + Zu$	linear	$\sigma_u^2 > 0$	log
Flexible	$X\beta + Zu$	any	$\sigma_u^2 > 0$	any

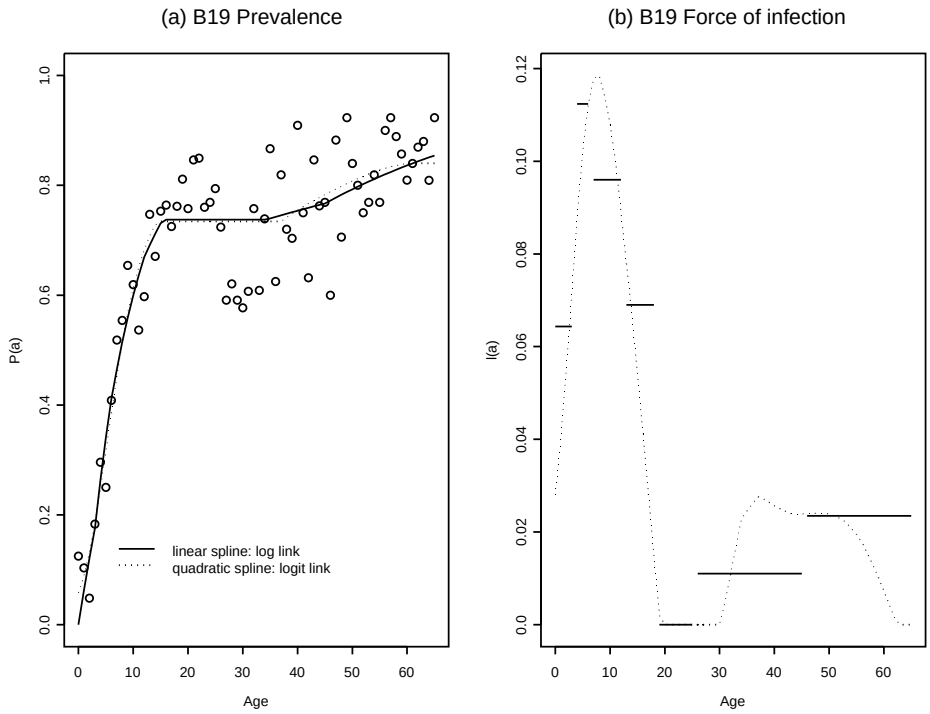

Figure 3.12: *Estimated prevalence and force of infection for the piecewise constant model (full line) and the quadratic spline model (dotted line)*

Table 3.8 presents different models for the force of infection and their GLMM representation. The model formulation for a model with constant force of infection is identical to the constant piecewise model (3.15). The main difference between the model is the variance of the random effects. The case with $\sigma_u^2 = 0$ implies that $\eta = \beta_0 + \beta_1 a$ and therefore a constant force of infection. A model with linear force of infection can be fitted using quadratic penalized spline and a log link function. Similar to the model with constant force of infection, one needs to test if $\sigma_u^2 = 0$ (i.e. $\eta(a) = \beta_0 + \beta_1 a + \beta_2 a^2$).

Figure 3.12 shows the constant piecewise model for B19 in Belgium. Seven age classes, [0,3), [3,6), [6,12), [12,18), [18,25), [25,45), and [45,65], according to the school structure of Belgium, were considered with the knots $t_1 = 3, t_2 = 6, t_3 = 12, t_4 = 18, t_5 = 25$ and $t_6 = 45$. The dotted line in the Figure corresponds to a quadratic spline model using the logit link function and 20 knots. The constant piecewise

model yielded a deviance of 3428.72 on 5.93 effective number of parameters giving an adjusted AIC of 1.1189. The deviance of the quadratic spline model is 3431.42 with 7.56 effective number of parameters yielding a slightly increased adjusted AIC of 1.12083.

3.6 Discussion and Conclusion

In this chapter we have discussed the model extension of the simple penalized spline model to accommodate a discrete covariate, country in this case. Among the various models considered, the model which allows country-specific intercepts provides useful interpretations. Using the logit link function the odds ratios between each country and the baseline country are proportional, where the exponent of the estimated country effect is the proportionality constant. Applying the complementary log-log link to the model yields proportional force of infection ratios between each country and the baseline country and a proportionality constant is the same as that mentioned for the logit link. An advantage of these additive models is that the effects are simple to summarize and interpret, requiring only a single parameter. In general this enables comparisons of the prevalences and forces of infection between country categories without having to look at the smooth function of age. However, when different smoothing functions are allowed for each country the proportionality feature does not hold and the odds ratio and the force of infection ratios become age-dependent.

The procedure for fitting the models had to be chosen. The method of likelihood estimation using the NLMIXED SAS procedure required excessive computation and demanded large amounts of storage for the estimation of the 20×20 diagonal covariance matrix corresponding to the 20 knots. A possible solution was to use the likelihood approach based on the full Bayesian analysis. The drawback encountered with the Bayesian approach was that convergence for all model parameters to their posterior distributions could not be achieved. However, convergence was considered satisfactory since the fixed effects parameters of the spline models attained convergence. An alternative strategy to fit the models was the pseudo-likelihood approach implemented with the GLIMMIX SAS procedure.

Another important point concerned model selection. The model fit statistics reported by the GLIMMIX SAS procedure are not recommended to compare between different models since the pseudo data change each time the mean structure changes. However, we used the pseudo-likelihood estimates to plug them into the hat matrix obtaining the effective number of parameters as the trace of the hat matrix. With the

trace and deviance given the pseudo-likelihood estimates, a model selection procedure based on an adjusted AIC discussed by Ruppert *et al.* (2003) was employed. The adjusted AIC model selection criterion was compared with the DIC criterion of the Bayesian approach. The two approaches to model selection yielded similar conclusions regarding the best model. Moreover, we have shown that the effective number of parameters are higher for the Bayesian technique than with pseudo-likelihood estimation, a finding that can be explained by the added variability into Bayesian models through prior distribution assignment for each parameter.

We have also shown that the piecewise constant force of infection can be formulated as GLMM using linear penalized splines with the log link function. Furthermore by imposing various constraints other models with constant and linear force of infection can be formulated as penalized spline models, leading to an appreciation of the flexibility of the penalized splines.

Chapter 4

Estimation of the Prevalence and Force of Infection of Hepatitis C Among Injecting Drug Users in Five European Countries

Unlike the previous chapters which use non- or semi-parametric modeling of the prevalence and force of infection, this chapter focuses primarily on modeling the prevalence and the force of infection using parametric models. In addition the estimated prevalence using the parametric model is compared with another nonparametric approach based on isotonic regression. The problem of interest in this chapter concerns hepatitis C virus (HCV) infection among injecting drug users. The hepatitis C virus is the leading cause of known liver diseases in most industrialized countries. It is a common cause of cirrhosis and hepatocellular carcinoma (HCC) as well as the most common reason for liver transplantation. At least 170 million people worldwide are believed to be infected with this virus. Following the identification of hepatitis A and hepatitis B, this disorder was categorized in 1974 as "non-A, non-B hepatitis." In 1989, the hepatitis C virus was discovered and was found to account for the majority of those patients with non-A, non-B hepatitis (Baker 2002). hepatitis C virus (HCV) is an

RNA virus of the Flaviridae family. There are 6 HCV genotypes and more than 50 subtypes. These genotypes differ by as much as 30 to 50 percent in their nucleotide sequences. The virus also has a high mutation rate. The extensive genetic heterogeneity of HCV has important diagnostic and clinical implications, causing difficulties in vaccine development and the lack of response to therapy (Baker 2002).

HCV transmission occurs primarily through exposure to infected blood. This exposure exists in the context of injection drug use (IDU), blood transfusion, solid organ transplantation from infected donors, use of unsafe medical practices, occupational exposure to infected blood, through birth to an infected mother, multiple heterosexual partners, and high-risk sexual practices (Baker 2002). Historically, in industrialized countries, blood transfusions and administration clotting factor concentrates were the most important mode of transmission. However, following the introduction of current blood screening strategies in the early 1990s, HCV infection via these routes has become a rare event in industrialized countries (Matheï *et al.* 2002). hepatitis C seems to be acquired rapidly after initiation of drug injection and many people may have been infected as a result of occassional experimentation with the drug (Matheï *et al.* 2002). Once infection has occurred, the virus replicates in the liver and can be detected in the serum using polymerase chain reaction (PCR) within 1-2 weeks. Detectable antibody to HCV is present in majority of cases by 12 weeks, though in small proportion this is delayed or does not occur. The majority of acute infections are asymptomatic or with minor symptoms. Matheï *et al.* (2002) reported that between 60% and 90% of people acutely infected develop chronic infection, of which an unknown but small proportion will clear the infection over a period of time.

In context of childhood infectious diseases, the epidemiological quantity of interest is the force of infection, which is the rate at which the susceptible become infected. Under the assumptions of life long immunity and that the disease is in a steady state the prevalence and the force of infection can be estimated from seroprevalence data (Grenfell and Anderson 1985). Parametric models for the prevalence and the force of infection of childhood infections estimated from seroprevalence data are discussed by Grenfell and Anderson (1985) who model the force of infection with a polynomial function of the host age. Other parametric models fitted within the framework of generalized linear models (GLM) with binomial error (McCullagh and Nelder 1989) were discussed by Becker (1989), Diamond and McDonald (1992) and Keiding *et al.* (1996) who model with complementary log-log link function in order to parameterize the prevalence and the force of infection as a Weibull model. Becker (1989) suggested to model a piecewise constant force of infection by fitting a model with a log link. In the case that other covariates, in addition to exposure time, are included in the model,

Jewel and Van Der Laan (1995) proposed, for current status data, a proportional hazards model with constant force of infection which can be fitted as a GLM with a complementary log-log link. Grummer-Strawn (1993) discussed two parametric models for current status data, the first one being a Weibull proportional hazards model with complementary log-log link and the second being the log logistic model with logit link function. For the latter, the proportionality in the model is interpreted as proportional odds. Farrington (1990) and Farrington *et al.* (2001) proposed a non linear model for which the force of infection is restricted to be non negative and applied the model for measles, mumps and rubella. For childhood infectious diseases, the exposure time is the host age and the prevalence and force of infection are assumed to be age dependent. In the context of hepatitis C, the exposure time is the length of the injecting career, i.e., it is the time interval from the age of entering into the risk group (the age at first injection) to the age at test, assuming uninterrupted exposure. In addition to the length of the injecting career other behavioral risk factors such as sharing syringes, sharing injecting paraphernalia and frequency of injecting can influence the transmission of HCV and therefore the impact of these risk factors on the transmission parameters is of primary interest. The aim of the analysis presented in this chapter is to investigate how the above risk factors, as measured by self report at the time of interview, are associated with the prevalence and force of infection of HCV across the six studies. In Section 4.1 we describe the study design, the main risk factors under investigation and discuss the statistical methodology used in order to investigate the associations of the risk factors on both the prevalence and the force of infection. Descriptive analysis and statistical models which were used to estimate both the prevalence and the force of infection are discussed in Section 4.2. The analyses of this chapter are available in Namata *et al.* (2008e).

4.1 Data and Methods

4.1.1 Study Design

Sample Size and Demographics

The data used for the analyses presented in this chapter consists of six seroprevalence samples of injecting drug users from five European countries: Belgium (N=335), two studies from the Czech Republic (N=237,754), Italy (N=947), Spain (N=511) and Sweden (N=310). Two IDUs sub populations were considered for the analysis: (1) ever injectors, i.e., IDUs who gave an affirmative answer to the question “did you ever

Table 4.1: *Data per study in the given study periods: Sample sizes and prevalence of HCV.*

Study	Study Period	Ever Injectors N(%HCV+)	Recent Injectors N(%HCV+)
Belgium	2006	335 (78.2)	97 (85.6)
Czech Republic I	1998-2001	237 (20.7)	185 (18.9)
Czech Republic II	2002-2003	754 (29.8)	661 (30.9)
Italy	2005	947 (76.6)	
Spain	2001-2003	511 (73.6)	427 (75.9)
Sweden	2004-2006	310 (86.5)	

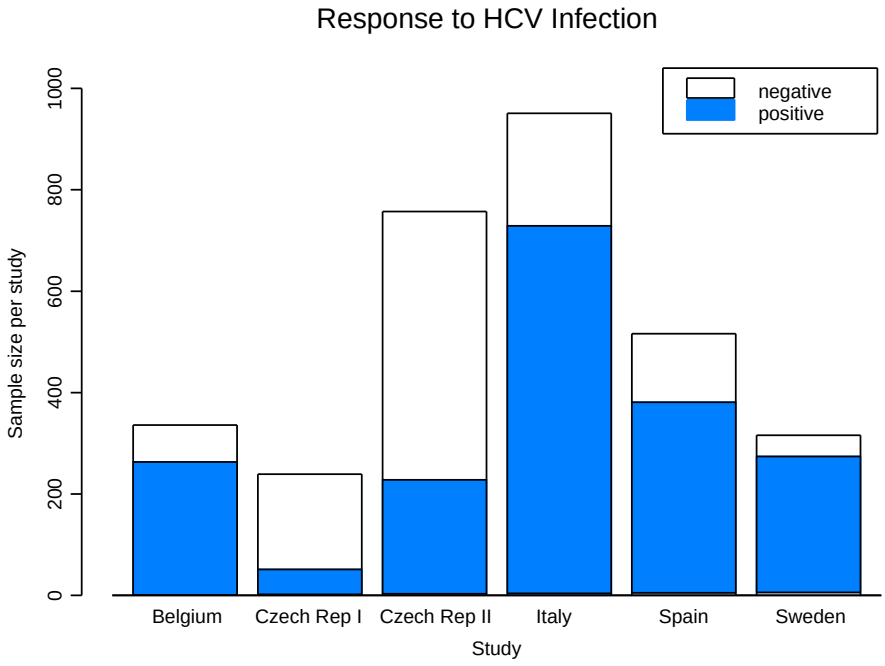

Figure 4.1: *Sample size and number of HCV seropositives per study for ever injectors.*

inject drugs" and (2) recent injectors, i.e, IDUs who injected drugs in the last month before the interview. All participants recruited in the studies were ever injectors which means that recent injectors are a subset of the ever injectors. It is worth noting that the Italian and the Swedish studies provided no information for recent injectors. For more details about these cross-sectional surveys we refer to Sutton *et al.* (2008). Table 4.1 presents the size of the data for analysis for each study, the periods in which they were collected and proportions of HCV seropositives for recent and ever injectors. The prevalence of HCV for ever injectors, presented in Table 4.1 and Figure 4.1, ranges between 20.7% in the first Czech Republic study to 86.5% in the Swedish study. In addition to the serological test, all IDUs who participated in the study were interviewed and information about demographic characteristics and injecting behavior was collected. Table 4.2 presents the demographic characteristics of the participants. The average age at interview ranges between 20.1 (SD=4.1) in the first study of Czech Republic to 39.2 (SD=7.9) in the Belgian study. The proportion of males among the participants ranges between 22.4% (first study of Czech Republic) to 85.8% in Sweden. Pearson chi square tests for independence (see second row of Table 4.3) reveal that the proportion of HCV positive across all studies (except the second study from the Czech Republic) is equal for males and females. The age at which the IDUs started their injecting career (the age at first injection) ranges between 17.9 (SD=3.2) for the first study of the Czech Republic to 22.7 (SD=7.5) in Belgium. Two analyses were considered. The first analysis includes ever injectors from all studies and in the second only IDUs who were classified as "recent" IDUs were included in the analysis (i.e., from the studies in Belgium, the Czech Republic and Spain). Finally, Figure 4.2 shows the proportion of HCV seropositive as a function of the length of the injecting career of the ever IDUs and reveals, as expected, an increasing trend of seropositive proportion with respect to the length of the injecting career. In the next section we will investigate this pattern in more detail.

Behavioral risk Factors

- **Sharing Syringes**
  In addition to the length of the injecting career, HCV transmission is associated with behavioral risk factors of the IDUs. It is well documented (Matheï *et al.* 2006) that IDUs who share syringes, other paraphernalia and have high frequency of injecting per day are exposed to a higher risk of infection. Information about ever syringe sharing is available in all studies except for Italy. Figure 4.3 and Table 4.3 (third row) show the distribution of HCV seropositive for ever

Table 4.2: *Descriptive statistics of HCV for demographic variables. Information about the age at first infection is not available in the second study from the Czech Republic*

	Study					
Demographic Variable	Belgium	Czech Republic I	Czech Republic II	Italy	Spain	Sweden
Gender						
Male						
N (%HCV+)	233 (80.7)	161 (22.4)	490 (32.7)	778 (76.5)	376 (74.7)	233 (85.8)
Female						
N (%HCV+)	102 (72.6)	76 (17.1)	264 (24.6)	164 (76.8)	135 (70.4)	77 (88.3)
Age at Interview						
mean (SD)	39.2 (7.9)	20.1 (4.1)	24.6 (6.7)	35.2 (7.6)	25.9 (3.2)	35.7 (10.1)
Age at 1st Injection						
mean (SD)	22.7 (7.5)	17.9 (3.2)		21.5 (5.2)	19.4 (3.9)	21.8 (8.7)

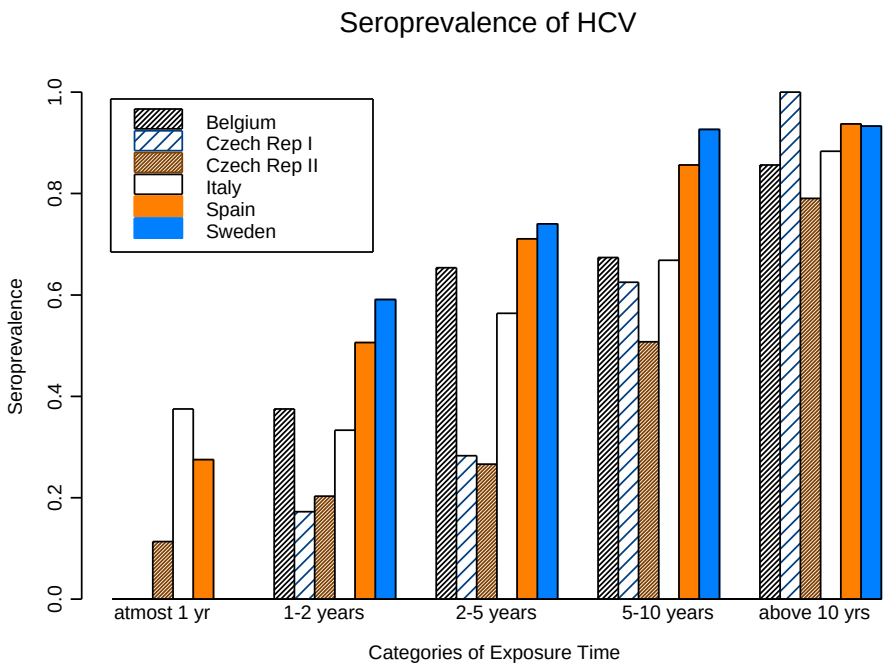

Figure 4.2: *Proportion of HCV seropositive by the length of the injecting career for ever injectors.*

injectors who share or do not share syringes. For all studies, the prevalence of HCV among IDUs who report ever sharing syringes is higher than the prevalence of IDUs who never shared syringes. Pearson chi-square tests indicate that self reported syringe sharing is a significant risk factor in Belgium, the second study from the Czech Republic and Spain. However, syringe sharing in the past month before the interview is borderline significant for Belgium.

- **Sharing Other Paraphernalia**
  Table 4.3 (fourth row) and Figure 4.3 show the distribution of HCV seroprevalence among IDUs who ever or never shared other paraphernalia than needles/syringes (for the second study in the Czech Republic, Spain and for Sweden). The prevalence of HCV for the IDUs who share other paraphernalia is equal to 30.6%, 77.2%, and 92.1% respectively in the three studies. Pearson chi-square tests reveal that self reported sharing of other paraphernalia is significant

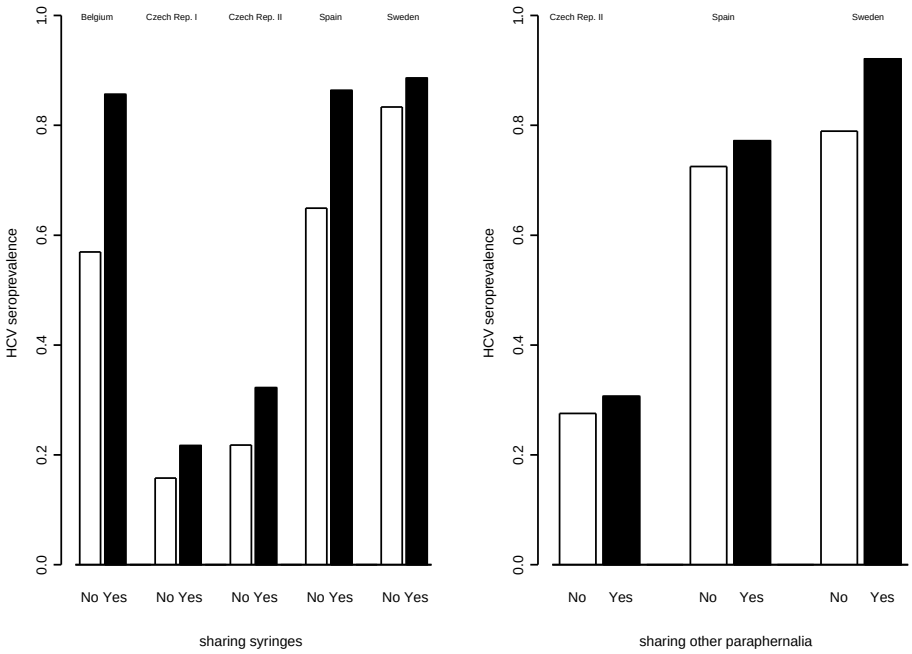

Figure 4.3: *Proportion of HCV seropositive by sharing status (No=IDUs who never shared syringes or other paraphernalia, Yes=IDUs who ever shared).*

in the Swedish study.

- **Frequency of Injections**
  Information about the self reported frequency of injections in the last month before the interview is available in the second study for the Czech Republic and in the Spanish study. In both studies the prevalence of HCV reveals an increasing trend when the number of injections per day increases (see Figure 4.4) and it was found to be a significant risk factor in both studies for ever injectors (last row, Table 4.3).

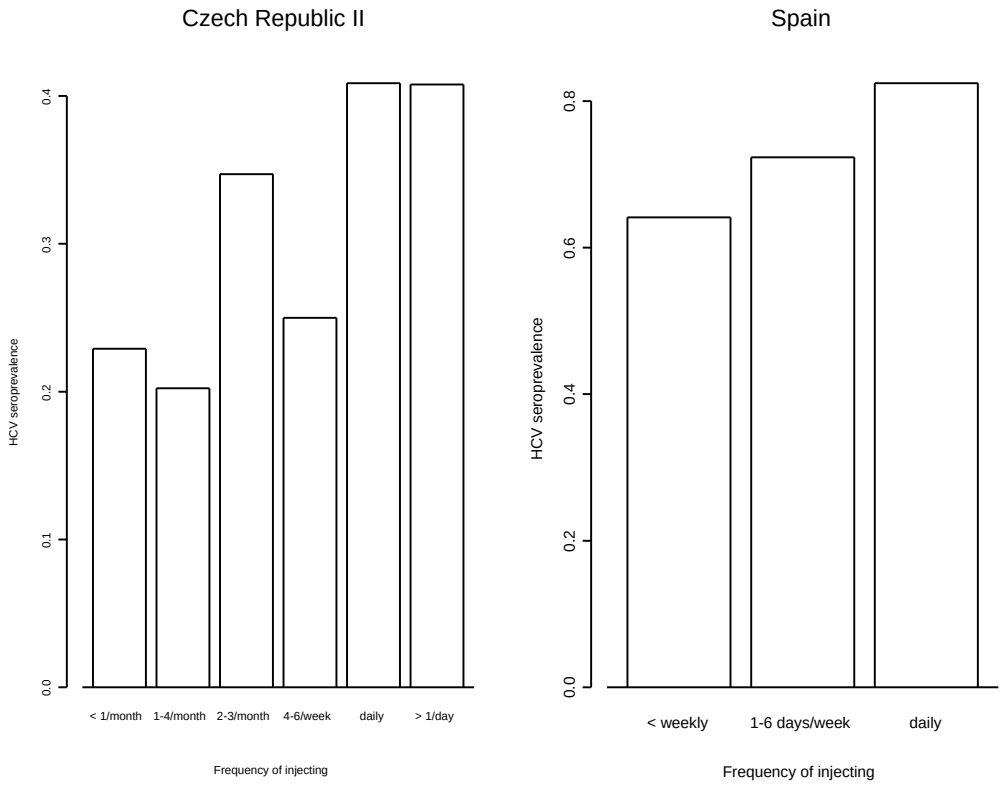

Figure 4.4: *Proportion of HCV seropositive by frequency of injecting in past month.*

4.1.2 The Exposure Time - The Length of the Injection Career

The time of exposure is the length of the injecting career and it is considered to be the length of time (in years) in which the IDUs are in the risk group. The exposure time is defined as the difference between the age at test and the age at first injection, i.e. we assume a continuous injecting career. For the second study from the Czech Republic the information about the age at first injection is not available. Therefore the length of the injecting career is taken as the midpoints of the grouped duration of injection, i.e, midpoints of 1-6 months, 6-12 months, 1-2 years, 2-5 years, 5-10 years and more than 10 years. For the last group the length of the injecting career is considered to be 10 years. Figure 4.5 presents the proportion of HCV seropositive (left panels) when, except for Czech Republic study two, the length of the injecting career is grouped in time intervals of 1 year. The right panels in Figure 4.5 show the sample

Table 4.3: *Distribution of hepatitis C virus seropositivity in relation to the predictor variables using Pearson chi square test of independence for five studies. Due to missing information on covariates, the sample sizes may not add up to total as in Table 4.1.*

	Study																	
	Belgium			Czech Rep I			Czech Rep II			Italy			Spain			Sweden		
Variable	hcv+	hcv-	Pvalue	hcv+	hcv-	Pvalue	hcv+	hcv-	Pvalue	hcv+	hcv-	Pvalue	hcv+	hcv-	Pvalue	hcv+	hcv-	Pvalue
Gender			0.0969			0.3511			0.0215			0.9232			0.3239			0.5822
male	188	45		36	125		160	330		595	183		281	95		200	33	
female	74	28		13	63		65	199		126	38		95	40		68	9	
Syringe Sharing			<.0001			0.5462			0.0110						<.0001			0.1843
never	41	31		3	16		34	122					185	100		105	21	
ever	179	30		46	166		188	394					190	30		163	21	
Sharing other Paraphenalia									0.4527						0.2444			0.0008
never							43	113					219	83		105	28	
ever							180	407					149	44		163	14	
Syringe Sharing in last month			0.0476			0.4022									0.4464			
no	200	61		20	65								313	119				
yes	20	1		28	120								38	11				
Frequency of Injection															0.0017			
daily													136	29				
1-6days/week													149	57				
less weekly													84	47				
Frequency of Injection									0.0001									
more times day							42	61										
daily							38	55										
4-6 times a week							19	57										
2-3 times/month							59	111										
1-4 times/month							36	142										
< once/month							30	101										

Table 4.4: *Mean Exposure time per study and analysis.*

Study	Exposure time	
	Ever	Recent
	Mean (SD)	Mean (SD)
Belgium	16.48 (8.78)	16.86 (8.81)
Czech Republic I	2.20 (2.45)	2.31 (2.37)
Czech Republic II	3.60 (3.05)	3.74 (3.08)
Italy	13.87 (7.92)	
Spain	6.49 (4.54)	6.47 (4.47)
Sweden	13.89 (12.00)	

size at each exposure time. The mean length of the injecting career is presented in Table 4.4. For ever injectors, the mean injecting time ranges from 2.2 years (SD=2.5) in the first Czech Republic study to 16.5 years (SD=8.8) in Belgium. For the recent injectors information is available for Belgium, the Czech Republic studies and for Spain with mean injecting time ranging from 2.3 (SD=2.4) in the first Czech study to 16.9 (SD=8.8) in Belgium.

4.1.3 Statistical Methodology

In the analysis discussed below, attention is placed on modeling the prevalence of seropositive HCV ($\pi(t)$) among the IDUs population and the force of infection $\ell(t)$, which is the rate at which susceptible individuals become infected, both as a function of the exposure time. The force of infection is defined as $\ell(t)dt$=P(infected between t and $t+dt|$ susceptible at t). Let $\Lambda(t) = \int_0^t \ell(t)dt$ and $\pi(t)$ the probability to be infected before exposure time t, then $\pi(t) = 1-\exp(-\Lambda(t))$. Consider a cross-sectional prevalence sample of size N and let t_i be the exposure time of the i'th subject. Instead of observing the exposure time at infection we observe a binary variable Y_i such that

$$Y_i = \begin{cases} 1 & \text{if subject } i \text{ had experienced infection before exposure time } t_i, \\ 0 & \text{otherwise.} \end{cases} \tag{4.1}$$

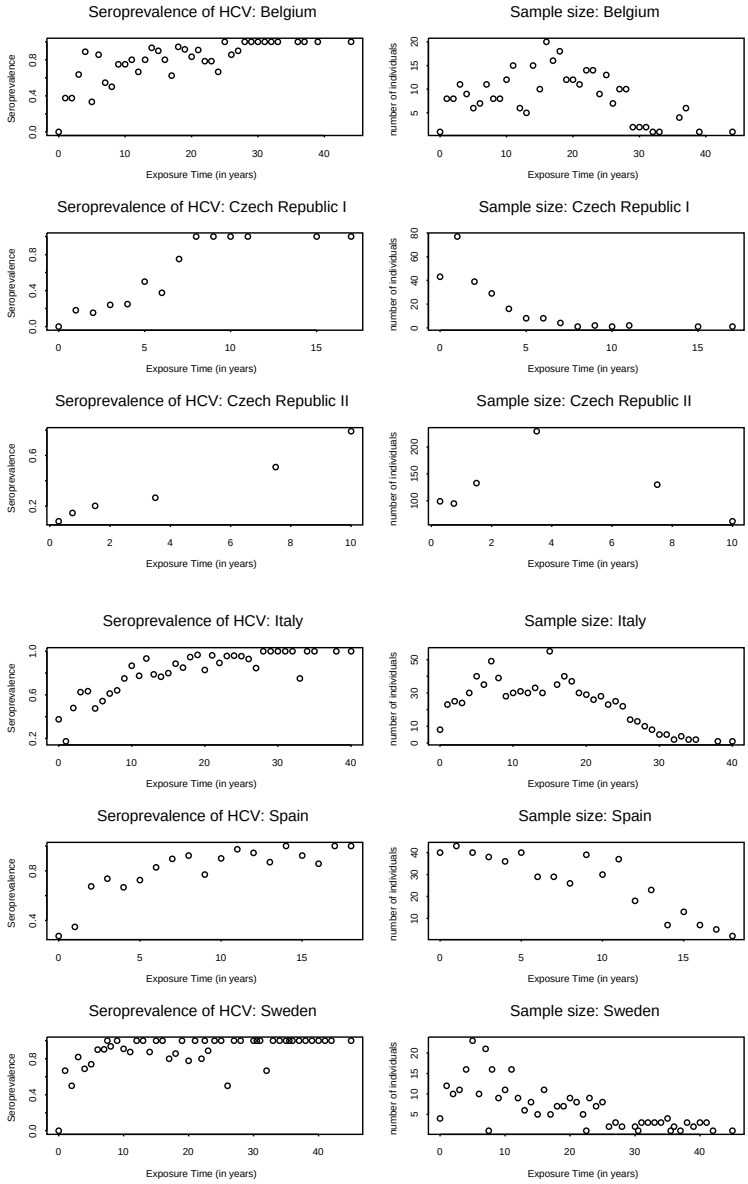

Figure 4.5: *Left panels: Proportion of HCV sero positive by exposure time among ever injectors, Right Panels: sample size by exposure time.*

the binomial log likelihood is given by

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{N} Y_i \log\{\pi(t_i)\} + (1 - Y_i) \log\{1 - \pi(t_i)\}. \tag{4.2}$$

Note that all observations in the sample are either left censored (if the IDU was infected before the test) or right censored (if the IDU was infected after the test).
In this chapter we follow the parametric approach of Diamond and McDonald (1992) and Keiding *et al.* (1996) who proposed generalized linear model (McCullagh and Nelder, 1989) for the binary data with a linear predictor given by

$$\eta(t) = \mu + \beta \log(t). \tag{4.3}$$

A model with complementary log-log link function implies that the prevalence is

$$\pi(t) = 1 - \exp(-\alpha t^{\beta}), \tag{4.4}$$

where $\alpha = \exp(\mu)$. Note that model (4.4) implies an underlying Weibull distribution in the susceptible class. The force of infection in this model is given by

$$\ell(t) = \alpha\beta t^{\beta-1}. \tag{4.5}$$

In case that other covariates are included in the model the linear predictor becomes

$$\eta(t) = \log(\alpha) + \beta \log(t) + \boldsymbol{Z}\boldsymbol{\gamma}, \tag{4.6}$$

where $\boldsymbol{Z}$ is a design matrix of the risk factors and $\boldsymbol{\gamma}$ is the parameter vector to be estimated. The probability to become infected before exposure time t is $\pi(t) = 1 - \exp(-\alpha t^{\beta} \exp(\boldsymbol{Z}\boldsymbol{\gamma}))$ and the Weibull force of infection in this case is

$$\ell(t|\boldsymbol{Z}) = \alpha \exp(\boldsymbol{Z}\boldsymbol{\gamma})\beta t^{\beta-1}. \tag{4.7}$$

4.2 Data Analysis

The analysis presented in this section was carried out in three parts. In the first part of Section 4.2.1 the association between HCV, HIV and HBV was investigated using Pearson chi-square tests for independence. In the second part of the analysis of Section 4.2.1 multiple logistic regression models were used in order to investigate the influence of the behavioral risk factors on the prevalence of HCV. In Section 4.2.2, the Weibull model discussed in Section 4.1.3 is used in order to model the prevalence and the force of infection for HCV and to assess the impact of the risk factors on both the prevalence and the force of infection.

4.2.1 Descriptive Analysis

The Prevalence of HCV, HBV and HIV

Typically three diseases: infection with hepatitis C virus, hepatitis B virus and Human Immunodeficiency virus were investigated among injecting drug users for the

Table 4.5: *Testing independence between HCV, HBV and HIV using Pearson chi square test. Chi square value (P value) are shown.*

Study	HCV vs HBV	HCV vs HIV	HBV vs HIV
Belgium		3.23 (0.0724)	
Czech Republic I	14.12 (0.0002)		
Italy	68.94 (<.0001)	17.35 (<.0001)	23.61 (<.0001)
Spain	16.86 (<.0001)	26.60 (<.0001)	6.88 (0.0087)
Sweden	39.63 (<.0001)		

six studies presented in this chapter. However, for the second study from the Czech Republic there was information only on HCV while for the first study from the Czech Republic and the Swedish study, information on HIV was not available. Figure 4.6 shows the proportions of seropositive IDUs for each of the viruses per study. Among the three viruses, HCV had higher prevalence in all studies than HBV and HIV. The seropositive percentages of both HCV and HBV were lowest for Czech Republic study one and highest for Sweden.

Using the Pearson chi-square tests of independence, the presence of association between the three diseases was examined. Test statistics and p-values are shown in Table 4.5. At a 5% significance level, we notice a highly significant association between all possible pairs of the three diseases (except in the Belgian study). This indicates that IDUs who are infected by one of the diseases are more likely to be infected with the other diseases as well. The measure of the magnitude of this association was studied using odds ratios. Table 4.6 presents the odds ratios and their 95% confidence intervals. Except in Belgium, the odds ratios are all highly significant (the lower confidence limits are all above one). For interpretation of results, let us consider the Swedish study and the relation between HCV and HBV. For an injecting drug user, the odds to be infected by both HBV and HCV are 19 times the odds of being not infected at all. Similar interpretations apply to the rest of the studies except Belgium. The association patterns will not be investigated further in this chapter. For an elaborate discussion about the association patterns and co-infection we refer to Shkedy *et al.* (2008).

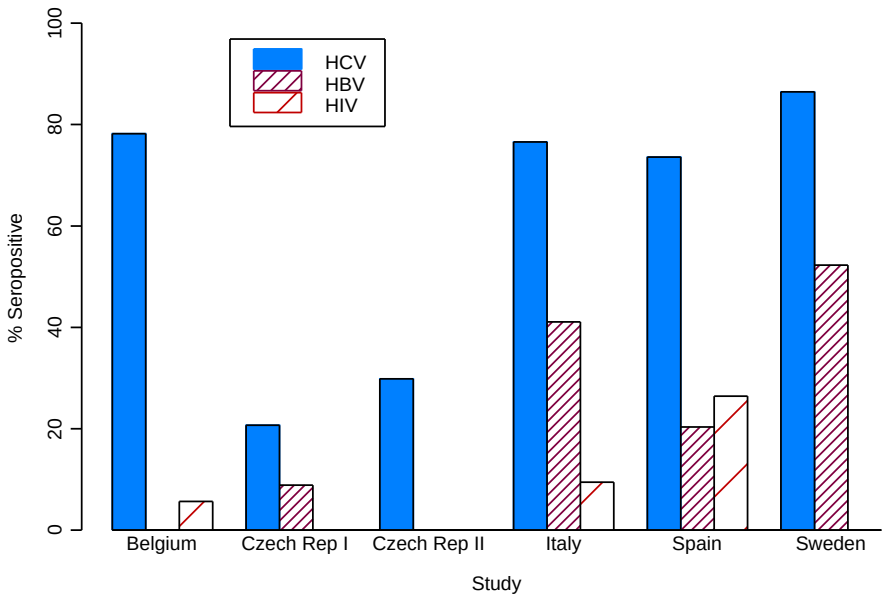

Figure 4.6: *Proportion of HCV, HBV and HIV seropositive by study for ever injectors.*

Descriptive Analysis of the behavioral risk Factors

The influence of the behavioral risk factors on the prevalence among both "recent" and "ever" injectors was explored using multiple logistic regression models in which the dependent variable is the seroprevalence status and the behavioral risk factors were included as covariates in the models. The odds ratios, their confidence intervals and p-values are shown in Table 4.7 and 4.8, respectively.

We first discuss the results obtained for the first analysis in which the six studies were included. For the second Czech study, gender, ever sharing of syringes and the frequency of injecting in the last month were found to be significant risk factors. The odds of males to be infected with HCV were 1.4 times the odds for women. The IDUs who ever shared syringes had 1.9 higher odds of infection relative to those who did not share syringes. IDUs who injected more than once per day had 1.9 odds of infection compared to those who injected less than once in a month. For the Italian study only age at first injection was significant with the odds of infection lower for

Table 4.6: *Odds Ratios and their confidence intervals between HCV, HBV and HIV .*

Study	HCV vs HBV	HCV vs HIV	HBV vs HIV
Belgium		5.31 (0.70 , 40.47)	
Czech Republic I	5.15 (2.04 , 13.00)		
Italy	5.74 (3.68 , 8.97)	5.68 (2.27 , 14.20)	3.80 (2.15 , 6.73)
Spain	3.70 (1.92 , 7.16)	4.51 (2.45 , 8.31)	1.84 (1.16 , 2.91)
Sweden	18.96 (5.72 , 62.92)		

a unit increase in age at first injection. The age at first injection, sharing syringes and the frequency of injecting were significant risk factors for the Spanish study. The IDUs who shared syringes were about 3 times more at risk for HCV than those who did not share syringes while those who injected drugs daily were about twice more at risk than those who injected less weekly. Similar patterns were observed in the Belgian study. IDUs who shared syringes were about 4.2 times higher at risk than IDUs who did not share syringes. For the Swedish study, sharing other paraphernalia was the significant risk factor with a risk 3.5 times higher for the IDUs who shared other paraphernalia than those who did not share other paraphernalia.
In the second analysis, only studies for which information about recent injectors is available were considered. The analysis shows that frequency of injecting in the last month prior to the interview and age at first injection, respectively, for the second Czech Republic study and for Spain, are significant risk factors.

4.2.2 Modeling the Prevalence and Force of Infection

Exposure Time

In the first step we consider the Weibull model (4.4) in which the only predictor is the exposure time. Parameter estimates are shown in Table 4.9 and the predicted models for the prevalence are shown in Figure 4.7. Note that the Weibull models reveal the same patterns as the non parametric isotonic regression models (Barlow *et al.* 1972 and Robertson *et al.* 1988). The force of infection curves are shown in Figure 4.8. The forces of infection for all studies are shown for the whole range of exposure time in the left panel. The right panel shows forces of infection up to 20 years in order to enable a closer look at earlier exposure times. For the second

Table 4.7: *Ever injectors. Estimated odds ratios, confidence intervals and significance pvalues for risk factors using Logistic regression model.*

	Belgium	Czech Rep I	Czech Rep II	Italy	Spain	Sweden
Gender						
males vs females	1.95	1.42	1.45	0.92	1.26	0.93
	(1.00,3.81)	(0.69,2.92)	(1.02,2.06)	(0.60,1.41)	(0.80,2.05)	(0.41,2.09)
	0.049	0.341	0.041	0.709	0.347	0.857
Age 1st Injection	0.95	1.03		0.97	0.89	1.00
	(0.92,0.99)	(0.94,1.14)		(0.94,0.99)	(0.85,0.95)	0.97,1.04)
	0.012	0.526		0.032	0.0002	0.906
Sharing Syringes						
ever vs never	4.26	1.57	1.92		2.72	0.78
	(2.24,8.11)	(0.43,5.79)	(1.19,3.09)		(1.64,4.52)	0.35,1.71)
	< .0001	0.498	0.007		0.0001	0.529
Sharing other paraph						
ever vs never			0.86		0.97	3.53
			(0.55,1.35)		(0.60,1.56)	(1.56,8.01)
			0.508		0.885	0.003
Frequency of Injection						
more×/day vs <1/month			1.997			
			(1.12,3.56)			
			0.019			
daily vs <1/month			1.94			
			(1.07,3.5)			
			0.028			
4-6×/week vs <1/month			0.900			
			(0.45,1.78)			
			0.763			
2-3×/month vs <1/month			1.591			
			(0.94,2.69)			
			0.084			
1-4×/month vs <1/month			0.732			
			(0.42,1.28)			
			0.277			
Frequency of Injection						
daily vs less/week					1.81	
					(1.01,3.23)	
					0.047	
1-6days vs less/week					1.25	
					(0.74,2.11)	
					0.401	

Table 4.8: *Recent injectors. Estimated odds ratios, confidence intervals and significance pvalues for risk factors using Logistic regression model.*

	Belgium	Czech Rep I	Czech Rep II	Spain
Gender				
males vs females	1.06	1.30	1.29	1.23
	(0.20 , 5.66)	(0.56 , 3.04)	(0.89 , 1.85)	(0.73 , 2.09)
	0.945	0.544	0.170	0.439
Age 1st Injection	0.94	1.03		0.89
	(0.86 , 1.01)	(0.92 , 1.15)		(0.83 , 0.94)
	0.103	0.615		0.0001
Syringe Sharing				
in past month				
yes vs no	4.19	1.05		1.01
	(0.48 , 36.80)	(0.42 , 2.68)		(0.48 , 2.14)
	0.196	0.912		0.978
Frequency of Injection				
more×/day vs <1/month			2.15	
			(1.06 , 4.36)	
			0.034	
daily vs <1/month			2.14	
			(1.05 , 4.38)	
			0.036	
4-6×/week vs <1/month			1.01	
			(0.46 , 2.23)	
			0.972	
2-3×/month vs <1/month			1.65	
			(0.85 , 3.21)	
			0.138	
1-4×/month vs <1/month			0.82	
			(0.41 , 1.63)	
			0.565	
Frequency of Injection				
daily vs less/week				1.93
				(0.99 , 3.73)
				0.053
1-6days vs less/week				1.10
				(0.61 , 2.00)
				0.748

Czech study, Belgium, Italy, Spain and Sweden, the force of infection is high in the beginning of the injecting career and decreases for IDUs with relatively long career. The opposite pattern is revealed for the first Czech study. However, for the first Czech Republic study, the parameter estimate for the exposure time (β in (4.3)) is equal to 1.1272 with 95% confidence interval of (0.75,1.50) which implies that the hypothesis that $\beta = 1$ can not be rejected. Note that a model for which $\beta = 1$ implies a constant force of infection. In this case the Weibull model (4.4) can be rewritten as a model with complementary log-log link function implying that the prevalence is

$$\pi(t) = 1 - \exp(-\alpha t) \quad \text{and} \quad \ell(t) = \alpha. \tag{4.8}$$

The constant force of infection for the first Czech Republic study is shown in Figure 4.8. Likelihood ratio tests, presented in the last row of Table 4.9, between the model with constant force of infection and the Weibull model indicate that in all other studies the Weibull model is to be preferred.

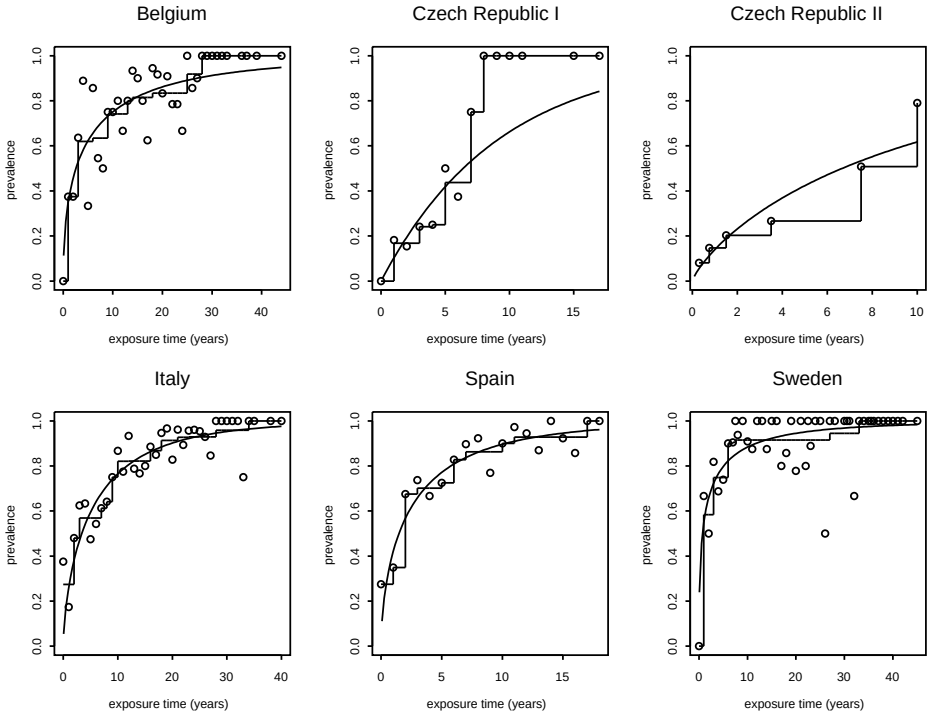

Figure 4.7: *Estimated prevalence of HCV by exposure time obtained from the Weibull model and isotonic regression (step function).*

Including the Risk Factors in the Model

As mentioned in Section 4.1.3 the proportional hazard model (4.6) allows us to include risk factors and to estimate the impact of a specific risk factor on the transmission of HCV among the IDUs population. In the second step of the analysis the basic model, discussed in the previous section, was adjusted for the risk factors introduced in Section 4.1.1, i.e., age at first injection and gender (all studies, except the second Czech study), ever sharing syringes (Belgium, Czech Republic, Spain, and Sweden), ever sharing other paraphernalia (Czech Republic second study, Spain and Sweden) and frequency of injecting in the past month before interview (Czech second study and Spain). The final model for each study is presented in Table 4.9. For all studies, no significant association was found between gender and the prevalence of HCV.

- **Age at First Injection**

  The age at first injection was found to be a non statistically significant risk

Table 4.9: *Ever injectors. Parameter estimates (Est) and standard errors (SE) using Weibull models. The contribution of log(duration) to the Weibull model is evaluated by the change in deviance (likelihood ratio test) with the χ^2 corresponding to the number of lost degrees of freedom. A p-value < 0.05 shows a significant contribution of log(duration) to the Weibull model.*

	Belgium		Czech Rep I		Czech Rep II		Italy		Spain		Sweden	
Basic model	Est (SE)	Pvalue	Est (SE)	Pvalue	Est (SE)	Pvalue	Est (SE)	Pvalue	Est (SE)	Pvalue	Est (SE)	Pvalue
Intercept	-0.85 (0.28)	0.002	-2.26 (0.25)	0.000	-1.89 (0.14)	0.000	-1.20 (0.17)	0.000	-0.61 (0.13)	0.000	-0.23 (0.20)	0.256
log Exposure	0.51 (0.10)	0.000	1.13 (0.19)	0.000	0.80 (0.08)	0.000	0.68 (0.07)	0.000	0.62 (0.07)	0.000	0.43 (0.08)	0.000
	Belgium		Czech Rep I		Czech Rep II		Italy		Spain		Sweden	
Full model	Est (SE)	Pvalue	Est (SE)	Pvalue	Est (SE)	Pvalue	Est (SE)	Pvalue	Est (SE)	Pvalue	Est (SE)	Pvalue
Intercept	-1.33 (0.63)	0.034	-2.96 (1.06)	0.005	-2.09 (0.29)	<.0001	-1.64 (0.32)	0.000	-1.82 (0.59)	0.002	-1.91 (0.56)	0.001
log Exposure	0.43 (0.14)	0.002	1.10 (0.19)	0.000	0.81 (0.09)	<.0001	0.73 (0.07)	0.000	0.71 (0.10)	<.0001	0.65 (0.11)	0.000
gender: m vs f	0.27 (0.17)	0.120	0.28 (0.33)	0.411	0.04 (0.16)	0.797	-0.11 (0.12)	0.342	-0.07 (0.15)	0.637	-0.11 (0.19)	0.581
age 1st Injection	-0.001 (0.01)	0.969	0.02 (0.04)	0.675			0.02 (0.01)	0.036	0.05 (0.02)	0.047	0.04 (0.01)	0.008
Syringe Sharing	0.67 (0.19)	0.001	0.25 (0.62)	0.683	0.54 (0.22)	0.012			0.30 (0.14)	0.034	0.15 (0.21)	0.478
Sharing other paraph					-0.23 (0.20)	0.249			-0.002 (0.14)	0.989	0.71 (0.21)	0.001
Frequency of Injection												
daily vs less/week									0.31 (0.17)	0.067		
1-6days vs less/week									0.02 (0.16)	0.912		
Frequency of Injection												
more×/day vs <1/month					-0.13 (0.25)	0.613						
daily vs <1/month					0.05 (0.26)	0.846						
4-6×/week vs <1/month					-0.26 (0.31)	0.397						
2-3×/month vs <1/month					0.18 (0.23)	0.427						
1-4×/month vs <1/month					-0.47 (0.26)	0.064						
	Belgium		Czech Rep I		Czech Rep II		Italy		Spain		Sweden	
Final model	Est (SE)	Pvalue	Est (SE)	Pvalue	Est (SE)	Pvalue	Est (SE)	Pvalue	Est (SE)	Pvalue	Est (SE)	Pvalue
Intercept	-1.15 (0.34)	0.001	-2.26 (0.25)	0.000	-2.22 (0.21)	0.000	-1.73 (0.31)	<.0001	-1.90 (0.58)	0.001	-1.88 (0.52)	0.000
log Exposure	0.44 (0.12)	0.000	1.13 (0.19)	0.000	0.82 (0.09)	0.000	0.73 (0.07)	0.000	0.70 (0.10)	0.000	0.64 (0.11)	0.000
age 1st Injection							0.02 (0.01)	0.039	0.05 (0.02)	0.040	0.04 (0.01)	0.010
Syringe Sharing	0.63 (0.19)	0.001			0.38 (0.19)	0.046			0.30 (0.13)	0.019		
sharing other paraph											0.80 (0.18)	0.000
Frequency of Injection												
daily vs less/week									0.35 (0.17)	0.038		
1-6days vs less/week									0.05 (0.16)	0.761		
Likelihood ratio Test												
Statistic & p-value	18.79	0.000	0.45	0.502	5.13	0.023	21.01	0.000	26.70	0.000	37.64	0.000

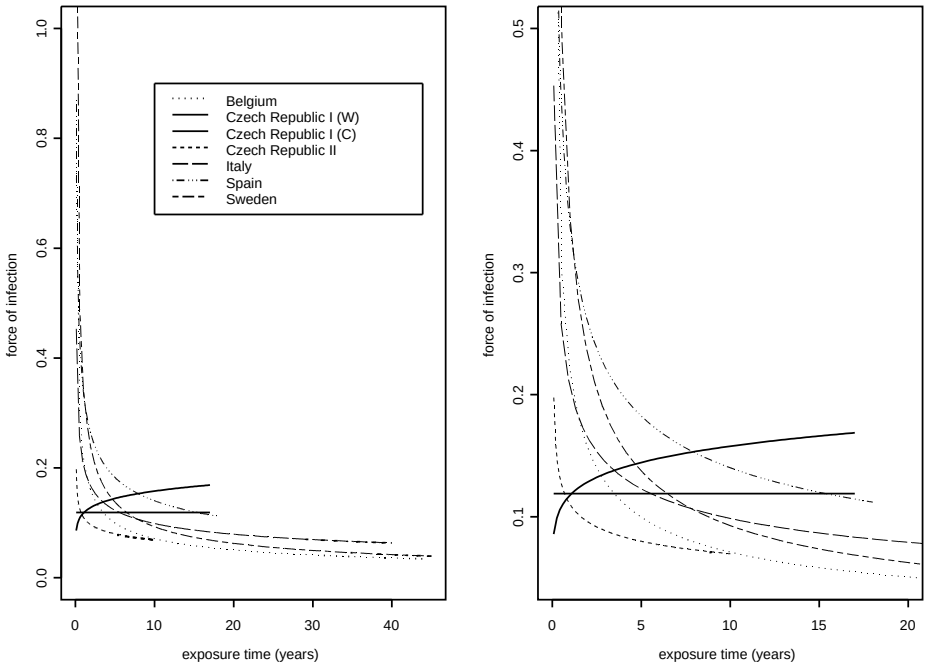

Figure 4.8: *Force of infection by exposure time for all studies. W– Weibull model, C–constant force of infection model. The left panel shows the force of infection against exposure time 0–45 years, the right panel gives the force of infection for 0–20 years of exposure.*

factor in the Belgian and the first Czech study. Parameter estimates for the age at first injection are equal to 0.019 (SE=0.0093), 0.047 (0.023) and 0.0375 (0.0145) for Italy, Spain and Sweden, respectively. This indicates that the force of infection for IDUs who started to inject at relatively older age is higher than the force of infection for IDUs who started to inject at relatively young age.

- **Sharing Syringes**
  Sharing syringes is found to be significant in Spain, Belgium and the second Czech study. For Spain, the parameter estimate is equal to 0.3 (SE=0.1276) which implies that the force of infection for IDUs who ever shared syringes is 1.34 (exp(0.3)) higher than the force of infection of IDUs who never share syringes. In Sweden the force of infection of IDUs who ever shared syringes is

1.16 times higher than the force of infection of IDUs who never shared syringes, although the parameter estimate is found to be non significant. In Belgium, the force of infection of IDUs who ever shared syringes is 1.87 (exp(0.6282)) times higher than the force of infection of IDUs who never shared syringes. In the second Czech study, the force of infection for IDUs who ever shared syringes is 1.45 (exp(0.3752)) times higher than the force of infection for IDUs who never share syringes.

- **Sharing other Paraphernalia**
  Sharing other paraphernalia is found to be a significant risk factor in Sweden. The force of infection for IDUs who share other injecting materials is 2.22 higher than the force of infection of IDUs who do not share other paraphernalia.

- **The Frequency of Injection**
  The frequency of injecting is available for Spain and the second Czech study. It was found to be a significant behavioral risk factor in Spain. No significant difference was found between IDUs who inject 1-6 days per week and those who inject less days per week. However, the force of infection for IDUs who inject on a daily basis is 1.4 times higher (exp(0.3473)) than the force of infection of IDUs who inject less days per week.

4.2.3 Second Analysis: IDUs With Recent Injecting Career

Information about recent drug injection (i.e, injecting drugs in the last month before the interview) is available in the studies from Belgium, the Czech Republic and Spain and allow us to compare the transmission parameters between the recent and the ever injectors (including recent injectors). The number of ever injectors who were recent injectors can be seen in Table 4.1, 97 out of 335 for Belgium for instance. The model parameter estimates for the four studies of recent injectors are given in Table 4.10. The final model, for the four studies, contains the intercept and the duration of injecting, which is a highly significant as similarly observed in the ever injectors. Since information on recent syringe sharing was not available for the second Czech study, its effect cannot be compared with that of the ever injectors. The full model for Spain shows marginal significance of the effects for age at first injection and IDUs who injected drugs on a daily basis relative to those who injected less days per week. The parameter estimate for recent syringe sharing changes sign in the recent analysis although this was non significant. Considering the length of the injecting career, Figure 4.9 shows the estimated models for both prevalence and force

of infection, revealing only minor differences between recent and ever injectors. For Belgium and Spain, we see slight increments of the force of infection for recent relative to the ever injectors at the beginning of the injecting career but hardly any difference between the two as the injecting career gets longer. While the force of infection in the second Czech study shows higher force of infection for recent than for ever injectors at short injecting times, the force of infection is higher for ever than the recent injectors at longer injecting times. The opposite trend is observed for the first czech study.

4.3 Discussion

Injecting drug users are divided into subgroups according to whether they were ever injectors or recent injectors. In this paper we investigated how the transmission of HCV is influenced by the risk behavior in the different subpopulations. We have shown that IDUs who share syringes experience a higher force of infection than IDUs who do not share syringes. Similar patterns were observed for all studies in which information about sharing syringes was available. It is important to mention that even when sharing syringes was found to be not significant (Czech Republic study one and Sweden) the same general pattern was observed. Sharing other paraphernalia (available in Sweden) increase significantly the force of infection as well. As expected, the frequency of injection has an impact of the transmission of HCV among IDUs. As the frequency of injection increases, the risk to be infected increases. The results obtained from the Spanish study indicate that the force of infection for IDUs with high frequency of injection indeed increases.

Table 4.10: *Recent Injectors. Parameter estimates (Est) and standard errors (SE) using Weibull models. A p-value* < 0.05 *shows a significant contribution of the variable to the model.*

	Belgium		Czech Rep I		Czech Rep II		Spain	
Basic model	Est (SE)	Pvalue	Est (SE)	Pvalue	Est (SE)	Pvalue	Est (SE)	Pvalue
Intercept	-0.44 (0.47)	0.345	-2.55 (0.33)	0.000	-1.82 (0.14)	0.000	-0.52 (0.14)	0.000
log Exposure	0.46 (0.17)	0.008	1.25 (0.24)	0.000	0.76 (0.09)	0.000	0.60 (0.08)	0.000

	Belgium		Czech Rep I		Czech Rep II		Spain	
Full model	Estimate (SE)	pvalue	Estimate (SE)	pvalue	Estimate (SE)	pvalue	Estimate (SE)	pvalue
Intercept	-1.15 (1.14)	0.316	-3.19 (1.06)	0.003	-1.84 (0.29)	0.000	-1.65 (0.64)	0.010
log Exposure	0.68 (0.26)	0.008	1.25 (0.25)	0.000	0.74 (0.09)	0.000	0.71 (0.10)	0.000
gender: m vs f	-0.17 (0.44)	0.701	0.29 (0.39)	0.465	-0.01 (0.16)	0.942	-0.02 (0.16)	0.909
age 1st Injection	0.01 (0.03)	0.844	0.02 (0.05)	0.703			0.04 (0.03)	0.091
SyringeSharing(past month)	1.22 (0.68)		0.13 (0.45)	0.779			0.05 (0.17)	0.765
Frequency of Injection								
daily vs less/week							0.37 (0.20)	0.067
1-6days vs less/week							-0.01 (0.19)	0.941
Frequency of Injection								
more×/day vs <1/month					0.03 (0.31)	0.935		
daily vs <1/month					0.22 (0.31)	0.489		
4-6×/week vs <1/month					-0.10 (0.35)	0.785		
2-3×/month vs <1/month					0.30 (0.29)	0.309		
1-4×/month vs <1/month					-0.30 (0.31)	0.332		

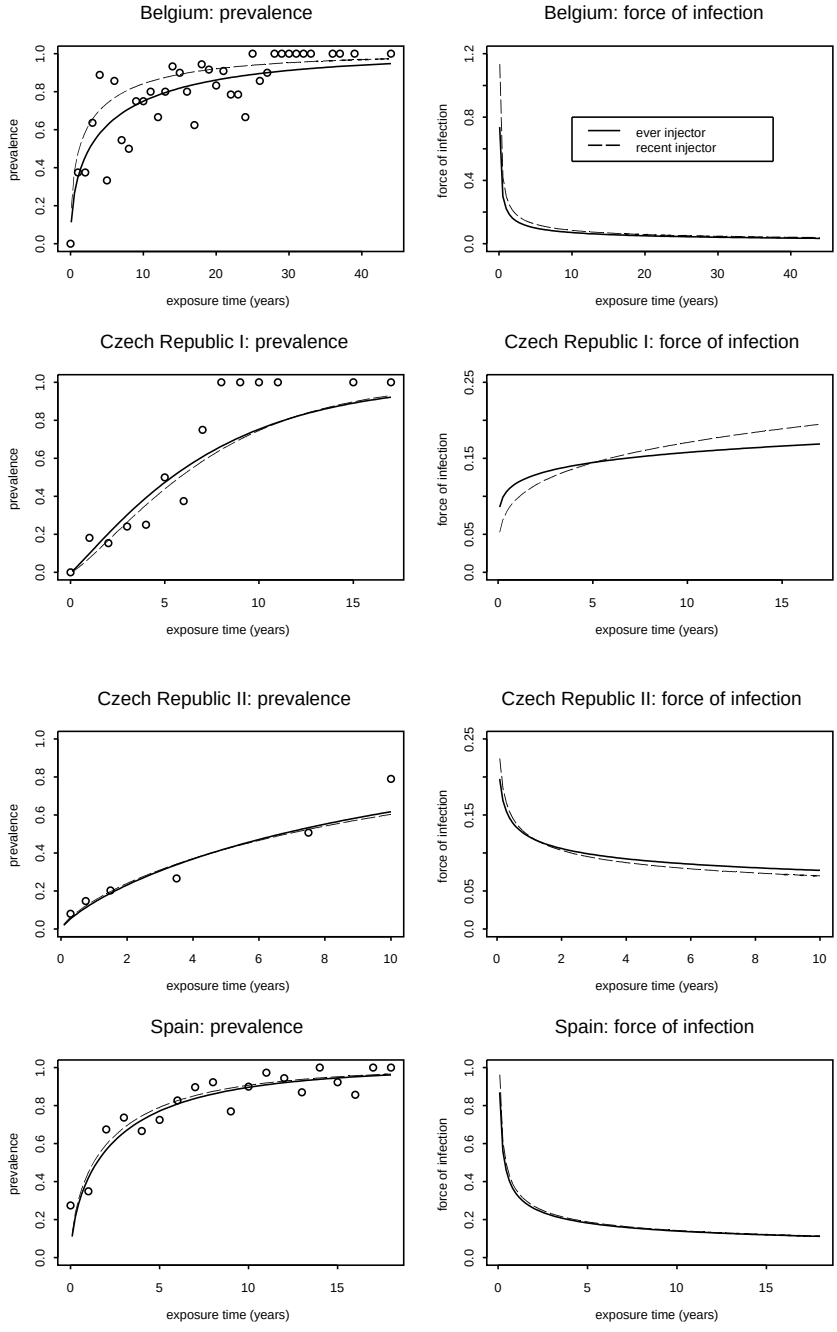

Figure 4.9: *Estimated models for the recent and ever injectors in Belgium (upper panels), the Czech Republic (middle panels) and Spain (lower panels). Left panels: Prevalence of HCV by exposure time, Right Panels: Force of infection by exposure time.*

Part II

Dose-Response Modeling of Food-Borne Infectious Diseases

Chapter 5

Model Averaging in Microbial Risk Assessment Using Modified Fractional Polynomials and Generalized Linear Mixed Models

The data that we have dealt with in the previous part have the common feature that they are transmitted from human to human. However, human exposure to microorganisms such as yeasts, molds, bacteria, protozoa, helminths (worms) through food and/or water consumption dictates potential risk of food-borne infection or illness if such microbes survive in the human system. The dose exposed to is a measurement of microorganisms in relation to their ability to cause infection or illness. The unit of measurement of the dose is the colony forming unit (cfu). Ensuring food safety is directly proportional to a reduction of infectious or toxic food-borne pathogens. This process involves the identification of the microorganisms, how much of the organism an individual is exposed to and the risk associated with a given dosage. Such relations have been studied using various dose-response models to estimate risks in diverse range of problems. It is indeed very important to provide unbiased risk estimates including uncertainty as much as possible, in order to allow stakeholders (policy

makers, producers, consumers) to assess risks as fairly as possible (given the current state of knowledge).

With regard to dichotomous responses, a dose-response model is a function describing the relationship between the dose administered and the probability of infection or disease. In general, any monotonic function which is bounded by zero and one is a possible function for such dose-response relationship. In the literature, several dose-response models have been proposed and they can be categorized into mechanistically and empirically oriented models. For example, Haas *et al.* (1999) developed a Beta-Poisson (BP) model from a mechanistic viewpoint where the biological process is readily taken into account. This model, owing to its biological plausibility, has received much attention in many pathogen dose-response studies (Teunis *et al.* 1996). It should be noted however that several assumptions are made in the biological process which can be questioned. Alternatively, so-called empirical models such as the log-logistic (LL), the log-normal (LN) and the extreme value (EV) models have also been used (Haas *et al.* 1999). Whichever origin the models may have, the conventional way to making inferences has largely been dependent on a single chosen model based on some model selection criterion. Then, one typically proceeds as though that was the only model and thus discards the other possible models and model uncertainty. Using a single selected model ignores variation that arises from other competing models and as a result leads to too small standard errors and narrow confidence intervals which are unrealistically optimistic (Burnham and Anderson, 2002).

Instead of proceeding with one single model, one could argue that each model is a possible candidate dose-response model. Buckland *et al.* (1997) proposed a way of incorporating the uncertainity that arises from other competitive models by model averaging using Akaike Information Criterion (AIC) weights. The better a model, according to a certain selection criterion, the larger the weight given to this model. This approach is further discussed by Burnham and Anderson (2002). In the discipline of microbial risk assessment, model averaging has been employed. Bailer *et al.* (2005) accounted for model uncertainty for experimental studies of quantal responses using a Bayesian approach and weighted the models using the Bayesian Information Criterion (BIC). In the frequentist approach model averaging has been used by Kang *et al.* (2000) to estimate microbial risk using AIC and by Moon *et al.* (2005) to estimate effective microbial doses for infection and illness using Kullback Information Criterion (KIC). In both papers, a mix of the four aforementioned mechanistically and empirically oriented models is considered. In this chapter, a further extension of dose-response modeling is investigated.

It is clear that a flexible dose-response model should be used to describe the data

well. Typically, in the empirically oriented models, a linear dose trend is assumed, which might be too restrictive in the setting of microbial risk assessment. Therefore, we propose to extend the dose-response models by considering fractional polynomials within the family of empirical models. While Royston and Altman (1994) proposed the fractional polynomials as a pre-defined set of (generalized) linear models, we propose a modified set according to a biological rationale. To account for the model-uncertainty in the extended set of possible dose-response models, we estimate the risk to *Salmonella typhi* and to *Campylobacter jejuni* data sets by model averaging.
The rest of this chapter proceeds with a set of plausible models considered in Section 5.1, a review of the model averaging approach in Section 5.2 followed by an application of the method to human volunteer data sets on *Salmonella typhi* and *Campylobacter jejuni* in Section 5.3 and simulation studies in Section 5.4. In Section 5.5 and 5.6 the model averaging approach is applied to several strains data on *Campylobacter jejuni* infection in chicken. A discussion and concluding remarks in Section 5.7 wind up the chapter. The contribution of this chapter to the application of model averaging on single-strain data can be found in Namata *et al.* (2008b).

5.1 Microbial Dose-Response Models

The data for microbial risk assessment can be obtained experimentally where dose groups are known or as a result of outbreaks in which the exact ingested dose is not known but can be approximated. The extracted information to study, for instance, dichotomous dose-response relations involves a total number n_i of individuals or animals to which a particular dose d_i of microbes was administered and out of these it is observed that X_i become infected or ill. To translate this process statistically, X_i is assumed to follow a binomial distribution with parameters n_i and $\pi(d_i)$, the latter being the probability that a subject becomes infected (or ill). Dose-response models refer to models for $\pi(d)$ as a function of d.
Different dose-response models with one, two, and three parameters have been proposed and studied in microbial risk assessment literature (Kodell *et al.* 2002; Moon *et al.* 2005). The inclusion of biological processes gave rise to so-called mechanistic models of which the Beta-Poisson (BP) model is the most popular and extensively used one (e.g., Haas *et al.* 1999). However, the adequacy of the BP model as automatic "default" model has been questioned (Marks *et al.* 1998). Alternative models such as the so-called log-normal (LN), log-logistic (LL), and the extreme value (EV) model have been suggested (see e.g. Pinsky, 2000). These latter models are standard

generalized linear models (GLM), used in statistics to analyse binary response data. At first sight, they however seem to lack any biological interpretation.
The purpose of this section is to indicate that many models, such as GLM's, could have a biological interpretation, similar to that of the BP model. A basic concept in applied statistics states that the true or correct model is never known (or does even not exist) and that all dose-response models are wrong; they are merely approximations (e.g., Burnham and Anderson, 2002). So, the best one can do is to use a good approximating model. Now, depending on the setting and the willingness to rely on assumptions or on data or on both, there are three options: i) to select the model in advance, possibly prior to data collection, only based on biological assumptions (e.g., the BP model); ii) to select a final model from a set of candidate models; and iii) to use as a final model a weighted average over all or some candidate models. For option ii) and iii) this set of candidate models may be (partly) inspired or based on biological knowledge and assumptions. In this chapter we opt for the model averaging approach, since it incorporates explicitly the process of model selection and thus reflects model uncertainty. A crucial aspect of both options ii) and iii) is the use of a well-defined and rich enough set of candidate models. For that purpose, we consider the family of fractional polynomials (FP) (Royston and Altman, 1994) and propose a modified version of FP's, such that they obey some basic biological constraints. These models together with the BP, LN, LL, and EV model will define our set of candidate models.

5.1.1 A Generic Mechanistic Dose-Response Model

So-called mechanistic dose-response models reflect underlying biological processes involved in the kinetics of microorganisms in the body of a human or animal host, in order to determine the proportion that develop an adverse event, owing to exposure to a source containing infectious microorganisms (Haas *et al.* 1999). The formulation of such models involves different subprocesses.
First of all, let f_1 be the probability of ingesting j organisms by an individual, from an exposure source of mean dose d. The second subprocess generates the event that k out of j organisms survive to initiate infection. Let f_2 be the probability of such an event, where r is the probability of survival of a single pathogenic organism in a human host. The total probability that k organisms survive to initiate infectious foci is then given by

$$f(k|r,d) = \sum_{j=k}^{\infty} f_1(j|d) f_2(k|j,r). \tag{5.1}$$

There is uncertainty about the minimal number of surviving organisms, k_{min}, that

are needed to initiate infection. For $k \geq 1$, denote by $f_{4,k}$ the probability that k surviving organisms initiate infection. The probability of infection, for a given mean dose d, can then be written as

$$\pi_I(d) = \sum_{k=1}^{\infty} \left(\sum_{j=k}^{\infty} f_1(j|d) f_2(k|j,r) \right) f_{4,k}$$

which, when $f_{4,k}$ equals 1 for $k \geq k_{min}$ and $f_{4,k}$ is 0 for $k < k_{min}$, simplifies to

$$\pi_I(d) = \sum_{k=k_{min}}^{\infty} \sum_{j=k}^{\infty} f_1(j|d) f_2(k|j,r). \tag{5.2}$$

One can expect different sources of heterogeneity in this approach. The variation in r between hosts and/or between pathogenic organisms can be described by a density $f_3(r)$ over the interval [0,1], leading to the marginal distribution

$$\pi_I(d) = \int_0^1 \left(\sum_{k=k_{min}}^{\infty} \sum_{j=k}^{\infty} f_1(j|d) f_2(k|j,r) \right) f_3(r) dr. \tag{5.3}$$

Model (5.3) can be considered as a generic mechanistic dose-response model. Different choices for f_1, f_2, f_3 and numerical values of k_{min}, and further assumptions, lead to different specific dose-response models. For example, taking f_1 to follow a Poisson distribution with mean d, a binomial distribution for f_2, a degenerate distribution for f_3 (so taking r as fixed), and assuming all parameters in each of these components are different (no shared parameters), we get the dose-response model

$$\pi_I(d) = \Gamma_{k_{min}}(rd) \tag{5.4}$$

where $\Gamma_{k_{min}}(rd)$ is the cumulative distribution of an incomplete gamma distribution with parameter k_{min}, evaluated at rd. Taking $k_{min} = 1$ leads to the exponential model

$$\pi_I(d) = 1 - \exp(-rd). \tag{5.5}$$

In some situations, choices and assumptions can be verified separately, such as the Poisson assumption in dose verification studies (see e.g. DuPont *et al.* 1995).
Choosing f_1 and f_2 and $k_{min} = 1$ as above, but now with f_3 a beta density, the dose-response relation expressed as a complement of the confluent hypergeometric function becomes

$$\pi_I(d) = 1 - {}_1F_1(\alpha, \alpha + \beta, -d) \tag{5.6}$$

which Furumoto and Mickey (1967) approximated to the popular BP model

$$\pi_I(d) = 1 - \left(1 + \frac{d}{\beta}\right)^{-\alpha} \tag{5.7}$$

provided $\beta \gg \alpha$. Teunis and Havelaar (2000) demonstrate that the BP model can produce results similar to the hypergeometric relation, as long as the conditions given by Furumoto and Mickey (1967) are fulfilled. But it is also shown that the BP model might lead to completely different results, which can even be misleading in case very little information is available. See also Teunis *et al.* (2004).

From the general mechanistic model (5.3) many other models can be derived. Choosing f_1 as a Poisson random variable with mean $\mu(d)$, $k_{min} = 1$, f_3 as the point mass distribution $I_{[r=1]}$, and f_2 as the point mass distribution $I_{[k=j]}$, we can recover the classical models and the fractional polynomials. Taking, respectively, $\mu(d) = \exp(\alpha + \beta \log d)$, $\mu(d) = \ln(1 + \exp(\alpha + \beta \log d))$ and $\mu(d) = \ln(1/(1 - \Phi(\alpha + \beta \log d)))$ leads to the EV, LL and LN models with Φ denoting the standard normal cumulative distribution function. In a similar way, by replacing the linear predictor in $\mu(d)$ the fractional polynomial models introduced in Section 5.1.2 can be reconstructed. Haas *et al.* (1999) and Kodell *et al.* (2002) motivated the biological plausibility of the LL, LN and EV model from a completely different angle, namely from their respective latent tolerance distribution.

We would like to emphasize that many other models can be derived from equation (5.3) and they can equally be useful in microbial risk assessment. The Poisson model and the binomial model for components f_1 and f_2 are based on an intrinsic assumption of independence and homogeneity. One could as well work with a zero-inflated Poisson, an overdispersed Poisson or even Poisson mixtures for f_1, and a similar story holds for f_2. Obviously many other distributions on [0,1] can replace f_3. Likewise, k_{min} can be a probability distribution as long as they are monotone increasing (since biologically that probability can not decrease as a function of k). In fact any cumulative distribution function on the integers, having value 0 at $k = 0$, can play the role. Many choices would however not lead to analytically tractable formula's. But any possible model, whatever choice of different distributions in (5.3), has some fundamental properties in common: no infection in the case when no pathogenic organisms are ingested; the more organisms ingested, the higher the probability of infection; and an extremely high dose exposure always results in infection; or equivalently, assuming the model is a differentiable function of dose d with derivative $\pi_I'(d)$,

$$\lim_{d \to 0} \pi_I(d) = 0, \tag{5.8}$$

$$\pi_I'(d) \geq 0, \tag{5.9}$$

$$\lim_{d\to\infty} \pi_I(d) = 1. \tag{5.10}$$

Properties (5.8) to (5.10) are exactly the properties of a cumulative distribution function on $[0, \infty)$, and therefore they can always be written as a GLM (such as the exponential) or a generalized nonlinear model (such as the approximate BP model, with identity link and predictor $\left(1 + \frac{d}{\beta}\right)^{-\alpha}$. Other examples include again the LL, LN, EV and the fractional polynomial models (Section 5.1.2). In Section 5.3 we illustrate that, depending on the application, additional constraints can or should be added to the minimal set of properties (5.8) to (5.10).

A crucial concept in the theory of multimodel inference is the fact that a "correct model" does not exist. Any model is incorrect and merely tries to approximate the true process. It is rather a matter of selecting a good approximating model or, in the approach of multimodel inference, to average over a certain subset of good approximating models, assigning higher weights to better approximating models. In this philosophy it is obviously important to define a rich enough family of candidate models. Here, as indicated in Section 5.1.2, the family of fractional polynomials have shown to be a well-defined and rich family of models. All models considered in this chapter as candidate models are listed in Table 5.1. They are the typical models used in the literature, the aforementioned BP, LL, LN, EV, extended with a new family of fractional polynomial models, as introduced in Section 5.1.2.

Infection versus Illness

Often the data refer to illness rather than to the infection status of the individual. This is the case for the *Salmonella typhi* data (Hornick *et al.* 1970) in Table 5.2. These are experimental data on healthy adult volunteers, not exposed previously. Infection was not reported separately, only illness, which was described as developing fever (higher than 103 ^{o}F) followed by headaches and abdominal pain. Nevertheless the BP model is typically used for analyzing such illness data. But since

$$P(\text{illness}|d) = P(\text{illness}|\text{infection}, d)\pi_I(d),$$

the BP model loses its direct biological interpretability for estimating the illness probability $P(\text{illness}|d)$. Moreover in most cases no data are available on $P(\text{illness}|\text{infection}, d)$ and little is known in order to build a biologically meaningful parametric model. In most cases however, one can assume the probability $P(\text{illness}|d)$ to share the same fundamental properties or constraints (5.8), (5.9) and (5.10), namely being monotone

increasing from 0 to 1, when dose d varies from 0 to ∞ (see e.g. Teunis *et al.* 1999). This again, even more strongly, motivates the point of view that many statistical models are biologically plausible.

5.1.2 Fractional Polynomials

Other more flexible models like the fractional polynomials can be competitors of the BP, LN, LL and EV commonly used models in microbial risk assessment. Unlike conventional polynomials that take on positive integer powers (up to the degree being considered), FP's use powers from a predefined set, $\mathcal{P} = \{-2, -1, -0.5, 0, 0.5, 1, 2, ..., \max(3, m)\}$, with m the degree of the FP (see below). In principle, other fractional (negative and positive) powers can be considered, but Royston and Altman (1994) illustrated that the above restricted set is sufficient for most practical purposes. The family of fractional polynomials has been shown to be useful in several other, somewhat related fields of application, see e.g. Faes *et al.* (2003, 2006ab), Shkedy *et al.* (2006), Hens *et al.* (2007).

Fractional polynomials of degree m with powers $p_1 \leq p_2 \leq \cdots \leq p_m$, and in the GLM framework for binary response data, are defined as

$$g(\pi(d)) = \beta_0 + \sum_{j=1}^{m} \beta_j H_j(d), \tag{5.11}$$

where for $j = 1, \ldots, m$,

$$\mathrm{H}_j(d) = \begin{cases} d^{p_j} & \text{if } p_j \neq p_{j-1}, \\ H_{j-1}(d)\log(d) & \text{if } p_j = p_{j-1} \end{cases}$$

with $p_0 = 0$ and $H_0(d) \equiv 1$. In (5.11), π denotes the probability on the adverse effect of interest (infection, illness) and g some link function (such as the logit or probit link). As shown by Royston and Altman (1994), FP's of degree m higher than 2 are rarely needed in practice.

The FP model (5.11) does not automatically satisfy properties (5.8) to (5.10). First of all, first order FP's seem to be less appropriate as a family of candidate models. For example, the model $g(\pi(d)) = \beta_0 + \beta_1 d^{p_1}$ with $p_1 > 0$, or with $p_1 < 0$ and $\beta_1 > 0$, does not obey property (5.8). Property (5.10) does not hold for $p_1 > 0$ and $\beta_1 < 0$, or for $p_1 < 0$. Similar issues for the model $g(\pi(d)) = \beta_0 + \beta_1 \log(d)^{p_1}$. Since no unambiguous or clear and consistent constraints on parameters β_k and powers p_1 guarantee properties (5.8) to (5.10), we do not further consider modifications of first degree FP's. Second degree FP's however can be modified to satisfy the fundamental

Table 5.1: *Set of candidate dose-response models: BP (approximate Beta-Poisson), LL (log-logistic), LN (log-normal), EV (extreme-value), FPL (FP with logit link), FPN (FP with probit link), FPEV (FP with complementary log-log link).* Φ *denotes the standard normal cumulative distribution function.* $\pi(d)$ *refers to the illness or infection probability. The column Subset* $\mathcal{P}$ *shows the admissible values for the powers* p_1 *and* p_2 *from the set* $\mathcal{P}$*, defining the FP candidate models.*

$\pi(d)$	Parameters	Subset $\mathcal{P}$	Model
$1-(1+d/\beta)^{-\alpha}$	$\alpha>0,\beta>0$		BP
$1/(1+\exp[-(\alpha+\beta\log(d))])$	$\alpha<0,\beta>0$		LL
$\Phi(\alpha+\beta\log(d))$	$\alpha<0,\beta>0$		LN
$1-\exp[(-\exp(\alpha+\beta\log(d)))]$	$\alpha<0,\beta>0$		EV
$1/(1+\exp[-(\beta_1(\log(d+1))^{p_1}+\beta_2(\log(d+1))^{p_2})])$	$\beta_1<0,\beta_2>0$	$p_1<0,p_2>0$	FPL
$\Phi(\beta_1(\log(d+1))^{p_1}+\beta_2(\log(d+1))^{p_2})$	$\beta_1<0,\beta_2>0$	$p_1<0,p_2>0$	FPN
$1-\exp[(-\exp(\beta_1(\log(d+1))^{p_1}+\beta_2(\log(d+1))^{p_2}))]$	$\beta_1<0,\beta_2>0$	$p_1<0,p_2>0$	FPEV

properties as follows

$$g(\pi(d))=\beta_1(\log(d+1))^{p_1}+\beta_2(\log(d+1))^{p_2} \quad \text{with} \quad \beta_1,p_1<0 \text{ and } \beta_2,p_2>0 \tag{5.12}$$

for a given link function g. So, as compared to the original definition of FP's, there is no intercept, d is replaced by $\log(d+1)$ and coefficients and powers of both terms have to be opposite in sign. The first term on the rhs of (5.12) guarantees property (5.8), while the second one guarantees property (5.10), and both terms are automatically monotone. We will use the typical GLM links: the logit, the probit and the complementary log-log link. In the applications and the simulations in the next sections, model (5.12) will be fitted by constrained maximum likelihood, to ensure that $\beta_1,p_1<0$ and $\beta_2,p_2>0$.

An overview of all candidate models considered in this chapter and the functions to derive them are displayed in Table 5.1.2 (for all admissible powers in the set $\mathcal{P}$). We include the "classical" models (Beta-Poisson, log-logistic, log-normal, and extreme-value) and the family (5.12) of modified fractional polynomials of degree 2 with logit, probit and complementary log-log link. Note that all models have the same degree of complexity (two parameters). This set of candidate models contains a total of $M=40$ models: 4 classical models (BP, LL, LN, EV) and three times 12 FP models (for three different link functions: 3 negative powers, each combined with 4 positive powers).

The SAS procedure NLMIXED has been used to fit all models. To improve computational stability, the BP model is reparameterized in terms of mean and variance related parameters, using

$$\frac{e^u}{1+e^u} = \frac{\alpha}{\alpha+\beta} \quad \text{and} \quad e^v = \alpha+\beta. \tag{5.13}$$

The approach used to calculate confidence intervals for the predicted probabilities was to calculate the confidence intervals on the scale of the underlying continuous variable (i.e., on the log-odds scale or the probit scale or the cloglog scale) and the resulting upper and lower confidence bounds are then converted into probabilities using the logistic distribution, the cumulative distribution function of the standard normal distribution or a Gompertz distribution. For the BP model, which does not use any of the three link functions, the confidence intervals for the estimated parameters are obtained using a logit-based transformation (shown in the next section).

5.2 Model Averaging Approach

This section recapitulates the necessary formulae for model averaging according to Burnham and Anderson (2002). The set of M plausible candidate models presented in Section 5.1 is fit to the data and the probability of infection (or illness) $\pi(d)$ at some low dose d is estimated for each of the models. Let us denote the estimate for, or better conditional on, model m as $\hat{\pi}_m(d)$. The 'unconditional' estimate for the probability of infection is defined as (suppressing the specification of the dose level d)

$$\hat{\pi}_a = \sum_{m=1}^{M} w_m \hat{\pi}_m,$$

with weights

$$w_m = \frac{\exp(-\frac{1}{2}\triangle_m)}{\sum_{h=1}^{M}\exp(-\frac{1}{2}\triangle_h)}.$$

So $\hat{\pi}_a$ is a weighted average of the conditional estimates with weights quantifying the relative importance of the different conditional models. These weights are based on the AIC differences $\triangle_m$, defined as the difference between the AIC value of model m and the AIC value of the 'best' model with the lowest AIC value.
From Buckland *et al.* (1997) the estimated variance around the averaged risk estimate is taken as:

$$\widehat{\text{var}}(\hat{\pi}_a) = \left[\sum_{m=1}^{M} w_m \sqrt{\widehat{\text{var}}(\hat{\pi}_m) + (\hat{\pi}_m - \hat{\pi}_a)^2}\right]^2.$$

This variance estimator is clearly the sum of two components: the conditional sampling variance $\widehat{\text{var}}(\hat{\pi}_m)$ of $\hat{\pi}_m$ given model m and a term for the variation in the estimates across the M models $(\hat{\pi}_m - \hat{\pi}_a)^2$. The square root of this sum is then weighted by the weights w_m. This formulation is useful since it accounts for both the within and between model variability in estimating the variance of the averaged risk estimate.

In order to guarantee that the confidence interval for $\pi(d)$ is part of the eligible (0,1) range, we first construct a confidence interval $[l, u]$ for the logit transformed parameter $\text{logit}(\pi(d)) = \log(\pi(d)/(1-\pi(d)))$ and next transform it back to the probability scale using the expit transformation, leading to $[e^l/(1+e^l), e^u/(1+e^u)]$. Since both transformations forth and back are monotone one-to-one, the coverage probability remains exactly the same (see also Burnham *et al.* 1987). Using the delta-method to calculate the standard error of $\text{logit}(\hat{\pi}_a)$ for the construction of the interval $[l, u]$, the final large sample confidence interval for $\pi(d)$ can be written as

$$\left[\frac{\hat{\pi}_a}{\hat{\pi}_a + (1-\hat{\pi}_a)C}, \frac{\hat{\pi}_a}{\hat{\pi}_a + (1-\hat{\pi}_a)/C}\right] \quad \text{with} \quad C = \exp\left[\frac{z_{\alpha/2}\widehat{\text{SE}}(\hat{\pi}_a)}{\hat{\pi}_a(1-\hat{\pi}_a)}\right].$$

5.3 Application to Single Strain Data

In this section, we illustrate the use of the 36 modified fractional polynomials (5.12) along with the four classical models and their model average in order to estimate the risk of illness or infection. In a first study, we use data from Hornick *et al.* (1970) where the volunteers ingested wild-type *Salmonella typhi* in 45 mL of milk. The data are presented in Table 5.2. It includes the dose of pathogenic *Salmonella typhi*, the total number of individuals exposed and the individuals who eventually became ill. Thus, in the first study we investigate the probability of illness due to *Salmonella typhi* in relation to the ingested dose. A second study, investigates the risk for infection due to *Campylobacter jejuni* (Black *et al.* 1988), with the data presented in Table 5.3. It includes the dose of *Campylobacter jejuni*, the total number of individuals exposed, the individuals infected and the individuals who eventually became ill. Both data sets are also discussed in Teunis *et al.* (1996).

Table 5.2: *Results of the dose-response experiment (Hornick* et al. *1970) for Salmonella typhi Quailes in healthy human subjects. Dose: number of organisms ingested. Total: number of subjects at a given dose. Ill: number of subjects with symptoms of typhoid fever.*

Dose (cfu)	Total	Ill
10^3	14	0
10^5	116	32
10^7	32	16
10^8	9	8
10^9	42	40

Table 5.3: *Results of the dose-response experiment (Black* et al. *1988) for Campylobacter jejuni in healthy volunteers. Dose: ingested number of* C. jejuni *A3249. Total: number of subjects exposed to a given dose. Infected: number of subjects infected (excretion of* C. jejuni*). Ill: number of subjects with gastro-enteric symptoms (fever, vomiting, diarrhea).*

Dose (cfu)	Total	Infected	Ill
8×10^2	10	5	1
8×10^3	10	6	1
9×10^4	13	11	6
8×10^5	11	8	1
1×10^6	19	15	2
1×10^8	5	5	0
1×10^8	4	4	2

5.3.1 Salmonella Typhi

Table 5.4 shows, for *Salmonella typhi*, the estimated risks at a dose of 100cfu, their standard errors and 95% confidence intervals for the 40 fitted models, ordered from best to least fitting according to AIC goodness-of-fit criterion. The estimated probability of illness due to *Salmonella typhi* ranges from 7.96×10^{-22} to 0.07407. The question is, on which model should inference be based. Using the Akaike information criterion, the relative importance of each model is calculated. One way is to use the estimates from a model that has AIC-weight greater or equal to 0.9 (Haas *et al.* 1999). However, none of our models meets that criterion and it is unlikely in reality that such a model would be found. For this data example, the 5 best fitting FPs have somewhat higher weights but still far below 0.9. Instead of selecting one final model, model averaging based on all available models (or a selection) can be used. The averaged risk estimate is shown on the last line of Table 5.4 together with its standard errors and 95% Wald confidence intervals. The confidence intervals are wider than for individual models but this is expected because model averaging incorporates variability between competing models. This uncertainty indicates the importance of model averaging, especially at low doses where we do not have data. This uncertainty is also clearly visualised in Figure 5.1. The curves are shown for the BP, LL, LN and EV models as well as the five best fitting FPs for reasons of clarity of the figure.

Table 5.4: *Salmonella typhi. Estimated probability of illness at dose 100cfu. For the Beta-Poisson and the model averaged(MA) estimates logit-back transformed CIs are used while the other model estimates use CIs back transformed according to their corresponding link function. Fractional polynomials are denoted as FPlink$_{(powers)}$.*

		Model			wald CIs	
Model	AIC	Weights	$\hat{\pi}(100)$	SE($\hat{\pi}(100)$)	lower	upper
FPEV$_{(-2,3)}$	20.4273	0.07474	0.00010	0.00012	1.04E-05	0.00103
FPN$_{(-2,3)}$	20.9307	0.05811	2.13E-09	1.09E-08	2.41E-14	1.27E-05
FPEV$_{(-1,3)}$	21.1171	0.05294	0.01986	0.00935	0.00787	0.04969
FPN$_{(-1,3)}$	21.1683	0.05160	0.00489	0.00492	0.00055	0.02851
FPEV$_{(-1,2)}$	21.2208	0.05026	0.00964	0.00532	0.00326	0.02829
FPEV$_{(-2,2)}$	21.6446	0.04067	2.87E-05	3.77E-05	2.18E-06	0.00038
FPL$_{(-2,3)}$	21.7401	0.03877	5.96E-05	8.70E-05	3.41E-06	0.00104
FPEV$_{(-0.5,2)}$	21.7845	0.03792	0.04254	0.01531	0.02091	0.08553
FPEV$_{(-0.5,1)}$	21.8671	0.03638	0.01506	0.00735	0.00577	0.03901
FPL$_{(-1,3)}$	21.9483	0.03494	0.01373	0.00825	0.00420	0.04394
FPN$_{(-1,2)}$	21.9491	0.03492	0.00058	0.00086	2.34E-05	0.00766
FPN$_{(-0.5,3)}$	22.0374	0.03341	0.04247	0.02072	0.01494	0.10136
FPEV$_{(-1,1)}$	22.0668	0.03293	0.00262	0.00187	0.00065	0.01055
EV	22.0954	0.03246	0.04779	0.01626	0.02442	0.09242
FPN$_{(-0.5,2)}$	22.1012	0.03237	0.01304	0.00948	0.00273	0.04729
FPEV$_{(-0.5,0.5)}$	22.2591	0.02991	0.00648	0.00386	0.00201	0.02079
FPEV$_{(-0.5,3)}$	22.4418	0.02730	0.07407	0.02189	0.04127	0.13108
FPL$_{(-0.5,3)}$	22.6063	0.02514	0.05436	0.02073	0.02541	0.11247
LN	22.7896	0.02294	0.01159	0.00852	0.00239	0.04275
FPEV$_{(-1,0.5)}$	22.8050	0.02276	0.00094	0.00079	0.00018	0.00492
FPN$_{(-2,2)}$	22.8445	0.02232	2.94E-13	2.16E-12	2.61E-20	7.74E-08
FPL$_{(-1,2)}$	22.9265	0.02142	0.00469	0.00344	0.00111	0.01956
FPL$_{(-0.5,2)}$	23.0040	0.02061	0.02479	0.01204	0.00948	0.06322
FPN$_{(-0.5,1)}$	23.0333	0.02031	0.00083	0.00112	0.00004	0.00890

Continued on next page

Table 5.4 – *continued from previous page*

		Model			wald CIs	
Model	AIC	Weights	$\hat{\pi}(100)$	SE($\hat{\pi}(100)$)	lower	upper
$FPEV_{(-2,1)}$	23.5374	0.01578	3.25E-06	5.16E-06	1.44E-07	7.31E-05
$FPN_{(-1,1)}$	23.6367	0.01502	4.17E-06	1.12E-05	1.16E-08	0.00044
$FPL_{(-2,2)}$	23.6483	0.01493	7.17E-06	1.22E-05	2.55E-07	0.00020
LL	23.7942	0.01388	0.02329	0.01128	0.00894	0.05930
$FPN_{(-0.5,0.5)}$	23.8865	0.01326	5.01E-05	0.00010	5.77E-07	0.00177
$FPL_{(-0.5,1)}$	24.0609	0.01215	0.00564	0.00389	0.00146	0.02160
$FPL_{(-1,1)}$	24.5806	0.00937	0.00066	0.00066	0.00009	0.00465
$FPEV_{(-2,0.5)}$	24.7196	0.00874	6.30E-07	1.15E-06	1.78E-08	2.23E-05
$FPN_{(-1,0.5)}$	24.8096	0.00836	2.87E-08	1.14E-07	4.49E-12	2.77E-05
$FPL_{(-0.5,0.5)}$	24.8644	0.00813	0.00168	0.00145	0.00031	0.00905
$FPN_{(-2,1)}$	25.4658	0.00602	7.96E-22	9.85E-21	9.87E-34	1.20E-12
$FPL_{(-1,0.5)}$	25.6564	0.00547	0.00014	0.00017	0.00001	0.00151
BP	26.0011	0.00461	0.00070	0.00030	0.00030	0.00162
$FPL_{(-2,1)}$	26.1635	0.00425	1.81E-07	3.95E-07	2.49E-09	1.31E-05
$FPN_{(-2,0.5)}$	26.9687	0.00284	6.38E-30	1.11E-28	1.33E-46	4.46E-17
$FPL_{(-2,0.5)}$	27.5996	0.00207	1.08E-08	2.79E-08	6.71E-11	1.72E-06
MA estimate			0.01285	0.01603	0.00109	0.13404

5.3.2 Campylobacter Jejuni

Applying model averaging over all 40 models on the *Campylobacter jejuni* data example, we obtained an averaged risk estimate, at the dose=10cfu, of 0.21281 with a standard error equal to 0.29038 and confidence intervals from 0.00896 to 0.88986. The five best fitting models extended with the "classical" models BP, LL, LN and EV (if not included in the top 5) are shown in Figure 5.2(a). This figure shows some curves with a peculiar pattern left from the data range. They can be characterised by a steep increase at low dose levels. Based on biological knowledge or expertise such models might be considered as not plausible. Extending the minimal set of criteria

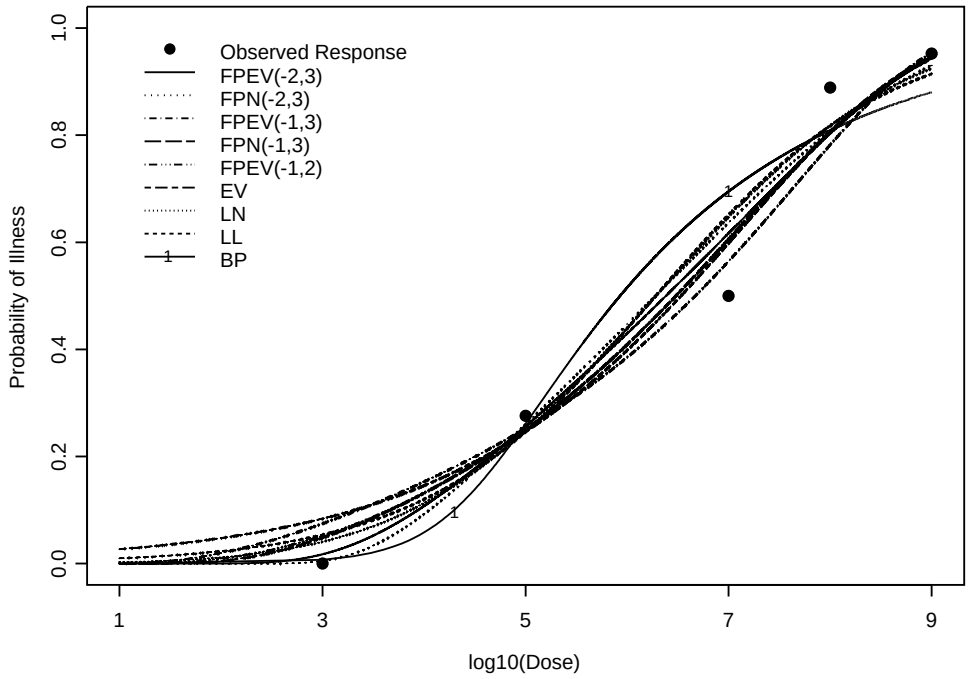

Figure 5.1: *Estimated probabilities of illness for the Salmonella typhi data of Table 5.2.*

(5.8), (5.9) and (5.10), such models might be excluded from the initial set of candidate models as follows.
In addition to (5.8), (5.9) and (5.10), the constraint

$$\pi_I'(d) \leq C, \text{ for } d \text{ in some low dose range,} \tag{5.14}$$

controls the increase of the models in the low dose range by excluding all models whose derivatives exceed a certain threshold C in the range. The choice of C depends on the particular application and should be governed by additional biological expertise. Here we illustrate this idea by applying the threshold $C = 0.2$ on the dose range from 1e-20 to 5 cfu (transformed to log base 10 scale). As a result 14 models are excluded and averaging is restricted to 26 models. Figure 5.2(b) shows the fitted curves of the five best fitting models, extended with the BP, LN, LL and EV model. Since now the EV model is ranked on the fourth position, there is one curve less as compared to Figure 5.2(a).
Table 5.5 shows the estimated risks at a dose of 10cfu, their standard errors and 95% confidence intervals for *Campylobacter jejuni* for the 26 fitted models after the rule-out, arranged from best to least fitting according to AIC. In this example, only small differences in AIC are seen, and all models get about the same weight. The probability of infection due to *Campylobacter jejuni* at a dose of 10cfu ranges from 4.52×10^{-11} to 0.3236. By model averaging we obtain an averaged risk estimate of 0.089 (last row of Table 5.5), which is about two times less than the averaged risk estimate over all 40 models. The confidence interval based on the model average is still very wide [0.00248,0.79301] (averaging over all 40 models led to [0.00896, 0.88986]), but it has to account for the (inevitable) high degree of model uncertainty when extrapolating at low dose levels.

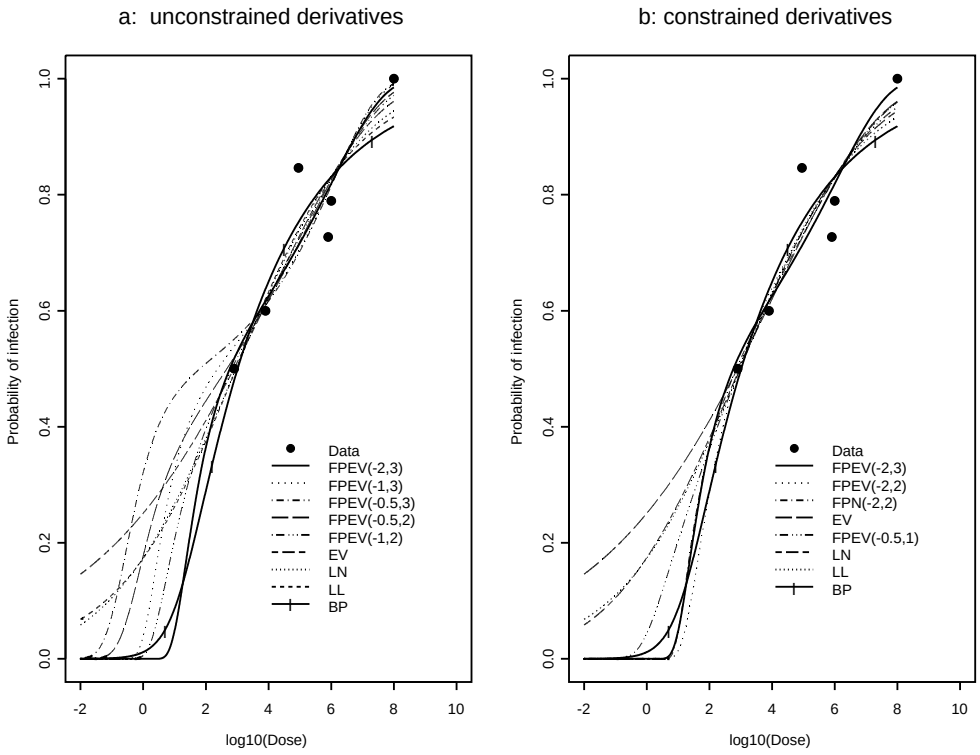

Figure 5.2: *Estimated probabilities of infection for the Campylobacter jejuni data of Table 5.3. Left panel: the fitted curves of the classical models (BP, LL, LN, EV) together with the five best fitting FP's for the full set of 40 candidate models. Right panel: the fitted curves of the five best models together with the classical models (BP, LL, LN) for the reduced set of 26 candidate models. The reduced set is based on the exclusion criterion: derivatives in the low dose region from dose=5cfu and below should not exceed 0.2.*

Table 5.5: *Campylobacter jejuni. Estimated probability of infection at dose 10cfu. For the Beta-Poisson and the model averaged(MA) estimates logit-back transformed CIs are used while the other model estimates use CIs back transformed according to their corresponding link function. Fractional polynomials are denoted as* $FPlink_{(powers)}$.

					Wald CIs	
Model	AIC	Weights	$\hat{\pi}(10)$	SE($\hat{\pi}(10)$)	lower	upper
$FPEV_{(-2,3)}$	19.8482	0.05985	0.04804	0.13982	0.00014	1.00000
$FPEV_{(-2,2)}$	20.1956	0.05031	0.01373	0.04392	2.5E-05	0.99951
$FPN_{(-2,2)}$	20.2796	0.04824	0.04627	0.28467	5.1E-14	0.99998
EV	20.3201	0.04727	0.32355	0.14222	0.12735	0.67428
$FPEV_{(-0.5,1)}$	20.4400	0.04452	0.20325	0.15167	0.04303	0.69081
$FPL_{(-2,2)}$	20.4464	0.04438	0.05259	0.23678	5.0E-06	0.99838
LN	20.5414	0.04232	0.26749	0.17085	0.05072	0.65434
$FPEV_{(-1,1)}$	20.5778	0.04156	0.08722	0.11266	0.00642	0.72542
$FPN_{(-0.5,1)}$	20.6301	0.04049	0.15863	0.19185	0.00532	0.71023
LL	20.6902	0.03929	0.26161	0.16551	0.06198	0.65514
$FPEV_{(-0.5,0.5)}$	20.6943	0.03921	0.12564	0.12056	0.01778	0.63389
$FPN_{(-1,1)}$	20.7200	0.03871	0.05069	0.13091	2.1E-05	0.79464
$FPL_{(-0.5,1)}$	20.7747	0.03766	0.15911	0.17442	0.01448	0.70895
$FPL_{(-1,1)}$	20.8612	0.03607	0.06192	0.11950	0.00117	0.78824
$FPEV_{(-2,1)}$	20.8765	0.03579	0.00173	0.00636	1.3E-06	0.90770
BP	20.8785	0.03576	0.08640	0.16325	0.00164	0.84489
$FPN_{(-2,1)}$	20.8927	0.03550	5.99E-6	9.10E-5	8.3E-28	0.98264
$FPEV_{(-1,0.5)}$	20.8967	0.03543	0.04249	0.06614	0.00192	0.62518
$FPN_{(-0.5,0.5)}$	20.8970	0.03543	0.05852	0.11247	0.00028	0.62542
$FPL_{(-0.5,0.5)}$	21.0254	0.03323	0.06983	0.10306	0.00334	0.62726
$FPL_{(-2,1)}$	21.0299	0.03315	0.00072	0.00390	1.9E-08	0.96585
$FPN_{(-1,0.5)}$	21.0638	0.03259	0.00602	0.02502	3.6E-08	0.64223
$FPL_{(-1,0.5)}$	21.1848	0.03068	0.01571	0.03731	0.00014	0.64367

Continued on next page

Table 5.5 – *continued from previous page*

					Wald CIs	
Model	AIC	Weights	$\hat{\pi}(10)$	SE($\hat{\pi}(10)$)	lower	upper
$\text{FPEV}_{(-2,0.5)}$	21.3155	0.02874	0.00036	0.00146	1.2E-07	0.66425
$\text{FPN}_{(-2,0.5)}$	21.3927	0.02765	4.52E-11	1.10E-9	7.9E-43	0.75888
$\text{FPL}_{(-2,0.5)}$	21.5031	0.02617	2.39E-5	0.00014	1.9E-10	0.75422
MA estimate			0.08900	0.15178	0.00248	0.79301

5.4 Simulation Study for Single Strain Data

In a small simulation study we explore and compare the performance of the estimated risk based on three model approaches: each of the 40 candidate models individually, the best fitting model according to AIC (varying from run to run), and by model averaging. The model chosen most often by AIC across the simulations was also monitored.

A total of $S = 1000$ datasets were generated. For each run, we assume the same dose levels (d) and total number of individuals exposed (n) as those in the *Salmonella typhi* data set, and generate the number of ill individuals based on the BP model or based on a fractional polynomial EV model, with parameters based on the estimates from these models fitted on the *Salmonella typhi* data. In the first setting, the data sets were generated using a BP model taking parameters $u = -11.8739$ and $v = 10.2810$ in formulas (5.13). In the second setting, the data sets were generated using the cloglog fractional polynomial with powers $(p_1, p_2) = (-1, 3)$, as one of the best models in the *Salmonella typhi* example, with parameters $\beta_1 = -18.1425$ and $\beta_2 = 22.5300 \times 10^{-5}$. The dose response curves corresponding to both settings are shown in Figure 5.1. To the 1000 generated samples, we fit the set of 40 candidate models described earlier, and estimate the risk to *Salmonella typhi* illness π at dose=100cfu, for each of the candidate models but also keeping track of the best fitting model according to AIC. Finally, an averaged risk over the candidate models is calculated. We summarize the performance of the different methods by reporting, across the simulations, the average variance of the estimated probability of *Salmonella typhi* illness $\overline{(SE(\hat{\pi}_*(100)))^2}$ using model approach $*$, the average length of the 95% confidence intervals and the coverage probability of these intervals. The variability of the risk estimate $\hat{\pi}_*$ about its average risk $\bar{\hat{\pi}}_* = \sum_{s=1}^{S} \hat{\pi}_{*s}/S$ over the S simulations

was calculated as $\hat{\sigma}^2(\hat{\pi}_*(100)) = \sum_{s=1}^{S}\{\hat{\pi}_{*s} - \bar{\hat{\pi}}_*\}^2/(S-1)$, while $bias = \{\bar{\hat{\pi}}_* - \pi\}$ gives the difference between the average risk estimate and the true risk based on the data generating model. The mean-squared error is given by $MSE = \text{bias}^2 + \hat{\sigma}^2(\hat{\pi}_*(100))$.

5.4.1 First Setting

Table 5.6 shows the results of the first setting, which uses the BP model as the true model. Not all 40 models are shown but the BP, LL, LN and EV, and the most frequently selected best fitting model. For this latter model it was observed which model was selected most often by AIC after all runs were finished. The results from the best fitting models (BFM) and by model averaging (MA) are shown as well. Note that the BFM changes from run to run. Estimates based on the BP model have very small bias and variance characteristics, as compared with the other models, and a coverage close to 95%, also illustrated by Figure 5.3. Since this model is the true simulation model, this is not unexpected. But the extremely poor behaviour of most other models is surprisingly low. Further note that the performances of the LL, LN and EV model are quite similar, indicating that the choice of link function has only minor influence on the estimated risk. Figure 5.3 helps to understand these results. The order of the simulation runs on the horizontal scale corresponds to the order of the corresponding point estimates. So runs leading to smaller estimates are more to the left side of the horizontal scale (as shown by the solid curve of estimates). The right upper panel of Figure 5.3 shows that the LN tends to overestimate the true value. The BFM (lower right panel of Figure 5.3) estimates show smaller bias and smaller variances, showing in some way its adaptive nature, but it also tends to underestimate the true value for most of the runs. At the right end of the horizontal scale, it exhibits an extremely high variable pattern. This all combines to a very low coverage of 6.6%. The model which was chosen most often as the best model is the fractional polynomial logit-model with powers (-2,0.5). Also this model performs surprisingly poorly in terms of coverage. The averaged risk estimate, based on all considered models, shows small bias but relatively high variance properties. But this model accounts for the variability introduced by the model selection procedure, resulting in wider confidence intervals and a coverage probability of 87% (see Figure 5.3). Actually this is also a somewhat disappointing result, but compared to all other misspecified models it is outstanding.

Table 5.6: *Simulation I results for estimated risk at dose 100cfu. For each model, columns (2) and (3) show the simulation-average of the variance-estimates of the estimated risk and the average confidence interval length respectively. The simulated coverage probability of the intervals is shown in column(4). Columns (5), (6) and (7) present the simulated squared bias, variance and mean squared error respectively. The direction of the bias shown by the sign beside the squared bias in column (4).* † *denotes the data generation model while* ‡ *shows the most frequently selected best AIC model. BFM denotes the best fitting models and MA denotes model averaging.*

(1)	(2)	(3)	(4)	(5)	(6)	(7)
Model	$\overline{(SE(\hat{\pi_*}))^2}$	$\overline{\text{CIlength}}$	$\widehat{\text{coverage}}$	(sign)$\widehat{\text{bias}^2}$	$\hat{\sigma}^2(\hat{\pi_*})$	$\widehat{\text{MSE}}$
†BP	2.2E-07	0.00172	95.60	(+)7.2E-09	1.5E-07	1.6E-07
LL	0.00020	0.05833	0.00	(+)0.00079	0.00017	0.00097
LN	0.00017	0.05206	5.50	(+)0.00031	0.00013	0.00045
EV	0.00042	0.08205	0.00	(+)0.00399	0.00037	0.00437
‡FPL$_{(-2,0.5)}$	1.5E-12	1.3E-05	0.00	(-)4.8E-07	2.8E-13	4.8E-07
FPL$_{(-2,1)}$	5.5E-11	6.1E-05	1.20	(-)4.8E-07	1.4E-11	4.8E-07
FPL$_{(-2,2)}$	9.3E-09	0.00061	25.20	(-)4.4E-07	3.5E-09	4.5E-07
FPL$_{(-2,3)}$	7.1E-07	0.00329	0.00	(+)0.25271	4.9E-07	0.25271
FPL$_{(-1,0.5)}$	2.1E-07	0.00245	78.40	(-)1.9E-07	1.1E-07	3.0E-07
FPL$_{(-1,1)}$	2.0E-06	0.00677	95.00	(+)1.7E-07	1.2E-06	1.4E-06
FPL$_{(-1,2)}$	3.3E-05	0.02515	20.10	(+)4.1E-05	2.4E-05	6.5E-05
FPL$_{(-1,3)}$	0.00015	0.05211	0.60	(+)0.00037	0.00012	0.00049
FPL$_{(-0.5,0.5)}$	6.3E-06	0.01165	76.60	(+)3.2E-06	4.3E-06	7.5E-06
FPL$_{(-0.5,1)}$	3.4E-05	0.02541	14.70	(+)4.9E-05	2.6E-05	7.5E-05
FPL$_{(-0.5,2)}$	0.00024	0.06421	0.00	(+)0.00098	0.00021	0.00119
FPL$_{(-0.5,3)}$	0.00064	0.10095	0.00	(+)0.00460	0.00057	0.00517
FPN$_{(-2,0.5)}$	1.1E-21	6.6E-08	0.00	(-)4.8E-07	2.0E-23	4.8E-07
FPN$_{(-2,1)}$	1.1E-17	5.7E-07	0.00	(-)4.8E-07	3.3E-19	4.8E-07
FPN$_{(-2,2)}$	4.0E-12	3.1E-05	0.80	(-)4.8E-07	3.5E-13	4.8E-07
FPN$_{(-2,3)}$	6.6E-07	0.00318	0.00	(+)0.25072	0.00101	0.25173
FPN$_{(-1,0.5)}$	1.0E-09	0.00030	11.10	(-)4.8E-07	1.8E-10	4.8E-07

Continued on next page

Table 5.6 – *continued from previous page*

(1) Model	(2) $\overline{(SE(\hat{\pi}_*))^2}$	(3) $\overline{\text{CIlength}}$	(4) $\widehat{\text{coverage}}$	(5) (sign)$\widehat{\text{bias}^2}$	(6) $\hat{\sigma}^2(\hat{\pi}_*)$	(7) $\widehat{\text{MSE}}$
$\text{FPN}_{(-1,1)}$	8.0E-08	0.00177	51.40	(-)4.0E-07	2.4E-08	4.3E-07
$\text{FPN}_{(-1,2)}$	1.1E-05	0.01511	94.50	(+)1.5E-06	6.0E-06	7.6E-06
$\text{FPN}_{(-1,3)}$	0.00011	0.04364	37.00	(+)9.3E-05	7.6E-05	1.7E-04
$\text{FPN}_{(-0.5,0.5)}$	6.7E-07	0.00436	80.60	(-)1.7E-07	2.7E-07	4.4E-07
$\text{FPN}_{(-0.5,1)}$	1.2E-05	0.01534	91.50	(+)2.4E-06	6.8E-06	9.2E-06
$\text{FPN}_{(-0.5,2)}$	0.00022	0.05914	3.30	(+)0.00042	0.00017	0.00060
$\text{FPN}_{(-0.5,3)}$	0.00070	0.10403	0.00	(+)0.00340	0.00061	0.00402
$\text{FPEV}_{(-2,0.5)}$	2.3E-10	9.7E-05	2.00	(-)4.8E-07	8.0E-11	4.8E-07
$\text{FPEV}_{(-2,1)}$	2.0E-09	0.00026	8.70	(-)4.6E-07	8.5E-10	4.6E-07
$\text{FPEV}_{(-2,2)}$	3.8E-08	0.00103	48.30	(-)3.5E-07	2.1E-08	3.7E-07
$\text{FPEV}_{(-2,3)}$	2.3E-07	0.00241	83.00	(-)1.5E-07	1.3E-07	2.8E-07
$\text{FPEV}_{(-1,0.5)}$	3.2E-06	0.00812	82.90	(+)1.5E-06	2.3E-06	3.7E-06
$\text{FPEV}_{(-1,1)}$	1.3E-05	0.01559	28.70	(+)1.7E-05	9.7E-06	2.6E-05
$\text{FPEV}_{(-1,2)}$	7.1E-05	0.03545	0.20	(+)0.00021	5.8E-05	0.00027
$\text{FPEV}_{(-1,3)}$	0.00018	0.05543	0.00	(+)0.00080	0.00015	0.00094
$\text{FPEV}_{(-0.5,0.5)}$	4.0E-05	0.02689	1.60	(+)9.8E-05	3.3E-05	0.00013
$\text{FPEV}_{(-0.5,1)}$	0.00012	0.04453	0.10	(+)0.00048	9.8E-05	0.00058
$\text{FPEV}_{(-0.5,2)}$	0.00039	0.07950	0.00	(+)0.00323	0.00034	0.00357
$\text{FPEV}_{(-0.5,3)}$	0.00069	0.10514	0.00	(+)0.00870	0.00060	0.00930
BFM	9.2E-06	0.00398	6.60	(+)8.5E-07	3.8E-05	3.9E-05
MA	7.8E-05	0.08390	87.10	(+)1.9E-05	2.2E-05	4.1E-05

5.4.2 Second Setting

In a second setting, we study the performance of the different models and of the model selected and averaged risk estimates again, now with the $\text{FPEV}_{(-1,3)}$ model being the true underlying simulation model. Results are summarized in Table 5.7.

In this setting, the MSE of the BP model is no longer the smallest, and also the coverage probability is very small as it underestimates the true value most of the times (see left upper panel of Figure 5.4). Again, as expected, the data generating model $\text{FPEV}_{(-1,3)}$ has the smallest MSE and its coverage probability nicely reaches the nominal as expected (see right upper panel of Figure 5.4). The best fitting model has an average bias similar to that of the BP model, but again, an additional source of variability enters the estimation process. The fractional polynomial EV model with powers (-0.5,3) is chosen most often as the best model but as in the previous setting it exhibits a very low coverage, partly also reflecting a similarly low coverage for the BFM. The averaged risk estimate has a small bias combined with a larger variance leading to a coverage probability of 97.90.

Both simulation settings show that model averaging has a beneficial effect in reducing bias, in accounting for the variability induced by the model selection process, and consequently in better coverage characteristics.

Table 5.7: *Simulation II results for estimated risk at dose 100cfu. For each model, columns (2) and (3) show the simulation-average of the variance-estimates of the estimated risk and the average confidence interval length respectively. The simulated coverage probability of the intervals is shown in column(4). Columns (5), (6) and (7) present the simulated squared bias, variance and mean squared error respectively. The direction of the bias shown by the sign beside the squared bias in column (3).* † *denotes the data generation model while* ‡ *shows the most frequently selected best AIC model. BFM denotes the best fitting models and MA denotes model averaging.*

(1)	(2)	(3)	(4)	(5)	(6)	(7)
Model	$\overline{(SE(\hat{\pi_*}))^2}$	$\overline{\text{CIlength}}$	$\widehat{\text{coverage}}$	(sign)$\widehat{\text{bias}^2}$	$\hat{\sigma}^2(\hat{\pi_*})$	$\widehat{\text{MSE}}$
BP	1.5E-06	0.00309	2.40	(-)0.00036	1.1E-06	0.00036
LL	0.00015	0.05118	95.00	(+)2.0E-05	0.00013	0.00015
LN	0.00011	0.04357	89.70	(-)3.2E-05	9.4E-05	0.00013
EV	0.00028	0.06764	34.50	(+)0.00077	0.00026	0.00104
$\text{FPL}_{(-2,0.5)}$	6.9E-10	0.00012	0.00	(-)0.00039	1.8E-10	0.00039
$\text{FPL}_{(-2,1)}$	2.7E-09	0.00022	0.00	(-)0.00039	9.2E-10	0.00039
$\text{FPL}_{(-2,2)}$	4.3E-08	0.00088	0.20	(-)0.00039	2.1E-08	0.00039
$\text{FPL}_{(-2,3)}$	7.3E-07	0.00334	0.00	(+)0.23402	3.6E-07	0.23402
$\text{FPL}_{(-1,0.5)}$	4.1E-07	0.00296	0.80	(-)0.00038	2.7E-07	0.00038
$\text{FPL}_{(-1,1)}$	2.2E-06	0.00672	4.30	(-)0.00035	1.6E-06	0.00035
$\text{FPL}_{(-1,2)}$	2.4E-05	0.02135	49.50	(-)0.00020	1.9E-05	0.00022
$\text{FPL}_{(-1,3)}$	0.00010	0.04246	91.20	(-)2.1E-05	8.6E-05	0.00011
$\text{FPL}_{(-0.5,0.5)}$	6.3E-06	0.01130	14.30	(-)0.00030	4.8E-06	0.00031
$\text{FPL}_{(-0.5,1)}$	2.7E-05	0.02271	56.40	(-)0.00017	2.2E-05	0.00019
$\text{FPL}_{(-0.5,2)}$	0.00017	0.05459	94.10	(+)3.6E-05	0.00015	0.00019
$\text{FPL}_{(-0.5,3)}$	0.00046	0.08608	33.80	(+)0.00119	0.00042	0.00161
$\text{FPN}_{(-2,0.5)}$	7.0E-12	3.9E-05	0.00	(-)0.00039	5.0E-13	0.00039
$\text{FPN}_{(-2,1)}$	3.4E-11	5.1E-05	0.00	(-)0.00039	3.1E-12	0.00039
$\text{FPN}_{(-2,2)}$	1.2E-08	0.00020	0.10	(-)0.00037	0.00026	0.00063
$\text{FPN}_{(-2,3)}$	7.0E-07	0.00328	0.00	(+)0.23410	3.7E-07	0.23410
$\text{FPN}_{(-1,0.5)}$	3.5E-08	0.00081	0.20	(-)0.00039	1.1E-08	0.00039

Continued on next page

Table 5.7 – *continued from previous page*

(1) Model	(2) $\overline{(SE(\hat{\pi}_*))^2}$	(3) $\overline{\text{CIlength}}$	(4) $\widehat{\text{coverage}}$	(5) (sign)$\widehat{\text{bias}^2}$	(6) $\hat{\sigma}^2(\hat{\pi}_*)$	(7) $\widehat{\text{MSE}}$
$\text{FPN}_{(-1,1)}$	2.9E-07	0.00232	1.00	(-)0.00039	1.3E-07	0.00039
$\text{FPN}_{(-1,2)}$	8.4E-06	0.01242	19.80	(-)0.00033	5.6E-06	0.00034
$\text{FPN}_{(-1,3)}$	6.6E-05	0.03346	69.30	(-)0.00016	5.1E-05	0.00021
$\text{FPN}_{(-0.5,0.5)}$	1.3E-06	0.00491	3.80	(-)0.00038	7.1E-07	0.00038
$\text{FPN}_{(-0.5,1)}$	9.9E-06	0.01357	21.60	(-)0.00032	6.9E-06	0.00033
$\text{FPN}_{(-0.5,2)}$	0.00014	0.04760	91.40	(-)1.8E-05	0.00012	0.00013
$\text{FPN}_{(-0.5,3)}$	0.00047	0.08590	70.70	(+)0.00058	0.00043	0.00101
$\text{FPEV}_{(-2,0.5)}$	1.0E-09	0.00014	0.00	(-)0.00039	4.4E-10	0.00039
$\text{FPEV}_{(-2,1)}$	3.5E-09	0.00025	0.00	(-)0.00039	1.9E-09	0.00039
$\text{FPEV}_{(-2,2)}$	3.2E-08	0.00079	0.00	(-)0.00039	2.1E-08	0.00039
$\text{FPEV}_{(-2,3)}$	1.6E-07	0.00178	0.10	(-)0.00039	1.1E-07	0.00039
$\text{FPEV}_{(-1,0.5)}$	2.0E-06	0.00619	2.00	(-)0.00034	1.5E-06	0.00034
$\text{FPEV}_{(-1,1)}$	7.1E-06	0.01155	15.30	(-)0.00028	6.0E-06	0.00028
$\text{FPEV}_{(-1,2)}$	4.0E-05	0.02651	73.50	(-)8.6E-05	3.6E-05	0.00012
$^{\dagger}\text{FPEV}_{(-1,3)}$	0.00011	0.04284	95.60	(+)8.7E-07	9.9E-05	9.9E-05
$\text{FPEV}_{(-0.5,0.5)}$	2.3E-05	0.02050	52.90	(-)0.00016	2.0E-05	0.00018
$\text{FPEV}_{(-0.5,1)}$	6.8E-05	0.03442	91.50	(-)1.5E-05	6.2E-05	7.7E-05
$\text{FPEV}_{(-0.5,2)}$	0.00025	0.06440	50.60	(+)0.00052	0.00024	0.00076
$^{\ddagger}\text{FPEV}_{(-0.5,3)}$	0.00049	0.08874	4.50	(+)0.00281	0.00047	0.00328
BFM	0.00023	0.05146	35.20	(+)0.00022	0.00101	0.00123
MA	0.00048	0.12179	97.90	(+)0.00002	0.00025	0.00028

In the philosophy that a true and correct model does not exist, or that you will never know it exactly (which seems the only realistic situation), this simulation also shows that not just one single model is appropriate to describe the dose-response relationship of microbial risks, but there exists a whole set of possible models. Fractional polynomials are very flexible to estimate the low-dose risk. However, the model selec-

tion procedure induces extra variability that should be accounted for. The averaged risk estimate gives larger weights to better-fitting models, resulting in a smaller bias. The model selection uncertainty is accounted for in this approach, and a better coverage probability obtained as illustrated in Figure 5.4. However, from both simulation settings it was evident that for this particular risk assessment application the coverage percentages for most of the model fall far below the chosen level of confidence (0.95). The magnitude of inflation of the coverage was certainly surprising to us and is in itself a clear warning to be very careful and thoughtful when restricting the analysis to one single model. It certainly motivates the use of different plausible models, at least as a type of sensitivity analysis, but even better in a multimodel approach such as model averaging.

5.4.3 To Include Fractional Polynomials or Not

As pointed out before, and clear from the nature of the model averaging approach, the set of candidate models should be rich enough. This section tries to illustrate this point by comparing the results from the data analyses in Section 5.3 and the simulation results in the previous sections to the corresponding results of model averaging over the four classical models only. Let us denote by set one the restricted set of classical models BP, LL, LN and EV and by set two the modified flexible fractional polynomials added to set one. For *salmonella typhi* example, the averaged risk estimates are 0.029 (SE 0.021 and confidence interval [0.007,0.114]) and 0.01285 (SE 0.016 and confidence interval [0.001,0.134]) respectively for set one and two. The risk estimate from set two is reduced to about half that of set one. For *campylobacter jejuni* example, the averaged risk estimates are 0.2429 (SE 0.180 and confidence interval [0.045,0.686]) and 0.089 (SE 0.152 and confidence interval [0.002,0.793]) for set one and two respectively. Including fractional polynomials reduces the estimate to about 2.5 times from not including them.

For the first simulation setting, with the restricted set one, the true risk estimate was captured in 38.5% of the runs by the confidence intervals of the averaged risk compared to 87.1% for set two. Also, the square bias and the variance were higher for set one models than for set two models. In the second simulation setting, averaging over the four models attained a coverage probability of 87.3% versus 97.9% for set two. To this end, we see that using the richer set two, which includes the flexible fractional polynomial models, yielded coverage probabilities closer to the nominal 95% level, less biased and more precise risk estimates than set one.

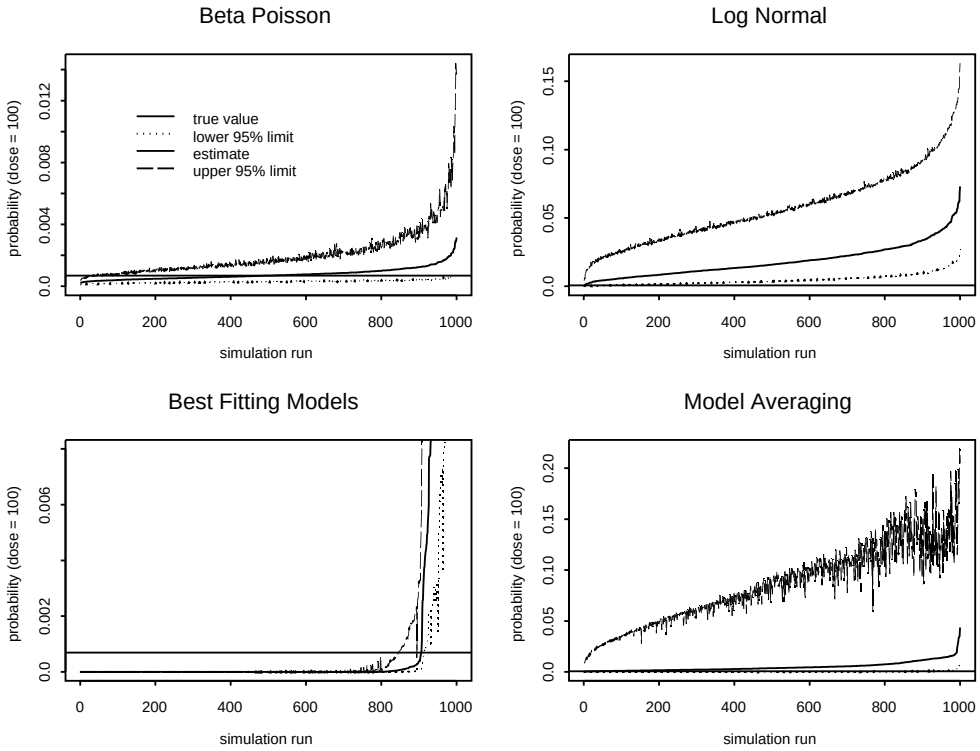

Figure 5.3: *Simulation I: True probability of S. Typhi based on the Beta Poisson generating model and the estimated probabilities for the Beta Poisson (data generation model), the log-normal model, the best fitting models across simulations and the averaged risk. The corresponding confidence limits of the estimated risks are shown to illustrate how best they capture the true value.*

5.5 Application to Multi-Strain Data

The model averaging approach extends to settings with additional sources of heteregeneity, within the framework of Generalized Linear Mixed Models (GLMM, see e.g. Molenberghs and Verbeke, 2005). Table 5.8 shows data on *Campylobacter jejuni* dose exposure for different strains, the total number of chicken exposed and the infected chicken (Chen *et al.* 2006). Additionally there is information about the sample, (F)resh or (L)aboratory, whether it originates from (C)hicken or (H)uman and whether its host is (W)ickman or (L)ohmann.

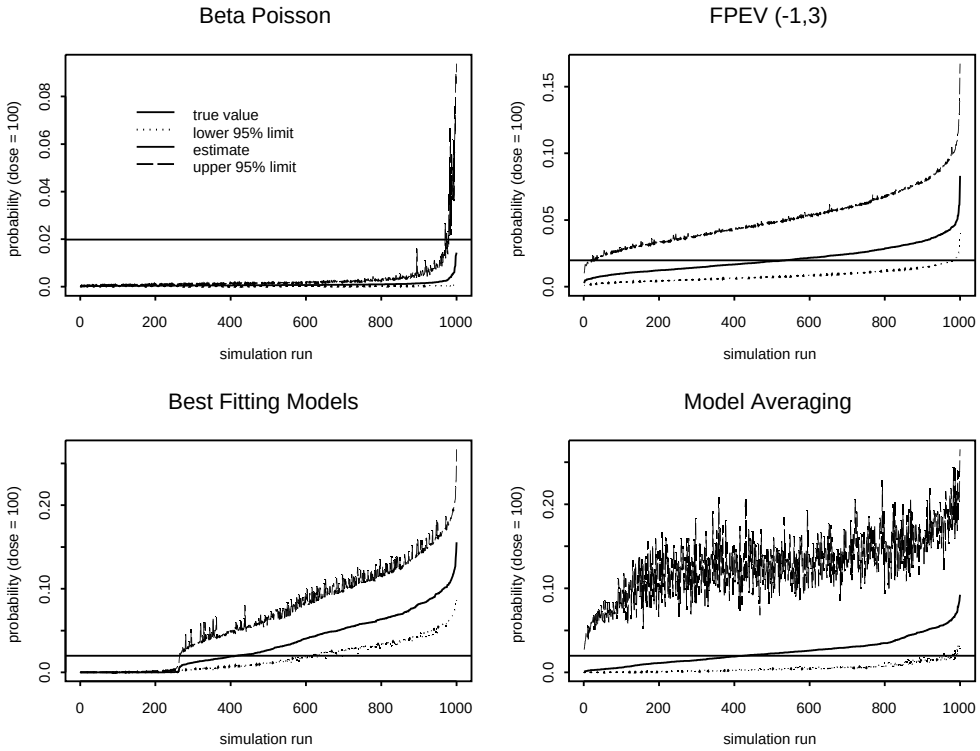

Figure 5.4: *Simulation II: True probability of S. Typhi based on FPEV(-1,3) the generating model and the estimated probabilities for the Beta Poisson (taken as standard model), the FPEV(-1,3) model, the best fitting models across simulations and the averaged risk. The corresponding confidence limits of the estimated risks are shown to illustrate how best they capture the true value.*

Table 5.8: *Campylobacter jejuni data in chicken. Isolates. Sample: F(resh)/L(aboratory). Origin: H(uman)/C(hicken). Host: W(ickham)/L(ohmann)*

Group	Isolate	Sample	Origin	Host	Dose	Exposed	Infected
1	480	F	H	W	10	10	0
	480	F	H	W	1000	10	0
	480	F	H	W	100,000	10	10
2	466	F	H	L	840	8	8
	466	F	H	L	8400	7	7
3	P5444	F	H	L	10,000	10	10
4	P5623	F	H	L	5000	10	9
5	119a	F	C	L	17,000	4	4
6	121a	F	C	L	11,000	4	4
7	123a	F	C	L	3500	4	2
8	UA585	F	H	W	95	10	10
	UA585	F	H	W	950	10	10
	UA585	F	H	W	9500	10	10
	UA585	F	H	W	95,000	10	10
9*	93/146	F	C	W	100	10	0
	93/146	F	C	W	1000	9	9
	93/146	F	C	W	100,000	10	10
10*	94/146	F	H	W	120	10	0
	94/146	F	H	W	1200	10	10
	94/146	F	H	W	12,000	10	10
11*	93175	F	C	W	28	10	0
	93175	F	C	W	280	10	10
	93175	F	C	W	2800	10	10
12*	0087	F	H	L	30	10	3
	0087	F	H	L	300	10	10
	0087	F	H	L	480	10	10
	0087	F	H	L	2000	6	6
	0087	F	H	L	4500	8	8
	0087	F	H	L	4800	10	10
	0087	F	H	L	200,000	10	10
13*	81116P	F	C	W	4	10	0
	81116P	F	C	W	40	10	10
	81116P	F	C	W	40	9	9
	81116P	F	C	W	400	10	10
	81116P	F	C	W	4000	9	9
	81116P	F	C	W	40,000	10	10
14	R3P	F	C	W	10	10	10
	R3P	F	C	W	100	10	10
15	81-176	L	C	W	2000	10	8
	81-176	L	C	W	2,000,000	10	10
16	81-176	L	C	L	800	9	8
	81-176	L	C	L	80,000	10	10

Continued on next page

Table 5.8 – *continued from previous page*

Group	Isolate	Sample	Origin	Host	Dose	Exposed	Infected
17	11168	L	C	W	200,000	10	10
18	11168	L	C	L	2000	10	4
	11168	L	C	L	20,000	10	5
	11168	L	C	L	200,000	10	7
	11168	L	C	L	2,000,000	10	9
19	R3	L	C	W	300	9	6
	R3	L	C	W	3000	10	10
20*	81116	L	C	W	54	8	3
	81116	L	C	W	540	6	3
	81116	L	C	W	5400	5	5
	81116	L	C	W	54,000	6	6
	81116	L	C	W	250,000	14	14
21*	81116	L	C	L	35	10	0
	81116	L	C	L	350	10	6
	81116	L	C	L	3500	10	8
	81116	L	C	L	100	10	4
	81116	L	C	L	1000	10	9
	81116	L	C	L	10,000	10	10
	81116	L	C	L	10,000	10	10
	81116	L	C	L	1,000,000	10	10
22	81117	L	C	L	35,000	10	10

The challenge here is to consider sensible candidate models which correctly account for the different sources of variability or heterogeneity (different types of isolate, sample, origin, and host) and to extend the model averaging approach over such a family of candidate models. In general one can extend models as discussed in the previous section, with fixed effects and/or random effects. It seems natural to incorporate sample, origin and host (all binary) in the model by fixed effects and isolate by using a random effect. But for the latter one could also consider a fixed effect. Models with all of the covariates (sample, origin, and host) included as fixed effects and with isolate as fixed or random effect are overparameterized. Indeed, the type of isolate covers most of the (sample, origin, host)-combinations (except for isolate 81-176, 11168, 81116). Moreover for several isolates, there are only observations for one single dose level, making dose response models with two or more parameters unestimable.

Therefore, to account appropriately for the heterogeneity resulting from different types of isolates, samples, origins and hosts, we considered two approaches: (i) a random effect for all different combinations of (isolate, sample, origin, host), leading to a total of 22 different "groups" (as indicated in the first column of Table 5.8); in this

approach it is possible to fit models with and without type of sample as an additional (fixed) effect in the model (ii) a selection of 7 groups (out of the 22 groups, those with group number marked by an asterisk $*$ in the first column), which allows a fully fixed effects approach (one parameter for each group), again with and without type of sample as an additional (fixed) effect in the model. Strains with only one or two dose levels together with those that exhibited outlying observations were eliminated. These four families of candidate models corresponds to the four parts of Table 5.9, as discussed further on. Because of computational problems and overparameterization, we did not consider fractional polynomial dose response models. Only predictors linear in dose were considered in what follows.

Table 5.9 shows, without and with *type of sample* (upper and lower part of the table respectively), the AIC and the weights of the models fitted to 7 strains (3rd and 4th columns) and also the AIC and weights of the models fitted to 22 groups (5th and 6th columns). Using the likelihood ratio test, the *type of sample* variable was significant in the models with an asterisk (*) fit on 7 strains while *type of sample* was only significant in the model with † sign fit on 22 groups. This data example shows that random effects (RE) models are more flexible than fixed effects models and can handle over-parametrization problems. As is clear from the empty cells in the right upper part of Table 5.9, fixed effects (FE) models could not be fit to the 22 groups. The random effects models with type *type of sample* show no substantial improvement compared to those without *type* regardless of the type of grouping used. Most importantly, they all point to the same or similar models.

As an illustration, consider the estimation of the probability to be infected at dose level $d = 1$, by the models listed in the left upper part of Table 5.9 (7 groups and without the *type of sample* variable). As the weights in the table reveal, some of the models have very little contribution and so we exclude them for the model averaging exercise. Table 5.10 shows the results for the best five individual models and the model averaged estimate for the probability of infection, together with Wald type confidence intervals. The table mainly shows that the best five models can lead to substantially different point estimates as well as confidence intervals. In general the model averaged estimates is situated somewhere between the individual model estimates, and the model averaged confidence intervals tend to be much wider, reflecting the model uncertainty.

For graphical illustration of the different models Figure 5.5 shows fitted dose response curves for the best fixed-effects model (FE log-logistic) and the best random-effects model (log-normal random $\alpha\&\beta$) together with all data of the 7 groups in the two upper panels, and fitted curves for the five best models, separately for the data of two

Table 5.9: *Multi-strain Campylobacter jejuni data: model fits and AIC-weights using data of only 7 groups, and of all 22 groups.*

without the *type of sample* variable					
Model	random	AIC^{7}	weight^{7}	AIC^{22}	weights^{22}
FE LL	-	77.32	0.43351		
BP FE on u	-	79.39	0.15434		
LN	α & β	79.72	0.13094	154.87	0.47784
FE LN	-	79.93	0.11796		
LL	α & β	80.37	0.09466	154.84	0.48299
EV	α & β	81.40	0.05650	159.88	0.03889
FE EV	-	84.58	0.01152		
LL	α	91.40	0.00038	188.65	2.2E-08
BP	u	94.14	9.7E-05	169.84	0.00027
LN	α	94.47	8.2E-05	194.15	1.4E-09
EV	α	100.00	5.2E-06	197.55	2.6E-10
BP	v	135.99	7.9E-14	253.30	2.0E-22
BP FE on v	-	141.92	4.1E-15		
BP	u & v	518.83	5.8E-97	177.26	6.5E-06
with the *type of sample* variable as fixed effect					
Model	random	AIC^{7}	weight^{7}	AIC^{22}	weights^{22}
*LN	α & β	76.85	0.60000	154.68	0.53379
*LL	α & β	77.67	0.39953	154.97	0.46359
LL	α	92.44	0.00025	188.80	2.1E-08
BP	u	94.04	0.00011	169.27	0.00036
LN	α	95.41	5.6E-05	194.32	1.3E-09
EV	α & β	95.72	4.8E-05	165.62	0.00225
EV	α	99.52	7.2E-06	196.97	3.5E-10
*†BP	v	129.66	2.1E-12	296.80	7.4E-32
BP	u & v	578.32	7.7E-110	182.52	4.8E-07

* models fit on 7 groups where the *type of sample* variable is significant using likelihood ratio test.

† models fit on 22 groups where the *type of sample* variable was significant.

groups (group 13 and 21) in the two lower panels.
The curves are extrapolated to a dose of 1cfu. The upper panels of Figure 5.5 show some considerable variability in the fits to the different groups for the log-normal model with random intercept and slope compared to the fixed effects model. This conclusion was also observed for the other logit random effects model and the log-normal and BP fixed effects models not shown here. In the lower panels of Figure 5.5, we observe that the response curves for the five models fit the data quite well, a conclusion also confirmed by the closeness of the model AIC's. However, despite the closeness of the AIC's there is a lot of variability at the lower dose of 1cfu (which is out of the observed range of dose levels). So, especially for extrapolation, model averaging stabilizes the highly variable point estimate, but variability in terms of model uncertainty is still reflected by the wide confidence intervals (see Table 5.10).

Table 5.10: *Multi-strain Campylobacter jejuni data. Model specific and model averaged estimated probability of Campylobacter jejuni: averaging over five models, the three best fixed effects models and the two best random effects models without type of sample. The last two letters added at the end of the strain indicate whether the strain originated from (h)uman or (c)hicken and its host is (l)ohmann or (w)ickham.*

						wald CIs	
Strain	Model	AIC	Weight	$\hat{\pi}(1)$	SE($\hat{\pi}(1)$)	lower	upper
0087hl	FE LL	77.32	0.46543	0.00109	0.00113	0.00014	0.00823
	BP FE on u	79.39	0.16571	0.01539	0.00594	0.00720	0.03261
	LN random α & β	79.72	0.14058	7.8E-11	1.9E-09	3.1E-42	0.77942
	FE LN	79.93	0.12665	9.7E-05	0.00021	7.3E-07	0.00418
	LL random α & β	80.37	0.10163	9.8E-06	6.0E-05	5.6E-11	0.63004
	Model Averaged			0.00307	0.00445	0.00018	0.05043
81116Pcw	FE LL	77.32	0.46543	0.01429	0.01178	0.00281	0.06946
	BP FE on u	79.39	0.16571	0.07314	0.02284	0.03917	0.13250
	LN random α & β	79.72	0.14058	1.0E-07	1.3E-06	1.5E-23	0.32164
	FE LN	79.93	0.12665	0.01153	0.01325	0.00087	0.07897
	LL random α & β	80.37	0.10163	9.5E-05	0.00047	6.0E-09	0.60125
	Model Averaged			0.02024	0.02259	0.00221	0.16152
81116cl	FE LL	77.32	0.46543	4.1E-05	5.4E-05	3.0E-06	0.00054
	BP FE on u	79.39	0.16571	0.00223	0.00073	0.00117	0.00425
	LN random α & β	79.72	0.14058	0.00028	0.00081	3.4E-07	0.02669
	FE LN	79.93	0.12665	9.0E-09	3.5E-08	1.9E-12	7.9E-06

Continued on next page

Table 5.10 – *continued from previous page*

Strain	Model	AIC	Weight	$\hat{\pi}(1)$	SE($\hat{\pi}(1)$)	wald CIs lower	wald CIs upper
	LL random α & β	80.37	0.10163	0.00244	0.00358	0.00014	0.04194
	Model Averaged			0.00068	0.00120	2.1E-05	0.02147
81116cw	FE LL	77.32	0.46543	7.7E-05	0.00011	4.8E-06	0.00124
	BP FE on u	79.39	0.16571	0.00284	0.00120	0.00124	0.00650
	LN random α & β	79.72	0.14058	0.00662	0.01964	2.7E-06	0.34367
	FE LN	79.93	0.12665	1.4E-07	4.8E-07	6.9E-11	5.7E-05
	LN random α & β	80.37	0.10163	0.01459	0.02560	0.00045	0.32680
	Model Averaged			0.00292	0.00756	1.8E-05	0.32243
93/146cw	FE LL	77.32	0.46543	2.8E-05	4.0E-05	1.6E-06	0.00049
	BP FE on u	79.39	0.16571	0.00290	0.00121	0.00129	0.00655
	LN random α & β	79.72	0.14058	1.3E-32	7.0E-31	1.0E-98	0.00229
	FE LN	79.93	0.12665	5.3E-09	2.3E-08	4.1E-13	9.0E-06
	LL random α & β	80.37	0.10163	5.7E-10	5.4E-09	5.6E-18	0.05498
	Model Averaged			0.00049	0.00085	1.7E-05	0.01405
93175cw	FE LL	77.32	0.46543	0.00030	0.00036	2.7E-05	0.00325
	BP FE on u	79.39	0.16571	0.00925	0.00363	0.00428	0.01990
	LN random α & β	79.72	0.14058	2.7E-25	1.1E-23	3.3E-77	0.01836
	FE LN	79.93	0.12665	4.8E-06	1.3E-05	8.6E-09	0.00065
	LL random α & β	80.37	0.10163	7.0E-09	5.8E-08	5.6E-16	0.07878
	Model Averaged			0.00167	0.00267	7.2E-05	0.03710
94/146hw	FE LL	77.32	0.46543	2.1E-05	3.2E-05	1.1E-06	0.00040
	BP FE on u	79.39	0.16571	0.00265	0.00106	0.00120	0.00582
	LN random α & β	79.72	0.14058	3.9E-33	2.0E-31	2.8E-96	0.00102
	FE LN	79.93	0.12665	2.3E-09	1.0E-08	1.2E-13	5.4E-06
	LL random α & β	80.37	0.10163	5.0E-10	4.6E-09	5.6E-18	0.04208
	Model Averaged			0.00045	0.00077	1.6E-05	0.01282

5.6 Simulation Study for Multi-Strain Data

This section summarizes the results of a limited simulation study for a multi-strain (or more general a multi-group) setting. We fix the dose levels and the total number of exposed chicken (n) to that of the *Campylobacter jejuni* data example and generate the number of infected chicken from a binomial distribution with parameters n and

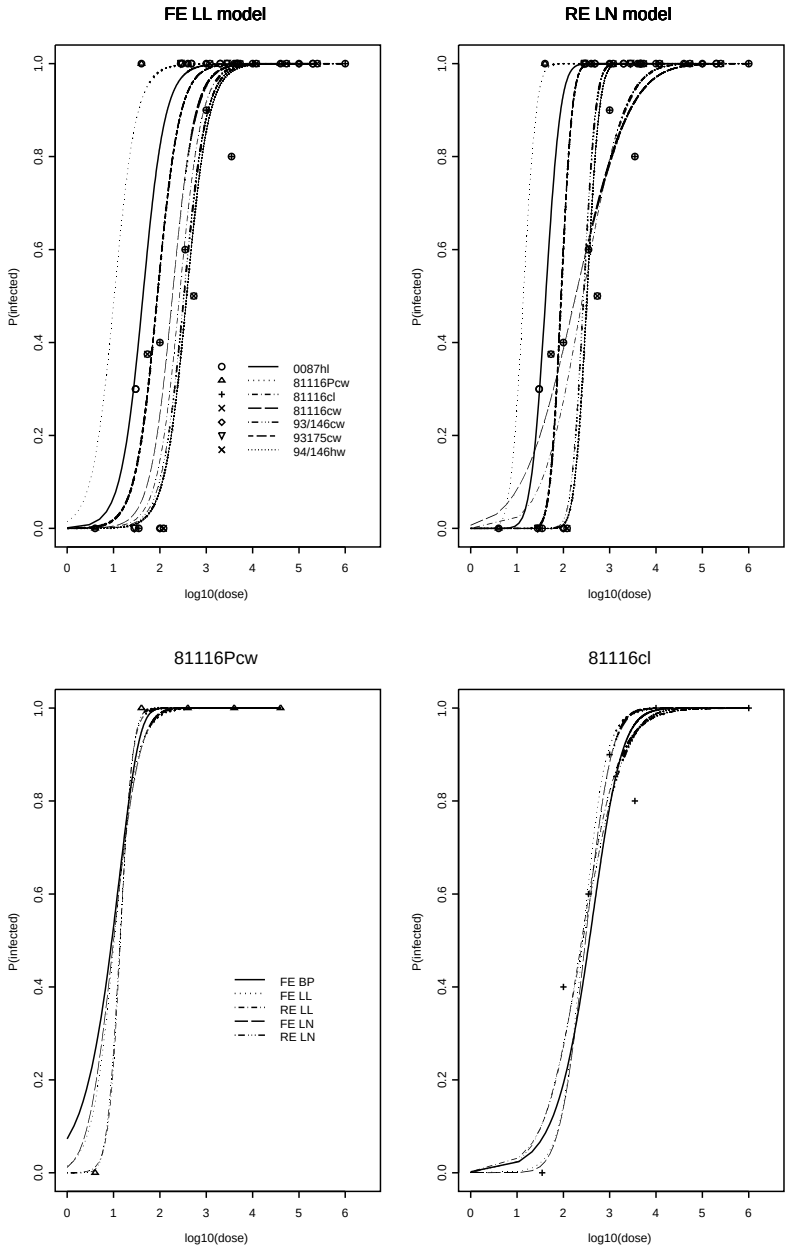

Figure 5.5: *Estimated Campylobacter jejuni dose-response curves: row 1 shows 7 strains using the FE log-logistic model and 2 REs log-normal model and row 2 for five models shown for the'fresh' isolate 81116Pcw and 'Laboratory' isolate 81116cl*

p. As probabilities p we take the predicted probabilities from the fixed effects log-logistic model, fitted to the *Campylobacter jejuni* data. To the simulated data we fit four models, the fixed and random effects, log-normal and log-logistic models. The BP model was excluded, due to computational problems. The probability of infection at dose 1 is estimated from each model, from the AIC selected model and also by model averaging. With the generation and fitting process repeated 1000 times, the average length and the coverage percentage of confidence intervals, are examined for each strain.

Figure 5.6 summarizes the simulation results on the confidence intervals. The performance characteristics for all models (4 single models, the AIC selected model, and the averaged model) are plotted against the strain numbers (on the horizontal axes), which are arranged in ascending order of the true probabilities. The graph with the coverage percentages in the right panel differentiates the different models markedly: the averaged model shows coverage percentages larger than but close to those of the true model, the latter as expected being close to 95%. The coverage probabilities of the other models, the fixed effects log-normal model and the AIC selected model are much lower, being low to extremely low on the left end of the scale and increasing to acceptable values at the right end of the scale (corresponding to higher values of the true probabilities). The left panel shows that, compared to the other models, the length of the confidence intervals produced by the two random effects models is much larger than those of the other models. Although this simulation study is very limited and many other settings could be considered, it shows that model selection and model averaging is a crucial issue, also in more complicated settings of multi-strain data and random effects models.

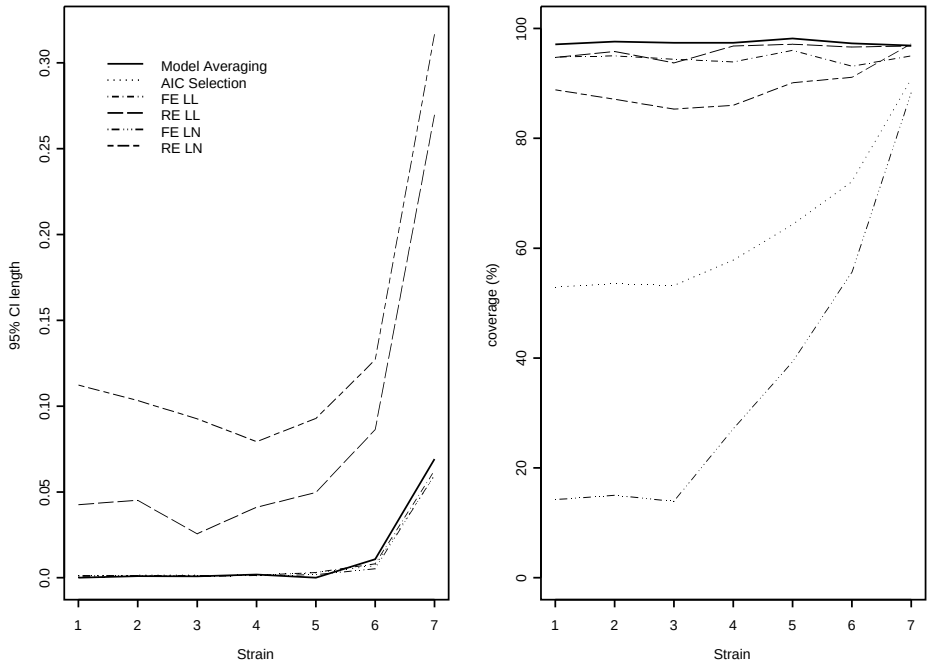

Figure 5.6: *Multi-strain simulation, assuming the fixed effects log-logistic model as the true model. From left to right: length of 95% confidence intervals and coverage percentage. The horizontal axes represent the 7 different strains in ascending order of the true probabilities.*

5.7 Discussion

In quantitative risk assessment for microbial pathogens, dose-response assessment is a critical issue. It must be included, health effects are the end-point in any risk assessment. Data are scarce, however. Only few pathogens have been used in clinical studies, and even these have been done on a small scale with few volunteers exposed per dose, due to the high costs involved. As a consequence, microbial dose-response data often do not contain a great deal of information on the shape of the dose-response relation, and model choice based entirely on experimental data is not feasible (Teunis and Havelaar 2000). This problem has raised may discussions and several solutions have been proposed. One could use animals, as these allow, at least in principle, collection of more extensive data sets. However, this chapter also has demonstrated

that this may merely give a better insight into the variability involved, not necessarily decreasing the uncertainty about the shape of the relation. Model averaging has been used before, weighting over an arbitrary collection of mathematical relations considered useful. Reasons for inclusion may range from biological plausibility to flexibility or mere tradition (like threshold models). Here it has been shown that the modified fractional polynomials can be derived from the general equation (5.3) thereby potentially extending the family of eligible models a great deal, and providing a natural choice for the collection of models to include in model averaging.

As observed in the simulation study, the coverage probabilities of the confidence intervals of the individual models, the best fitting models and the model average highly depend on the generating (true) model. When bias values are high relative to the variance, the coverage of these intervals considerably fall below the nominal level of confidence. The bias values for the averaged risk were 10 times higher when only averaging over the four commonly used dose-response models than when fractional polynomials were included and as a result the coverage probabilities in the former case were always below nominal. To this end we recommend model averaging including the proposed set (or biological more plausible subset) of fractional polynomials in the set of candidate models.

An essential improvement in dose-response assessment is its extension to a hierarchical framework: like most biological problems, data can frequently be organized hierarchically. For example several isolates (pure strains of microorganisms separated from a mixed bacterial culture) of a single pathogen species in the same host species, different but related pathogen species in the same host, or a single pathogen isolate in hosts with different levels of immunity. As the purpose of risk assessment usually is to predict risks for an exposed population, dose-response relations should be translated from the special, experimental setting to a more general, unspecified situation. For instance, given the different responses to a limited collection of isolates of a pathogen, what would be the response to a newly isolated specimen, of the same type? Fixed effects models are easy to implement and represent a first solution, but the random models, although more complex, use all available data and are more suitable for explicative studies. The generalized linear mixed models (Agresti, 2002; Molenberghs and Verbeke, 2005) allow prediction by using the (joint) distribution of the random parameters to make exactly this generalization. Our future interest is to extend these proposed fractional polynomials to investigate dose-response relationships between isolates of a single pathogen using generalized linear mixed models and full Bayesian models.

Further research is needed to go into some issues, like the estimation at very low

dose levels, especially in case of limited information. Since low dose extrapolation is crucial in risk assessment and the AIC method selects a model regardless of its intended use, we intend to further investigate selecting a model looking at the (lower dose) region of interest, by using e.g. the focused information criterion (Claeskens and Hjort, 2003). Another interesting avenue for further research is to investigate bootstrap based alternatives as proposed by Faes *et al.* (2006b) and Wheeler and Bailer (2007).

Part III

Modeling Data on *Salmonella* Infection in Broiler and Layer Chicken Flocks

Chapter 6

Risk Factor Identification for *Salmonella* in Belgian Laying Hens

The consumption of foods of animal origin, particularly poultry eggs, can lead to transmission of zoonotic diseases (*Salmonella*, in this case) from the animals to humans. Salmonellosis constitutes a major public health burden and represents a significant cost in many countries. In Belgium, the disease ranks high among the reported food borne illnesses (Collard *et al.* 2004). Even if the incidence of human salmonellosis has diminished since 1999, in 2004, 9545 cases were reported in the country (EFSA, 2006a). As in most of the countries around the world, Belgian *Salmonella* outbreaks in humans are very often linked to the consumption of contaminated eggs (Davies and Breslin, 2001; Van Immerseel *et al.* 2005; Collard *et al.* 2007). The most frequently isolated serotype in layer flocks in the EU as well as in Belgium is *Salmonella* Enteritidis which is a non-typhoid non-host adapted serotype with a very wide host range (Baird-Parker, 1990; Gast *et al.* 2005; Quinet, 2005; VAR, 2005; EFSA, 2004). The bacterium infects the eggs by two processes: first by vertical transmission during the development of the egg within the ovary or its passage through the oviduct and secondly by horizontal transmission through trans-shell contamination (Kinde *et al.* 2000; WHO FAO, 2002; Davies and Breslin, 2003a; Van Immerseel *et al.* 2005). Vertical transmission is considered to be the major route of egg contamination and should be controlled by applying sanitary measures at the breeders level (that is, hy-

giene practices and eventually vaccination) while horizontal transmission should be reduced by preventing contacts between the layer hens and by cleaning and disinfecting the flock's environment. *Salmonella* is known for its ability to asymptomatically infect the hen's oviduct (De Buck *et al.* 2004a; 2004b). Therefore detection of infected flocks depends entirely on laboratory analysis. An infected hen may contaminate one egg out of 200 (Quinet, 2005). Reducing *Salmonella* flock prevalence results in a directly proportional reduction in human health risk (Altekruse *et al.* 1993). This suggests that sanitary measures at the flock level contribute to a significant reduction of the risk for salmonellosis due to egg consumption. In Belgium, the layer breeders are not significantly infected, probably due to the many years' efforts of control at this level and therefore, it is reasonable to assume that most day-old chicks are free from *Salmonella* when placed on farms (Davies and Breslin, 2001; AFSCA, 2004). The majority of the infections in layer hens seem to be attributed to the persistent contamination of the farm. Indeed, the presence of *Salmonella* in the laying house environment has been strongly correlated with the probability that hens will lay contaminated eggs. Chicken are infected after oral ingestion of the bacteria from the environmental sources (for example, contaminated fluff, dust, feed etc) invasion of the mucosal epithelial cells, which leads to systemic dissemination and colonization of the ovary and oviduct (Henzler *et al.* 1998; Davies and Breslin, 2003b). The primary control should focus at farm level. Control measures include preventing contacts with contaminated feed and visitors, wearing house-specific clothing, thorough cleaning and disinfection of the layer houses, vaccination, rodent control programs. In Belgium, every holding housing more than 5000 hens is required to be sampled for *Salmonella* diagnosis 3 weeks before slaughter time. This measure probably has contributed to a reduction of the risk for food-borne salmonellosis. However, in 2004, still 27% of the layer flocks analysed remained positive for *Salmonella* (AFSCA, 2004). Several risk factors have been described, but in order to advise the Belgian competent authority (Federal Agency for the safety of the Food Chain) with detailed, practical guidance, an understanding of possible causal factors is essential. The objective of the study reported here was to investigate the risk factors which are associated with the occurrence of *Salmonella* in laying hens in Belgium using data collected for the Baseline Study on the Prevalence of *Salmonella* in laying flocks of Gallus gallus f. domestica in the European Union (SANCO/34/2004 and Commission decision 2004/665/EG). Although it would be worthwhile to utilize data from earlier years, the 2005 data set contained flock information, particularly on some demographic factors and *Salmonella* vaccination status, which were unavailable for earlier databases. Section 6.1 describes how the data were obtained and gives the methods used to analyse the data. In Sec-

tion 6.2, the exploratory results and model fitting results are presented. Section 6.3 concludes the chapter with a discussion. A part of the analyses of this chapter are published in Namata *et al.* (2008a).

6.1 Material and Methods

6.1.1 Data Collection

The Belgian part of the Baseline Study on the Prevalence of *Salmonella* in egg laying flocks of Gallus gallus in the European Union consisted of a cross-sectional study that covered the year 2005 from February to September in Belgium. The primary sample size providing the number of holdings which had to be tested was calculated on the basis of a target prevalence of 20%, a confidence level of 95% and an accuracy of 3% (Commission decision 2004/665/EG). The population of laying hens was stratified according to holding size (below 1000, 1000-2999, 3000-4999, 5000-9999, 10000-29999, 30000 and more). The number of holdings to be sampled was subsequently distributed proportionally to the number of holdings in each class. In all cases, only one flock per holding was sampled. Seven different samples, two dust samples and five faecal samples were collected from each selected flock. The dust samples were any of these types: 1) dust from different places in case of barn or free range flocks, 2) dust from egg belts, 3) dusty material from beneath cages. Faecal samples were any of these types: 1) boot swabs which are socks placed over the boots and are sufficiently absorptive to collect faecal or moist litter samples from the floor surfaces (SANCO/34/2004 and Commission decision 2004/665/EG), 2) pooled faecal samples from dip pits, 3) pooled faecal samples from dropping belts, 4) pooled faecal samples from scrapers. The collection of these samples was as follows: There had to be five pooled faecal samples taken per selected flock. For the pooled faecal samples in cages, there are normally several stacks of cages within a henhouse. The material from each stack picked up using a new pair of plastic gloves for each individual sample was included in each of the five pooled faecal samples of 200-300 grams. For the boot swabs in barns and free range flocks, each henhouse was divided in sectors of at least 100m that were walked on with new boot swabs, five pairs of boot swabs per henhouse. Each of the five pooled samples comprised of faecal material fixed to a pair of boot swabs. The dust material from beneath cages was obtained from 20 separate locations within a henhouse using a new pair of plastic gloves for each sample. Finally for the dust from different places from barns and free range, each dust sample was collected in a 250ml plastic jar or bag ensuring that all parts of the henhouse like from exhaust fan, ledges,

beams etcetera were covered. In order to maximise sensitivity both faecal material (5 out of 7) and dust material from the environment (2 out of 7) were sampled, depending on whether the birds were reared in cages or barns or free-range, in such a way that the complete farm was represented. The hens were sampled at the end of their laying period, within a maximum of 9 weeks before depopulation. Samples were sent within 24 hours to the laboratory. The detection method was as recommended by the Community Reference Laboratory for *Salmonella* in Bilthoven, The Netherlands, that is, a modification of ISO 6579:2002. *Salmonella* isolates were serotyped following the Kaufmann-White scheme (Popoff, 2001; VAR, 2005).

The explanatory variables recorded include: region (1= Walloon or 0= Flanders), sampling time (month the flock was sampled: February to September), production type (cage or barn/ free range), age (in weeks), flock size (number of hens in the flock considered) and vaccination status against *Salmonella* (yes, unknown, or no). The flocks were vaccinated against *Salmonella* enterica, serovar Enteritidis during the rearing period (one day to 18-20 weeks) with either a live or inactivated vaccine type although for some flocks the vaccine type was not known. The last dose was administered a few weeks before the onset of laying eggs. The pullets were kept in separated installations on the laying farm considering special conditions like temperature and light among others.

6.1.2 Single-Level Analysis

In the initial analysis of these *Salmonella* data we ignore the repeated design of the data and assume that all observations act independently. With this assumption a logistic regression model

$$\text{logit}(P(Y_j = 1) = \mathbf{X^T}\boldsymbol{\beta} \tag{6.1}$$

can be fitted and risk factors investigated. The probability $P(Y_j = 1)$ that the jth sample was positive for *Salmonella* was predicted as a function of the explanatory variables contained in the $\mathbf{X}$ design matrix using the logit link function. The estimates of the model parameters, $\boldsymbol{\beta}$ were obtained using maximum likelihood estimation.

6.1.3 Two-Level Analysis

This analysis adjusts for the flock level by collapsing observations over samples in each sample type. The dust response was defined to be *Salmonella* positive (outcome=1) if at least one the dust samples was positive otherwise when all of the dust samples were negative the dust response was negative (outcome=0). Likewise the faeces response

was taken to be positive (outcome=1) if at least one of the faecal samples was positive and negative (outcome=0) otherwise. This thus reduces the seven responses recorded per flock to two responses: the 'dust response' and the 'faeces response'.

To explore the data, the frequencies of infected flocks were obtained using the two responses separately. Also, since the two outcomes occurred on each flock, it was important to examine the association between them. This was done using the Pearson Chi-square test of independence (FREQUENCY procedure in SAS). The measure of this association was examined using the Pearson correlation coefficient using the SAS CORRELATION procedure. The existence of an association signals the necessity for the two outcomes to be modeled jointly. The associations between each of the outcomes and each of the categorical variables were investigated using the Pearson chi-Square test of independence (FREQUENCY procedure in SAS). For the continuous explanatory variables, the mean values were estimated and compared for the positive and negative outcome categories.

The data can be analysed by performing separate analyses for the two outcomes, for example, by fitting a logistic model for the dust outcome and another logistic model for the faecal outcome. These separate analyses, however, would ignore the correlation between the two outcomes. Therefore it is appropriate to use the approaches which account for the correlation between the outcomes. We use the two general approaches, introduced in chapter one: generalized linear mixed models (GLMM) and marginal models such as the generalized estimating equations (GEE) and the alternating logistic regression model (ALR).

The two-level GLMM to model the probability $\pi_{is} = E(Y_{is}|u_i)$ that the sth sample type for the ith flock is *Salmonella* positive conditional on the random effect, u_i, for the ith flock is

$$\text{logit}(\pi_{is}) = \mathbf{X^T}\boldsymbol{\beta} + u_i \tag{6.2}$$

where $u_i \sim N(0, \sigma_u^2)$. Based on the underlying continuous variable coming from a logistic distribution, with a variance of $\pi^2/3$, which we substitute for the level 1 variance leads to a formulation of the intra-class correlation (ICC) (Browne *et al.* 2005) across flocks as

$$\frac{\sigma_u^2}{\sigma_u^2 + \pi^2/3}.$$

The marginal or population-averaged logistic model is expressed by

$$\text{logit}(\pi_{is}) = \mathbf{X^T}\boldsymbol{\beta} \tag{6.3}$$

where $\pi_{is} = E(Y_{is})$ is the marginal probability that the sth sample type for the ith flock is a positive case for *Salmonella*.

The parameters in (6.3) can be estimated using generalized estimating equations, introduced by Liang and Zeger (1986), by quasi-likelihood fitting. Instead of assuming a bivariate binomial distribution for (Y_{i1}, Y_{i2}), the quasi-likelihood method specifies a model for the means of the marginal distributions of Y_{i1} and Y_{i2}; a variance function describing how the variance of Y_{i1} and Y_{i2} depend on their means; and a pairwise correlation, $\text{corr}(Y_{i1}, Y_{i2}) = \rho$ between the outcomes. We assume an exchangeable working correlation structure. Essentially the correlation between the outcomes was estimated and then used to re-estimate the regression parameters and adjust the standard errors. An advantage of the GEE model is that the estimates are valid even if one misspecifies the variance-covariance structure (Agresti, 2002; Molenberghs and Verbeke, 2005).

Alternatively, instead of using the correlation to model the association between the repeated responses, odds ratios can be used and this is accomplished by the ALR model.

The GLMM model was fitted using the SAS GLIMMIX procedure while the the marginal models were fitted by the SAS GENMOD procedure. A parsimonious model was built based on the single-level logistic model by including one explanatory variable (two continuous and four categorical) at a time and the variables that had a p-value less than 0.25 were introduced in the multiple logistic regression models. A stepwise automatic selection procedure was also used to supplement the model selection. The two criteria led to the same model. Along with the selected main factors, their two-way interactions were added to the model. Higher interactions were not considered in order to keep a reasonable number of parameters in regard to estimation. However, two-way interactions between categorical variables, for instance, production type by vaccination status resulted into observations with only one type of the outcomes causing difficulties in estimation. The interactions between the categorical and continuous variables posed no estimation problems but were found to be non-significant. Therefore the final model considered eliminated the 'region' variable and the interactions.

6.1.4 Three-Level Analysis

The final analysis considers the seven responses by taking into account the samples (level 1) for each sample type (level 2) from each flock (level 3) that were tested for *Salmonella*. Thus we model the probability, $\pi_{jis} = E(Y_{jis}|u_i, u_{is})$ that the jth sample of sample type s for flock i was positive for *Salmonella*, as

$$\text{logit}(\pi_{jis}) = \mathbf{X}^{\mathbf{T}}\boldsymbol{\beta} + u_i + u_{is} \tag{6.4}$$

where $u_i \sim N(0, \sigma_u^2)$, and $u_{is} \sim N(0, \sigma_s^2)$. We assume the sample type and flock random effects are statistically independent. The formulation of the intra-class correlation (ICC) across flocks and sample type, respectively, are estimated as

$$\frac{\sigma_u^2}{\sigma_u^2 + \sigma_s^2 + \pi^2/3} \quad \text{and} \quad \frac{\sigma_u^2 + \sigma_s^2}{\sigma_u^2 + \sigma_s^2 + \pi^2/3}.$$

The model was fitted using the SAS GLIMMIX procedure.

6.2 Results

6.2.1 Data Exploration

In total, data were recorded for 148 flocks. In Figure 6.1, we show the number of flocks that were positive or negative for *Salmonella* for dust and faecal samples. The numbers at the top of the bars indicate the number of flocks in each category on the horizontal axis. Specific to the dust sample type, panels (a) show that in 102 flocks none of the dust samples were *Salmonella* positive whereas 22 flocks had one positive dust sample and 24 flocks had both dust samples positive. A similar interpretation follows for the faecal sample type. Grouping the results from panels (a) into *Salmonella* positive flocks (if at least one sample was *Salmonella* positive) and *Salmonella* negative flocks (if all samples were *Salmonella* negative) produced panels (b). Considering the dust sample type, for instance, 102 out of 148 flocks were *Salmonella* negative while the 46 were positive for *Salmonella*. The frequencies for the faecal sample type are interpreted in a similar manner. The Pearson chi-square statistic for the association between the two outcome variables was estimated as 66.60 ($p < 0.001$) which rejects the null hypothesis of no association between the dust and faecal outcomes. The Pearson correlation coefficient between the two outcomes was obtained as 0.6708 giving an indication of moderate to strong positive association. Table 6.1 shows the distribution of the number of *Salmonella* positive and negative flocks for each categorical explanatory variable. Also shown in the table are: the percentages of all flocks that were positive or negative and the association of each categorical variable with the presence of *Salmonella* using Pearson chi-Square test of independence. For both sample types there seems to be significant (p-values < 0.05) associations of production type and *Salmonella* vaccination status on the occurrence of *Salmonella*.

The boxplots in Figure 6.2 show the distributions of the continuous variables with the responses. The diamond in the box indicates the mean of the variables, the lower and upper hinges of the box show the 25% and 75% percentiles of the variables, while

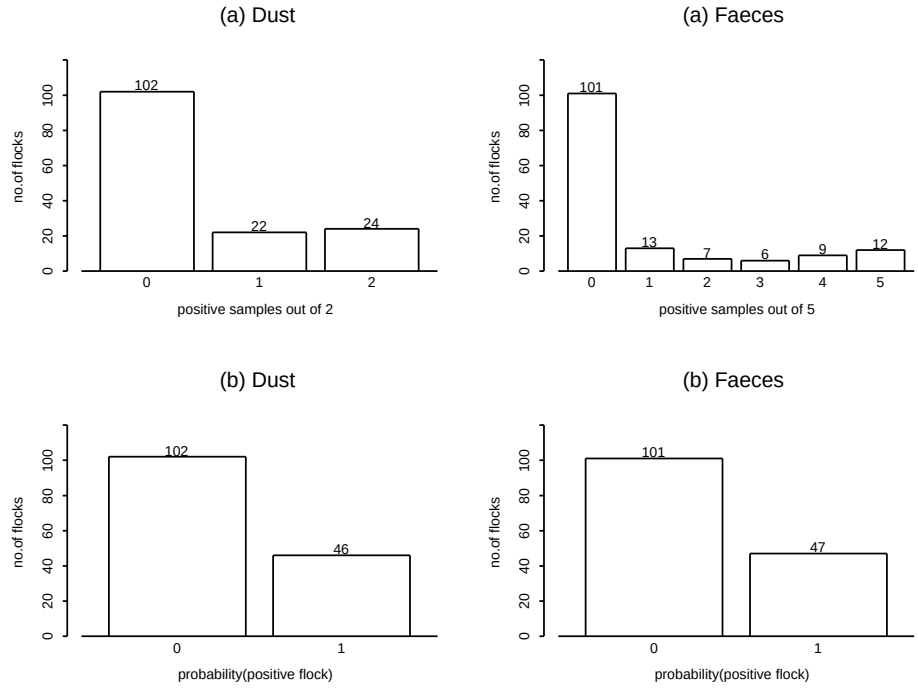

Figure 6.1: *Summary of data: Upper panels show Salmonella positive samples out of the 2 dust samples and out of the 5 faeces samples. Lower panels define a positive flock if at least one of the samples in the upper panel was positive.*

the line in the box is the median value, the ends of the vertical lines indicate the minimum and maximum variable values. From the figure we observe that for both age and flock size, the 25% percentile, the median (50% percentile), and the 75% percentile as well as the mean values are higher where *Salmonella* is present with the difference more evident for the flock size variable. For the *Salmonella* positive group, the flocks' mean age (in weeks) was 74.87 and 76.15 while the mean flock size was 21929.22 and 22156.6 for dust and faecal materials respectively. Similarly, for the *Salmonella* negative group, the mean age was 70.75 and 70.11 while the mean flock size was 13912.28 and 13727.1 for dust and faecal materials respectively. The mean age and mean flock size were higher for the *Salmonella* infected flocks than for the uninfected ones, suggesting an increase in risk for *Salmonella* as the hens get older and as the flock size increases. The points outside the ends of the vertical lines show

Table 6.1: *Frequency of Salmonella positive/negative (+ve/-ve) flocks (percentage of all 148 flocks) by categorical independent variables and sample type. Association P-values between each categorical variable and the presence/absence of Salmonella using Pearson Chi-Square test are shown.*

Variable	Dust sample type			Faecal sample type		
	+ve (%)	-ve (%)	χ^2 P-value	+ve (%)	-ve (%)	χ^2 P-value
Region			0.9698			0.5598
Flanders	38 (25.68)	84 (56.76)		40 (27.03)	82 (55.41)	
Walloon	8 (5.41)	18 (12.16)		7 (4.73)	19 (12.84)	
Sampling Month			0.6570			0.4347
February	2 (1.35)	4 (2.70)		3 (2.03)	3 (2.03)	
March	4 (2.70)	14 (9.46)		7 (4.73)	11 (7.43)	
April	5 (3.38)	16 (10.81)		4 (2.70)	17 (11.49)	
May	7 (4.73)	15 (10.14)		9 (6.08)	13 (8.78)	
June	12 (8.11)	16 (10.81)		7 (4.73)	21 (14.19)	
July	7 (4.73)	10 (6.76)		8 (5.41)	9 (6.08)	
August	4 (2.70)	8 (5.41)		3 (2.03)	9 (6.08)	
September	5 (3.38)	19 (12.84)		6 (4.05)	18 (12.16)	
Production Type			<0.0001			0.0002
Cage	45 (30.41)	69 (46.62)		45 (30.41)	69 (46.62)	
barn & free range	1 (0.67)	33 (22.30)		2 (1.35)	32 (21.62)	
Vaccination Status			0.0260			0.0573
vaccinated	22 (14.86)	68 (45.95)		22 (14.86)	68 (45.95)	
unvaccinated	22 (14.86)	26 (17.57)		21 (14.19)	27 (18.24)	
status unknown	2 (1.35)	8 (5.41)		4 (2.70)	6 (4.05)	

some extreme age and flock size values.

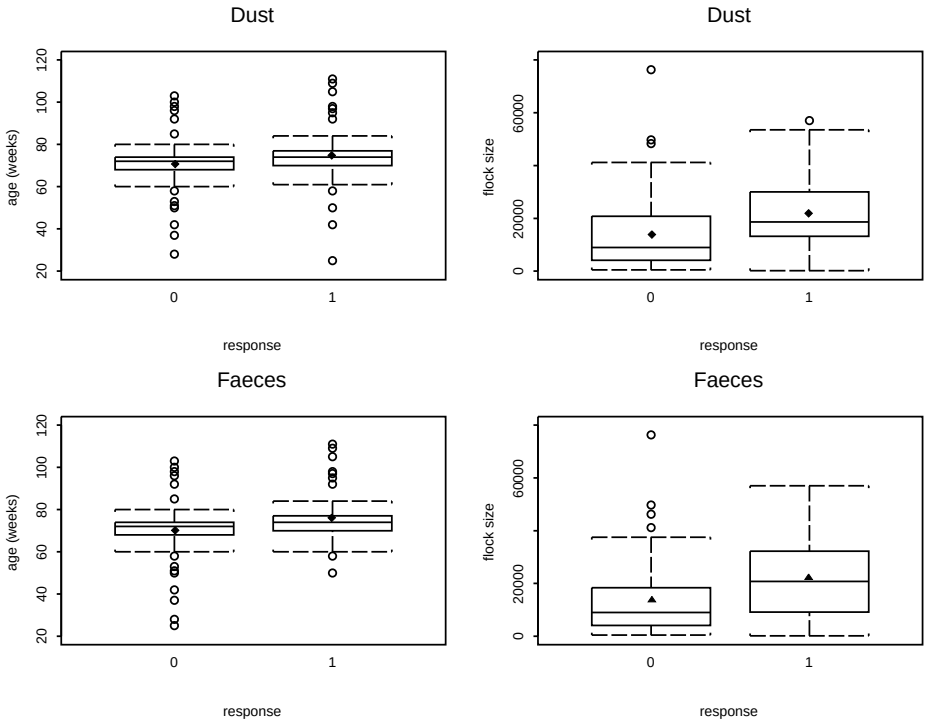

Figure 6.2: *Boxplots showing the presence (1) and absence (0) of Salmonella as related to the continuous independent variables, age and flock size for the dust samples (row 1) and faeces samples (row 2).*

6.2.2 Data Analysis

Results from single-level logistic regression model are compared with the GEE and ALR models (Table 6.2) and the two-level and three-level GLMM models (Table 6.3). Relative to the multilevel models, the standard errors from the single-level logistic regression are much smaller and lead all the predictors to be statistically significant risk factors for *Salmonella*. In contrast, neither the GEE nor the ALR nor the GLMM models identified a statistically significant vaccination status effect while from the GEE and ALR models there was no significant effect of the sampling month. Results from the GLMM models reveal borderline significant effect for the month of July relative to September. From the two-level models the the effect of age becomes borderline significant but nonsignificant with the three-level GLMM. Except for the two-level GLMM, which shows a significant effect of flock size, the other two-level and

three-level model results show borderline significant effects for flock size. Generally, from the three approaches accounting for correlated nature of data, a statistically significant effect for rearing hens in cages versus barns and free-range systems was found. The estimated population-averaged odds ratio of *Salmonella* to cage systems relative to barns and free-range was $e^{2.3434} = 10.41$ for GEE and 11.79 for ALR models.

The estimated flock-specific odds ratio of *Salmonella* to cage systems versus barns and free range was 14.69 and 14.30 for the two-level and three-level GLMMs respectively. The estimated variance of the flock-specific random effects was $\hat{\sigma}_u = 2.56$ for the two-level GLMM giving an estimated ICC of 0.437. In contrast the estimated exchangeable correlation (ICC) was $\hat{\rho} = 0.618$ for the model based on GEE. Compared to the exploratory measure of the Pearson correlation coefficient of 0.6708 that did not account for other factors, $\hat{\rho}$ was slightly lower. Parameters from mixed models and those from GEE can be compared via the relationship $\beta_{GLMM} \approx \beta_{GEE}\sqrt{1 + 0.346\hat{\sigma}_u^2}$ (Schukken *et al.* 2003; Kim *et al.* 2006). This is reflected in the estimates with the two-level GLMM estimates being greater in absolute value than the GEE estimates. The three-level analysis shows that the variation in the probability of *Salmonella* infection attributable to flocks is ten times greater than the variation attributable to sample type. This finding raises further issues for studying more explanatory variables to attempt to explain the variation.

Table 6.2: *Comparison of a naive (incorrect) single-level logistic regression model and two level logistic regression models fitted with GEE and ALR adjusting for the flock level, predicting probability of response.*

	Logistic		GEE		ALR	
Level(s)	Single		Two		Two	
Adjustment	None		Flock		Flock	
	$\hat{\beta}$ (SE)	P-value	$\hat{\beta}$ (SE)	P-value	$\hat{\beta}$ (SE)	P-value
Intercept	-3.59 (0.51)	<.0001	-5.39 (1.55)	0.0005	-5.52 (1.56)	0.0004
Sampling month						
February	-0.21 (0.34)	0.540	0.23 (1.03)	0.825	0.24 (1.02)	0.814
March	-0.56 (0.24)	0.019	-0.04 (0.74)	0.957	-0.02 (0.74)	0.979
April	-0.89 (0.25)	0.001	-0.68 (0.80)	0.400	-0.66 (0.80)	0.411
May	0.30 (0.21)	0.162	0.57 (0.78)	0.469	0.57 (0.78)	0.461
June	0.06 (0.18)	0.755	0.39 (0.69)	0.572	0.41 (0.69)	0.557
July	1.03 (0.21)	<.0001	1.21 (0.83)	0.144	1.25 (0.84)	0.134
August	0.31 (0.26)	0.228	0.12 (0.93)	0.899	0.14 (0.92)	0.883
[b]September						
Production type						
Cages	0.96 (0.19)	<.0001	2.34 (0.71)	0.001	2.47 (0.71)	0.001
[b]barns&free range						
Vaccination status						
vaccinated	-0.45 (0.14)	0.001	-0.52 (0.45)	0.247	-0.52 (0.45)	0.244
status unknown	0.31 (0.21)	0.147	-0.17 (0.61)	0.773	-0.20 (0.61)	0.745
[b]unvaccinated						
Age						
Age	0.02 (0.01)	0.010	0.03 (0.02)	0.090	0.03 (0.02)	0.097
Flock Size						
Flock Size	2.1E-5	0.001	3.2E-5	0.076	3.0E-5	0.079
	(6.1E-6)		(1.7E-5)		(1.7E-5)	
$\hat{\rho}$			0.6176			
pairwise log odds					2.88 (0.50)	<.0001

[b] denotes the baseline category. The corresponding effects are interpreted relative to the baseline.

Table 6.3: *Two-level GLMM and Three-level GLMM models predicting probability of response.*

Levels	Two		Three	
Adjustment	Flock		Flock and Sample type	
	$\hat{\beta}$ (SE)	P-value	$\hat{\beta}$ (SE)	P-value
Intercept	-6.14 (1.72)	0.001	-6.27 (1.77)	0.001
Sampling month				
February	0.23 (1.17)	0.841	-0.18 (1.24)	0.887
March	0.01 (0.84)	0.991	-0.46 (0.92)	0.618
April	-0.77 (0.85)	0.369	-1.06 (0.92)	0.252
May	0.71 (0.87)	0.420	0.54 (0.92)	0.552
June	0.49 (0.78)	0.531	0.30 (0.82)	0.714
July	1.52 (0.88)	0.087	1.71 (0.911)	0.061
August	0.19 (0.98)	0.849	0.54 (1.038)	0.605
[b]September				
Production type				
Cages	2.69 (0.92)	0.004	2.66 (0.92)	0.004
[b]barns&free range				
Vaccination status				
vaccinated	-0.63 (0.51)	0.215	-0.89 (0.55)	0.103
status unknown	-0.28 (0.94)	0.768	-0.065 (0.99)	0.947
[b]unvaccinated				
Age				
Age	0.03 (0.02)	0.079	0.02 (0.02)	0.234
Flock Size				
Flock Size	3.6E-5 (1.7E-5)	0.034	3.5E-5 (1.8E-5)	0.0530
$\hat{\sigma}_u^2$(Flock)	2.56 (0.72)		4.21 (0.85)	
$\hat{\sigma}_u^2$(Sample type)			0.41 (0.34)	

[b] denotes the baseline category. The corresponding effects are interpreted relative to the baseline.

6.3 Discussion

The prevalence of *Salmonella* in commercial holdings of laying hens in Belgium is relatively high, especially when compared to the northern European countries (EFSA, 2006b). However, it should be mentioned that Belgium has many laying hens compared to neighbouring countries (Quinet, 2005). The European survey was based on environmental sampling which is considered to be an accurate and representative indicator for the presence of *Salmonella* in layer flocks and for the probability that hens would lay contaminated eggs (Henzler *et al.* 1994; Kinde *et al.* 2005). The persistence of the pathogen in the intestinal tract is more important when infection occurs in young chicks, since bacterial clearance occurs more efficiently in adults. Genetically distinct lines of hens and various breeds can also be responsible for differences in the presence of *Salmonella* in the faeces of a contaminated animal. It is important to take these factors into account as the duration of this shedding can influence the detection of *Salmonella* in the threatening flocks (Kinde *et al.* 2000; Gast *et al.* 2005). Environmental sampling is not entirely reliable as it can miss flocks which passed the peak of infection but which are still producing contaminated eggs (Kinde *et al.* 1996; Davies and Breslin, 2004; Van Immerseel *et al.* 2005). The fact that one specific type of sample would be more contaminated than others helped identify risk factors, for example, a high level of the bacteria in dust (two dust samples positive instead of one) could point out a problem due to the ventilation system in the hen house or may be associated with cleaning and disinfection of the house, or with insufficient rodent control. A study from Gast *et al.* 1998 suggested that infection could, among other things, occur by oral ingestion of external surfaces contaminated by airborne movement of *Salmonella* during the feeding or pecking. From our findings, we saw differences in the statistical relations between the response variable and the predictors.

The major risk factor identified from the analyses was rearing flocks in cages compared to rearing in barns and free-range systems. The risk of contamination with *Salmonella* is thought to be higher when eggs are produced in non-cage systems, because of the greater exposure of layers to environmental contamination (Kinde *et al.* 1996; EFSA, 2004). However, in practice, control is not easier in cage layer houses; due to the difficulty to efficiently disinfect the cages and the higher densities of birds which produce a larger volume of contaminated faeces and dust (Davies and Breslin, 2004). The result of the current study clearly corroborates this finding. In addition, a clear difference was noticed in the proportions of vaccinated hens in the two types of production systems: 88% of the barn and free-range birds were vaccinated, while

only 53% for the cage system poultry. The vaccination variable can act here as a confounding factor on the apparent association between production type and *Salmonella* status. However, in the description of the sampled population of this present study, we noticed that the proportion of the "barn and free-range" category is relatively small (23%). Moreover, the very wide confidence intervals suggest that there might be a problem due to sample size.

Most of the studies have proven vaccination to be an important aid to reduce or possibly eliminate *Salmonella* Enteritidis from laying flocks (Davies and Breslin, 2001; 2003b; 2004). In the United Kingdom for instance, most of the laying flocks which have been implicated in the recent outbreaks of *Salmonella* Enteritidis in human beings were unvaccinated (Davies and Breslin, 2001). In the present analysis vaccination seemed not to have a significant protective effect. In the cases when *Salmonella* serovars other than *Salmonella* enteritidis are present concurrently in flocks vaccinated for *Salmonella* enteritidis, then considerably more contamination with these other *Salmonella* serovars may occur (Davies and Breslin, 2004). Another explanation why vaccination was less effective than expected, is that hens might have been infected before the vaccination was completed. Therefore it would have been interesting to exploit the period when the flock had been vaccinated as an explanatory variable. Such a variable was indeed available in the initial database but we chose to leave it aside for two main reasons. First, since the variables "vaccination status" and "vaccination period" were related to each other, we used only one of them to avoid multicollinearity problems. Second, from the description of the "vaccination period" variable, we had 88 holdings where vaccination was performed at rearing out of the 90 holdings where hens were vaccinated, leaving us with nothing to properly compare these findings with. Furthermore, effective protection owed to vaccination might occur only when the challenge dose is low. It is crucial to keep in mind that for vaccination to work effectively, an efficient cleaning and disinfection of laying houses between successive flocks is compulsory (Davies and Breslin, 2003b; Van den Bosch 2003). In this study, other factors like hygiene practices or pest control and their potentially confounding effects on the association between vaccination and the probability of being infected by *Salmonella*, were not taken into account.

The influence of temperature on the growth of *Salmonella* in food has been well documented. It is known that in all countries the incidence of human salmonellosis is highest during the summer (Baird-Parker, 1990; CNRSS, 2004; Kovats *et al.* 2004). Even though a statistically significant effect of the "month" variable is reported from our study, it is difficult to show the direction of the influence as only the month of July had borderline significance. Mollenhorst *et al.* 2005 came to the same conclusion.

During the summer season of the year 2003, a large increase of *Salmonella* infections was observed in Belgium and in The Netherlands. This increase could probably be attributed to the extremely hot weather during the summer of 2003. The Dutch study (Van Pelt *et al.* 2004) showed that a concomitant outbreak of *Salmonella* and avian influenza led to a shortage of eggs on the Dutch market, which was to be compensated for with imports, providing a reasonable explanation for this apparent seasonal trend. This present study showed no evidence of significant differences in the distribution of *Salmonella* among laying flocks according to regional repartition. Again we should note that the sample repartition is not really equitable, the Walloon holdings representing only 18%. On the other hand, the number of human salmonellosis cases across the country is clearly much higher in Flanders. Although the eggs produced in Belgium do not necessarily tend to be consumed locally, the food practices vary between both regions (CNRSS, 2004; AFSCA, 2006).
The impact of the age factor on the occurrence of *Salmonella* among egg laying flocks cannot really be established here, as it ranged from borderline to nonsignificant for the two and three level analysis.

Finally, other risk factors which were not considered in the present study are important to mention. For example, it could be useful to build a model taking into account flock characteristics (type of breed, number of flocks on the farm, multi-age farm or not), farm management (control of pest access, visitors allowed or not, feed composition and feeding practices, drinking water), cleaning and disinfecting practices related with the contamination status of the previous flock in the same hen house (Henzler & Opitz, 1992; Kinde *et al.* 1996; Shirota *et al.* 2000; Garber *et al.* 2003; Liebana *et al.* 2003; Kinde *et al.* 2005). Knowing that non-typhoid *Salmonella*e have very wide host ranges, it is important to take into consideration all various potential vectors surrounding the flock.

Chapter 7

Prevalence and Persistence of *Salmonella* in Belgian Broiler Chicken Flocks: An Identification of Risk Factors.

Broilers are an important source of salmonellosis after eggs, and pork. Salmonellosis is still one of the main causes of infectious food-borne gastroenteritis in humans worldwide (Bouwknegt *et al.* 2004; Collard *et al.* 2007; EFSA (European Food Safety Authority), 2007). In addition to the health consequences, *Salmonella* infection also has a severe economical impact (Collard *et al.* 2007; World health organisation (WHO), 2005). *Salmonella* spp with about 2600 existing serovars (Coburn *et al.* 2007), is responsible for human illness and thus causes a real public health issue (Altekruse *et al.* 2006; Collard *et al.* 2007; Van Immerseel *et al.* 2005). The symptoms in humans are most often characterised by the "non typhoid syndrome" which consists of an acute onset of fever, abdominal pain, nausea, and sometimes vomiting. These symptoms are self limiting in time. Humans become most often infected after consumption of contaminated eggs, poultry meat, or pork, or, less frequently bovine meat. In order to manage the risk to human health it is essential to tackle the problem at the farm level to reduce the cross contamination which can occur throughout the food chain process (Collard *et al.* 2007; Van Immerseel *et al.* 2005). Because animals most often are sub-clinically infected the disease tends to spread easily within a herd or flock,

and because animals can become intermittent or persistent carriers, it is not easy to detect the prevalence of *Salmonella* other than by routine sampling for bacteriology testing (EFSA (European Food Safety Authority), 2007).

Belgium has implemented a *Salmonella* eradication programme in poultry in accordance with the European legislation 2160/2003 (EU, 2003) in which a vaccination programme has been implemented in breeders and in layers but not in broilers because of the short life expectancy of broilers (42 days) (Anonymous, 2007; EFSA (European Food Safety Authority), 2004a). In broilers, a compulsory sampling, at least 3 weeks before slaughter, is requested from all farms with more than 5000 birds, as well as from farms who wish to trade their meat. A sanitary certificate is provided to the farm based on the results for *Salmonella* isolation (Anonymous, 1998; EFSA (European Food Safety Authority), 2004b). These results are requested by the slaughterhouses in order to programme their slaughter process, i.e., positive flocks must be slaughtered at the end of the day after slaughtering all negative flocks in order to avoid cross contamination within the slaughterhouse. Afterwards, the slaughterhouse is then thoroughly cleaned and disinfected. All positive farms are recorded in a notification system which exists since the 1st of January 2004 (Anonymous, 2007; EFSA (European Food Safety Authority), 2004b). One day-old chicks are sampled in the breeding house before being brought to the broiler farm. In Belgium, after a peak of infection in 1999, cases of salmonellosis in humans has been decreasing constantly, probably following vaccination and other sanitary measures implemented in poultry breeders and layers. In 2005 a total of 4872 human cases caused by *Salmonella* spp were registered (AFSCA (Agence Fédérale pour la sécurité de la chaîne alimentaire), 2007; Collard *et al.* 2007; EFSA (European Food Safety Authority), 2004b; Van Pelt *et al.* 2004).

Both vertical and horizontal transmissions play an important role in the contamination of flocks with *Salmonella*. Introducing only *Salmonella*-free chicks, e.g. by vaccinating the parental flocks against *Salmonella*, is an effective way to control the vertical transmission but will not prevent contamination of the birds with the environment if in addition no hygienic measures are taken simultaneously (Van Immerseel *et al.* 2005). Measures to reduce the horizontal transmission include: ensuring *Salmonella*-free feed and water, effective cleaning and disinfection of the farm, the use of food additives, applying all in all out procedures, appropriate biosecurity measures against animated or unanimated vectors, etcetera. (Anonymous, 2006; Davies and Breslin, 2001, 2004; Garber *et al.* 2003; Gradel and Rattenborg, 2003; Hald *et al.* 1998; Renwick *et al.* 1992; Skov *et al.* 1999; Skov *et al.* 2004; Van Immerseel *et al.* 2005; Wales *et al.* 2007; Wales *et al.* 2006). A detailed description of the hy-

gienic requirements for the farms in Belgium are described in the Belgian legislation (Anonymous, 1998). Even though these sanitary measures have been implemented as mentioned above, the burden of *Salmonella* infections on farms (mainly broilers) still exist, probably through contamination of the environment. For this purpose, further investigation of *Salmonella* on the broiler farms with different flocks in time was essential. The first objective of the study was to examine the potential risk factors contributing to *Salmonella* infection of the current broiler flock on the farm given the *Salmonella* status of the previous flock. The other objective was to investigate the risk factors associated with the persistence (positive test result for the previous and current flocks) of *Salmonella* infection on the farm.

Section 7.1 gives the data origins and the methods used to analyse these data. In Section 7.2, the exploratory results and the results from model fitting are presented and finally follows the discussion in Section 7.3. This work is submitted in Namata *et al.* (2008c).

7.1 Materials and Methods

7.1.1 Data Collection

The database of the 2005-2006 Belgium *Salmonella* control programme carried out by the Federal Agency for the Safety of the Food Chain was used to investigate the *Salmonella* status at the entrance of one-day old broiler chicks and the *Salmonella* status three weeks before slaughter (exit status). All the farms with more than 5000 birds and those willing to trade their meat must follow compulsory *Salmonella* sampling. Samples were taken by the farm owner. The epidemiological unit was a broiler flock. A flock is defined as a group of chicken belonging to the same herd, with the same sanitary and immune status, reared in the same room or barn, and having the following common characteristics: species, category (breeders, production), type (laying, broiler), stage of production (age), sanitary status (Anonymous, 1998, 2007; EFSA (European Food Safety Authority), 2004b). Each flock was sampled on the entrance day and three weeks before slaughter. To obtain the *Salmonella* status at entrance, day old chicks arriving from the reproduction holding were sampled by collecting specimens (20 pieces/flock, 5cm/5cm) of the inner lining of their transport boxes. The specimens were taken to the regional laboratory and tested for *Salmonella.* To obtain the *Salmonella* status at exit at about three to two weeks before slaughter, faeces samples were sampled by one of the three following sampling methods: 1) a pooled sample (60 x 1g) taken with swabs, 2) 60 pooled faecal samples (300 to 600

grams), 3) a pooled sample collected with 2 pairs of overshoes by walking in the barn. The samples were taken from different places of the barn where the flocks are kept and they were sent to an accredited laboratory within 48 hours according to standard norm ISO6579:2002 (Anonymous, 2007; ISO (Commité international de normalisation AW/9), 2002). A flock was considered positive when *Salmonella* was isolated from at least one sample and a farm was considered to be persistently positive if two consecutive flocks were positive on exit occasions. The information on the potential risk factors was obtained from a checklist questionnaire that was submitted to the different farmers during the 2003 Avian Influenza epidemic and answered on a voluntary basis. The risk factors which were investigated in our study are summarized in the data description part (Table 7.1). The information in the data set of the 2005-2006 Belgian *Salmonella* national control program in broilers and that of the data set identifying the risk factors were linked together using the farms identification numbers. The risk factors as well as the entrance Salmonella status for day-old chicks comprise the explanatory variables while the response variable refers to Salmonella status at exit. A more elaborate definition of the response variable follows in later sections.

7.1.2 Data Description

The design of the study was longitudinal with multiple observations collected on the same farms giving rise to correlated data. Table 7.1 shows the description of the variables that were recorded for the study. The response variables are binary outcomes of presence (outcome=1) or absence (outcome=0) of *Salmonella*. To get started, the data were re-structured to have the entrance outcome at a current occasion, the entrance outcome at the previous occasion, the exit outcome at a current occasion and the exit outcome at the previous occasion as separate variables. This implies that at least two flocks had to come on a farm thus eliminating farms that had one flock because they had no previous outcome. The interval in days between the consecutive flocks was calculated, thus creating a new variable ("duration"), split into three categories: less or equal to 6 weeks, between 6 and 12 weeks and over 12 weeks. The first objective of this chapter used the exit outcome at a current occasion as the response variable. The previous exit outcome along with the current entrance outcome and other explanatory variables were used as predictor variables. The current entrance outcome was considered as baseline. For the second objective a new binary variable was created and denoted 1 if the current and previous exit outcomes were both positive and 0 otherwise. The explanatory variables included continuous and

categorical variables.

Table 7.1: *Variable descriptions. The binary variables take the value of 1 for a 'yes' reply to the question and 0 for a 'no' reply.*

Variable Name	Description	Variable Type
FarmID	Identifier for a farm.	as given
Broiler houseID	Identifier for broiler house.	as given
Samplingdate	Date the sample was taken.	as given
Analysedate	Date the sample was analysed.	as given
ReferenceID	Identifier for a sample.	as given
Sampletype	Type of sample.	categorical
Entrance Result	positive Salmonella status for one-day old chicks?	binary
Exit Result	positive Salmonella status for adult broilers before going for slaughter?	binary
Province	Province the data was obtained.	categorical
Numberbroiler-houses	Number of broiler houses on a farm.	continuous
Numberbroilers (Nbroilers)	Number of broilers at the time of sampling.	continuous
Distance	The distance to the nearest poultry holding	continuous
Production Type	Place where the broilers are reared.	categorical
Shared materials	Are there shared materials in broiler houses?	binary
Species separation	Is there separation between birds of different species on a holding?	binary
Protection Net	Is there a net protecting broilers from wild birds when there is an open air production type?	binary
Pre-broilerhouse Disinfection	Is there one bucket to put in feet before entering the broiler house?	binary
Pre-broilerhouse-hygiene place (HP)	Is there a place for changing clothes before entering the broiler house?	binary
Broilerhouse HP	Is there one place for hygiene per broiler house?	binary
Hand-wash place	Is a place available to wash hands per HP?	binary
Undress place	Is a place available to undress per HP?	binary

Continued on next page

Table 7.1 – *continued from previous page*

Variable Name	Description	Variable Type
HP Disinfection	Is a bucket for disinfection per HP available?	binary
Visitors dress	Are clean clothes for visitors available?	binary
Ventilation	Is the ventilation mechanic?	binary
All-in All-out principle	Is the flock taken out and the place cleaned for about 3 days before the next flock is brought in?	binary
Cleaning Firm	Does an external firm clean the farm?	binary
Feed producer	Is the feed from an accredited producer?	binary
Town water	Is the water for drinking and cleaning from town?	binary
Outside feeding	Do the broilers feed outside?	binary
Temporary Workmen	Are temporary workmen present on the farm?	binary
International contacts	Do poultry or farmers have contact with foreign poultry or persons?	binary
External contacts	Do poultry or farmers have contact with external poultry or persons?	binary

The frequencies of the data variables were explored as a check for sparseness. Moreover, the associations between each of the categorical predictor variables with the responses were examined using the Pearson chi square test of independence. A probability value of less than 0.05 leads to rejection of the null hypothesis of independence. Because of the availability of many independent variables per place of hygiene in a broiler house, a study of multicollinearity was imperative. Multicollinearity refers to the fact that independent variables are correlated with one another (Agresti, 2002; Neter *et al.*, 1996). To check this, Pearson chi square test was used to investigate the presence of association between any two classification variables while Pearson correlation coefficient was used to give an indication of the magnitude of this association. The variables were considered highly associated if their Pearson correlation coefficient was greater than 0.7. Multicollinearity can have some serious effects on the values of the model coefficients. When there is a high degree of multicollinearity, small changes in the data can cause large changes in the values of the coefficients and some variables may appear to be completely redundant and may be excluded from the model. The relation of the continuous variables with the responses was investigated by estimating the difference in means between positive and negative *Salmonella* outcomes at exit.

7.1.3 Data Analysis

Like in the previous chapter we considered three statistical methods (also discussed in Chapter one), the random intercept generalized linear mixed model with a logistic link function and the marginal models: generalized estimating equations (GEE) and alternating logistic regression (ALR), to incorporate and study the clustered type of data on *Salmonella* in Belgian broiler chicken flocks' farms. The analyzes were twofold: the first analysis uses the current exit outcome conditional on the previous exit outcome as the response variable while the second uses the joint outcome that the current and previous exit outcomes were both positive for *Salmonella*, as response variable.

Conditional Analysis

For each farm i, we distinguished the previous entrance response Y_{it}^{e} as the entrance response at time t, the current entrance response Y_{it+1}^{e} as the entrance response at time $t+1$, the previous exit response Y_{it}^{o} as the exit response at time t, and the current exit response Y_{it+1}^{o} as the exit response at time $t+1$. The statistical methods simultaneously account for clustering and the influence of covariates. For particular values of the explanatory variable, $\mathbf{X_i} = (x_{i1} \dots x_{ip})$, we modeled the current exit probability of *Salmonella* adjusting for the previous exit outcome and the current entrance response for one-day old chicks as baseline, using the logistic model

$$\text{logit}[P(Y_{i\,t+1}^{o} = 1 | y_{it}^{o}, y_{i\,t+1}^{e})] = \beta_0 + \beta_1 y_{it}^{o} + \beta_2 y_{i\,t+1}^{e} + \sum_{p} \boldsymbol{\beta}_p \mathbf{X_{ip}} \qquad (7.1)$$

where $\boldsymbol{\beta}_p$ are effects of the p explanatory variables. Here, the GEE method solves score equations of a marginal formulation of the likelihood function and uses a working correlation matrix (for our case, the exchangeable structure) to adjust for the correlation within clusters. The estimation using ALR is via iterative recalculation of *Salmonella* clustering in the form of a pairwise odds ratio (assuming an exchangeable log odds structure) and logistic regression on the outcomes (Agresti, 2002; Carey *et al.* 1993).

In the GLMM we allowed the intercepts to vary for each farm and modeled the current exit probability of *Salmonella* adjusting for the previous exit outcome and the current entrance response Y_{it+1}^{e} using the extended logistic model

$$\text{logit}[P(Y_{i\,t+1}^{o} = 1 | u_i)] = \beta_0 + \beta_1 y_{i\,t}^{o} + \beta_2 y_{i\,t+1}^{e} + \sum_{p} \boldsymbol{\beta}_p \mathbf{X_{ip}} + u_i, \quad u_i \sim N(0, \sigma_u^2). \quad (7.2)$$

The model describes farm-specific intercepts instead of farm-averaged intercepts. Like in the previous chapter, the intra-class correlation across farms can be calculated as

$\sigma_u^2/(\sigma_u^2 + \pi^2/3)$. Note that these models are formulated as two-state discrete time Markov chains. We refer the reader to Agresti (2002) and Lindsey (1997) for more details.

Joint Analysis

The same models as in the conditional analysis are adapted to model the persistence of *Salmonella* infection on a farm. We modelled the probability that at two consecutive occasions a farm was infected using the marginal models

$$\text{logit}[P(Y_{it}^o = 1, Y_{it+1}^o = 1)] = \beta_0 + \sum_p \boldsymbol{\beta}_p \mathbf{X_{ip}} \tag{7.3}$$

and the farm-specific model

$$\text{logit}[(P(Y_{it}^o = 1, Y_{it+1}^o = 1)|u_i] = \beta_0 + \sum_p \boldsymbol{\beta}_p \mathbf{X_{ip}} + u_i, \quad u_i \sim N(0, \sigma_u^2). \tag{7.4}$$

Model Selection

The data constituted more than 20 potential predictor variables (Table 7.1). Selecting a model from all main effects and their two-way or higher interactions often leads to a selection from a very large number of effects and produces a model that overfits the data. Moreover, when these effects include classification variables with several levels, the number of parameters available for selection is even larger. To determine what main effects and interactions to allow, we considered the dependence of each of the variables on the response and the presence and magnitude of associations between predictor variables in order to avoid multicollinearity problems (see data description in Section 7.1.2). If multicollinearity existed, the choice of the variable to be included in the model was based on how strong it was related to the responses.

The model was constructed in a way that the response variable depends on the continuous variables and classification variables as well as on some two-way interactions of these effects. The pre-selected variables from above and their two way interactions were entered in the multiple logistic regression model which selected the parsimonious model using the automatic backward selection procedure implemented with the SAS LOGISTIC procedure. The selected variables using the automatic procedure were then entered in the multiple logistic GEE model, in the alternating Logistic model and in the random intercept GLMM model and the models fitted using the GENMOD and NLMIXED SAS procedures. The significance of each variable in the models was examined and if a variable appeared non-significant it was removed

from the model and the model was refitted. The reduced model was compared with the previous model using Akaike Information Criterion (AIC) for the GLMM model and using the Quasi under Independence model Criterion (QIC) for the marginal models. The smaller the criteria value the better a particular model fits. The QIC criterion proposed by Pan (2001) and further discussed by Hardin and Hilbe (2003) is an analogue to the AIC extending its applicability to quasi-likelihood models. Like the AIC, the QIC adds a penalty term of twice the number of parameters in the model to the quasi-likelihood. The final GLMM model was fitted with the GLIMMIX SAS procedure. In the next Section we present the estimated effects of the fitted models.

7.2 Results

This section presents the descriptive results and the results from model fitting. However, it is worth mentioning that when interpreting model fitting results caution must be taken with those risk factors involved in higher order interactions since the interpretation of effects related to interaction terms involves the description of the effects of one variable depending on the value of the other variable.

7.2.1 Data Description

A description of all variables used in this chapter is presented in Table 7.1. The frequencies, response rates and chi-square association probability values corresponding to the predictor variable categories in regard to both the conditional response (Table 7.2) and joint response (Table 7.3) are presented. Because of sparseness of data in some categories of province, we combined the provinces of Brabant Wallon, Hainaut, Liége, Luxembourg and Namur into the Walloon region (denoted 1) and the provinces of Antwerpen, Limburg, Oost-Vlaanderen, Vlaams Brabant and West-Vlaanderen to form the Flanders region (denoted 0). The upper part of the table includes binary predictor variables. During the period considered (2005-2006) 6824 broilers flocks on 723 farms were sampled. Of the 41 one-day old chicks which were positive for *salmonella* at the current entry, 19.51% (Table 7.2) resulted positive at the current exit occasion. Given the 404 flocks that were infected at the previous exit occasion, 27.97% were also infected at the current exit occasion (Table 7.2). None of the one-day old chicks were infected at two consecutive entrance occasions. The proportion of broiler flocks that were infected at two consecutive exit occasions was 1.66%.

For the conditional response (Table 7.2), the following variables with chi-square p-values less than 0.05 were observed to be associated with the probability of salmonella

infectivity of a current exit flock: a previous positive *Salmonella* status at exit, a positive *Salmonella* status of one-day old chicks of the current flock during entrance, availability of shared materials in broiler houses, having a separation between the different bird species, presence of a hand-wash place, use of an external cleaning firm, having temporary workmen, having poultry or farmers in contacts with foreign poultry or persons and rearing birds in the Walloon versus Flanders region. For the joint response (Table 7.3), we observed the following variables to influence the probability that farms were infected at two consecutive exit occasions: availability of shared materials in broiler houses, separation between different bird species, applying the all-in all-out principle, using an external cleaning firm, having temporary workmen, having poultry or farmers in contact with foreign poultry or persons, having poultry or farmers in contact with external poultry or persons, rearing birds in Walloon versus Flanders region and the duration in between consecutive flocks. For the distributions of conditional and joint responses with the continuous variables (lower panels of Table 7.2 and Table 7.3), number of broilers and number of broilerhouses, we see that the mean predictor values were higher for the infected flocks relative to the non-infected ones suggesting these variables to be possible risk factors. Also the mean distance to the nearest poultry holding was smaller for the infected groups than for the non-infected indicating that reduced distance to the nearest poultry holding might be a potential risk factor.

The findings on multicollinearity using Pearson chi square test for independence showed highly significant (pvalue $< .0001$) associations between the pairs of the following variables: having a hand-wash place per hygiene place (HP), having an undressing place per HP, availability of a disinfection bucket per HP, presence of visitors special clothing and feed from accredited producers. Table 7.4 presents their Pearson correlation coefficients and they range from 0.72 to 0.84. Feed from accredited producers and use of town water for drinking and cleaning were also highly associated with a correlation of 0.75. The presence of a hygiene place per broiler house was found to be associated with presence of the visitors' special clothing with a correlation of 0.70. Because the presence of a hand wash place per HP was more related to the responses (see χ^2 p-values, Table 7.2 and Table 7.3), it was used in substitute of the others to avoid multicollinearity. Some variables like production type and the number of broiler houses were not considered further for the analyzes due to a large portion of missingness. Observations for feed from accredited producers and use of town water for drinking and cleaning and outside feeding existed for one category of the joint response and thus could not be considered for analysis as they would be inestimable.

Table 7.2: *Distribution of the conditional response with the study variables based on 6824 flocks from 723 farms. Percentages (%) of positive flocks out of the flock observations for the designated categorical variables are shown along with their chi–square association p-values with the response. p-values < 0.05 show significant association. The mean values of the continuous predictors are estimated for positive and negative salmonella status.*

Binary	category 0		category 1		
Variable	flock observations	Positive flocks(%)	flock observations	Positive flock(%)	χ^2 p-value
Previous exit response	6420	3.99	404	27.97	<.0001
Current entry response	6783	5.32	41	19.51	<.0001
Shared materials	3702	4.38	3122	6.63	<.0001
Species separation	6501	5.26	323	8.36	0.0163
Protection Net	567	5.64	6257	5.39	0.7950
Pre-broilerhouse disinfection	354	4.80	6470	5.44	0.6052
Pre-broilerhouse hygieneplace	1096	4.84	5728	5.52	0.3611
Broilerhouse HP	447	7.16	6377	5.28	0.0903
Handwash place/HP	354	8.47	6470	5.24	0.0088
Undressplace/HP	343	7.58	6481	5.29	0.0679
HP disinfection	324	4.94	6500	5.43	0.7021
Visitors clothing	261	6.13	6563	5.38	0.5985
Mechanic ventilation	737	5.56	6087	5.39	0.8431
All-in All-out principe	1117	6.09	5707	5.27	0.2716
Cleaning firm	5137	5.88	1687	3.97	0.0027
Feed producer	232	3.02	6592	5.49	0.1015
Town water	326	3.99	6498	5.48	0.2455
Outside feeding	6803	5.41	21	4.76	0.8958
Temporary workmen	6287	5.22	537	7.64	0.0174
International contacts	6652	5.31	172	9.30	0.0222
External contacts	5926	5.30	898	6.12	0.3078
Region					
Walloon(1) vs Flanders(0)	5878	5.78	910	3.08	0.0008

Continued on next page

Table 7.2 – *continued from previous page*

Categorical Variable	flock observations	positive flocks(%)	χ^2 p-value		
Duration (in weeks)			0.2924		
up to 6	1181	6.18			
6 to 12	4537	5.11			
more than 12	1106	5.79			
Production type			0.9096		
Bio	1	0.00			
Cage	50	6.00			
Free range	45	4.44			
Barn	1789	4.02			
	Overall		Salmonella negative		Salmonella positive
Continuous Variable	Mean(SD)	Flock observations	Mean(SD)	Flock observations	Mean(SD)
Number of broilers	35657.27	6388	35160.57	367	44302.76
	(23404)		(22498.27)		(34572.74)
Number broilerhouses	1.8671	4941	1.8438	297	2.2559
	(1.2630)		(1.2463)		(1.4618)
Distance to poultryFarm	2.1289	5764	2.1657	317	1.4606
	(3.2293)		(3.2718)		(2.2238)

Note: 36, 4939, 69, 1586, 743 respectively, were missing data for region, production type, number of broilers, number of broilerhouses and distance to nearest holding.

Table 7.3: *Distribution of the joint response with the study variables based on 6824 flocks from 723 farms. Percentages (%) of positive flocks out of the flock observations for the designated categorical variables are shown along with their chi–square association p-values with the response. p-values < 0.05 show significant association. The mean values of the continuous predictors are estimated for positive and negative salmonella status.*

Binary	category 0		category 1		
Variable	flock observations	Positive flocks(%)	flock observations	Positive flock(%)	χ^2 p-value
Shared materials	3702	1.19	3122	2.21	0.001
Species separation	6501	1.46	323	5.57	<.0001
Protection Net	567	0.88	6257	1.73	0.1314
pre-broilerhouse disinfection	354	0.85	6470	1.7	0.2209
pre-broilerhouse hygieneplace	1096	1.19	5728	1.75	0.1834
Broilerhouse HP	447	2.68	6377	1.58	0.0779
Handwash place/HP	354	3.95	6470	1.53	0.0005
Undressplace/HP	343	3.21	6481	1.57	0.0209
HP disinfection	324	1.54	6500	1.66	0.8706
Visitors clothing	261	1.92	6563	1.65	0.7374
Mechanic ventilation	737	0.81	6087	1.76	0.0579
All-in All-out principle	1117	2.78	5707	1.44	0.0013
Cleaning firm	5137	1.95	1687	0.77	0.001
Feed producer	232	0.00	6592	1.71	0.0443
Town water	326	0.00	6498	1.74	0.0164
Outside feeding	6803	1.66	21	0.00	0.5515
Temporary workmen	6287	1.48	537	3.72	<.0001
International contacts	6652	1.56	172	5.23	0.0002
External contacts	5926	1.50	898	2.67	0.0104
Region					
Walloon(1) vs Flanders(0)	5878	1.80	910	0.77	0.0233
Categorical Variable	Flock observations	positive flocks(%)	χ^2 p-value		
Duration (in weeks)			<.0001		
up to 6	1181	3.81			
6 to 12	4537	1.28			

Continued on next page

Table 7.3 – *continued from previous page*

more than 12	1106	0.90			
Production type			0.0609		
Bio	1	0.00			
Cage	50	6.00			
Free range	45	0.00			
Barn	1789	1.45			
	Overall		*Salmonella* negative		*Salmonella* positive
Continuous	Mean(SD)	Flock	Mean(SD)	Flock	Mean(SD)
Variable		observations		observations	
Number of broilers	35657.27	6642	35370.45	113	52516.27
	(23404)		(22821.77)		(43082.64)
Number broilerhouses	1.8671	5137	1.8491	101	2.7822
	(1.2630)		(1.2405)		(1.9058)
Distance to poultryFarm	2.1289	5985	2.1445	96	1.1563
	(3.2293)		(3.2467)		(1.5783)

Note: 36, 4939, 69, 1586, 743 respectively, were missing data for region, production type, number of broilers, number of broilerhouses and distance to nearest holding.

It should be noted that all these results should be considered as indicative though not as formal inferential results, as they did not account for the clustered nature of the data. In the next section, models and methods for clustered data as introduced earlier on, will be used to identify risk factors for *Salmonella*.

7.2.2 Conditional Analysis

The results from the conditional analysis, which investigated the risk factors associated with the probability of *Salmonella* infection of a current flock at exit from the farm given the *Salmonella* status of the previous flock using generalized estimating equations, alternating logistic regression models and logistic-normal random intercept model (GLMM) are presented in Table 7.5. From the three approaches, 15 predictors were shown to be associated with *Salmonella* infection of the current broiler flock. One-day old chicks at entrance infected with *Salmonella* was a highly significant risk factor for *Salmonella* to the current flock on the farm. The estimated farm-averaged odds ratios of *Salmonella* to one-day old chicks were $e^{1.658} = 5.24$ and $e^{1.503} = 4.50$, respectively for GEE and ALR models while the estimated farm-specific odds ratio was $e^{1.481} = 4.4$ using the GLMM model. Generally, the three approaches produced

Table 7.4: *Pearson correlation coefficients for testing independence between any two of the designated covariates. the pvalue was <.0001 for all combinations rejecting the null hypothesis of independence. Covariates with a correlation greater than 0.70 were considered to be highly correlated pointing to multicollinearity.*

Variable	var1	var2	var3	var4	var5	var6	var7
var1: Broilerhouse HP	1.00	0.65	0.68	0.62	0.70	0.63	0.52
var2: Handwash place		1.00	0.78	0.78	0.80	0.80	0.59
var3: Undress place			1.00	0.72	0.81	0.73	0.60
var4: HP disinfection				1.00	0.84	0.84	0.62
var5: Visitors cloth					1.00	0.84	0.70
var6: Feed producer						1.00	0.75
var7: Town water							1.00

similar results in terms of statistical significance. Except for the one-day old chicks' predictor variable, the other predictors were found to interact with each other as they influenced *Salmonella* infection of the current broiler flock on the farm.

The impact of the *Salmonella* status of the previous flock on the probability of *Salmonella* for the current flock was found to depend, pair wise, on five other factors. From GEE and ALR models, while having a hygiene place for changing clothes before entering the broiler house increased the odds for *Salmonella* for the current flock when the previous flock was infected with *Salmonella*, the existence of the hygiene place decreased the risk when a previous flock was uninfected. With the GLMM model, the presence of a hygiene place decreased the odds for *Salmonella* when the previous flock was infected, but decreased further when the previous flock was uninfected. Also from the GLMM model, the use of mechanic ventilation decreased the odds for *Salmonella* when the previous flock was infected, but the risk decreased further when the previous flock was uninfected. Still, applying the all-in all-out principle or using an external cleaning firm or introducing a new flock on a farm at least six weeks after the previous flock, decreased the odds for *Salmonella* when the previous flock was infected, with further decrease when the previous flock was not infected.

The effect of the number of broilers on the occurrence of Salmonella to the current flock given the *Salmonella* status of the previous flock interacted with five other predictors. Separating between birds of different species or having a hygiene place for

changing clothes before entering the broiler house or region of location or employing an external cleaning firm to clean or using temporary workmen, decreased the odds for *Salmonella* when the number of broilers was less or equal to 2SDs from the mean number of broilers (Nbroilers $\leq$ 82465) (see Tables 7.2 or 7.3). With this number of broilers, the odds for *Salmonella* decreased further when in the Walloon region than with the Flanders region. Using the GLMM model, a larger number of broilers (Nbroilers = mean + 3SDs) increased the odds for *Salmonella* when an external firm cleaned, while from the ALR model this larger number of broilers increased the risk when there were temporary workmen. For illustration purposes, the interaction effect of the number of broilers and the external cleaning firm using the GEE model was derived as

Log(odds) = -3.985 + 3.1E-5*Nbroilers - 3.289*(0) + 3.4E-5*Nbroilers*(0) for Firm = 0

Log(odds) = -3.985 + 3.1E-5*Nbroilers - 3.289*(1) + 3.4E-5*Nbroilers*(1) for Firm = 1

The three models also revealed that separating between birds of different species or having a hand wash place in the hygiene place; decreased the odds for *Salmonella* with a unit increase in the distance to the nearest poultry holding. The GEE and ALR models showed that using mechanic ventilation reduced the odds for *Salmonella* when the distance to the nearest poultry holding was increased. While using temporary workmen increased the risk for *Salmonella* when there was a separation between birds of different species, the odds decreased when birds of different species were separated and there were no temporary workmen. Similarly, farms located in the Walloon region had increased odds for *Salmonella* when there was a separation between birds of different species, but the odds decreased when there was a separation between birds of different species for farms located in the Flanders region.

Using an external cleaning firm decreased the odds for *Salmonella* when there was a protection net sheltering the broilers from wild birds, but the odds went down further when the external firm was employed and the protection net was not available. The presence of a hand wash place decreased the odds for *Salmonella* when there were poultry or farmers in contact with external poultry or persons, but absence of a hand wash place and presence of external contacts led to an increase in the odds for *Salmonella*. Finally, using an external cleaning firm decreased the odds for *Salmonella* regardless of the existence of temporary workmen, but the odds decreased further when there were temporary workmen than when they did not exist.

The GLMM model estimated the variance of the farm-specific intercepts as $\sigma_u^2 = 0.6526$ giving an estimated intra-class correlation of 0.165. In contrast, the estimated exchangeable correlation based on GEE was $\hat{\rho} = 0.032$. The pairwise exchangeable odds ratio using the ALR was 0.758 and it is highly significant.

Table 7.5: *Parameter estimates and their standard errors and their significance p-values from the conditional analysis using GEE and ALR Marginal Models and the Random intercepts model (GLMM).*

	Marginal model GEE		Marginal model ALR		GLMM	
	$\hat{\beta}$ (SE)	pvalue	$\hat{\beta}$ (SE)	pvalue	$\hat{\beta}$ (SE)	pvalue
Intercept	-3.99(0.44)	<.0001	-3.93(0.47)	<.0001	-3.99(0.77)	<.0001
Main Effects						
Previous exit Y_t^o	1.80 (0.78)	0.0210	1.52(0.85)	0.073	0.32(1.00)	0.7469
Current entry Y_{t+1}^e	1.66 (0.52)	0.0015	1.50(0.57)	0.008	1.48(0.46)	0.0014
Number of broilers	3.1×10^{-5}	<.0001	3.1×10^{-5}	<.0001	3.1×10^{-5}	0.0003
	(3.7×10^{-6})		(4.5×10^{-6})		(8.5×10^{-6})	
Distance poultryFarm	-2.92(0.87)	0.0008	-2.92(0.78)	0.0002	-2.58(1.08)	0.0175
Species separation	2.45 (0.68)	0.0003	2.56(0.65)	0.0001	2.68(0.97)	0.0057
Protection Net	-0.77 (0.33)	0.0203	-0.65(0.41)	0.1111	-	-
Pre-broilerhouse HP	1.24 (0.32)	0.0001	1.27(0.32)	0.0001	0.97(0.41)	0.0190
Handwash place	-0.02 (0.31)	0.9415	-0.04(0.37)	0.9219	-0.70(0.62)	0.2566
Mechanic ventilation	-0.03 (0.26)	0.9045	-0.13(0.27)	0.6320	-0.17(0.27)	0.5286
All-in All-out Principle	0.06 (0.24)	0.8101	0.02(0.23)	0.9489	0.01(0.26)	0.983
Cleaning Firm	-3.29 (0.91)	0.0003	-2.88(0.97)	0.0029	-0.90(0.37)	0.0143
Temp workmen	-0.87 (0.48)	0.0728	-0.83(0.52)	0.1101	-0.48(0.34)	0.1537
External contacts	5.30 (1.73)	0.0021	5.24(1.52)	0.0006	4.38(2.03)	0.0311
Region: Wal vs Fla	-0.52 (0.30)	0.0797	-0.49(0.29)	0.0949	0.48(0.50)	0.339
Duration (weeks)						
dur1: 6 to 12	0.49 (0.24)	0.0375	0.49(0.24)	0.0406	0.58(0.25)	0.0188
dur2: > 12	0.86 (0.28)	0.0019	0.85(0.28)	0.0022	0.97(0.28)	0.0005
Interaction Effects						
PrevY*Pre-broh'seHP	1.90 (0.64)	0.0029	1.47(0.68)	0.0311	1.76(0.69)	0.0103
PrevY*principle	-1.36 (0.51)	0.0070	-1.21(0.53)	0.0228	-1.23(0.49)	0.0128
PrevY*Clean'gFirm	-1.04 (0.52)	0.0457	-	-	-1.03(0.47)	0.0268
PrevY*duration						

Continued on next page

Table 7.5 – *continued from previous page*

	Marginal model GEE		Marginal model ALR		GLMM	
	$\hat{\beta}$ (SE)	pvalue	$\hat{\beta}$ (SE)	pvalue	$\hat{\beta}$ (SE)	pvalue
*prevY*dur1*	-0.74 (0.41)	0.0690	-0.73(0.46)	0.1082	-0.97(0.39)	<.0129
*prevY*dur2*	-1.73 (0.69)	0.0118	-1.66(0.81)	0.0399	-1.91(0.58)	<.0010
PrevY*MechanicV	-	-	-	-	1.54(0.70)	0.0267
Nbros*Species	-8.3×10^{-5}	0.0012	-8.3×10^{-5}	0.0005	-8.0×10^{-5}	0.0051
	(2.6×10^{-5})		$(2.4\times10^{-}5)$		$(3.0\times10^{-}5)$	
Nbros*Pre-broh'seHP	-3.1×10^{-5}	<.0001	-3.1×10^{-5}	<.0001	-2.0×10^{-5}	0.0100
	(5.5×10^{-6})		(6.1×10^{-6})		(9.4×10^{-6})	
Nbros*Clean'gFirm	3.4×10^{-5}	<.0001	2.8×10^{-5}	0.0008	2.1×10^{-5}	0.0066
	(8.0×10^{-6})		(8.4×10^{-6})		(7.7×10^{-6})	
Nbros*region	-	-	-	-	-3.0×10^{-5}	0.0339
					(1.2×10^{-5})	
Nbros*Workmen	1.4×10^{-5}	0.0114	1.5×10^{-5}	0.0140	-	-
	(5.4×10^{-6})		(5.9×10^{-6})			
Distance*Species	-1.43 (0.36)	0.0001	-1.38(0.35)	0.0001	-1.38(0.51)	0.0070
Distance*Handwash	2.41 (0.85)	0.0046	2.43 (0.76)	0.0014	2.52(1.09)	0.0202
Distance*MechanicV	0.47 (0.18)	0.0070	0.46 (0.17)	0.0067	-	-
Species*Workmen	7.41 (1.48)	<.0001	7.12(1.44)	<.0001	6.98(1.84)	0.0002
Species*Region	5.05 (1.36)	0.0002	4.86 (1.31)	0.0002	4.21(1.87)	0.0245
Nets*Clean'gFirm	2.02 (0.68)	0.0030	1.71 (0.79)	0.0296	-	-
Handwash*External	-5.19 (1.73)	0.0027	-5.19(1.53)	0.0007	-4.31(2.05)	0.0352
Clean'gFirm*Workmen	-2.39 (0.68)	0.0004	-2.36(0.70)	0.0008	-	-
Association Estimates						
pairwise $\hat{\rho}$	0.03					
pairwise $\widehat{\text{OR}}$			0.76(0.17)	<.0001		
$\hat{\sigma_b^2}$ (Farm)					0.65(0.15)	

7.2.3 Joint Analysis

In the joint analysis, we investigated risk factors impacting the probability that two consecutive flocks at exit (previous and current) were positive for *Salmonella*. The results are shown in Table 7.6. The persistence of *Salmonella* on a farm by having two consecutive flocks with positive test results was associated with four variables in addition to five interaction terms.

Employing an external cleaning firm led to a decrease in the risk for persistent *Salmonella*. The estimated farm-averaged odds ratios of *Salmonella* were $e^{-1.645} = 0.23$ and $e^{-1.222} = 0.29$, respectively for GEE and ALR models while the estimated farm-specific odds ratio was $e^{-1.076} = 0.34$ using the GLMM model. Also, the duration between the consecutive flocks of at least six weeks led to a significant decrease of the risk for *Salmonella*. Furthermore, applying the all-in all-out principle decreased the risk for persistent *Salmonella* infection on a farm. In the GLMM model the effect of the number of broilers did not interact with other variables and it was found to increase the risk by a small magnitude but statistically significant (odds ratio=1.000014 and confidence interval [1.000001,1.000027]).

The odds for *Salmonella* decreased with an increase in the number of broilers (Nbroilers $\leq$ mean + 2SDs) when there were poultry or farmers in contact with external poultry or persons, but the odds further decreased when there were no external contacts. Also, the odds for *Salmonella* decreased with an increase in the number of broilers for farms located in the Flanders region, but decreased more for farms in the Walloon region. Although the risk for *Salmonella* decreased with the presence of a hand washing place whether or not there were temporary workmen, the risk decreased further when there were temporary workmen. While having poultry or farmers in contact with foreign poultry or persons increased the risk for *Salmonella* when there were temporary workmen, the risk decreased when there were international contacts but no temporary workmen. Likewise while external contacts increased the odds for *Salmonella* when there were temporary workmen, external contacts decreased the odds for *Salmonella* when there were no temporary workmen.

The estimated variance of the farm-specific random effects was $\sigma_u^2 = 3.178$ for the GLMM model giving an estimated intra-class correlation of 0.491. In contrast, the estimated exchangeable correlation was $\hat{\rho} = 0.009$ for the model based on GEE. The difference in the magnitudes of the parameter estimates from the GLMM models and those from GEE depends on the estimated random effects variance as shown by the relationship in the previous chapter.

Table 7.6: *Parameter estimates and their standard errors and their significance p-values from the joint analysis using GEE and ALR Marginal Models and the Random intercepts model (GLMM).*

	Marginal model GEE		Marginal model ALR		GLMM	
Effect	$\hat{\beta}$ (SE)	pvalue	$\hat{\beta}$ (SE)	pvalue	$\hat{\beta}$ (SE)	pvalue
Intercept	-4.84 (0.87)	<.0001	-5.18 (0.99)	<.0001	-5.42(1.08)	<.0001
Main Effects						
Number of broilers	1.4×10^{-5}	0.0082	2.2×10^{-5}	<.0001	1.4×10^{-5}	0.0369
	(5.4×10^{-6})		(5.5×10^{-6})		(6.8×10^{-6})	
Handwash place	2.51 (0.94)	0.0075	2.13 (0.98)	0.0286	0.99(1.03)	0.3408
All-in All-out Principle	-1.65 (0.61)	0.0071	-1.31 (0.46)	0.0048	-	-
Cleaning Firm	-1.46 (0.43)	0.0007	-1.22 (0.47)	0.0098	-1.08(0.51)	0.0350
Temp Workmen	3.29 (0.87)	0.0002	3.14 (1.29)	0.0151	4.42(1.57)	0.0048
Int'l contacts	-0.23 (0.67)	0.7353	-0.02 (0.83)	0.9774	0.11(1.07)	0.9202
External contacts	-2.52 (1.37)	0.0650	0.49 (0.50)	0.3294	-	-
Region: Wal vs Fla	2.08 (1.44)	0.1502	1.67 (1.35)	0.2138	-	-
Duration(weeks)						
6 to 12	-1.01 (0.26)	0.0001	-0.84(0.26)	0.0013	-1.04(0.27)	0.0001
>12	-1.36 (0.34)	0.0001	-1.28(0.36)	0.0003	-1.45(0.43)	0.0008
Interaction Effects						
Nbros*ExtContacts	6.4×10^{-5}	0.0011	-	-	-	-
	(1.9×10^{-5})					
Nbros*Region	-12.0×10^{-5}	0.0001	-12.0×10^{-5}	<.0001	-	-
	(3.1×10^{-5})		(2.5×10^{-5})			
Handwash*Workmen	-11.07 (2.87)	0.0001	-6.81(1.93)	0.0004	-6.31(1.98)	0.0014
Workmen*IntContacts	9.35 (2.44)	0.0001	4.92(1.68)	0.0034	4.670(1.96)	0.0171
Workmen*ExtContacts	4.960 (1.81)	0.0062	3.37(1.32)	0.0107	-	-
Association Estimates						
pairwise $\hat{\rho}$	0.01					
pairwise $\widehat{\text{OR}}$			3.19(0.41)	<.0001		
$\hat{\sigma}_b^2$ (Farm)					3.18 (0.43)	

7.3 Discussion

The investigations from this study showed that *Salmonella* infection in broiler chicken flocks involves several risk factors and their interactions. Multivariable logistic regression is a valuable tool to study risk factors in broilers (Chriel *et al.* 1999; Henken *et al.* 1992; Skov *et al.* 1999). In the analyzes presented here we have used generalized estimated equations (GEE), alternating logistic regression models (ALR) and random intercept GLMM, extensions of the ordinary logistic regression model to model correlated data, to determine risk factors based on the variables shown in Table 7.1. *Salmonella* prevalence for current broiler flocks conditional on the *Salmonella* status of the previous flock, according to 2005-2006 data, was estimated as 27.9% which is rather close to the community observed prevalence of 23.7% in the year 2005 (EU, 2005). Using the three modeling approaches, the conditional analysis revealed one-day old chicks infected with *Salmonella* as an important risk factor to a farm, as also observed in other studies by Kim *et al.* (2007) and Van Immerseel *et al.* (2004, 2005). Positive chicks can spread the infection through their faeces and quickly contaminate the farm. The boxes in which they arrive may constitute a way for introducing the infection as well (Kim *et al.* 2007; Renwick *et al.* 1992; Van Immerseel *et al.* 2004). Thus the first control measure is having *Salmonella* free breeding flocks (Bailey, 1993; Bouwknegt *et al.* 2004; Breytenbach, 2004; Collard *et al.* 2007; Garber *et al.* 2003; Skov *et al.* 1999; Van Immerseel *et al.* 2005; Van Immerseel *et al.* 2004). This can be easily achieved for instance through vaccination of parental lines. In Belgium since a few years, hatcheries have managed to obtain a good control of *Salmonella* infection even though at the time of this study, vaccination was only performed on a voluntary basis in breeders and in layers. Vaccination on broiler farms is never considered due to the short life expectancy of broilers and a diverse range of *Salmonella* serovars implicated. Thus vaccination of the broiler breeders is important, and has proven being effective in reducing the possibility of human infection through contaminated poultry products consumption (Cogan and Humphrey, 2003; Van Immerseel *et al.* 2005). Vaccination is now since June 2007 a legal obligation in Belgium in breeders and in layers (Anonymous, 2007).

The above risk factor is associated with the vertical transmission of *Salmonella*, but other factors associated with the horizontal transfer of *Salmonella*, mainly through the environment (Breytenbach, 2004; Davies and Breslin, 2003c; Kim *et al.* 2007; Renwick *et al.* 1992; Van Immerseel *et al.* 2004; Wales *et al.* 2007), were found in this study. *Salmonella*'s capability of resisting desiccation, allows it to survive for long periods in the environment. It has been found to remain for several months in dust

of ventilation filters (Davies and Wray, 1994; Kim *et al.* 2007; Renwick *et al.* 1992). A proper cleaning and disinfecting procedure conducted by external firms, especially trained for that purpose, seemed to be a major decreasing risk factor in our study and as proven before (Davies and Breslin, 2001, 2003a; Huneau-Salaun *et al.* 2007). A sanitary break (i.e. the duration between the previous and current flock of at least 6 weeks), or applying the all-in all-out procedure, or using mechanic ventilation, all contributed as well to reducing the risk of *Salmonella* to the current flock when the previous flock was infected.

An increase in risk was observed, according to the conditional analysis using GEE and ALR models, with having a hygiene place to change clothes prior to entering the broiler house when the previous flock was infected. This suggests that having proper biosecurity measures such as a clean hygiene place before entering a unit is probably not sufficient enough if a proper maintenance of those rooms is not ensured. Equipping the barns with individual hygiene places would only be effective if in addition the barns are equipped with their own individual ventilation systems, and extending biosecurity measures to all entering objects such as vehicles, litter, feed, water in order to be fully effective (Anonymous, 2006; De Zutter *et al.* 2001; Hald *et al.* 2000; Heyndrickx *et al.* 2002; Huneau-Salaun *et al.* 2007; Renwick *et al.* 1992; Wales *et al.* 2007). Management of those places, such as the cleaning and disinfecting procedure applied to them must be taken into account as well. Not only is it important to have an effective cleaning and disinfecting procedure, but also controlling its efficacy (Barker *et al.* 2003; Wales *et al.* 2007; Wales *et al.* 2006). A crucial element is the choice of the right products. Bacteria can persist in biofilms, which is organic matter accumulating, for instance, in water pipes (Garber *et al.* 2003; Morgan-Jones, 1980; Renwick *et al.* 1992; Van Immerseel *et al.* 2004). Chlorine which is often used to disinfect those systems does not remove organic matter. Therefore a possible cause of *Salmonella* presence could be due to these biolfilms (Alchalabi, 2007; Davies and Breslin, 2003c; Renwick *et al.* 1992; Ziggity Systems Inc, 2006).

From the joint analysis, studying the persistence of *Salmonella* on a farm, the main factors influencing this outcome were as previously seen in the conditional analysis, i.e., a cleaning and disinfection procedure conducted by an external firm, as well as applying all-in all-out procedure and at least a period of six weeks of sanitary break decreased the risk. A possible explanation for the increase in risk due to the interactions of external contacts and international contacts with temporary workmen could be an introduction of bacteria through contaminated tools or persons, as previously seen in other studies (Hald *et al.* 2000; Huneau-Salaun *et al.* 2007).

In conclusion, although a lot of risk factors have been investigated in this study,

due to sparseness of data, some of them, had to be omitted such as water, feed, and litter supply and the storing of these supplies. It would be interesting to also include in future studies the performance of the different ventilation systems, the temperature in the houses, as these factors have been recognized to greatly influence the poultry sector performance (Morrow, 2007). Measures against rodents, flies and manure disposing are key points in controlling the infection and avoiding persistence (Anonymous, 2006; Breytenbach, 2004; Davies *et al.* 1997; Henken *et al.* 1992; MacKenzie and Bains, 1976) and therefore should be considered to further enrich the Belgian database for studying potential risk factors contributing to *Salmonella* infection.

It is worth noting that this study was not explicitly designed for the study of risk factors associated with *Salmonella* infection but data on risk factors were obtained from the 2003 Avian Influenza check list, filled in on a voluntary basis by the farm owner. Nevertheless, risk factors recognized to play a critical role in avian influenza infection appear to be the same as those triggering *Salmonella* infection. Also, the fact that the farm owner was responsible for the filling in of the questionnaire and to collect samples might highlight a problem of bias in the data. To avoid such bias in the future it is important to have an independent person filling in the questionnaire in a standardized way as well as an independent standardized sampling method in order to have reliable good quality data.

Accounting for interactions leads to an improved determination of the risk factors that propagate the susceptibility to *Salmonella*. The epidemiological studies of *Salmonella* or other diseases should be designed with interactions in mind. The consistency in the results with the three modeling approaches is encouraging and strengthens their usefulness in identifying risk factors for *Salmonella* when faced with many variables and repeated data. These techniques can also handle higher order interactions than two-way interactions but these are seldom investigated due to small sample sizes.

Chapter 8

Concluding Remarks and Future Research

In this thesis we have focused on modeling binary response data of infectious diseases transmitted from human to human, from food/water to humans/animals but also from animals to humans. The sections below discuss some important issues pertaining to each part of the thesis.

8.1 Modeling Human to Human Infectious Diseases Data

In Part I of this research we used an approach to nonparametric modeling based on penalized splines using truncated power basis (Chapter 2 and 3). The penalized splines can be formulated as GLMM where the coefficients of the truncated power basis are allowed to vary randomly with an equal variance for each coefficient. It is important to realize that although GLMMs are used, the data analysed in Chapter 2 and 3 are cross-sectional, that is we have data at only individual level. However, in order to allow us to estimate the variance of the coefficients of the truncated power functions, we declare a second level with one unit which spans the entire data set. Note that this second level is a non-hierarchical classification.

Future research will be directed to extend the penalized spline model to public health problems that take place in the context of a hierarchical structure. Individuals may belong to one grouping at a given level of hierarchy and the grouping can be a

source of random variation. For example, for the study of Parvovirus B19 in Chapter 3, we can consider these data as clustered by country and instead of having a fixed effect for country we could treat the country effect as random thus yielding a country-specific P-spline model of the form

$$g(\boldsymbol{\pi}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + b_j$$

where g is the link function and b_j is the random intercept for country j.

GLMM models can be computationally intensive regarding direct calculations of the intractable integral over random effects in order to maximize the likelihood. The integration process can be more difficult with the GLMM fitting of the P-spline model because the dimension of the integral is equivalent to the number of knots and these can be as many as even 40. However, while marginal models like generalized estimating equations can provide a remedy to the maximum likelihood problem, they are specific to hierarchical data settings rather than the type of mixed models that arise in smoothing. Therefore, for GLMM fitting of the spline models we used a simple remedy to the maximum likelihood problem, which uses penalized quasilikelihood, implemented in the SAS procedure GLIMMIX.

The nonparametric modeling presented in Chapter 2 to 4 are known for their flexibility to capture hidden features of the data but this does not guarantee non decreasing estimates for prevalence. A nonmonotonic estimate of the prevalence leads to negative force of infection, which is nonsensical in epidemiology. To ensure monotonicity on the probability scale we have used the pool adjacent violater algorithm (PAVA) (Robertson *et al.* 1988) in the order of 'smooth then constrain'. That is, we applied the PAVA to the estimated prevalences and the force of infection is set to zero whenever it was negative. However, when model selection is an issue of concern, caution must be taken since the unconstrained predicted probabilities are used to compute the model deviance and the model selection criteria and these could influence the results. The impact can be removed by carrying out a constrained estimation in the order of 'constrain then smooth', which we recommend for future research.

8.2 Dose-Response Modeling for Food-borne Infectious Diseases

8.2.1 Single Strain

In this part we have shown that several models can be derived from the generic mechanistic model (5.3). The derivation of the Beta-Poisson model given by Haas *et*

al. (1999) was recapitulated in Chapter 5. Furthermore we derived the log-logistic, log-normal, extreme value models and the modified fractional polynomials. We have compared all these models for their ability to describe human dose-response data for *Salmonella Typhi* and *Campylobacter jejuni* and have found all of them to fit the data well. That is, none of the models was found to be substantially better than the others. The risk to *Salmonella Typhi* for an ingested dose of 100cfu microorganisms estimated from the set of 40 models ranged from 7.96 $\times 10^{-22}$ to 0.07407 while the risk to *Campylobacter jejuni* for a dose of 10cfu microorganisms was estimated from the set of models ranged from 4.52×10^{-11} to 0.3236. Clearly, the models vary a lot at low doses. This extreme variation between models at low doses illustrates the need for more data and the difficulty to generalize the risk estimate in order to develop regulations to protect public health. In Chapter 5 we have presented an illustration of incorporating model uncertainty into the risk estimation process based on the model averaging approach suggested by Buckland *et al.* (1997), where the model-specific risk estimate is weighted using Akaike weights to obtain an average risk estimate. Furthermore we found that averaging across a set of models that includes the modified fractional polynomials yields less biased and more precise risk estimates and attains coverage probabilities closer to the nominal 95% level compared to the set that does not include these fractional polynomials. Moreover model averaging performed better than selection of a single model.

Future research will focus on studying the performance of the proposed modified fractional polynomials in relation to the commonly used dose-response models using full Bayesian analysis.

8.2.2 Several Strains

In Chapter 5, sections 5.5 and 5.6, we have illustrated extrapolation to low doses for several strains of *Campylobacter jejuni* data for chicken using fixed effects and random effects models. These models are very important because data often fall into categories such as strains and one normally wants to control for characteristics of those categories that might affect the response variable. However, fixed effects models are not without their drawbacks. The fixed effects models may have many cross-sectional units of observations requiring many dummy variables for their specification. Too many dummy variables may sap the model of sufficient number of degrees of freedom for adequately powerful statistical tests. Moreover, a model with many such variables may be plagued with multicollinearity, which increases the standard errors and thereby drains the model of statistical power to test parameters. With

the *Campylobacter jejuni* data example, the models adjusting for all variables (host, origin, type) contained variables that do not vary within the strains and therefore parameter estimation was precluded. To over come this problem we partitioned the original strain data into groups according to possible combinations of the variables (origin and host) and proceeded to make analyses based on the formulated groups as the subjects. Because fixed effects models rely on within-group variation, we need repeated observations for each group, and a reasonable amount of variation of the predictor variables within each group. As a result groups with one or two observations were eliminated from the analysis. However, with the groups having at least three observations we were faced with the quasi-separation problem, which means that some linear combination of the predictor variables can be used to separate the dependent variable's 1's versus 0's and lead maximum likelihood estimates to go to infinity. As noted by Heinze (1999), separation primarily occurs in small samples with several unbalanced and highly predictive covariates, and this can be seen with our current study data set. Firth (1993) developed a procedure to reduce the bias of maximum likelihood estimates and this has proven to provide an ideal solution to monotone likelihood (Heinze and Schemper, 2002). Heinze and Ploner (2004) present the SAS macro, an S-PLUS library and an R package to apply Firth's procedure to logistic regression. We recommend a deeper look into the separation problem for future research.

One potentially significant limitation of fixed effects models is that we cannot assess the effect of variables that have little within-group variation. To be able learn about the effect of a variable that does not show much within-group variation, alternative models such as random effects models can be used. Random effects models can handle all the available data per group. Their disadvantage is that they can be computationally intensive and thus limit sensitivity analyses via simulations. Fitting these models using GLIMMIX or NLMIXED SAS procedures requires good starting values, which cannot be guaranteed within simulation runs. Random-effects models are desirable when there is no a priori knowledge on the strains or when only a few strains are expected to comprise a partition of interest.

Another alternative is to use fully Bayesian models where a priori information on strains is accounted for and summarised by the prior distributions assigned to each parameter in the model. This will be the focus for further research in order to elaborately compare between fresh isolates and laboratory isolates. A comparison of these isolates has been done using only the Beta Poisson model by Chen *et al.* (2006) but we will extend it to incorporate random effects models and model averaging.

An important aspect worth noting is that we have estimated the probability of

infection for each strain at a common dose. An alternative, which seems to be more reasonable regarding this data set, would be to fix the probability of infection and estimate the infectious dose for each strain or group. The latter estimation appears to be intriguing because of the widely varying scale of doses for each strain in the current study.

8.3 Modeling Data on *Salmonella* Infection in Broiler and Layer Chicken Flocks

In Part III we have studied risk factors for *Salmonella* in Belgian laying hens and broiler chicken flocks using three models: the marginal models, generalized estimating equations and alternating logistic regression model; and the random intercept model. This section presents some points to be considered in the analysis of repeated binary data. First, depending on the objective of the study, one can choose whether to use marginal models or random-effects models. If the goal is to study risk factors for Salmonella to a group of farms or at slaughterhouse the marginal models are suitable. However, when the association between repeated measurements is of interest the alternating logistic model is more appropriate than generalized estimating equations which treat the association as a nuisance. In contrast, if the goal is to study the farm risk factors for Salmonella, the random-effects model is more suitable because it allows adjustment on non-observed farm characteristics.

The choice of the structure of the covariance between the repeated responses is important. In the marginal models, the inferences on the parameter estimates are asymptotically valid under any assumed structure but it is better to choose a structure corresponding to the data. In contrast, in the random model, the fixed and random parameters are simultaneously estimated and the choice of the covariance structure influences the final results.

The estimates for the marginal models and the random models differ. Moreover, the interpretation of the estimated parameters is different. In the marginal models, the exponential of an estimated regression parameter is a farm-averaged odds ratio for Salmonella and concerns the sub-group of farms that shares a characteristic relative to the sub-group of farms not sharing this characteristic. In the random model, the exponential of an estimate is an odds ratio for a farm that has a characteristic relative to this same farm if it were free of this characteristic.

Marginal models are easy to implement and represent a first solution, but the random models, although more complex, use all available data and are more suitable

for explicative studies.

References

Aerts, M., Geys, H., Molenberghs, G., and Ryan, L.M. (2002). Topics in modelling of Clustered Data. Chapman & Hall.

AFSCA, Agence Fédérale de la Sécurité de la Chaîne Alimentaire. (2004). Salmonella, rapport annuel 2004.

AFSCA, Agence Fédérale pour la Sécurité de la Chaîne Alimentaire. (2006). Bulletin de l'Agence Fédérale pour la Sécurité de la Chaîne Alimentaire, Mai 2006, 4-5.

AFSCA (Agence Fédérale pour la sécurité de la chaîne alimentaire). (2007). Trends and Sources report on zoonotic agents in Belgium in 2005.

Agresti, A. (2002). Categorical Data Analysis. Second edition. Wiley, New York.

Alchalabi, D. (2007). Setting up on a small scale: how to design a broiler research unit. *Poultry International*, 8–11.

Altekruse, S., Koehler, J., Hickman-Brenner, F., Tauxe, R.V., and Ferris, K. A. (1993). Comparison of Salmonella enteritidis phage types from egg-associated outbreaks and implicated laying flocks. *Epidemiology and Infection* **110(1)**, 17-22.

Altekruse, S.F., Bauer, N., Chanlongbutra, A., DeSagun, R., Naugle, A., Schlosser, W., Umholtz, R., and White, P. (2006). Salmonella enteritidis in broiler chickens, United States, 2000-2005. *Emerging Infectious Disease* **12**, 1848-1852.

Anderson, L.J. (1987). Role of parvovirus B19 in human disease. *The pediatric Infectious Disease Journal,* **6**, 711-718.

Anderson M.J., Higgins P.G., Davis L.R., Willman J.S., Jones S.E., Kidd I.M., Pattison J.R. and Tyrrell D.A. (1985). Experimental parvoviral infection in humans. *J Infect Dis.* **152(2)**, 257-265.

Anderson, R.M., and May, R.M. (1982). Directly transmitted infectious diseases: control by vaccination. *Science* **215(4536)**, 1053-1060.

Anderson, R. M., and May, R. M. (1991). Infectious Diseases of Humans. Oxford, U.K. Oxford University Press.

Anonymous. (1998). Arrêté ministériel concernant les modalités d'application de l'arrêté royal du 10 août 1998 établissant certaines conditions pour la qualification sanitaire des volailles. 19.08.1998 In Moniteur belge, 1-4.

Anonymous. (2006). Salmonella Control in poultry from feed to farm. Workshop, Uppsala. Sweden.

Anonymous. (2007). Arrêté Royal relatif à la lutte contre les Salmonelles chez les volailles. In Moniteur belge.

Bailer, A.J, Noble, R.B., and Wheeler, M.W. (2005). Model uncertainty and risk estimation for experimental studies of quantal responses. *Risk Analysis* **25**, 292-299.

Bailey, J.S. (1993). Control of *Salmonella* and *Campylobacter* in poultry production. A summary of work at Russell Research Center. *Poult. Sci.***72**, 1169-1173.

Baird-Parker, A.C. (1990). Foodborne salmonellosis. Lancet **336**, 1231-1235.

Baker, R. (2002). *Natural history of hepatitis C.* NIH consensus report on hepatitis C Bethesda, Maryland.

Barker, J., Naeeni, M., and Bloomfield, S.F. (2003). The effects of cleaning and disinfection in reducing Salmonella contamination in a laboratory model kitchen. Journal of Applied Microbiology **95**, 1351-1360.

Barlow, R.E., Bartholomew, D.J., Bremner, M.J. and Brunk, H.D. (1972). *Statistical inference under order restriction*, New York: Wiley.

Becker, N.G. (1989). *Analysis of infectious disease data.* London: Chapman and Hall.

Black, R.E., Levine, M.M., Clements, M.L., Hughes, T.P., and Blaser, M.J. (1988). Experimental *Campylobacter jejuni* infections in humans. *Journal of Infectious Diseases*, **157(3)**, 472-479.

Bobashev, G.V, and Anthony, J.C. (1998). Clusters of Marijuana Use in the United States. American Journal of Epidemiology, **148**, 1168-1174

Bouwknegt, M., Dam Deisz, W.D.C., W.J.B., W., Van Pelt, W., Visser, G., and Van de Giessen, A.W. (2004). Surveillance of zooonotic bacteria in farm animals in the Netherlands. Results from January 1998 until December 2002. RIVM (Netherlands National Institute for Public Health and the Environment), 1-13.

Breytenbach, J.H. (2004). Salmonella control in Poultry. Intervet International b.v., 1-4.

Buckland, S. T, Burnham, K. P, and Augustin, N. H. (1997). Model selection: An integral part of inference. *Biometrics*, **53**, 603-618.

Burnham, K. P, and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach.* Second Edition. Springer Verlag, New York.

Burnham, K. P., Anderson, D. R., White, G. C., Brownie, C., and Pollock, K. H. (1987). Design and analysis methods for fish survival experiments based on release-recapture. American Fisheries Society, Monograph **5**.

Browne, W.J., Subramanian, S.V., Jones, K., and Goldstein, H. (2005). Variance partitioning in multilevel logistic models that exhibit overdispersion. J.R Statist. Soc.A, **168**, 599-613.

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

Chen, L., Geys, H., Cawthraw, S., Havelaar, A., and Teunis, P. (2006). Dose response for infectivity of several strains of Campylobacter jejuni in chickens. *Risk Analysis*, **26(6)**, 16131621.

Chriel, M., Stryhn, H., and Dauphin, G. (1999). Generalised linear mixed models analysis of risk factors for contamination of Danish broiler flocks with Salmonella typhimurium. *Preventive Veterinary Medicine* **40**, 1-17.

Claeskens, G. and Hjort, N.L. (2003). The Focussed Information Criterion, *Journal of the American Statistical Association*, **98**, 900-916.

CNRSS, Centre National de Référence des Salmonella et Shigella, 2004. Rapport Annuel 2004. Institut scientifique de Santé Publique.

Coburn B., Grassl G.A., and Finlay B.B. (2007). Salmonella, the host and disease: a brief review. Immunol. Cell Biol. **85**, 112-118.

Cogan T.A., and Humphrey T.J. (2003). The rise and fall of Salmonella Enteritidis in the UK. J. Appl. Microbiol. **94** Suppl, 114S-119S.

Coleman, M. and Marks, H. (1998). Topics in Dose-Response Modelling. *Journal of Food Protection*, **61(11)**, 1550-1559.

Collard, J.M., Bertrand, S., Dierick, K., Godard, C., Wildemauwe, C., Vermeersch, K., Duculot, J., F, V.A.N.I., Pasmans, F., Imberechts, H., and Quinet, C. (2007). Drastic decrease of Salmonella Enteritidis isolated from humans in Belgium in 2005, shift in phage types and influence on foodborne outbreaks. *Epidemiology and Infection*, 1-11.

Collard, J.M., Bertrand, S., Willems, L., Baeyens, D., De Cooman, F., Steenhaut, H., Lattuca, M., Mairiaux, E., Dupont, Y., Godard, C., Wildenauwe, C., and Vrints, M. (2004). Human Salmonellosis in Belgium: recent trends and outbreaks in 2003. Proceedings of Belgian symposium on Salmonella research and control in poultry.

Cossart, Y.E., Field, A.M., Cant, B., and Widdows, D. (1975). Parvovirus-like particle in human sera. *lancet*, **1**, 72-73.

Crainineanu, C.M., Ruppert, D., and Wand, M.P. (2004). Bayesian Analysis for Penalized Spline regression Using WinBUGS.

Davies, R., and Breslin, M. (2001). Environmental contamination and detection of *Salmonella* enterica serovar enteritidis in laying flocks. Veterinary Record, **149(23)**, 699-704.

Davies, R., and Breslin, M. (2003a). Observations on Salmonella contamination of commercial laying farms before and after cleaning and disinfection. *Veterinary Record* **152(10)**, 283-287.

Davies, R., and Breslin, M. (2003b). Effects of vaccination and other preventive methods for Salmonella enteritidis on commercial laying chicken farms. *Veterinary Record* **153(22)**, 673-677.

Davies, R.H., and Breslin, M. (2003c). Investigation of Salmonella contamination and disinfection in farm egg-packing plants. *Journal of Applied Microbiology* **94,** 191-196.

Davies, R. and Breslin, M. (2004). Observations on Salmonella contamination of eggs from infected commercial laying flocks where vaccination for Salmonella enterica serovar Enteritidis had been used. *Avian Pathology* **33(2)**, 133-144.

Davies R.H., and Wray C. (1994). An approach to reduction of Salmonella infection in broiler chicken flocks through intensive sampling and identification of cross-contamination hazards in commercial hatcheries. Int. J. Food Microbiol. **24**, 147-160.

Davies, R.H., Nicholas, R.A., McLaren, I.M., Corkish, J.D., Lanning, D.G., and Wray, C. (1997). Bacteriological and serological investigation of persistent *Salmonella* enteritidis infection in an integrated poultry organisation. *Veterinary Microbioly* **58**, 277-293.

De Buck, J., Pasmans, F., Van Immerseel, F., Haesebrouck, F. and Ducatelle, R. (2004a). Tubular glands of the isthmus are the predominant colonization site of Salmonella Enteritidis in the upper oviduct of laying hens. *Poultry Science* **83**, 352-358.

De Buck, J., Van Immerseel, F., Haesebrouck, F., and Ducatelle, R. (2004b). Recent insights on egg contamination and control. Proceedings of Belgian symposium on Salmonella research and control in poultry.

De Zutter, L., Herman, L., Heyndrickx, M., Butzler, J.P., and Vanderkerckhove, D. (2001). Studie van Salmonella en Campylobacter kringlopen bij de productie van braadkuikens. In Wetenschappelijke ondersteuning van een prenorminatief onderzoek in de voedingssector in het kader van een duurzame ontwikkeling.

Diamond, I.D., and McDonald J.M. (1992). Analysis of current-status data. In Demographic Application of Event History Analysis, Trussel J , Hankinson R , Tiltan J (eds), Chapter 12. Oxford University Press.

Edmunds, W.J., Gay, N.J., Kretzschmar, M., Pebody, R.G and Wachmann, H. (2000). the pre vaccination epidemiology of measles, mumps and rubella in Europe: implications for modelling studies, *Epidemiol. Infect.*, **125**, 635-650.

EFSA, European Food Safety Authority. (2004). Opinion of the AHW Panel related to the Welfare aspects of various systems of keeping laying hens. EFSA Journal **197**, 1-23.

EFSA (European Food Safety Agency). (2004a). Opinion of the scientific panel on biological hazards on the request from the commission related to the use of vaccines for the control of Salmonella in poultry. EFSA journals.

EFSA (European Food Safety Agency). (2004b). Trends and sources of zoonoses and zoonotic agents in humans, foodstuffs, animals and feedingstuffs. In Zoonoses monitoring, 1-30.

EFSA, European Food Safety Authority. (2006a). Trends and sources of zoonoses, zoonotic agents and antimicrobial resistance in the European Union in 2004. The EFSA Journal 2005, 23-95.

EFSA, European Food Safety Authority. (2006b). Preliminary Report. Analysis of the baseline study on the prevalence of Salmonella in laying hen flocks of Gallus gallus. EFSA Journal **81**, 1-71.

EFSA (European Food Safety Agency). (2007). Salmonella. In The community summary report on trends and sources of zoonoses, zoonotic agents, antimicrobial resistance and foodborne outbreaks in the European Union in 2005 pp. 27-81, 215-236/288.

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

Erkanali, A., Soyer, R. and Costello, E.J. (1999). Bayesian inference for prevalence in longitudinal two-phase studies, *Biometrics* **55**, 1145-1150.

EU (European Union). (2003). Reglement 2160/2003 Control of Salmonella and other foodborne zoonotic agents.

EU (European Union). (2005). Baseline survey on the prevalence of Salmonella in Broiler flocks of Gallus in the EU.

Faes, C., Geys, H., Aerts, M. and Molenberghs, G. (2003). On the use of fractional polynomial predictors for quantitative risk assessment in developmental toxicity studies. *Statistical Modelling*, **3**, 109-126.

Faes, C., Hens, N., Aerts, M., Shkedy, Z., Geys, H., Mintiens, K., Laevens, H. and Boelaert, F. (2006a). Random-effects models for clustered binary data using monotone fractional polynomials; with application to the estimation of a herd-specific force of infection. *Applied Statistics*, **55**, 595-613.

Faes, C., Aerts, M., Geys, H., and Molenberghs, G. (2006b). Model Averaging using Fractional Polynomials to Estimate a Safe Level of Exposure. *Risk Analysis*, **27**, 111-123.

FAO/WHO. (2003). Hazard characterization for pathogens in food and water: guidelines. *Microbiological risk assessment series*, no.3. WHO, Geneva, and FAO, Rome. Available at ftp://ftp.fao.org/docrep/fao/006/y4666e/y4666e00.pdf

Farrington, C.P. (1990), Modelling Forces of infection for measles, mumps and rubella. *Statist. Med.,* **9**, 953-967.

Farrington, C.P., Kanaan, M.N., and Gay, N.J. (2001). Estimation of the basic reproduction number for infectious diseases from age-stratified serological survey data (with discussion). *Appl. Statist.,* **50**, 251-292.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. Biometrika **80**, 2738.

Furumoto, W. A, and Mickey, R. (1967). A mathematical model for the infectivity-dilution curve of tobacco mosaic virus: Theoretical considerations. Virology, **32**, 216-223.

Garber, L., Smeltzer, M., Fedorka-Cray, P., Ladely, S., and Ferris, K. (2003). Salmonella enterica serotype enteritidis in Table Egg Layer House Environments and in Mice in U.S. Layer Houses and Associated Risk Factors. Avian Diseases **47**, 134-143.

Gast, R.K., Guard-Bouldin, J., and Holt, P.S. (2005). The relationship between the duration of fecal shedding and the production of contaminated eggs by laying hens infected with strains of Salmonella enteritidis and Salmonella Heidelberg. Avian Disease **49,** 382-386.

Gast, R.K., Mitchell, B.W., Holt, P.S. (1998). Airborne transmission of Salmonella enteritidis infection between groups of chicks in controlled-environment isolation cabinets. Avian Diseases **42(2)**, 315-320.

Gelfand, A.E., Ecker, M.D.,Chrisiansen, C., Mclaughlin, T.J. and Soumerai, S.B. (2000). Conditional categorical response with application to treatment of acute myocardial infraction, *Applied Statistics* **49**, 171-186.

Gilbert, G. L. (2000). Parvovirus B19 infection and its significance in pregnancy. Commun. Dis. Intell. **24**(Suppl.), 69-71.

Gilks, W.R., Richardson, S., and Spiegelhalter, D.J. (1996). *Markov Chain Monte Carlo in Practice.* Chapman and Hall, London.

Gradel, K.O., and Rattenborg, E. (2003). A questionnaire-based, retrospective field study of persistence of Salmonella Enteritidis and Salmonella Typhimurium in Danish broiler houses. Preventive Veterinary Medicine **56,** 267-284.

Grenfell, B.T. and Anderson, R.M. (1985). The estimation of age-related rates of infection from case notification and serological data, *J. Hyg. Camb.* **95**, 419-436.

Griffiths, D. (1974). A catalytic model of infection for measles. *Appl. Statist.*, **23**, 330-339.

Grummer-Strawn, L.M. (1993). Regression analysis of current status data: an application to breast feeding, *Biometrika* **72**, 527-537.

Haas, C. N, Rose, J. B, and Gerba, C. P. (1999). *Quantitative Microbial Risk Assessment.* Wiley, New York.

Hald, B., Olsen, A., and Madsen, M. (1998). Typhaea stercorea (Coleoptera: Mycetophagidae), a carrier of Salmonella enterica serovar Infantis in a Danish broiler house. Journal of Economic Entomology **91,** 660-664.

Hald, B., Wedderkopp, A., and Madsen, M. (2000). Thermophilic Campylobacter spp. in Danish broiler production: a cross sectional survey and a retrospective analysis of risk factors for the occurence in broiler flocks. Avian pathology **29,** 123-131.

Hardin, J.W. and Hilbe, J.M. (2003). Generalized Estimating Equations, Chapman & Hall/CRC: New York.

Heinze, G. (1999). Technical Report 10: The application of Firths procedure to Cox and logistic regression. Department of Medical Computer Sciences, Section of Clinical Biometrics, Vienna University, Vienna.

Heinze, G. and Schemper, M. (2002). A solution to the problem of separation in logistic regression. Statistics in Medicine **21**, 24092419.

Heinze, G. and Ploner, M. (2004). Technical Report 2: A SAS macro, S-PLUS library and R package to perform logistic regression without convergence problems. Department of Medical Computer Sciences, Section of Clinical Biometrics, Vienna University, Vienna.

Henken, A.M., Frankena, K., Goelema, J.O., Graat, E.A., and Noordhuizen, J.P. (1992). Multivariate epidemiological approach to salmonellosis in broiler breeder flocks. Poultry Science 838-843.

Hens, N.; Aerts, M.; Shkedy, Z.; Theeten, H.; Van Damme, P. and Beutels, P. (2008). Modelling multi-sera data: the estimation of new joint and conditional epidemiological parameters. Statistics in Medicine, **14**, 2651-2664.

Hens, N., Faes, C., Aerts, M., Shkedy, Z., Mintiens, K., Laevens, H. and Boelaert, F. (2007). Handling missingness when modelling the force of infection from clustered zeroprevalence Data. *Journal of Agricultural, Biological and Environmental Statistics*, **12**, 1-16.

Henzler, D.J., Ebel, E., Sanders, J., Kradel, D., and Mason, J. (1994). Salmonella enteritidis in eggs from commercial chicken layer flocks implicated in human outbreaks. Avian Diseases **38(1)**, 37-43.

Henzler, D.J., Kradel, D.C. and Sischo, W.M. (1998). Management and environmental risk factors for Salmonella enteritidis contamination of eggs. American journal of veterinary research **59**, 824-829.

Henzler, D. J., and Opitz, H. M. (1992). The role of mice in the epizootiology of Salmonella enteritidis infection on chicken layer farms. Avian Diseases **36(3)**, 625-631.

Heyndrickx, M., Vandekerchove, D., Herman, L., Rollier, I., Grijspeerdt, K., and De Zutter, L. (2002). Routes for salmonella contamination of poultry meat: epidemiological study from hatchery to slaughterhouse. Epidemioly and Infection **129,** 253-265.

Holcomb, D.L., Smith, M.A., Ware, G.O., Hung, Y.C., Brackett, R.E., and Doyle, M.P. (1999). Comparison of six dose-response models for use with food-borne pathogens. Risk Analysis, **19**, 10911100.

Hornick, R.B., Greisman, S.E., Woodward, T.E., DuPont, H.L., Dawkins, A.T., and Snyder, M.J. (1970). Typhoid fever: pathogenesis and immunological control. *The New England Journal of Medicine* **283(13)**, 686-691.

Huneau-Salaun, A., Denis, M., Balaine, L., and Salvat, G. (2007). Risk factors for Campylobacter spp. colonization in French free-range broiler-chicken flocks at the end of the indoor rearing period. Preventive Veterinary Medicine **80,** 34-48.

ISO (Commité international de normalisation AW/9). (2002). ISO 6579:2002 Microbiology - General guidance on methods for the detection of Salmonella - Amendment 1: Annex D: Detection of Salmonella spp. in animal faeces and in environmental samples from the primary production stage.

Jewell, N.P., and Van Der Leen, M. (1995). Generalizations of current status data with applications, *Lifetime data analysis*, **1**, 101-109.

Kang, S., Kodell, R. L, and Chen, J. J. (2000). Incorporating Model uncertainties along with Data Uncertainties in Microbial Risk Assessment. *Regulatory Toxicology and Pharmacology*, **32**, 68-72.

Keiding, N. (1991). Age-specific incidence and prevalence: a statistical perspective (with discussion). *J. R. Statist. Soc.* A, **154**, 371-412.

Keiding, N., Begtrup. K., Scheike. H.T., and Hasibeder, G. (1996). Estimation from current-status data in continuous time. *Lifetime Data Analysis*, **2**, 119-129.

Kim, A., Lee, Y.J., Kang, M.S., Kwag, S.I., and Cho, J.K. (2007). Dissemination and tracking of Salmonella spp. in integrated broiler operation. Journal of Veterinary Science **8**, 155-161.

Kim, H-Y., Preisser, J.S., Rozier, R.G., and Valiyaparambil, J.V. (2006). Multilevel analysis of group-randomized trials with binary outcomes. Community Dent Oral Epidemiol, **34**, 241-251.

Kinde, H., Read, D.H., Chin, R.P., Bickford, A.A., Walker, R.L., Ardans, A., Breitmeyer, R.E., Willoughby, D., Little, H.E., Kerr, D., and Gardner, I.A. (1996). Salmonella enteritidis, phase type 4 infection in a commercial layer flock in southern California: bacteriologic and epidemiologic findings. Avian Diseases **40(3)**, 665-671.

Kinde, H., Shivaprasad, H.L., Daft, B.M., Read, D.H., Ardans, A., Breitmeyer, R., Rajashekara, G., Nagaraja, K.V., and Gardner, I.A. (2000). Pathologic

and bacteriologic findings in 27-week-old commercial laying hens experimentally infected with Salmonella enteritidis, phage type 4. Avian Disease **44(2)**, 239-248.

Kinde, H., Castellan, D.M., Kerr, D., Campbell, J., Breitmeyer, R., and Ardans, A. (2005). Longitudinal Monitoring of Two Commercial Layer Flocks and Their Environments for Salmonella Enterrica Serovar Enteritidis and Other Salmonellae. Avian Diseases **49**, 189-194.

Koch, W.C. and Adler, S.P. (1989). Human parvovirus B19 infections in women of childbearing age and within families. *The Prediatric infectious Diseases Journal,* **8**, 83-87.

Kodell, R.L., Kang, S., and Chen, J.J. (2002). Statistical models of health risk due to microbial contamination of foods. *Environmental and Ecological statistics*, **9**, 259-271.

Kovats, R.S., Edwards, S.J., Hajat, S., Armstrong, B.G., Ebi, K.L., and Menne, B. (2004). The effect of temperature on food poisoning: a time-series analysis of salmonellosis in ten European countries. Epidemiology and Infection **132(3)**, 443-453.

Kunkel Dennis Microscopy, Inc. (2007). Electron Microscopy Science Stock Photography. Available at http://www.denniskunkel.com/.

Liang, K.Y., and Zeger, S.L. (1986). Longitudinal Data Analysis Using Generalized Linear Models, Biometrika **73**, 13-22.

Liebana, E., Garcia-Migura, L., Clouting, C., Clifton-Hadley, F.A., Breslin, M., and Davies, R. H. (2003). Molecular fingerprinting evidence of the contribution of wildlife vectors in the maintenance of Salmonella Enteritidis infection in layer farms. Journal of Applied Microbiology **94(6)**, 1024-1029.

Lindsey, J.K. (1997). Applying Generalized Linear Models. Springer, New York.

MacKenzie, M.A., and Bains, B.S. (1976). Dissemination of Salmonella serotypes from raw feed ingredients to chicken carcases. Poultry Science **55,** 957-960.

Marks, H.M., Coleman, M.E., Lin, C.J., and Roberts, T. (1998). Topics in microbial risk assessment;dynamic flow tree process. *Risk Analysis*, **18,** 309-328.

Matheï, C., Buntix, F., and Van Damme, P. (2002). Seroprevalence of hepatitis C markers among intravenous drug users in western European countries. *A systematic review J Viral Hepatitis.* 9,1-17.

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

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models.* Chapman and Hall. New York.

McCulloch, C.E. and Searle, S.R. (2001). *Generalized, Linear, and Mixed Models.* New York: Wiley.

Microsoft® Encarta® Online Encyclopedia. (2008). Salmonella, Available at http://encarta.msn.com/salmonella.html

Molenberghs, G., and Verbeke, G. (2005). *Models for Discrete Longitudinal Data.* Springer Verlag, New York.

Mollenhorst, H., van Woudenbergh, C.J., Bokkers, E.G., and De Boer, I.J.M. (2005). Risk factors for Salmonella enteritidis infections in laying hens. Poultry Science **84(8)**, 1308-1313.

Moon, H., Kim, H., Chen, J. J, and Kodell, R. L. (2005). Model averaging using Kullback information criterion in estimating effective doses for microbial infection and illness. *Risk Analysis*, **25(5)**, 1147-1159.

Morgan-Jones, S.C. (1980). The occurrence of salmonellae during the rearing of broiler birds. Poultry Science **21**, 463-470.

Morrow C. (2007). Diseases, Poisons and Toxins.

Mossong, J., Hens, N., Friederichs, V., Davidkin, I., Broman, M., Litwinska, B., Siennicka, J., Trzcinska, A., Van Damme, P., Beutels, P., Vyse, A., Shkedy, Z., Aerts, M., Massari, M. and Gabutti, G. (2008). Parvovirus B19 infection in five European countries: seroepidemiology, force of infection and maternal risk of infection. *Epidemiology and infection*, **00**, 000-000

Muench, H. (1959). Catalytic models in epidemiology. Boston: Harvard University Press.

Muench, H. (1934). Derivation of rates from summation data by the catalytic curve, *Journal of the American statistical association*, March Edition, 25-38.

Nagelkerke, N., Heisterkamp, S., Borgdorff, M., Broekmans, J., and Houwelingen, H.V. (1999). Semi-parametric estimation of age-time specific infection incidence from serial prevalence data. *Statistics in Medicine*, **18**, 307-320.

Namata, H., Aerts, M., Faes, C., and Teunis, P. (2008b). Model Averaging in Microbial Risk Assessment Using Fractional Polynomials. Risk Analysis, **28(4),** 891-905.

Namata, H., Méroc, E., Aerts, M., Faes, C., Cortiñas, A.J., Imberechts, H., and Mintiens, K. (2008a). Salmonella in Belgian laying hens: An identification of risk factors. *Preventive Veterinary Medicine*, **83(3-4)**, 323-336.

Namata, H., Shkedy, Z., Aerts, M., Faes, C., Hens, N., Van Damme, P., and Beutels, P. (2008d). Modeling the Force of Infection for Parvovirus B19 in Europe Using Penalized Spline Models. *Technical Report.*

Namata, H., Shkedy, Z., Aerts, M., Faes, C., Mathei, C., Kretzschmar, M., Wiessing, L., Mravcik, V., Suligoi, B., Nordén, L. and Vallejo, F. (2008e). Estimation of the Prevalence and Force of Infection of Hepatitis C among Injecting Drug Users in Five European Countries. *Technical Report.*

Namata, H., Shkedy, Z., Faes, C, Aerts, M., Molenberghs G., Theeten, H., Van Damme, P. and Beutels, P. (2007). Estimation of the Force of Infection from Current Status Data Using Generalized Linear Mixed Models. *Journal of Applied Statistics*, **8**, 1-17.

Namata, H., Welby, S., Aerts, M., Faes, C., Cortiñas, A.J., Imberechts, H., Vermeersch, K., Hooyberghs, J., and Mintiens, K. (2008c). Identification of Risk Factors for the Prevalence and Persistence of Salmonella in Belgian Broiler Chicken Flocks. *Preventive Veterinary Medicine*, Submitted.

Neter, J., Kutner, M.H., Nachtsheim, C.J., and Wasserman, W. (1996). Applied Linear Statistical Models (4th Edition). Burr Ridge, IL: Irwin.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. Biometrics, **57**, 120–125.

Pinsky, P.F. (2000). Assessment of risk from long term exposure to waterborne pathogens. *Environmental and Ecological statistics*, **7**, 155-175.

Popoff, M.Y. (2001). Antigenic Formulas of the Salmonella Serovars. 8th edition. WHO Collaborating Centre for Reference and Research on Salmonella. Institut Pasteur. Paris, France.

Quinet, C. (2005). La Salmonellose aviaire, état des lieux. Arsia infos **20**, 1-2.

Rahman, H.J., Wafield, J.C., Stephens, D.A and Falcoz, C. (1999). The bayesian analysis of pivotal pharmacokinetic study, *Statistical methods in medical research* **28**, 195-216.

Renwick, S.A., Irwin, R.J., Clarke, R.C., McNab, W.B., Poppe, C., and McEwen, S.A. (1992). Epidemiological associations between characteristics of registered broiler chicken flocks in Canada and the Salmonella culture status of floor litter and drinking water. Canadian Veterinary Journal **33**, 449-458.

Robertson, T., Wright, F.T., and Dykstra, R.L. (1988). *Order Restricted Satistical Inference.* NewYork: Wiley.

Royston, P, and Altman, D. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion). *Applied Statistics*, **43**, 429-467.

Ruppert, D., Wand, M.P. and Carrol, R.J. (2003). *Semiparametric Regression.* NewYork: Cambridge University Press.

SAS Institute Inc. (2004). The GLIMMIX *Procedure (Experimental).* Cary, NC: SAS Institute Inc.

Schukken Y.H., Grohn Y.T., McDermottb B., and McDermottc J.J. (2003). Analysis of correlated discrete observations: background, examples and solutions. Preventive Veterinary Medicine, **59**, 223-240.

Shirota, K., Katoh, H., Ito, T., and Otsuki, K. (2000). Salmonella Contamination in Commercial Layer Feed in Japan. Journal of Veterinary Medicine Science **62(7)**, 789-791.

Shkedy, Z., Aerts, M., Molenberghs, G., Beutels, P., and Damme, P.V. (2006). Modelling age-dependent force of infection from prevalence data using fractional polynomials. *Statistics in Medicine*, **25**, 1577-1591.

Shkedy, Z., Aerts, M., Molenberghs, G., Beutels, P., and Damme, P.V. (2003). Modelling forces of infection by using monotone local polynomials. *Appl. Statist.*, **52**, 469-485.

Shkedy, Z., Namata, H., Kasim, A., Maringwa, J.T., Aerts, M., Wiessing, L., Kretzschmar, M. (2008). Cross Sectional and Longitudinal Evidence for Co-Infection of HCV and HIV. *Technical Report.*

Skov, M.N., Angen, O., Chriel, M., Olsen, J.E., and Bisgaard, M. (1999). Risk factors associated with Salmonella enterica serovar typhimurium infection in Danish broiler flocks. Poultry Science **78**, 848-854.

Skov, M.N., Spencer, A.G., Hald, B., Petersen, L., Nauerby, B., Carstensen, B., and Madsen, M. (2004). The role of litter beetles as potential reservoir for Salmonella enterica and thermophilic Campylobacter spp. between broiler flocks. Avian Disease **48**, 9-18.

Spiegelhalter D.J., Best N.G., and Carlin B.P. (1998). Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models, *Research report 98-009, Division of Biostatics, University of minisota.*

Spiegelhalter D.J., Best N.G., Carlin B.P., and Van der linde, A. (2002). Bayesian measures of model complexity and fit. *J. R. Statistical Soc. B* **64**, 1-34.

Sutton, A.J., Hope, V.D., Ncube, F., Mathei, C., Mravcik, V., Sebakova, H., Vallejo, F., Sullgoi, B., Wiessing, L. and Kretzschmar, M. (2008). A comparison between the force of infection estimates for blood-borne viruses in injecting drug user population across the European union - A modeling study. Journal of Viral Hepatitis. accepted.

Teunis, P. F. M., and Havelaar, A. H. (2000). The Beta Poisson model is not a single hit model. *Risk Analysis*, **20(4)**, 511-518.

Teunis, P. F. M., Nagelkerke, N. J. D., and Haas, C. N. (1999). dose-response models for infectious gastroenteritis. *Risk Analysis.* **19(6)**, 1251-1260.

Teunis, P. F. M., Takumi, K. and Shinagawa, K. (2004). dose-response for infection by Escherichia coli O157:H7 from outbreak data. *Risk Analysis*, **24(2)**, 401-408.

Teunis, P. F. M., Van der Heijden, O. G, Van der Giessen, W. B, and Havelaar, A. H. (1996). The Dose-Response Relation in Human Volunteers for Gastro-intestinal Pathogens, Report Nr. 284550002. National Institute of Public Health and the Environment (RIVM), Bilthoven, The Netherlands.

Valeur-Jensen, A., Pedersen, C., Westergaard, T., Jensen, I., Lebech, M., Andersen, P., Aaby, P., Pedersen, B. and Melbye, M. (1999). Risk factors for parvovirus

B19 infection in pregnancy. *Journal of the American Medical Association*, **281**, 1099-1105

Van den Bosch., G. (2003). Vaccination versus treatment: How Europe is tackling the eradication of Salmonella. Asian Poultry Magazine, 2-4.

Van Effelterre, T.; Shkedy, Z.; Aerts, M.; Molenberghs, G.; Van Damme, P. and Beutels, P. (2008). Contact patterns and their implied basic reproductive numbers: an illustration for varicella-zoster virus. *Epidemiology and Infection*, **00**, 000-000

Van Immerseel, F., De Buck, J., Boyen, F., Pasmans, F., Bertrand, S., Collard, J.M., Saegerman, C., Hooyberghs, J., Haesebrouck, F., and Ducatelle, R. (2005). Salmonella dans la viande de volaille et dans les oeufs: Un danger pour le consommateur qui demande la mise en place d'un programme de lutte efficace. Annales Mdecine Vtrinaire **149**, 34-48.

Van Immerseel, F., De Buck, J., Pasmans, F., Bohez, L., Boyen, F., Haesebrouck, F., and Ducatelle, R. (2004). Intermittent long-term shedding and induction of carrier birds after infection of chickens early posthatch with a low or high dose of Salmonella enteritidis. Poultry Science **83**, 1911-1916.

Van Pelt W., Mevius D., Stoelhorst H.G., Kovats S., Van de Giessen A.W., Wannet W., and Duynhoven Y.T.H.P. (2004). A large increase of Salmonella infections in 2003 in the Netherlands: hot summer or side effect of the avian influenza outbreak? Eurosurveillance **9(7)**, 17-19.

VAR, Veterinary and Agrochemical Research Centre, Laboratory of General Bacteriology. (2005). Salmonella Serotypes analysed at the CODA-CERVA in 2005. Evolution among Poultry, Cattle and Pig isolates from 1992 to 2005 with results of Antimicrobial Resistance Testing. VAR Report data 2005.

Verbyla, A.P., Cullis, B.R., Kenward, M.G. and Welham, S.J. (1999). The analysis of designed experiments and longitudinal data by using smoothing splines. *Appl. Statist.*, **48**, 269-311.

Wales, A., Breslin, M., Carter, B., Sayers, R., and Davies, R. (2007). A longitudinal study of environmental salmonella contamination in caged and free-range layer flocks. Avian Pathology **36**, 187-197.

Wales, A., Breslin, M., and Davies, R. (2006). Assessment of cleaning and disinfection in Salmonella-contaminated poultry layer houses using qualitative and semi-quantitative culture techniques. *Vet. Microbiol.* **116**, 283-293.

Whitaker, H.J. and Farrington, C.P. (2004). Estimation of infectious disease parameters from serological survey data: the impact of regular epidemics. *Statistics in Medicine*, **23**, 2429-2443.

Wheeler, M.W., and Bailer, A.J. (2007). Properties of Model-Averaged BMDLs: A study of Model Averaging in Dichotomous Response Risk Estimation. *Risk Analysis*, **27(3)**, 659-670.

WHO FAO, World Health Organization, Food and Agriculture Organization of the United Nations. (2002). Risk Assessments of Salmonella in eggs and broiler chickens, an interpretative summary. Microbiological Risk Assessment Series 1.

World health organisation (WHO). (2005). Salmonelles multirésitantes, WHO website media centre, reference 139, Available on the web at http://www.who.int/mediacentre/factsheets/fs139/fr/ (november 2007).

Ziggity Systems Inc. (2006). Five common poultry watering myths. World Poultry **22**, 18.

Samenvatting

Gezondheid is een kostbaar goed voor elk levend wezen, in het bijzonder voor mens en dier. Maar deze gezondheid worden continu blootgesteld aan risico's. Micro-organismen kunnen schadelijk zijn voor mens en dier, in infectieziekte-terminologie de zogenaamde "gastheer" voor deze organismen. Een infectieziekte kan veroorzaakt worden door verschillende soorten micro-organismen: virussen, bacterin, parasieten en fungi. Ziekten door infectie kunnen variëren van alledaagse, vrij onschadelijke kwaaltjes, zoals verkoudheden, tot dodelijke ziektes zoals AIDS. Afhankelijk van vele verschillende factoren kan een infectieziekte zich verspreiden op verschillende manieren. Infecties kunnen seksueel overdraagbaar zijn, overdraagbaar via de lucht in elkaars directe omgeving, via direct huidcontact, via contact met besmet bloed, maar infecties kunnen ook via insecten verspreid worden, alsook via het consumeren van besmet voedsel en water. In deze thesis worden methoden voor het modeleren van infectieziekten voorgesteld en bestudeerd voor infectieziekten die i) overgedragen worden door de lucht via bijvoorbeeld druppeltjes vocht, die bij het niezen en hoesten door een genfecteerd persoon verspreid worden, ii) via genfecteerd bloed dat druggebruikers aan elkaar overdragen, door eenzelfde injectienaald te delen, en iii) het gevolg zijn van het consumeren van gecontamineerd voedsel.

Deel I van de thesis spitst zich toe op virale infecties zoals rodehond (rubella), bof (mumps), varicella, parvo B19 en hepatitis C. Centraal hierbij staat de zogenaamde infectiedruk als functie van bijvoorbeeld de leeftijd van het individu. De infectiedruk is de kans dat een nog vatbaar persoon met een bepaalde leeftijd ogenblikkelijk geïnfecteerd wordt. Mathematisch kan de infectiedruk afgeleid worden uit de leeftijdsafhankelijke prevalentie van de ziekte. De prevalentie geeft het percentage van de populatie dat al genfecteerd en niet langer vatbaar is. Deze prevalentie kan op verschillende manieren gemodelleerd worden als functie van de leeftijd. Parametrische modellen veronderstellen een vooropgesteld functioneel verband tussen de prevalentie en de leeftijd (of een andere maat voor de duur van blootstelling zoals de tijd sinds het

begin van het delen van injectienaalden bij drukgebruikers). In Hoofdstuk 4 wordt een zogenaamd Weibull model toegepast voor het modelleren van de prevalentie en de afgeleide infectiedruk voor het hepatitis C virus. Dit Weibull model, gecombineerd met een logistisch regressiemodel, wordt gebruikt om inzicht te krijgen in factoren die het risico op hepatitis C verhogen (risicofactoren). Bij niet-parametrische modellen wordt het functioneel verband niet vooraf vastgelegd. Dergelijk model is in staat zich aan te passen aan onverwachte patronen in de gegevens. In Hoofdstuk 4 wordt het resultaat van een isotoon niet-parametrisch regressiemodel vergeleken met een parametrisch Weibull model. Beide methoden geven eenzelfde trend aan. In Hoofdstuk 2 wordt een andere populaire niet-parametrische methode toegepast op prevalentie-gegevens: de zogenaamde "penalized splines" gerepresenteerd als een veralgemeend lineair gemengd model. Een simulatiestudie toont aan dat kwadratische splines gebaseerd op 20 knots een betere performantie biedt dan de lineare en cubische splines of splines van graad vier. Hoofdstuk 4 wordt de methode uitgebreid tot een semi-parametrische methoden, die parametrische componenten combineert met niet-parametrische splines. Toegepast op serologische gegevens van vijf Europese landen, leidt een semiparametrisch model met een logit link tot proportionele odds ratios voor elk land in combinatie met een referentieland. Als echter de cloglog link wordt gebruikt krijgt men een proportionele infectiedruk voor een land in vergelijking met een referentieland. Tenslotte wordt aangetoond dat een constante infectiedruk, een stuksgewijs constante infectiedruk en een lineaire infectiedruk kan geformuleerd worden als een penalized spline model.

Het tweede en derde deel van de thesis handelt over modellen voor bacteriële infecties en ziekten. Het is de doelstelling van het tweede deel om te tonen hoe statistische modellen de response op een nadelig effect van het consumeren van gecontamineerd voedsel relateren met de hoeveelheid organismen (de dosis). Haas *et al.* (1999) stellen terecht dat een experiment om rechtstreeks een aanvaardbare kleine dosis bij een aanvaardbaar klein risico te bepalen praktisch niet haalbaar is, omdat het risico op een enkele blootstelling zo klein is dat de bepaling ervan een heel groot aantal studiesubjecten zou vereisen. Voor dergelijke dosis-respons modellen zijn parametrische modellen nodig om extrapolatie naar lage dosis met een laag risico mogelijk te maken. In dit deel van de thesis worden twee belangrijke deelgebieden van risico-beoordeling behandeld: (1) de klasse van mathematische modellen die gebruikt kunnen worden voor extrapolatie van hoge naar lage dosis, (2) methoden om de onzekerheid in de schattingen in rekening te brengen. Hoofdstuk 5 draagt bij tot het eerste deelgebied in de vorm van nieuwe aangepaste fractionele veeltermmodellen. Fractionele veeltermen, in hun oorspronkelijk geïntroduceerd door Royston en Altman (1989) toegepast

op dosis-respons modellen, zijn niet bruikbaar omdat ze kunnen leiden tot een niet-geschikt functioneel verband voor een binaire response (geïnfecteerd ja of neen). Meer bepaald dienen dosis-repons modellen aan biologische basisvoorwaarden te voldoen: monotoon stijgend als functie van de dosis en begrensd door 0 en 1. Om aan deze voorwaarden te voldoen, werd een nieuwe klasse van gemodificeerde fractionele veeltermen gedefinieerd. In Hoofdstuk 5 worden een hele reeks modellen (40 in totaal) voor het schatten van risico's bij lage dosissen voorgesteld. De allereerste vraag is welke modellen echt gebruikt kunnen worden voor deze doeleinden. Traditioneel wordt één beste eindmodel geselecteerd op basis van zogenaamde "goodness-of-fit" criteria, statistische maten die aangeven hoe goed de modellen aansluiten bij de data, en nadien wordt inferentie (verklarende statistiek) gebaseerd op dit finaal model; Maar een dergelijke procedure houdt geen rekening met de voorafgaande modelselectie procedure en bijgevolg ook niet dat andere competitieve modellen quasi dezelfde goodness-of-fit kunnen vertonen, wat leidt tot een misleidende onderschatting van de standaardfouten op de schatters. Buckland *et al.* (1997) introduceerden een manier om de onzekerheid van dergelijke selectie uit een familie competitieve modellen in rekening te brengen, aan de hand van het uitmiddelen van de modellen op basis van Akaike verschillen en Akaike gewichten. Deze benadering van het uitmiddelen over verschillende modellen werd ook bestudeerd in Burnham and Anderson (2002). Deze methode brengt zowel de onzekerheid omtrent het gebruikte model als de variabiliteit van het schatten op basis van gegevens in rekening. Hoofdstuk 5 bestudeert en illustreert hoe de gemodificeerde fractionele veeltermen als mogelijke geschikte dosis-respons modellen samen met traditionele dosis-respons modellen zoals het Beta-Poisson model, het log-logistisch, het log-normaal en het extreme waarde model kan toegepast worden om het risico op ziekte door Salmonella Typhi en op infectie door Campylobacter jejuni infectie te modelleren in gebieden van lage dosissen en hoe de techniek van het uitmiddelen over al deze modellen in deze context kan toegepast worden.

Daar waar Deel II handelt over de consumptie van gecontamineerd voedsel (bv. gevogelte), richt Deel III zich tot de preventie van de overdracht van bacteriën in voedsel naar de mens door de betreffende infectie op het niveau van de boerderijen te controleren. Dit laatste deel van de thesis betreft het statistisch ondersteunen van controle en preventiemaatregelen tegen infectieziekten bij kippen, in dit geval Salmonella. Het onderzoekt vooral statistische modellen voor de identificatie van potentiële factoren die het risico op Salmonella vergroten bij kippen voordat zij bij de consument belanden. De gegevens uit dergelijke controle en bewakingssystemen zijn hiërarchisch gestructureerd. Gegevens binnen een groep kippen zijn gecorreleerd, vervolgens zijn gegevens binnen dezelfde boerderij gecorreleerd. Statistische risicomodellen moeten

deze geclusterde structuur in de gegevens correct in rekening brengen. In Hoofdstuk 6 en 7 worden drie soorten risiciomodellen gebruikt voor de identificatie van risicofactoren voor Salmonella bij legkippen en bij braadkippen: marginale modellen, veralgemeende schattingsvergelijkingen waaronder alternerende logistische regressie, en veralgemeende lineaire gemengde modellen. De belangrijkste risicofactor die voor legkippen werd geïndentificeerd is de factor die aangeeft of de groep kippen wordt gehouden in een kooi, of een schuur of in een open ruimte (met een hoger risico voor de kooi). Andere significante factoren waren leeftijd van de groep en omvang van de groep (in beide gevallen een stijgend effect). Een andere analyse in dit laatste deel betreft factoren die de persistentie van Salmonella bepalen. De aanwezigheid van werkkrachten op de boerderij, tijdelijk aanwezig en ook in contact me vreemde en externe dieren en mensen, is de belangrijkste factor die hiervoor werd geïdentificeerd.