DOCTORAATSPROEFSCHRIFT

2011 | Faculteit Wetenschappen

Statistical and Mathematical Models to Estimate the Transmission of Airborne Infections from Current Status Data

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

Nele GOEYVAERTS

Promotor: prof. dr. Niel Hens
Copromotoren: prof. dr. Marc Aerts
prof. dr. Philippe Beutels

D/2011/2451/2

universiteit hasselt

Cover figure – upper left panel: electron micrograph of a varicella zoster virus (1982), identification number 1878, Public Health Image Library, Centers for Disease Control and Prevention (CDC). Content Providers: CDC/ Dr. Erskine Palmer, B.G. Partin.

Jury Members:

Prof. Dr Niel Hens (main supervisor, Hasselt University and University of Antwerp, Belgium)
Prof. Dr Marc Aerts (co-supervisor, Hasselt University, Belgium)
Prof. Dr Philippe Beutels (co-supervisor, University of Antwerp, Belgium)
Prof. Dr Jean Manca (chairperson, Hasselt University, Belgium)
Prof. Dr John Edmunds (London School of Hygiene and Tropical Medicine, United Kingdom)
Prof. Dr Christel Faes (Hasselt University, Belgium)
Prof. Dr Paddy Farrington (The Open University, United Kingdom)
Prof. Dr Ziv Shkedy (Hasselt University, Belgium)
Prof. Dr Jacco Wallinga (University Medical Center Utrecht and National Institute for Public Health and the Environment, The Netherlands)

February 25, 2011

Dankwoord

This was supposed to be the easiest part of my thesis :-)

I would like to thank everyone who contributed to the realization of this thesis. First of all, I would like to acknowledge my supervisor, Niel, for giving me the opportunity to prepare a PhD thesis in the very interesting field of modeling infectious diseases, and for his guidance throughout these four years. Niel, thank you for the stimulating discussions, for your valuable ideas and advice. I would also like to thank my co-supervisors, Marc and Philippe, for their great ideas and suggestions that helped to shape the content of this thesis. I am grateful for the opportunity to prepare my PhD dissertation at two institutions, first and foremost at the Center for Statistics as the UHasselt side of the Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat), and second at the Centre for Health Economics Research and Modelling Infectious Diseases (CHERMID) at the University of Antwerp.

I am indebted to the 'Strategisch Basis-Onderzoek' of the IWT and the SIMID-project for funding my PhD. I was fortunate to conduct research within an interdisciplinary context together with health economists, mathematical modelers and virologists, mostly in collaboration with the Vaccine & Infectious Disease Institute (VAXINFECTIO) at the University of Antwerp. I would like to acknowledge VAXINFECTIO and the Health Protection Surveillance Centre in Dublin, for kindly providing the serological data. My gratitude also goes out to my co-authors: it was very nice to collaborate with you all. I would like to thank the colleagues I met at conferences, workshops and meetings, for providing valuable comments, having useful discussions or encouraging me in my research. I am grateful for the feedback and suggestions I received from the jury members to an earlier version of this thesis.

Laten we even teruggaan naar de start van mijn doctoraat op 1 maart 2007, halverwege mijn 'master of biostatistics'-studies hier in Diepenbeek. Ik ben goed terechtgekomen in de 'Heilige Rita'-gang waar Adam en co voor een fijne sfeer zorgden. An heeft me wegwijs gemaakt binnen CenStat en daar is een mooie vriendschap uit ontstaan. Bedankt, An, voor de ontelbare vragen die ik jou heb mogen stellen en omdat ik bij jou altijd terechtkon voor een lach en een traan. Tegenwoordig gaat dat nog gemakkelijker want sinds oktober delen we samen een bureau :-) Ik wil mijn UHasselt collega's bedanken voor de leuke tijd in de JOSS board (voorheen A^2B^2D) en tijdens de vele activiteiten met onze internationale studenten, de gezellige koffiepauzes, de verkwikkende loopuurtjes, etc. Ook mijn UA collega's wil ik bedanken voor het creëren van een toffe werkssfeer en in het bijzonder Joke voor de motiverende gesprekjes.

Mijn vrienden en vriendinnen en in het bijzonder Leen, Gitte, Anneke, Els en Hanne, ben ik dankbaar voor de momenten van ontspanning (gaande van babbelen tot singstarren) en inspanning (zumba :p), de leuke uitstapjes en hun interesse en steun. Mama en papa, jullie hebben mij altijd gesteund en gemotiveerd van jongs af aan, ook in mijn studies en tijdens mijn doctoraat: dank jullie wel! Jullie staan altijd liefdevol klaar voor de 'kindjes' en zonder jullie goede zorgen en raadgevingen was ik zeker niet zo ver geraakt. Mijn zus en broer en ook de rest van mijn familie wil ik bedanken voor hun sympathie en aanmoedigende woorden. Mijn laatste woorden zijn voor Wouter, al acht jaar mijn steun en toeverlaat. Het is moeilijk om uit te drukken hoe belangrijk je bent voor mij. Dankzij jouw liefde en steun heb ik het nooit opgegeven en heb ik doorgezet op de momenten dat ik het echt niet meer zag zitten. Je bent er altijd geweest voor mij en je hebt veel van mij moeten verdragen en daarvoor wil ik je bedanken vanuit het diepste van mijn hart.

Nele Goeyvaerts
Diepenbeek, 2 februari 2011

Contents

List of Abbreviations

AIC	Akaike's Information Criterion
AP	Age-dependent Proportionality
AW	Age-dependent Waning
BAH	Bahadur
BE	Belgium
BIC	Bayesian Information Criterion
CI	Confidence Interval
CIL	Critical Immunization Level
CP	Constant Proportionality
CV	Cross-Validation
CW	Constant Waning
DE	Germany
ELISA	Enzyme-Linked Immunosorbent Assay
EPI	Expanded Program on Immunization
ESEN2	European Sero-Epidemiological Network 2
EW	England and Wales
FD	Fetal Deaths
FI	Finland
GB	Great Britain
GEE	Generalized Estimating Equations
IgG	Immunoglobulin G
IT	Italy
LU	Luxembourg
LR	Likelihood Ratio

ML	Maximum Likelihood
MMR	Measles, Mumps, Rubella
MSE	Mean Squared Error
NL	The Netherlands
ODE	Ordinary Differential Equation
PDE	Partial Differential Equation
PL	Poland
PVB19	Parvovirus B19
RCS	Restricted Cubic Splines
RMSE	Root Mean Squared Error
VZV	Varicella Zoster Virus
WAIFW	Who Acquires Infection From Whom
WGEE	Weighted Generalized Estimating Equations

Chapter 1

Introduction

1.1 Infectious Diseases: Transmission and Immunology

Infectious diseases are illnesses in humans, animals or plants resulting from the presence of microbial pathogens, such as viruses, bacteria, parasites, etc. There are several routes by which these pathogens can be transmitted from one host to the other: e.g. airborne, droplet contact, direct or indirect physical contact, fecal-oral (contaminated food or water), sexual contact or vector-borne (e.g. via a mosquito). In this dissertation, the focus is on models for human viral infectious diseases for which the main transmission route is through non-sexual, social contacts such as airborne transmission, droplet contact or direct physical contact.

The transmission route is called airborne when viruses travel on small respiratory droplets that may become aerosolized when individuals sneeze, cough or talk (see Figure 1.1, left panel). These infectious particles hang invisibly in the air and can remain there for a long period of time and even travel over considerable distances. Droplet contact occurs when an individual sneezes or coughs on someone else such that infectious agents may enter the latter person's body through his or her respiratory system. Non-sexual, direct physical contact mainly involves skin-to-skin touching such as shaking hands, kissing and so on. Important examples of viruses which are spread through these social interactions are influenza, smallpox, varicella zoster virus, measles (Figure 1.1, right panel), mumps, rubella and parvovirus B19. Some ideas and methods discussed in this thesis may be useful for bacterial infections as well,

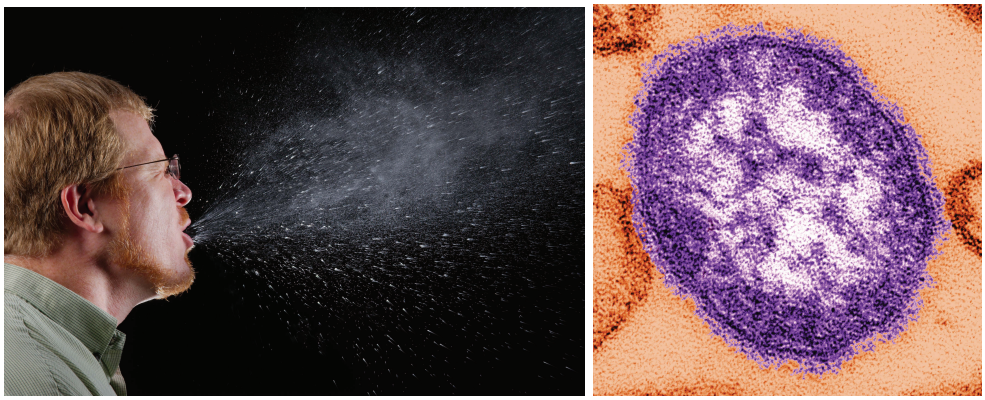

Figure 1.1: Left panel: aerosolized droplets resulting from a sneeze (2009), identification number 11162, content provider: CDC/ Brian Judd, photo credit: James Gathany. Right panel: micrograph of a single measles virus particle, identification number 10707, content providers: CDC/ Cynthia S. Goldsmith and William Bellini. Public Health Image Library, Centers for Disease Control and Prevention (CDC).

for example for pertussis and tuberculosis. Note that a distinction is made between horizontal and vertical transmission of infectious diseases, the former being from any individual to another one, independent of blood-relationships, and the latter being from parent to child, often during pregnancy or at birth.

In general, when an individual is infected with a viral infectious disease, the adaptive immune system activates complex mechanisms to induce protective immunity to the host. The adaptive immune system consists of two types of defence mechanisms: cell-mediated immunity and humoral immunity. T cells such as cytotoxic T cells and T helper cells, are responsible for the cell-mediated immunity. T cells are able to detect suspicious viral fragments on a cell's surface and to destroy the cell if necessary. Cell-mediated immunity also involves the activation of macrophages and cytokines.

Humoral immunity is more relevant to our research since it is related to the production of virus-specific antibodies, which are responsible for long term protection and can be detected in the blood of the host. Antibodies are produced by plasma cells that originate from B cells, and the two main types of antibodies are Immunoglobulin M (IgM) and Immunoglobulin G (IgG). IgM antibodies are produced rapidly upon infection to neutralize viral agents, but they are only present for a short period of time. It takes a longer while until IgG antibodies are produced, but they can persist for years after the infection. The IgG antibody response may entail lifelong immunity, protecting the individual each time he or she is re-exposed to the infectious disease.

In pregnant women, IgG antibodies can be transferred to protect the fetus and infant until the infant's immune system has matured. The principle behind vaccination is to introduce an antigen from a pathogen in order to stimulate the immune system and develop specific immunity against that particular pathogen without causing disease associated with that organism (Alberts *et al.*, 2002).

In the absence of vaccination, the presence of IgG antibodies in blood serum thus indicates past infection with a specific virus. In this thesis, the primary source of data are cross-sectional sets of serum samples. The IgG titer values obtained by testing the blood samples give rise to serological data and provide information with respect to the immunity status of the individuals. Our focus is on dichotomized serological test results, also called current status data. For a more detailed description of the collection and interpretation of serological data and for an introduction to the data sets we use in this thesis, we refer to Chapter 3.

1.2 Historical Overview

This thesis has been made and founded as part of 'Simulation models of infectious disease transmission and control processes (SIMID), with applications to five major health policy issues in Flanders', a strategic basic research project (SBO) that is funded by the Institute for the Promotion of Innovation by Science and Technology in Flanders (project 060081). The SIMID project's aim was to further develop the expertise in infectious disease modelling in Flanders and to improve the overall quality of health economic evaluation capacity as applied in the Belgian context. From a public health perspective, modelling of infectious diseases is far more complex than that of non-infectious diseases, as interventions aimed at controlling the spread of infectious diseases also affect individuals who are not targeted by these interventions. The objective of this thesis was to develop statistical models funded on mathematical modelling equations, to estimate specific parameters related to the transmission of infectious diseases, either endemic or actively immunized, from current status data. By 'endemic' it is meant that disease incidence may undergo cyclical epidemics, however fluctuating around a stationary average over time. Some concepts of statistical inference used throughout this thesis, and an introduction to mathematical modelling of age related infectious disease dynamics, are given in Chapter 2.

The estimation of infectious disease parameters from current status data is an important research topic, since it helps to infer and understand age-specific patterns of disease spread at the population level. Further, it aids in planning and monitoring

of universal immunization programs for vaccine-preventable infections, and of control measures (school closure, antivirals, vaccines, etc.) in the event of an epidemic outbreak. One of the key parameters is the 'force of infection', the rate at which an individual acquires infection, which is the infectious disease analogue of the hazard rate in survival theory. Hens *et al.* (2010a) recently presented a review paper on 75 years of estimating the force of infection from current status data. Muench (1934) formulated the first catalytic model to estimate a constant force of infection from serological data. Grenfell and Anderson (1985) generalized Muench (1934)'s model to a polynomial age-dependent force of infection and derived a stepwise maximum-likelihood approach for parameter estimation. A few years later, Farrington (1990) was the first to consider a nonlinear model for the age-dependent force of infection and to use constrained optimization to ensure a monotonically increasing seroprevalence profile. He illustrated his method using pre-vaccination serological data on measles, mumps and rubella.

Another important, related parameter in infectious disease modeling is the basic reproduction number R_0, defined as the average number of secondary infections generated by a single typical infective individual in a totally susceptible population. Using the mass action principle, which relates the force of infection to age-dependent transmission rates, R_0 can be estimated from serological data as well, and this approach has been popularized by the book of Anderson and May (1991). Anderson and May (1991) labelled the matrix containing the age-dependent transmission rates as the 'Who Acquires Infection From Whom' or WAIFW matrix. The transmission rates were estimated by imposing various mixing patterns on the WAIFW matrix, which were based on prior knowledge of social mixing behavior. Greenhalgh and Dietz (1994) and Farrington *et al.* (2001) elaborated on this approach by Anderson and May (1991) and highlighted the effect on $\widehat{R}_0$ of the mixing pattern assumed. Whitaker and Farrington (2004a) assessed the sensitivity of the Anderson and May (1991) approach with respect to the assumption of endemic equilibrium, more specifically, they looked at the impact of regular epidemic cycles. The same authors also extended the traditional method to situations in which transmission rates vary over time, by augmenting the serological data with case reports data (Whitaker and Farrington, 2004b). Kanaan and Farrington (2005) used Bayesian methods to select the most plausible mixing patterns for rubella and mumps from all published models in the literature. Finally, a continuous parametric contact surface was proposed as an alternative to the low dimensional mixing matrices for the age-dependent transmission rates (Farrington and Whitaker, 2005).

Also at Hasselt University, a variety of methodological work has been done recently

in the area of modelling infectious diseases, mostly in collaboration with the University of Antwerp. Shkedy *et al.* (2003), Shkedy *et al.* (2006) and Namata *et al.* (2007), proposed a non-parametric, parametric and semiparametric approach to model the seroprevalence and the force of infection by using local polynomials, fractional polynomials, and penalized splines within the generalized linear mixed model framework, respectively. Marginal and conditional models to model bivariate current status data were studied by Hens *et al.* (2008), hereby exploiting the association between the transmission routes of the two infectious diseases. Hens *et al.* (2010b) extended the model of Shkedy *et al.* (2006) to change-point fractional polynomials to detect distortions with respect to monotonicity in the seroprevalence, which could for instance be due to maternally-derived immunity, violation of time-homogeneity (epidemics), or waning immunity. Bollaerts *et al.* (2011) investigated the effect of serological test misclassification on the estimation of the prevalence and the force of infection, and proposed a mixture-model based approach for continuous antibody titers to avoid the use of thresholds. These methods have now been joined into the book of Hens *et al.* (2011).

1.3 Outline of the Thesis

The mixing patterns imposed on the WAIFW matrix in the Anderson and May (1991) approach are based on prior knowledge of age-related social mixing behavior rather than observations. The choices of the parametric structure and the age classes are often ad hoc and may heavily influence $\widehat{R}_0$ (Greenhalgh and Dietz, 1994). Further in practice, the method may entail non-realistic discontinuities due to the low dimension of the mixing patterns imposed. An alternative approach has been initiated by the work of Wallinga *et al.* (2006), who assumed that transmission rates for infections transmitted predominantly through non-sexual social contacts, are directly proportional to rates of conversational contact, which can be estimated from a contact survey. Recently, a large social contact survey was conducted in eight European countries as part of the POLYMOD project (Mossong *et al.*, 2008b), which allowed us to elaborate on the method of Wallinga *et al.* (2006). In Chapter 3, an elaborate description of the POLYMOD contact survey and two preliminary analyses are presented. We specifically focus on the Belgian contact survey in Chapter 4, following Hens *et al.* (2009a), to discuss the recording and estimation of professional contacts, to look for associations between different contact characteristics, to model the number of contacts using generalized estimating equations to account for the two days of

contact recording, and to illustrate the spread of an epidemic.

In Chapter 5, a comparison is made between the traditional Anderson and May (1991) approach and the new method of Wallinga *et al.* (2006) to estimate age-specific transmission rates for the varicella zoster virus in Belgium (Goeyvaerts *et al.*, 2010a; Ogunjimi *et al.*, 2009). We use a flexible, bivariate smoothing model to estimate a continuous contact surface from the social contact survey data. In general, however, contacts reported in such surveys are proxies of those events by which transmission may occur and there may exist age-specific characteristics related to susceptibility or infectiousness which are not captured by the contact rates. Therefore, in Chapter 5, we propose to model the transmission rates as the product of two age-specific variables: the contact surface and an age-specific proportionality factor. Furthermore, we address the impact on the estimation of R_0, using non-parametric bootstrapping to account for different sources of variability and using multimodel inference to deal with model selection uncertainty.

In Chapter 6, we explore the hypothesis of waning IgG antibodies for parvovirus B19 (PVB19) (Goeyvaerts *et al.*, 2010b), motivated by the decrease or plateau observed in the seroprevalence profiles between the ages of 20 and 40, in each of 5 European countries. We investigate whether secondary infections are plausible, and whether natural boosting of immunity by exposure to infection may occur. Several immunological scenarios are tested for PVB19 by fitting different compartmental dynamic transmission models to serological data using data on social contact patterns. We assess whether different views on the evolution of the immune response to PVB19 infection may lead to altered estimates of R_0, the age-specific force of infection and the associated risk in pregnancy.

Finally, in Chapter 7, we discuss the work of Gay (2000) and Altmann and Altmann (2000) on the estimation of trivalent vaccination coverage from trivariate serological data. We extend the estimation method of Gay (2000) to incorporate the dependency between the probabilities of acquiring natural infection with each of the three diseases for the non-vaccinated population. The method is developed within a likelihood-based marginal model framework, and applied to trivariate current status data for measles, mumps and rubella in Belgium and Ireland (Goeyvaerts *et al.*, 2011).

Chapter 2

Basic Concepts

In this chapter, basic terminology and fundamental concepts which are used in the field of mathematical modelling of infectious disease dynamics, are introduced. Further in Section 2.2, the main statistical methods used throughout the thesis for parameter estimation, construction of confidence intervals, model selection and multimodel inference, are briefly described.

2.1 Mathematical Models for Infectious Disease Transmission

Deterministic models which describe infectious disease dynamics by partitioning the population into different disease states or compartments, were already used centuries ago. The first model dates back to 1760 when Bernoulli aimed to demonstrate the benefits of variolation against smallpox for the population of France (Bernoulli, 1760). The flow between disease states is typically represented by a set of (partial) differential equations, which yield an explicit solution for the distribution of the population with respect to the infection status. Important contributions to epidemic theory have been made by Kermack and McKendrick (1927), Bailey (1975), Dietz (1975) and Anderson and May (1991), amongst others. Though deterministic models are very insightful to study infectious disease spread in large populations, they are however less useful for small or isolated populations. To this purpose, stochastic models were developed, with the chain binomial model proposed by Reed and Frost in their lectures given in 1928, as the most well-known. Stochastic models make up a second important branch

in infectious disease modelling and are rooted in the theory of random processes (Becker, 1989). They are usually defined at the level of the individual and aim to describe the stochasticity seen in real-life disease outbreaks, for instance making use of temporal or final outcome data (O'Neill, 2010). For two more recent accounts of stochastic epidemic modelling we refer to Daley and Gani (1999) and Andersson and Britton (2000). In this dissertation, however, we only make use of deterministic models since the aim is to develop methods to estimate the age-related heterogeneity inherent to the spread of airborne infections, either endemic or actively immunized, in large populations. In the following description of some basic concepts, we mainly follow Anderson and May (1991).

2.1.1 Basic Deterministic Model

In general, deterministic models represent age- and time-dependent mathematical models describing the flow of individuals through different mutually exclusive infection states. One of the most basic compartmental models for infectious disease transmission is the MSIR model. Suppose we consider a large population and let $\mu(a,t)$ denote the mortality rate depending on age a and calendar time t. The MSIR model, which is presented in Figure 2.1, assumes that all newborns are protected by maternal antibodies (first stage, denoted 'M') until waning results in loss of passive immunity, and the infants become susceptible to infection (second stage, denoted 'S'). As they age from then on, they may become infected and infectious to others (third stage, denoted 'I'). After the infectious stage, individuals are removed and no longer able to transmit the disease to others (fourth stage, denoted 'R'). Depending on the disease, the R stage may for example correspond to recovery, immunity, isolation or death (O'Neill, 2010). The corresponding number of individuals in each stage or compartment of the MSIR model can be expressed as a function of age and time by $M(a,t)$, $S(a,t)$, $I(a,t)$ and $R(a,t)$, respectively. As illustrated in Figure 2.1, $\alpha(a,t)$ denotes the rate of losing maternal antibodies, $\lambda(a,t)$ the rate of acquiring infection or the 'force of infection', $\gamma(a,t)$ the removal rate, and $\upsilon(a,t)$ the disease-induced mortality rate.

The total number of individuals of age a at time t is then defined as $N(a,t) = M(a,t) + S(a,t) + I(a,t) + R(a,t)$. The set of partial differential equations (PDEs)

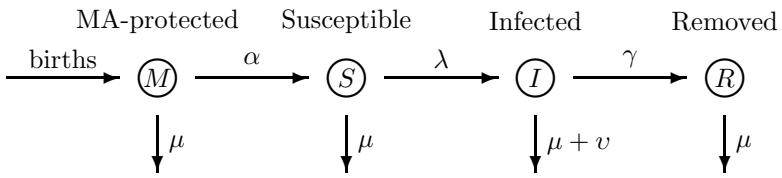

Figure 2.1: Illustration of the MSIR compartmental model.

describing the dynamics of the MSIR model is given by:

$$\begin{cases} \frac{\partial M(a,t)}{\partial a} + \frac{\partial M(a,t)}{\partial t} & = -\{\alpha(a,t) + \mu(a,t)\}M(a,t), \\ \frac{\partial S(a,t)}{\partial a} + \frac{\partial S(a,t)}{\partial t} & = \alpha(a,t)M(a,t) - \{\lambda(a,t) + \mu(a,t)\}S(a,t), \\ \frac{\partial I(a,t)}{\partial a} + \frac{\partial I(a,t)}{\partial t} & = \lambda(a,t)S(a,t) - \{\gamma(a,t) + \mu(a,t) + \upsilon(a,t)\}I(a,t), \\ \frac{\partial R(a,t)}{\partial a} + \frac{\partial R(a,t)}{\partial t} & = \gamma(a,t)I(a,t) - \mu(a,t)R(a,t), \end{cases} \tag{2.1}$$

with the boundary condition that all individuals are born with protective maternally derived immunity:

$$M(0,t) = N(0,t) = \int_0^\infty f(a)N(a,t)da \equiv B(t),$$

where $f(a)$ denotes the fertility in the population as a function of age a.

Demographic and Endemic Equilibrium

Throughout this work, we will make a number of simplifying assumptions to come up with a set of equations which are computationally more feasible in order to estimate age-specific transmission dynamics. First, we assume that mortality due to infection can be ignored ($\upsilon \approx 0$). This is plausible for the infectious diseases studied in this thesis, which are primarily childhood infections in developed countries. Further, we assume that the population has reached a demographic equilibrium such that the population age distribution is stationary. We also assume that the number of births and deaths are constant over time and exactly balanced, entailing a constant total population size N. Finally, we assume that the disease is in an endemic steady state at the population level, meaning that disease incidence may undergo cyclical epidemics, however fluctuating around a stationary average over time.

Under these assumptions, the time dependency cancels out and the set of partial differential equations in (2.1) simplifies to a set of ordinary differential equations

(ODEs):

$$\begin{cases} \frac{dM(a)}{da} &= -\{\alpha(a)+\mu(a)\}M(a), \\ \frac{dS(a)}{da} &= \alpha(a)M(a) - \{\lambda(a)+\mu(a)\}S(a), \\ \frac{dI(a)}{da} &= \lambda(a)S(a) - \{\gamma(a)+\mu(a)\}I(a), \\ \frac{dR(a)}{da} &= \gamma(a)I(a) - \mu(a)R(a). \end{cases} \tag{2.2}$$

Adding the four equations in (2.2) then yields the following differential equation for the stationary age distribution $N(a)$ of the population size:

$$\frac{dN(a)}{da} = -\mu(a)N(a),$$

so that

$$N(a) = N(0)\exp\big(-\Omega(a)\big) \text{ where } \Omega(a) = \int_0^a \mu(u)du. \tag{2.3}$$

The monotone decreasing function $g(a) \equiv \exp\big(-\Omega(a)\big)$ corresponds to the well-known survival function in survival analysis and reflects the probability of surviving to age a: $g(a) = P(T > a)$, where T is the time of death. $1 - g(a)$ is the lifetime distribution function, such that the expected age a of dying or 'life expectancy' equals

$$L = \int_0^\infty -ag'(a)da = -ag(a)\big|_0^\infty + \int_0^\infty g(a)da = \int_0^\infty \exp\big(-\Omega(a)\big)da. \tag{2.4}$$

Making use of the boundary condition on the number of newborns and equation (2.3), it can be easily seen that births and deaths are indeed exactly balanced, since

$$N(0) = B = \int_0^\infty \mu(a)N(a)da, \tag{2.5}$$

is equivalent to $\int_0^\infty \mu(a)\exp\big(-\Omega(a)\big)da = 1$, which is satisfied. Since the total population size N equals $\int_0^\infty N(a)da$, it follows from (2.3) and (2.4) that the number of newborns equals $M(0) = N(0) = N/L$.

In the following, as we use upper case letters to denote the total number of individuals by compartment, we will use lower case letters to denote age-specific proportions or fractions, e.g. $s(a) = S(a)/N(a)$. It is convenient to divide each dependent variable by $N(a)$, as this eliminates the terms involving $\mu(a)$ from the set of differential equations (2.2):

$$\begin{cases} \frac{dm(a)}{da} &= -\alpha(a)m(a), \\ \frac{ds(a)}{da} &= \alpha(a)m(a) - \lambda(a)s(a), \\ \frac{di(a)}{da} &= \lambda(a)s(a) - \gamma(a)i(a), \\ \frac{dr(a)}{da} &= \gamma(a)i(a). \end{cases}$$

By solving the above set of differential equations, the following expressions for the fraction of infants protected by maternal antibodies and the fraction of susceptibles are obtained:

$$\begin{aligned} m(a) &= \exp\left(-\int_0^a \alpha(u)du\right), \\ s(a) &= \int_0^a \alpha(u)m(u)\exp\left(-\int_u^a \lambda(t)dt\right)du. \end{aligned} \tag{2.6}$$

Mortality and Maternal Antibodies

In some applications, it is convenient to make simplifying assumptions with respect to the mortality rate and the protective period of maternal antibodies. One of the assumptions often made is that the mortality rate is an age-independent constant μ, which is referred to as 'type II mortality'. In this case, the survival function is of the form $g(a) = \exp(-\mu a)$, which is generally not realistic for developed countries as illustrated in Figure 2.2. The solid line in Figure 2.2 represents $\hat{g}(a) = \exp\big(-\int_0^a \hat{\mu}(u)du\big)$, with the mortality rate estimated from demographical data on the number of deaths and population sizes per age class for Belgium anno 2003, using a Poisson generalized additive model. From the mortality rate $\hat{\mu}(a)$, a life expectancy estimate of $\hat{L} = 78.8$ years is computed using (2.4). The survival curve for type II mortality (dotted line) is set to entail the same life expectancy as the survival function estimated from the demographical data. From (2.5) it can be seen that type II mortality implies $N/L = \mu N$, and therefore a constant mortality rate $\hat{\mu} = 1/\hat{L}$ is chosen. As is clear from Figure 2.2, a better approximation of the population age distribution is obtained for 'type I mortality' (dashed line):

$$g(a) = \begin{cases} 1, & \text{if } a < L \\ 0, & \text{if } a \geq L, \end{cases} \quad \text{or} \quad \mu(a) = \begin{cases} 0, & \text{if } a < L \\ \infty, & \text{if } a \geq L, \end{cases} \tag{2.7}$$

which implies that all individuals survive up to age L and then promptly die.

Similar to type I mortality, one can assume that all newborns are protected through maternally-derived passive immunity until a certain age A and then instantaneously become susceptible to infection. We refer to this assumption as 'type I maternal antibodies':

$$m(a) = \begin{cases} 1, & \text{if } a \leq A \\ 0, & \text{if } a > A, \end{cases} \quad \text{or} \quad \alpha(a) = \begin{cases} 0, & \text{if } a \leq A \\ \infty, & \text{if } a > A. \end{cases} \tag{2.8}$$

Making use of the Dirac delta property of $\alpha \cdot m$ under the assumption of type I maternal antibodies (see Proposition 1) and equation (2.6), the proportion of susceptibles is

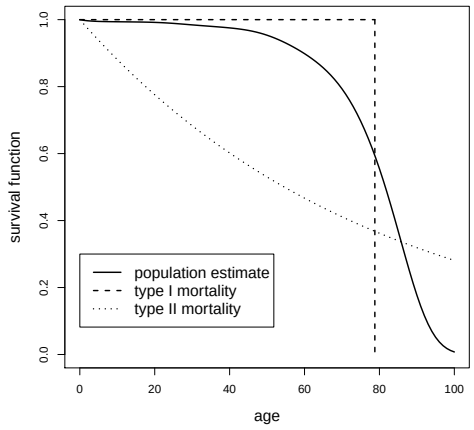

Figure 2.2: Survival probability as a function of age, estimated from demographical data using the exponential model in (2.3) (solid line), assuming type I mortality (dashed line), and type II mortality (dotted line). Parameters are chosen such that the three models entail the same life expectancy (78.8 years).

given by

$$s(a) = \exp\left(-\int_A^a \lambda(u)du\right), \quad \text{if } a > A, \tag{2.9}$$

and $s(a) = 0$ if $a \leq A$.

Proposition 1 (Dirac delta property of $\alpha \cdot m$ for type I maternal antibodies) *For every differentiable function f on $[0, +\infty)$, the following equation holds when $\alpha(a)$ and $m(a)$ are defined as in (2.8):*

$$\int_0^a \alpha(u)m(u)f(u)du = \begin{cases} 0, & \text{if } a \leq A \\ f(A), & \text{if } a > A. \end{cases} \tag{2.10}$$

Proof. We will only give the proof for the case $a > A$, since the result for $a \leq A$ follows directly from (2.8). Define $\alpha_\varepsilon(a)$ and $m_\varepsilon(a)$ as follows, for $\varepsilon > 0$ small:

$$\begin{aligned} \alpha_\varepsilon(a) &= \begin{cases} 0, & \text{if } a \leq A \\ \frac{1}{\varepsilon}, & \text{if } a > A, \end{cases} \\ m_\varepsilon(a) &= \begin{cases} 1, & \text{if } a \leq A \\ \exp\left(-\frac{a-A}{\varepsilon}\right), & \text{if } a > A. \end{cases} \end{aligned}$$

The integral in (2.10) can now be solved using the fact that $\alpha(a) = \lim_{\varepsilon \downarrow 0} \alpha_\varepsilon(a)$ and $m(a) = \lim_{\varepsilon \downarrow 0} m_\varepsilon(a)$:

$$\begin{aligned}
\int_0^a \alpha(u)m(u)f(u)du &= \int_0^a \lim_{\varepsilon \downarrow 0} \left[\alpha_\varepsilon(u)m_\varepsilon(u)\right] f(u)du \\
&= \lim_{\varepsilon \downarrow 0} \int_0^a \alpha_\varepsilon(u)m_\varepsilon(u)f(u)du \\
&= \lim_{\varepsilon \downarrow 0} \int_A^a \frac{1}{\varepsilon} \exp\left(-\frac{u-A}{\varepsilon}\right) f(u)du \\
&= \lim_{\varepsilon \downarrow 0} \left[-\exp\left(-\frac{u-A}{\varepsilon}\right) f(u)\Big|_A^a + \int_A^a \exp\left(-\frac{u-A}{\varepsilon}\right) f'(u)du\right] \\
&= f(A),
\end{aligned}$$

where Lebesgue's dominated convergence theorem is used to justify the second step.■

Model Extensions

The MSIR model is just one, though fundamental, example of a deterministic model used to describe infectious disease transmission. In practice, many extensions exist with different numbers of compartments having various interpretations. The SIR-model, which omits the M-compartment from the MSIR model, is the most frequently used in the literature. In some cases, it may be of interest to account for the fact that once infected, individuals are generally not immediately infectious to others. After having acquired infection by exposure (E) to the disease, individuals often go through a 'latent period' in which they are not yet infectious to others, thus not yet able to transmit the disease. The SEIR-model is an extension of the SIR-model which includes an extra E-compartment to account for this latent period. Another example is the SIS-model which is suitable for certain sexually transmitted infections such as gonorrhea. For these diseases, individuals become susceptible again after infection and multiple infections during one lifetime may occur. The removed state R at the end of the dynamical cycle is therefore implausible and replaced by a flow from the infected state I back to the susceptible state S. In Chapter 6, we will consider other model extensions involving waning, boosting and reinfections, to specifically infer on parvovirus B19 transmission.

2.1.2 Who Acquires Infection From Whom

The force of infection $\lambda(a)$ can be generally written as (e.g. Anderson and May, 1991):

$$\lambda(a) = \int_0^\infty \beta(a, a')I(a')da', \tag{2.11}$$

where $\beta(a, a')$ denotes the transmission rate: the average per capita rate at which an individual of age a' makes effective contacts with a person of age a, per unit time. The contact is called effective when the person of age a is 'successfully' infected by the person of age a', given that the first individual is susceptible and the other one infectious. The average rate at which a susceptible of age a acquires infection per unit time, thus roughly equals the sum of the average rates at which he/she makes effective contact with all infectious individuals present in the population, per unit time. This formula reflects the so-called 'mass action principle' and probably arose from the law stating that the rate of any given chemical reaction is proportional to the product of the concentrations of the reactants.

The mass action principle implicitly assumes that susceptible and infectious individuals mix completely with each other and move randomly within the population. This brings us immediately to the main drawback of this mean-field definition (2.11): it does not account for the fact that contacts are often directed and clustered in e.g. households, schools or workplaces. Network-based approaches to infectious disease dynamics allow to model these aspects of social mixing behavior (e.g. Keeling and Eames, 2005; Bansal *et al.*, 2007). Two individuals (nodes) in a contact network model are for example connected with an edge if they interact in such a way that an infection could be transmitted. Each individual has its own 'degree', defined as the number of edges or contacts, and the network model thus explicitly captures the degree variability observed in real life (degree distribution). Since network models have an individual-based interpretation, however, it seems less straightforward to derive population-based estimates such as the force of infection.

Taking into account recovery ($\gamma(a)$) and mortality of the infected individuals, formula (2.11) can be rewritten as (see Farrington *et al.*, 2001):

$$\lambda(a) = \int_0^\infty \left\{ \int_0^\infty \beta(a, a'+t) e^{-\int_0^t \gamma(u)du} e^{-\int_{a'}^{a'+t} \mu(u)du} dt \right\} \lambda(a')S(a')da'.$$

If the infectious period is short compared to the timescale on which transmission and mortality rates vary, making use of (2.3), the force of infection can be approximated by:

$$\lambda(a) = \frac{ND}{L} \int_0^\infty \beta(a, a')\lambda(a')s(a') \exp\left(-\int_0^{a'} \mu(u)du \right) da', \tag{2.12}$$

where D denotes the mean infectious period:

$$D = \int_0^\infty \exp\left(-\int_0^a \gamma(u)du\right) da.$$

The main difficulty in estimating $\beta(a, a')$ from the force of infection, is that $\lambda(a)$ is a one-dimensional projection of the transmission process, while the transmission rates make up a two-dimensional matrix. The matrix elements β_{ij} range over age class i of the susceptible (rows) and age class j of the infected (columns).

Anderson and May (1991) proposed to impose certain mixing patterns on this β_{ij} matrix, which is called the 'Who Acquires Infection From Whom' (WAIFW) matrix, thereby constraining the number of distinct elements for identifiability reasons, and to estimate the mixing parameters from serological data. The mixing patterns from the traditional Anderson and May (1991) approach are based on prior knowledge of social mixing behavior, and typically have low dimensions which may induce non-realistic discontinuities in the estimated β_{ij} matrix. In Chapter 5, we present an illustration of this traditional approach for the varicella zoster virus. Further in Chapter 5, we elaborate on the new method of using social contact surveys to infer on transmission rates, firstly applied by Wallinga *et al.* (2006). Unlike the traditional approach, strong parametric assumptions about social mixing in the population are avoided when using data on social contacts. An introduction to social contact surveys is given in Section 3.2.

Basic Reproduction Number

One of the most popular parameters in infectious disease epidemiology, which is directly related to the WAIFW matrix, is the basic reproduction number R_0 (sometimes called the basic reproductive ratio). R_0 represents the average number of secondary cases produced by one typical infected person during his or her entire period of infectiousness, when introduced into an entirely susceptible population. For a nice historical overview of the development of R_0, which has an analogous interpretation in demography and ecology as the expected number of female offspring born to one female during her entire life, we refer to Heesterbeek (2002).

Figure 2.3(a) presents an illustration of the epidemiological definition for a simplistic situation where $R_0 = 3$. R_0 is referred to as a 'threshold parameter' since the following criterion holds for large populations: if $R_0 > 1$, the infection can invade the population and may become endemic, while if $R_0 \leq 1$, the infection will eventually die out with probability 1. The value of R_0 thus reflects the epidemic potential of an infection, and encapsulates the interaction between the infectious organism, its

(a) (b)

Figure 2.3: Illustration of the basic reproduction number R_0 and the critical immunization level. (a) One infected person (black circle) is introduced into a fully susceptible population, and infects $R_0 = 3$ others (gray circles). (b) Same situation as (a), only now 2/3 of the population was immunized (dotted circles) prior to introduction of the infected person. In this case, the infection is only transmitted to $R_v = 1$ other person.

(human) host and the environment. As formulated by Dietz (1993), R_0 depends on the following three factors: the duration of the infectious period, the probability that a contact between an infected and a susceptible individual leads to an infection, and the contact rate. Additionally, the stationary age distribution of the population also determines the value of R_0.

When an infected individual of age a' is introduced into an entirely susceptible population, the average number of individuals of age a he/she infects during the infectious period, equals:

$$G(a, a') \equiv \frac{ND}{L} \exp\left(-\int_0^a \mu(u)du\right) \beta(a, a'). \tag{2.13}$$

The function $G(a, a')$ is called the next generation operator, since it expresses the age distribution (as well as the size) of the next generation of infectious cases. Diekmann *et al.* (1990) showed that the basic reproduction number R_0 is the dominant eigenvalue of this next generation operator. The leading right eigenfunction is proportional to the distribution of infected individuals during the initial exponential growth phase of an epidemic, and represents the 'typical' infected person used in the definition of R_0.

Although R_0 is a theoretically interesting summary measure of the WAIFW matrix, its value is rarely observed in practice. The effective reproduction number R_{eff} is a reflection of the actual average number of secondary cases, which takes into account the fact that, when an infection emerges in a population, not all individuals are susceptible to the infection. This indeed depends on historical immunity, the con-

Table 2.1: Typical values of R_0 and CIL for selected infections (from: Farrington, 2003).

Infectious Disease	R_0	CIL (%)
measles	10-20	90-95
chickenpox	5-10	80-90
mumps	5-10	80-90
rubella	4-7	75-85
smallpox	3-5	65-80

trol measures taken (such as active immunization), and the depletion of susceptible individuals when the epidemic progresses. Hence the value of R_{eff} is always smaller than or equal to the value of R_0. In our endemic equilibrium setting, each infectious individual will on average infect one other individual, and thus R_{eff} must be equal to 1 (Farrington, 2003).

Critical Immunization Level

Another interesting parameter related to the basic reproduction number is the critical immunization level (CIL): the minimal proportion of the population that must be immunized by vaccination to eliminate the infection from the population. Consider the simplistic situation where a proportion v of the population is immunized at birth by a fully protective vaccine. The expected number of secondary cases produced by a typical infected person when introduced into the population, assuming that all immunity is vaccine-derived, equals $R_v = (1-v)\cdot R_0$. In order for R_v to be ≤ 1, v must be $\geq 1 - (1/R_0)$, and the latter value is then referred to as the CIL. The larger the value of R_0, the larger the CIL hence the more effort required to globally eradicate the infection. Returning to our simple example in Figure 2.3, we observe that by immunizing $1 - (1/3) = 2/3$ of the population (dotted circles in (b)), on average $R_v = 1$ secondary case is produced which will ultimately lead to elimination of the disease due to chance fluctuations. Table 2.1 presents some values of R_0 and CIL for a selection of well-known infectious diseases, obtained from Farrington (2003). Note that the largest value of R_0 in the list is for measles, while the smallest is for smallpox; a serious, sometimes fatal disease which induced worldwide epidemics during thousands of years. Smallpox is the only human viral infection which has been eradicated up till now by means of vaccination.

2.2 Statistical Inference

In this dissertation, the analyses were carried out within a frequentist framework. In the next sections, we briefly introduce the method of maximum likelihood for parameter estimation and the non-parametric bootstrap method which we use to assess variability. The concept of multimodel inference as a tool to conduct inference from an entire set of candidate models, is described at the end of this chapter. Alternatively, the applications presented in this thesis could have been approached from a Bayesian perspective, though we have not explored any Bayesian methods.

2.2.1 Maximum Likelihood Estimation

Throughout this work, we will use the standard estimation method of maximum likelihood (ML) to infer on unknown parameters for a given model. Suppose we want to estimate k unknown parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ from a set of observed values $\boldsymbol{y} = (y_1, \ldots, y_n)$ of a random sample $Y_1, \ldots, Y_n$, from a distribution subject to heterogeneity. Denote by $f_i(y_i|\boldsymbol{\theta})$ the density function of the random variable Y_i. Since $Y_1, \ldots, Y_n$ are independent, the likelihood function is given by:

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = \prod_{i=1}^{n} f_i(y_i|\boldsymbol{\theta}).$$

An estimate for $\boldsymbol{\theta}$ can be obtained by maximizing the likelihood function over the entire parameter space Θ, and this ML-estimate is denoted by $\hat{\boldsymbol{\theta}}$. It is more convenient, however, to maximize the log-transformed likelihood function $\ell(\boldsymbol{\theta}|\boldsymbol{y}) = \log(L(\boldsymbol{\theta}|\boldsymbol{y}))$ or 'loglikelihood', since this implies calculating the derivative of a sum instead of a product. As the natural logarithm is a monotone increasing function, this optimization is entirely equivalent. The score vector is defined as $S(\boldsymbol{\theta}|\boldsymbol{y}) = \left(\frac{\partial}{\partial\theta_1}\ell(\boldsymbol{\theta}|\boldsymbol{y}), \ldots, \frac{\partial}{\partial\theta_k}\ell(\boldsymbol{\theta}|\boldsymbol{y})\right)$, and the ML-equations which need to be solved are

$$\begin{cases} S_1(\boldsymbol{\theta}|\boldsymbol{y}) = 0 \\ \ldots \\ S_k(\boldsymbol{\theta}|\boldsymbol{y}) = 0. \end{cases}$$

In order for the solution $\hat{\boldsymbol{\theta}}$ to be a maximum, the information matrix $I(\boldsymbol{\theta}|\boldsymbol{y})$ which contains the second order partial derivatives of the loglikelihood, should be positive definite. There are several numerical techniques which can be used to solve the ML-equations, such as Newton-Raphson or Fisher's scoring, the EM-algorithm, and so on. Although an ML-estimator is not necessarily unique or unbiased, it is weakly consistent and asymptotically normal under certain regularity conditions.

Further, we will mainly perform criterion-based model selection, and hereby focus on two information criteria: Akaike's information criterion (Akaike, 1973),

$$\text{AIC} = -2\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) + 2 \cdot k, \tag{2.14}$$

and the Bayesian information criterion,

$$\text{BIC} = -2\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y}) + \log(n) \cdot k, \tag{2.15}$$

as proposed by Schwarz (1978). The AIC is an estimate of the expected, relative Kullback-Leibler (K-L) distance, whereas the K-L distance embodies the information lost when an approximating model is used instead of the unknown, true model. The BIC originates from a Bayesian perspective and more strongly penalizes the number of parameters in the model (factor $\log(n)$ instead of 2). Given a set of candidate models, the 'best' model is the one with the smallest value for AIC or BIC. One should keep in mind that there may exist other models, which are closer to the true underlying model, but were not considered as a candidate. Model selection is always conditional on the set of candidate models considered, so this set should be chosen thoughtfully.

The profile likelihood method can be used to compute approximate confidence intervals (CIs) for each of the parameters (e.g. Cox, 1970). This method reduces the loglikelihood to a function of a single parameter by treating the other parameters as nuisance and maximizing over them. The construction of the confidence interval (CI) is then based on the asymptotic χ^2_1 distribution of the likelihood ratio (LR) test statistic. Suppose we would like to calculate a profile likelihood CI for the parameter θ_j. Then define the 'restricted' parameter space $\Theta_j(x) = \{\boldsymbol{\theta} \in \Theta | \theta_j = x\}$, and the profile likelihood function:

$$\ell^*_j(x) = \max_{\boldsymbol{\theta} \in \Theta_j(x)} \ell(\boldsymbol{\theta}|\boldsymbol{y}),$$

which maximizes the loglikelihood over the other $k-1$ parameters while keeping θ_j fixed at x. The set

$$\{x| - 2(\ell^*_j(x) - \ell(\hat{\boldsymbol{\theta}}|\boldsymbol{y})) \leq \chi^2_{1,(1-\alpha)}\} \tag{2.16}$$

now defines a $(1-\alpha)\%$ profile likelihood CI for θ_j, where $\chi^2_{1,(1-\alpha)}$ denotes the $(1-\alpha)$ quantile of the χ^2_1 distribution. In case $\alpha = 0.05$, the quantile used is $\chi^2_{1,(0.95)} \approx 3.84$. An alternative, distribution-free method to compute confidence intervals is described in the next section.

2.2.2 Bootstrap Inference

In 1979, Bradley Efron introduced the bootstrap (Efron, 1979) which is now a widely used method to estimate the standard error or bias of a given parameter estimate, or to calculate an approximate CI. For a comprehensive account of bootstrap methods, we refer to Efron and Tibshirani (1993). The idea is that further information on the variability of an estimator can be obtained by drawing samples with replacement from the observed data $y_1, \ldots, y_n$, i.e. one sample giving rise to many others. These independent bootstrap samples, denoted by $\boldsymbol{y}^*(1), \ldots, \boldsymbol{y}^*(B)$, usually have the same sample size n as the original data set. Now denote $\hat{\boldsymbol{\theta}}^*(b)$ the bootstrap replication of $\hat{\boldsymbol{\theta}}$ obtained by maximizing the loglikelihood $\ell(\boldsymbol{\theta}|\boldsymbol{y}^*(b))$ for bootstrap sample $\boldsymbol{y}^*(b)$. Note that more generally, bootstrap estimates can be obtained for any statistic of interest from $\boldsymbol{y}$. The bootstrap estimate for the standard error of the ML-estimate $\hat{\theta}_j$ equals:

$$\widehat{\text{se}}_B(\hat{\theta}_j) = \sqrt{\frac{\sum_{b=1}^{B}\left(\hat{\theta}_j^*(b) - \bar{\theta}_j^*\right)^2}{B-1}}, \qquad \text{where} \qquad \bar{\theta}_j^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_j^*(b).$$

There are several bootstrap methods to calculate approximate confidence intervals for model parameters. Efron and Tibshirani (1993) first present the bootstrap-t method, which can be seen as a bootstrap version of the Student's t CI for finite samples. The percentiles of the Student's t distribution are then replaced by percentiles of the B bootstrap versions of the statistic $Z = (\hat{\theta}_j - \theta_j)/\widehat{\text{se}}(\hat{\theta}_j)$. In practice, however, the bootstrap-t method can give somewhat erratic results, and can be heavily influenced by outlying data points (Efron and Tibshirani, 1993, p.160). We will therefore use the more reliable percentile-based bootstrap intervals. Let $\hat{\theta}_{j,B}^{*(\alpha)}$ denote the $100 \cdot \alpha$th empirical percentile of the $\hat{\theta}_j^*(b)$ values, then the approximate $(1-\alpha)\%$ percentile CI for θ_j is

$$[\hat{\theta}_{j,B}^{*\left(\frac{\alpha}{2}\right)}, \hat{\theta}_{j,B}^{*\left(1-\frac{\alpha}{2}\right)}]. \tag{2.17}$$

Percentile intervals, however, may lead to undercoverage and therefore Efron and Tibshirani (1993) proposed an extension called the bias-corrected and accelerated method (BC_a) which has better theoretical coverage properties. Since the latter method involves an additional jackknife procedure, its computation would be much more time consuming for the large data sets we are dealing with (see Chapter 3). Furthermore, it is unclear how the theoretical formula for the BC_a interval could be extended to account for the multiple sources of variability we face in our applications (see e.g. Chapter 5).

Up till now we described the non-parametric bootstrap approach, when sampling is based on the empirical distribution function. Alternatives are the parametric bootstrap when samples are drawn from a specific parametric model, and the semiparametric bootstrap e.g. by resampling model-based residuals in a regression setting. The latter methods require parametric assumptions about the form of the 'true' underlying population, and are therefore generally less useful compared to the non-parametric bootstrap.

2.2.3 Multimodel Inference

We already referred to model selection criteria such as AIC and BIC to select a model from a set of candidate models. By selecting one model and base inferences on that particular model, one implicitly discards the information contained in the other models. Multimodel inference comprises methods for making formal statistical inference from all the models in an a priori set, and for a comprehensive account of the topic we refer to Burnham and Anderson (2002). Suppose we consider a set of m candidate models, and list them according to their AIC value. Let $\text{AIC}_{\min}$ correspond to the model with the smallest AIC value from the set of candidate models considered. The AIC differences $\Delta_i = \text{AIC}_i - \text{AIC}_{\min}$ $(i = 1, \ldots, m)$, estimate the expected relative K-L differences. The best model has $\Delta_{\min} \equiv 0$, and the larger the AIC difference Δ_i, the less empirical support of model i.

These AIC differences can be used to calculate Akaike weights $(i = 1, \ldots, m)$:

$$w_i = \frac{\exp(-\frac{1}{2}\Delta_i)}{\sum_{\ell=1}^{m} \exp(-\frac{1}{2}\Delta_\ell)}, \tag{2.18}$$

which can be interpreted as the weight of evidence in favor (or the probability) of a model i being the actual K-L best model for the situation at hand, given the data and the set of candidate models considered. Thus, the Akaike weights are model probabilities that directly quantify the uncertainty associated with model selection. The evidence ratio ER is then defined as $w_{\min}/w_i$ $(\text{ER} \geq 1)$, and is the relative amount of evidence favoring one model over another. The closer ER is to 1, the more support for that particular model.

Model averaging allows to incorporate model selection uncertainty, quantified by the Akaike weights, into parameter estimates. Suppose that each of the m candidate models allows for an estimate of θ, the parameter of interest. If there is no single model which is clearly superior to the others and if the parameter estimates $\hat{\theta}_i$ differ widely between the models, it is not sensible to base prediction only on the selected

best model. In that situation, it is more feasible to compute a model averaged estimate of θ:

$$\hat{\bar{\theta}} = \sum_{i=1}^{m} w_i \hat{\theta}_i, \tag{2.19}$$

weighting the model estimates by the corresponding Akaike weights w_i.

Chapter 3

Data Sources and Initial Analyses

In our applications, we will make use of two main data sources. The first are serological data sets that, in the absence of an immunization program, represent the age-specific prevalence of past infection in a population. In Section 3.1, the serological data sets for varicella zoster virus in Belgium and parvovirus B19 in five different European countries, and the trivariate data on measles, mumps and rubella from Belgium and Ireland, are introduced. In some applications, we 'augment' the serological data with contact rates obtained from social contact surveys to estimate the WAIFW matrix (cf. Section 2.1.2). Motivational statements to use social contact data, an introduction to the largest social contact survey conducted up till now (as part of the POLYMOD project), and two preliminary analyses highlighting the various informational facets of the POLYMOD contact survey, are given in Section 3.2.

3.1 Serological Data

The serological data presented here consist of cross-sectional sets of residual blood samples collected from hospital laboratories and adult blood donors, which are tested for infection-specific immunoglobulin G (IgG) antibodies using a biochemical technique called 'enzyme-linked immunosorbent assay' (ELISA). The IgG antibody level is an indicator of past infection or vaccination, and the manufacturer of the ELISA-test determines a cut-off value (or range) above which the individual is classified as being seropositive, and below as seronegative. Provided that a serological correlate of protection is agreed upon, the serological status is a direct measure of immunity against the disease. An ELISA-test, however, is subject to diagnostic uncertainty and

misclassification may occur, including both false negatives (seronegative individuals with protective immunity) as well as false positives (seropositive individuals without protective immunity). Bollaerts *et al.* (2011) investigated the effect of test misclassification on the estimation of the prevalence and the force of infection, and proposed a mixture-model based approach for continuous antibody titers to avoid the use of thresholds.

In this thesis, the focus is on dichotomized serological data (binary response variable: 1 = seropositive and 0 = seronegative). These are in fact type I interval-censored data also known as current status data: a person is only observed once and the only information available is his/her current status with respect to the infection, i.e. whether the individual has experienced infection before the sample was taken or not. The seroprevalence is then the proportion of seropositives in the serological sample, and if representative, provides crucial information with regards to disease dynamics in the population. In this respect, it should be kept in mind that the sera we are able to use here are residual samples, i.e. not randomly sampled from the population. However, to avoid bias due to immunocompromised patients, blood samples from patients with specific known infections or disorders (e.g. HIV) or from hospital wards such as oncology, were excluded.

Unlike incidence data such as case reports which are counts of disease notified through passive or active surveillance systems, or laboratory reports which are cases confirmed through laboratory identification, serological data do not suffer from bias originating from changes in clinical awareness or underreporting e.g. due to non-specific symptoms or asymptomatic infection. Grenfell and Anderson (1985) discussed the advantages and disadvantages of using serological data and incidence data and stressed that availability is the key criterion for which type of data to use.

3.1.1 Varicella Zoster Virus

The varicella zoster virus (VZV), also known as human herpes virus 3, is one of the eight herpes viruses known to affect humans. Primary infection with VZV results in varicella (or chickenpox) and mainly occurs in childhood. In general, the disease is benign, however, symptoms may be more severe in adults and complications may occur when varicella is acquired during pregnancy. Infected individuals are highly contagious and transmission of the virus occurs through direct contact with lesions or aerosol contact by saliva and sneezing. The incubation period varies from 13 to 18 days, and a person infected with chickenpox is able to transmit the virus for about 7 days. Antibody response following primary infection with VZV induces lifelong

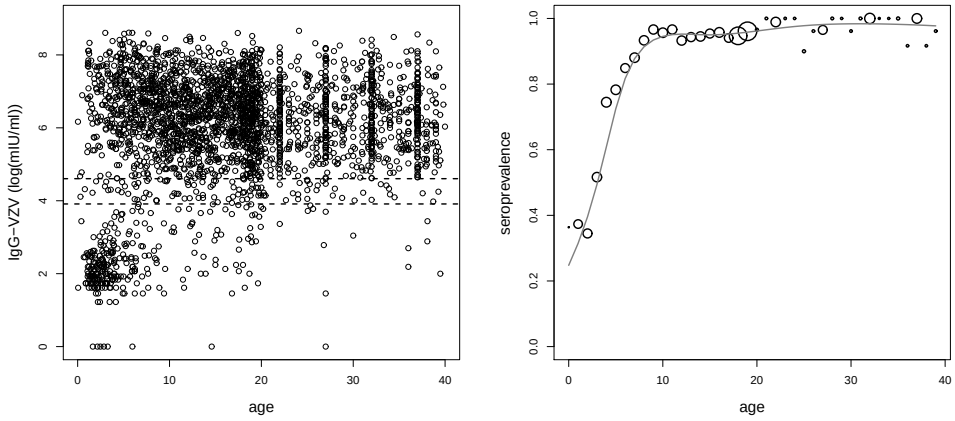

Figure 3.1: Log-transformed IgG antibody titers for VZV in Belgium plotted against age (in years) with equivocal range in dashed lines (left panel) and corresponding seroprevalence profile (right panel) with cubic regression spline model fit (gray solid line).

protective immunity against chickenpox. Upon recovery from varicella, the virus becomes dormant in the body and may reactivate years to decades later when humoral and cell-mediated immunity levels have waned, resulting in herpes zoster (or shingles). Several authors among which Brisson *et al.* (2002) and Thomas *et al.* (2002), have shown that exposure to varicella boosts immunity to herpes zoster, which has important implications for the evaluation of vaccination.

Two types of VZV vaccine are available: one for infants and pre-adolescents to prevent chickenpox, and an adaptation for elderly to prevent herpes zoster. There is no mass vaccination for VZV in Belgium, though in some countries VZV vaccination is part of the universal childhood immunization programme, which is the case for the United States since 1995. Colleagues from the Centre for Health Economics Research and Modeling Infectious Diseases (CHERMID) at the University of Antwerp, are currently investigating which (if any) universal vaccination strategy for VZV would be (cost-)effective in a Belgian context. It is important to assess the impact of active immunization on VZV dynamics, and in particular on the average age at primary infection and on herpes zoster incidence, since both are associated with disease-related burden (Brisson *et al.*, 2003). To this purpose, pre-vaccination epidemiological data are required.

In a period from November 2001 until March 2003, 2760 serum samples were collected in Belgium and tested for the presence of IgG antibodies against VZV as

part of the European Sero-Epidemiological Network 2 (ESEN2) project (Nardone *et al.*, 2007). Together with the ELISA test results, gender and age of the individuals were recorded, the latter which ranged from 0 to 40 years. The resulting antibody titers were standardized to common units and classified into positive, equivocal or negative using an equivocal range of 50-100 mIU/ml (de Ory *et al.*, 2006). The log-transformed standardized titers are displayed in Figure 3.1 (left panel) and the equivocal range is marked with dashed lines. The age-specific transmission of VZV is studied profoundly in Chapter 5 by augmenting the serological data with data on social contacts. The analyses described in Chapter 5 are based on the 2655 univocally dichotomized results (seroprevalence profile in Figure 3.1 on the right), thus excluding the 105 equivocal titers (3.8%). The results from this study served as partial input for an extensive deterministic model evaluating the impact of mass vaccination for VZV in Belgium.

3.1.2 Parvovirus B19

Parvovirus B19 (PVB19) was the first human parvovirus to be discovered in 1975, causing a range of diseases among which erythema infectiosum, commonly known as fifth disease of childhood or slapped cheek syndrome (Anderson and Cherry, 2004). In children and teenagers, the disease is usually mild, but in adults, especially women, it is often complicated by acute arthritis which may persist in some cases (Cohen, 1995). PVB19 can cause transient aplastic crisis in patients with increased erythropoiesis and chronic anemia in immunocompromised patients due to persistent infection (Young and Brown, 2004). Infection with PVB19 during pregnancy has been associated with intrauterine fetal death, fetal anemia and hydrops fetalis (Tolfvenstam *et al.*, 2001). Exposure to children, particularly in the household, has been identified as the main risk factor in pregnant women (Valeur-Jensen *et al.*, 1999). From the onset of rash or arthralgia, after a period of about 17 days, the infected individual is usually no longer contagious, which complicates the detection and control of the virus. Furthermore, subclinical PVB19 infection is a common finding in both children and adults (Heegaard and Brown, 2002). Although it is under development, there is currently no vaccine available for PVB19.

In Belgium (BE), England and Wales (EW), Finland (FI), Italy (IT) and Poland (PL), a seroprevalence survey was conducted totalling 13449 serum samples collected between 1995 and 2004 (Mossong *et al.*, 2008a). The serum samples were tested for the presence of IgG antibodies against PVB19, and the same batch of a commercial immunoassay test was used for each country (Mikrogen recomWell, Martinsried, Ger-

Table 3.1: Summary of the PVB19 serological data collection for each country, and some figures obtained from demographic data: life expectancy L, total population size N, and total number of live births B.

Country	Serological Data			Demographic Data		
	year of collection	age range	sample size	L	N	B
BE	2001-2003	0-65	3075	79	10355197	114001
EW	1996	1-79	2822	77	51125400	649034
FI	1997-1998	1-79	2499	78	5146965	57108
IT	2003-2004	1-79	2514	81	57880478	562603
PL	1995-2004	1-79	2495	73	38651893	382002

many). The resulting antibody titers are depicted on the log scale in Figures 3.2 and 3.3 (left panels). Note that the sampled age range for Belgium was smaller than for the other countries (X-axis). The few equivocal results, located within the cut-off range specified by the manufacturer (dashed lines in Figures 3.2 and 3.3, left panels), are spread over all age groups and excluded from the analyses. The univocal serological data, of which a short summary is presented in Table 3.1, were analyzed before using monotone local polynomials (Mossong *et al.*, 2008a). The study indicated substantial epidemiological differences in Europe regarding PVB19 infection.

It is generally assumed that the IgG antibodies persist for a lifetime (Young and Brown, 2004). However, after an initial monotone increase with age, the seroprevalence profiles for PVB19 in these five European countries show a decrease or plateau between the ages of 20 and 40 (Figures 3.2 and 3.3, right panels), which does not support the assumption of lifelong immunity if the infection is at endemic equilibrium. In Chapter 6, we explore different compartmental dynamic transmission models to investigate whether this phenomenon is induced by waning antibodies for PVB19 and, if this is the case, whether secondary infections are plausible or whether boosting by re-exposure may occur.

3.1.3 Modelling the Seroprevalence

As a preliminary exploration, we simply fit a flexible model to the serological data sets for VZV and PVB19, without taking into account the underlying disease transmission process. Let y_i denote the binary variable indicating whether subject i had experienced the infection before age a_i or not ($i = 1, \ldots, n$). We use a semiparametric model with cubic regression splines (Hastie and Tibshirani, 1990) to relate the current

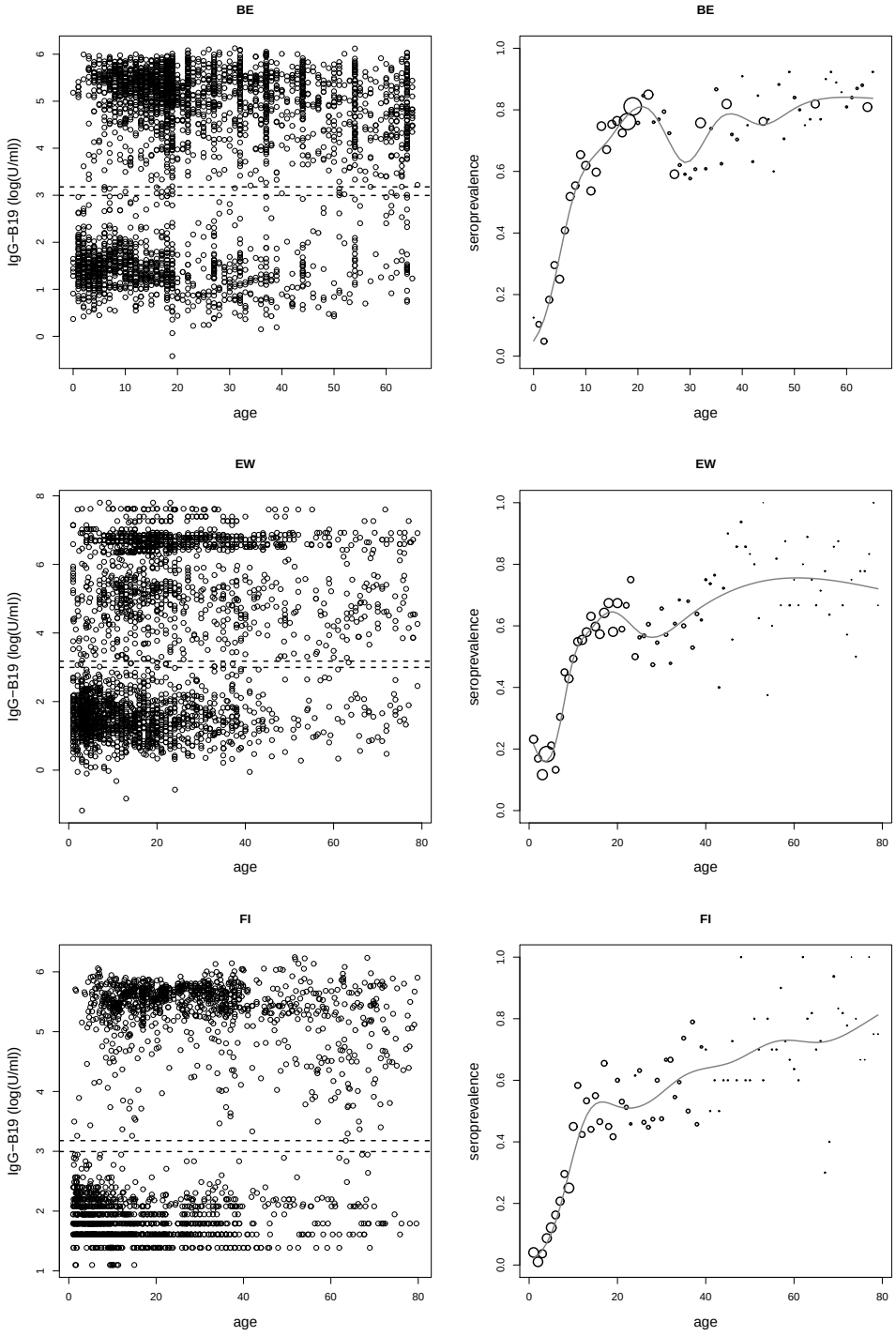

Figure 3.2: Log-transformed IgG antibody titers for PVB19 plotted against age (in years) with equivocal range in dashed lines (left panels) and corresponding seroprevalence profile with cubic regression spline model fit as a gray solid line (right panels), for BE, EW and FI.

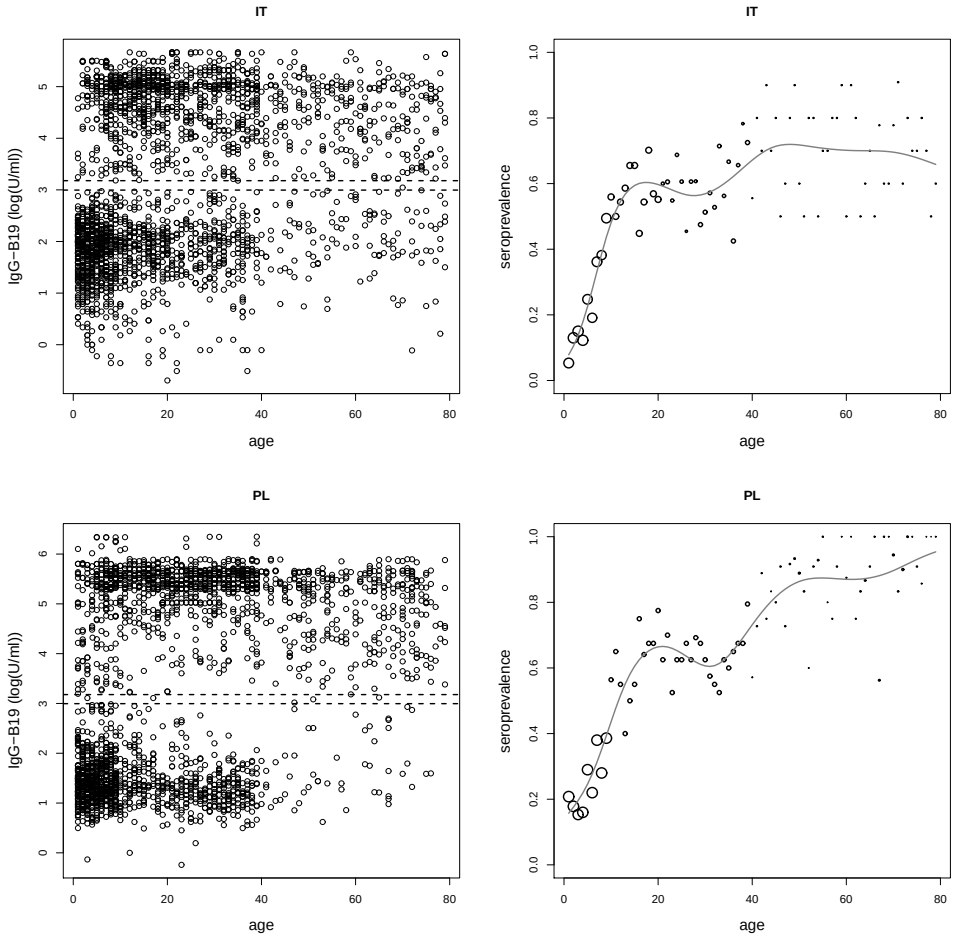

Figure 3.3: Log-transformed IgG antibody titers for PVB19 plotted against age (in years) with equivocal range in dashed lines (left panels) and corresponding seroprevalence profile with cubic regression spline model fit as a gray solid line (right panels), for IT and PL.

status to age in a smooth non-linear way:

$$g\left(P(Y_i = 1|a_i)\right) = \eta(a_i) = \beta_0 + \beta_1 a_i + \beta_2 a_i^2 + \beta_3 a_i^3 + \sum_{j=1}^{K} \beta_{3j}(a_i - \kappa_j)_+^3 \qquad (3.1)$$

with

$$(a_i - \kappa_j)_+ = \begin{cases} a_i - \kappa_j & \text{if } a_i \geq \kappa_j \\ 0 & \text{if } a_i < \kappa_j \end{cases},$$

where g is some link function, η is the linear predictor, and $\kappa_1, \ldots, \kappa_K$ are the knots for the independent variable, i.e. age. Cubic regression splines join cubic polynomials at the knots of the spline to ensure continuity and differentiability up to degree two. For a summary of other penalized spline models, and for an overview of different parametric and non-parametric techniques to model the seroprevalence, such as local polynomials, we refer to Hens *et al.* (2011).

We fit cubic regression spline models with a logit link to the serology, by using the `gam` function from the `mgcv` package in `R` (Wood, 2006). As a default, the knots are spread evenly throughout the covariate values, so according to the quantiles of the age distribution. The smoothing parameter is automatically selected by either unbiased risk estimation (UBRE) or generalized cross-validation (GCV). Figures 3.1, 3.2, and 3.3, display the fit of the cubic regression spline models to the VZV and PVB19 seroprevalence profiles (gray solid lines). For VZV in Belgium, we observe that the proportion of seropositives increases very rapidly in young children up till the age of 10 years. After that age, nearly everyone has experienced infection with VZV and the prevalence stabilizes, reflected by the plateau-shape of the fitted curve. The fitted profiles for PVB19 look very different compared to VZV. There is a less steep increase of the prevalence in children, and the estimated proportion of seropositive teenagers and adults is not as high as for VZV in Belgium. For each country, the spline models estimate a decrease in the seroprevalence between the ages of 20 and 40 years, an observation we already made directly from the data. We note substantial variability for PVB19 prevalence between the different countries. The spline models also accentuate two limitations of the serology for adults and elderly: digit preference due to the use of age classes for BE, and small sample sizes for the four other countries.

3.1.4 Measles, Mumps and Rubella

Measles, mumps and rubella (MMR) are three highly contagious viral diseases that are transmitted from person to person through direct or aerosol contact. If mumps is acquired by a teenage or adult male, a possible complication is painful testicular

inflammation that in rare cases may lead to infertility. Rubella is usually mild and the disease may pass by unnoticed, however, infection during pregnancy can cause the potentially severe congenital rubella syndrome in the newborn. Before the introduction of mass vaccination, measles, mumps and rubella were common childhood infections spread worldwide. The level of protection required to fully prevent circulation is supposed to be the highest for measles (Anderson and May, 1990). In Chapter 7, we analyze two post-vaccination serological data sets for MMR from Belgium and Ireland, which were collected as part of the ESEN2 project.

Universal, combined MMR vaccination was introduced in Belgium in 1985, and in Ireland in 1988. In both countries, a first MMR dose is given in the second year of life. In Belgium, from 1995 onwards, a second MMR dose was recommended to be given at the age of 10-13 years. Similarly in Ireland in 1992, a second dose of MMR was recommended at the age of 10-14 years, however, the target age was reduced to 4-5 years in 1999. Since the start of universal vaccination, the incidence of MMR declined rapidly, however, local outbreaks of measles and mumps still occur due to suboptimal coverage. The circulating MMR vaccines are highly immunogenic, with at least 95% of the individuals developing protective antibodies for each of the three diseases upon vaccination. In clinical trials, the highest seroconversion rate is observed for rubella. While for measles and rubella, vaccine-induced immunity seems long-lasting, the vaccine efficacy for mumps declines to 75% after 10 to 20 years (Cochi *et al.*, 1994; Davidkin *et al.*, 2008). The infections with mumps notified through the PediSurv surveillance network in Belgium are often in MMR vaccinated individuals. This may be attributed to primary (failure to seroconvert) or secondary (waning immunity) failure of the mumps vaccine (Briss *et al.*, 1994; Vandermeulen *et al.*, 2004).

From November 2001 until March 2003, a total of 3378 serum samples were collected in Belgium and tested for IgG antibodies against measles, mumps and rubella. Similarly, in Ireland, a total of 2537 serum samples were collected and tested in 2003. To adjust for laboratory and assay differences which could compromise the comparison between countries, we use the standardized serological results obtained by the method of Kafatos *et al.* (2005). We restrict our analyses in Chapter 7 to the age groups who have been targeted for universal MMR vaccination. For the Belgian data set, these individuals are 1-18 years old and belong to the birth cohorts of 1984-2001. For the Irish data set these individuals are 1-16 years old and belong to the birth cohorts of 1987-2002. Since the sparse structure of the data for the oldest Irish birth cohort causes convergence issues during the analyses, the individuals aged 16 years are left out of consideration. For each age group, the number of individuals in each

Table 3.2: Number of Belgian 2002 sera in each category (seropositive or seronegative to each of measles, mumps and rubella) by age.

Age	+++	++−	+−+	+−−	−++	−+−	−−+	−−−	Total
[1, 2)	25	0	4	3	1	1	3	26	63
[2, 3)	58	0	11	2	1	0	1	7	80
[3, 4)	60	1	10	0	0	0	1	5	77
[4, 5)	57	1	9	0	2	0	1	10	80
[5, 6)	55	0	8	0	0	2	2	8	75
[6, 7)	46	1	11	2	0	0	2	9	71
[7, 8)	51	2	10	2	0	0	4	4	73
[8, 9)	49	0	15	1	1	0	1	7	74
[9, 10)	37	0	12	6	0	1	3	6	65
[10, 11)	56	0	13	4	0	0	1	3	77
[11, 12)	56	1	5	2	2	1	1	3	71
[12, 13)	55	1	3	4	6	1	2	2	74
[13, 14)	49	1	4	2	3	1	3	3	66
[14, 15)	54	1	1	0	4	1	3	2	66
[15, 16)	54	1	5	1	4	0	3	1	69
[16, 17)	61	0	4	0	6	1	1	1	74
[17, 18)	57	3	5	1	1	1	2	2	72
[18, 19)	103	1	15	2	4	0	6	5	136
Total	983	14	145	32	35	10	40	104	1363
	72%	1%	11%	2%	3%	1%	3%	8%	

of the 8 different immunity states, classified based on the univocally dichotomized results for each of measles, mumps and rubella, are presented in Tables 3.2 and 3.3, for Belgium and Ireland, respectively. The serum samples with at least one equivocal test result, which are excluded because of their ambiguous interpretation, constitute 22% and 21% of the Belgian and Irish data, respectively.

3.2 Social Contact Data

Mathematical models of person to person infectious disease spread require assumptions regarding the underlying transmission process (cf. Section 2.1). For infections transmitted by air, respiratory droplets, or direct non-sexual contacts, these assump-

Table 3.3: Number of Irish 2003 sera in each category (seropositive or seronegative to each of measles, mumps and rubella) by age.

Age	+++	++−	+−+	+−−	−++	−+−	−−+	−−−	Total
[1, 2)	28	1	3	7	1	2	0	31	73
[2, 3)	28	3	7	0	0	0	2	12	52
[3, 4)	26	0	10	2	0	0	2	5	45
[4, 5)	36	2	8	0	0	0	3	2	51
[5, 6)	40	2	11	3	0	1	1	5	63
[6, 7)	34	0	7	0	0	0	1	6	48
[7, 8)	42	2	9	0	0	0	2	5	60
[8, 9)	22	0	4	2	0	1	0	3	32
[9, 10)	30	1	5	1	2	0	2	6	47
[10, 11)	21	2	7	2	2	0	0	4	38
[11, 12)	23	0	8	1	0	0	3	6	41
[12, 13)	34	0	2	0	0	0	0	3	39
[13, 14)	45	0	4	2	0	0	4	3	58
[14, 15)	34	0	3	0	0	1	1	1	40
[15, 16)	30	0	6	0	1	0	0	1	38
Total	473	13	94	20	6	5	21	93	725
	65%	2%	13%	3%	1%	1%	3%	13%	

tions are related to human social interactions of which the frequency and intensity typically depend on age. In the traditional approach of Anderson and May (1991), preconditioned mixing patterns in combination with age-specific incidence or serological data, are used to estimate the WAIFW matrix. Recall that the WAIFW matrix represents the age-specific average per capita rate at which two individuals make an effective contact, per time unit (Section 2.1.2). The WAIFW matrix is central to dynamic transmission models, as are closely related parameters like the basic reproduction number R_0 and the age-specific force of infection. Many authors have elaborated on this approach of Anderson and May (1991), among which Greenhalgh and Dietz (1994); Farrington *et al.* (2001); Van Effelterre *et al.* (2009). However, estimates of important epidemiological parameters such as R_0 turn out to be sensitive with respect to the choice of the imposed mixing pattern (Greenhalgh and Dietz, 1994). An alternative method was proposed by Farrington and Whitaker (2005), where contact rates were modeled as a continuous contact surface and estimated from serological

data.

3.2.1 The Quest for Mixing Patterns

The aforementioned methods to estimate age-dependent transmission rates clearly involve a somewhat ad hoc and uncertain choice, namely the structure for the WAIFW matrix or the parametric model for the contact surface. Since this choice is highly influential for quantitative model projections, authors have explored several ways to empirically inform the estimation of age-specific contact rates. Over the last decade, several small scale social contact surveys in which participants had to record information on conversational (and physical) contacts, were conducted to gain more insight in social mixing behavior relevant to the spread of close contact infections: Edmunds *et al.* (1997); Beutels *et al.* (2006); Edmunds *et al.* (2006); Wallinga *et al.* (2006); Mikolajczyk *et al.* (2008). Assuming that transmission rates for infections transmitted predominantly through non-sexual social contacts, are directly proportional to rates of conversational contact, Wallinga *et al.* (2006) were the first to contrast social contact data against seroprevalence profiles (for mumps and pandemic influenza). Alternatively, Del Valle *et al.* (2007) obtained information on mixing patterns from a simulated social network, and Zagheni *et al.* (2008) used time-use surveys to estimate 'time-of-exposure' matrices.

While the contact survey approach directly measures contact rates and assumes that talking with or touching another person constitute the main at-risk events by which an infection can be transmitted, the latter two methods are more indirect and assume that being in the same location at the same time (time of exposure) is an appropriate proxy for at-risk events. Zagheni *et al.* (2008) make the specific assumption of 'proportionate time mixing' which means that, for single activity/location and small time intervals, people allocate their time to the other participants in the activity proportionally to their relative participation in the activity. These methods all have their advantages and limitations, for example, contact surveys do not record events of being in close physical proximity to other individuals and not talking to them (e.g. on public transport), whereas the location-based approaches from Del Valle *et al.* (2007) and Zagheni *et al.* (2008) may underestimate the level of mixing assortativeness since the locations are generally not stratified by age. It could be interesting to compare and combine the different methods to estimate the WAIFW matrix, however, that was beyond the scope of this thesis.

Since the social contact surveys carried out until that time were in small or non-representative populations, the European commission project 'POLYMOD' conducted

large population-based surveys between May 2005 and September 2006 (Mossong *et al.*, 2008b). These prospective surveys of social contacts were held in eight European countries: Belgium (BE), Germany (DE), Finland (FI), Great Britain (GB), Italy (IT), Luxembourg (LU), The Netherlands (NL) and Poland (PL). For an extensive description of the survey methodology and results from exploratory data analysis, we refer to Mossong *et al.* (2008b). In the next section, a brief outline of the survey design and the main findings from Mossong *et al.* (2008b) are summarized. Two initial applications of the POLYMOD survey from Hens *et al.* (2009b) are described in Sections 3.2.3 and 3.2.4. In Chapter 4, the Belgian contact survey is studied in more detail by means of a data mining analysis.

3.2.2 POLYMOD Contact Survey

Survey participants were recruited in such a way as to be broadly representative of the whole population in terms of age, sex, and geographical spread. Children and adolescents were deliberately oversampled, because of their important role in the spread of infectious agents. Only one person in each household was asked to participate in the study. Paper diaries (excerpt in Figure 3.4) were sent by mail or given face to face, and participants were explained by telephone or in person how to complete the diary. For young children, a parent or exceptionally another adult caregiver filled in the diary. Teenagers filled in a simplified version of the diary and were closely followed up to anticipate interpretation problems. A total of 7 290 participants to the study completed the diary, recording a total of 97 904 contacts made during one randomly assigned day. A short summary of the survey methodology and sample sizes for each country are provided in Table 3.4.

Participant-related information such as age, gender and occupation had to be recorded in the diary as well as details about each contact made (Figure 3.4): age and gender of the person made contact with, location or circumstance of the contact (multiple options possible), total duration of the contact (over the entire day), and frequency or habitual nature of the contact. In case the exact age of the contacted person was unknown, participants had to provide an estimated age range. In all analyses requiring the age of the contacted person, the median of the age range was used as a surrogate. Further, participants had to distinguish between two types of contact: non-close contacts, defined as two-way conversations of at least three words in each others proximity, and close contacts which involved any sort of physical skin-to-skin touching. Using EUROSTAT census data on population sizes of different age by household size combinations for the year 2000, post-stratification weights were given

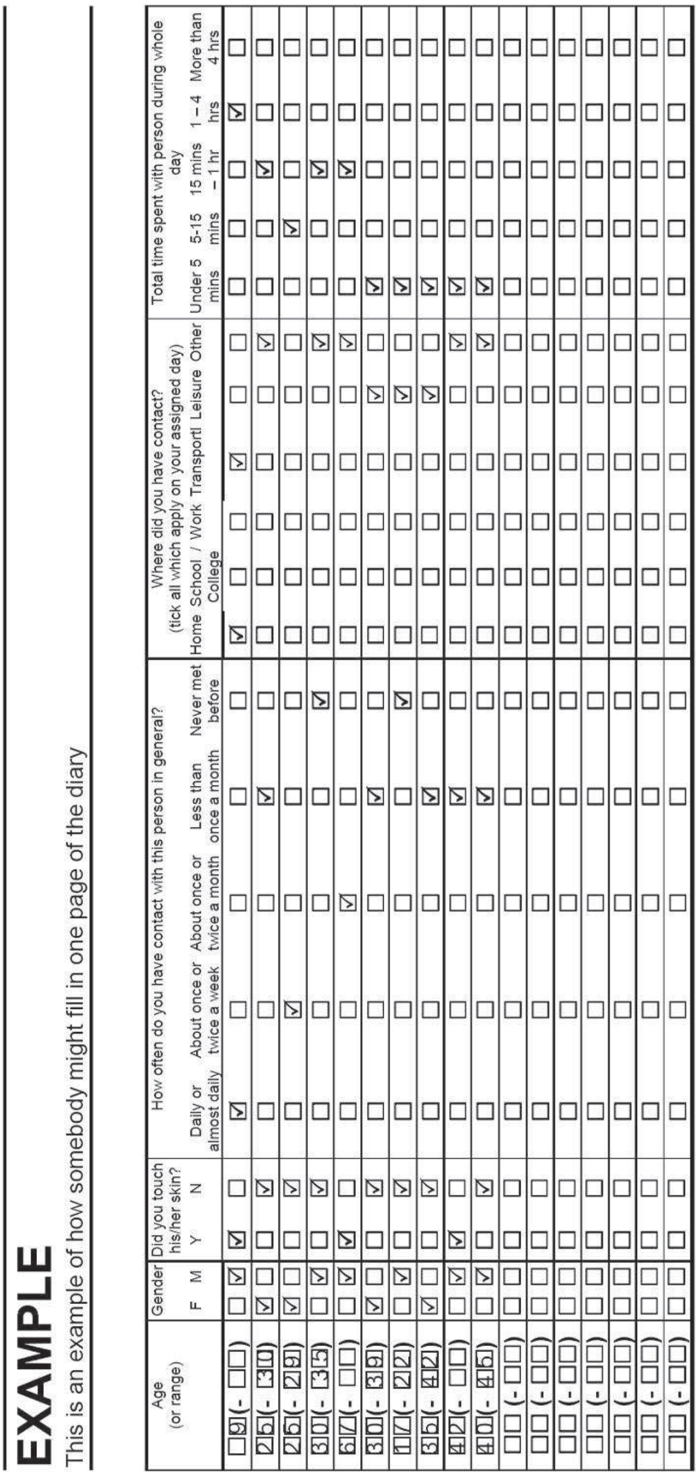

Figure 3.4: Fictive example of one diary page as collected in the POLYMOD Contact Survey.

Table 3.4: Details of the survey methodology, the total number of participants, and the total number of recorded contacts, for each country.

Country	Recruitment (max # contact entries)	No Professional Contact Recording	# Participants (missing age)	# Contacts (missing age)
BE	random digit dialling (90)	if > 20	750 (0)	8880 (3)
DE	face-to-face interview (73)	if > 10*	1341 (46)	10659 (107)
FI	population registers (34)	if > 10	1006 (0)	11128 (0)
GB	face-to-face interview (29)	not instructed	1012 (0)	11876 (3)
IT	random digit dialling (45)	not instructed	849 (7)	16784 (0)
LU	random digit dialling (55)	not instructed	1051 (0)	18352 (0)
NL	population registers (45)	if > 10	269 (12)	3726 (8)
PL	face-to-face interview (45)	not instructed	1012 (0)	16501 (2)
		Total	7290 (65)	97906 (123)

* Note that for DE no participants recorded more than 10 professional contacts.

to the participants in order to make the data representative of the different populations. In all countries except DE, single-person households were underrepresented in the sample, which can be partially explained by the oversampling of children and adolescents who tend to live in larger households.

The analyses conducted by Mossong *et al.* (2008b) showed that age-specific mixing patterns and contact characteristics were very similar across different European countries, even though the average number of contacts recorded differed. Figure 3.5 displays the contact rate matrices estimated by Mossong *et al.* (2008b) using a bivariate smoothing approach with 5 year age bands whilst incorporating post-stratification weights (technical details on bivariate smoothing are provided in Chapter 5). The plot revealed a strong diagonal component: contact patterns were highly assortative with age; particularly schoolchildren and young adults tended to mix with people of the same age. Two secondary off-diagonals presented parent-child mixing, though the contact rates were an order of magnitude lower than the main assortative diagonal. A wider 'plateau' of adults mixing with other adults was apparent as well, and primarily due to low-intensity contacts of which many occurred at work.

Interestingly, the mixing patterns displayed in Figure 3.5 are qualitatively similar to the matrices obtained by Del Valle *et al.* (2007) and Zagheni *et al.* (2008) who used simulated social network data and time-use data, respectively. While this is already a first step towards validation of the survey, future research should assess whether the results are reproducible by using other data collection methods, and whether similar contact patterns arise when conducting the survey in other countries or parts

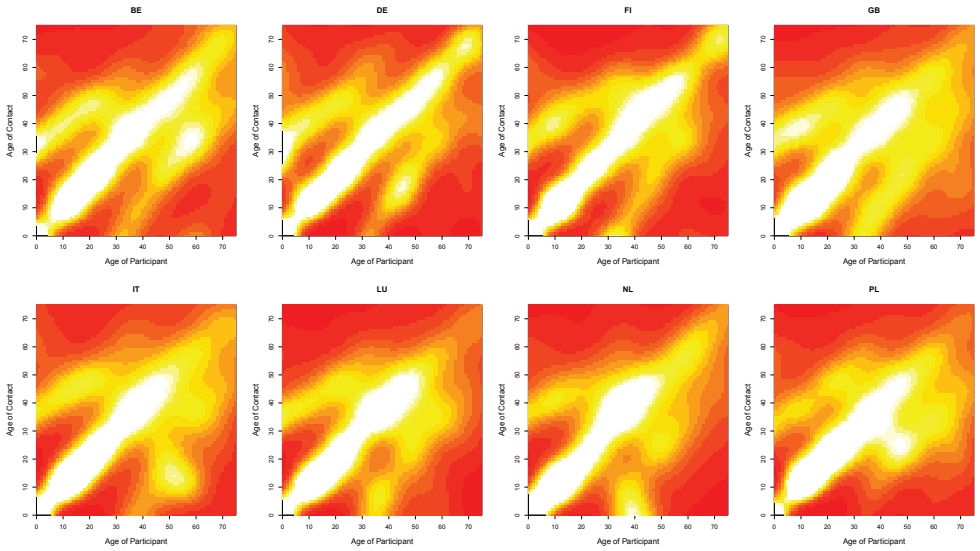

Figure 3.5: Smoothed contact matrices based on all reported contacts for BE, DE, FI, GB, IT, LU, NL, and PL, respectively. White indicates high contact rates, yellow intermediate contact rates, and red low contact rates, relative to the country-specific contact intensity.

of the world. In a small Australian study, McCaw *et al.* (2010) compared three different methods to collect contact data using a location-based reporting design: a pre-entry questionnaire, a paper diary, and an electronic recording device. Similar to the POLYMOD survey, McCaw *et al.* (2010) found high levels of assortative mixing and adult-child mixing within families. The participants rated the use of the paper diary higher than the electronic recording device, an more encounters were captured using the paper diary compared to the other methods.

Further, Mossong *et al.* (2008b) found that contacts lasting at least one hour or occurring on a daily basis mostly involved close contact, while short duration and infrequent contacts tended to be non-close. Contacts at home, school, or leisure were more likely to be close than contacts at work or while traveling. A high degree of association was established between close contact and other measures of more intimate contact (e.g. of long duration, occurring frequently,. . .), suggesting that close contacts may serve as a proxy for high intensity contacts. Preliminary modelling indicated that in a completely susceptible population, 5- to 19-year-olds are expected to suffer the highest incidence during the initial epidemic phase of an emerging infection transmitted through social contacts. Ultimately, the contact data should lead to improved parametrization of mathematical models used to design intervention strategies.

Nevertheless, it is good to bear in mind some limitations of the POLYMOD survey. First, underrepresentation of certain groups in the population is partially accounted for by means of post-stratification (e.g. for single-person households), however, a small percentage of the population is not represented by the survey. In BE, IT and LU, sampling was done through random digit dialing using land lines (Table 3.4), thus automatically excluding people who do not have land line telephones in their household. Considering the rise of cell phone use in recent years, this sampling method may compromise the representativeness of similar future surveys. Further, the data are self-reported which may be a potential source of bias, particularly with regard to underreporting of contact encounters. In countries where respondents were asked to indicate whether they encountered problems to fill in the diary, only a small percentage of participants indicated that they had so. This suggests that the questionnaire and recording instructions were readily accepted and understood by responding participants.

Parental proxy reporting for young children may have been poor when they were not spending time together, for instance when the child was in daycare or kindergarten. However, it is a practical and feasible method, and alternatives such as direct observation by an independent observant are likely to influence the child's (contact) behaviour. As long as there is no important trend of contact underreporting with age, the use of the contact data to infer on mixing patterns is still relevant since it is most important to grasp the age-specific relative differences (heterogeneities) in contact intensities. Also, the data are egocentric and clustering of contacts (relationships or contacts between contacted persons etc.) was not recorded. Therefore, one cannot directly infer on the underlying social networks, though Potter *et al.* (2011) developed a latent variable method to estimate within-house contact networks from the POLYMOD data. Finally, comparisons between countries should be made with caution due to the variations of survey design, recruitment, and follow-up. Harmonizing survey methods should be a point of attention for future multi-country contact surveys.

3.2.3 Modelling the Number of Contacts

In Hens *et al.* (2009b), we studied the effect of participant characteristics on the total number of reported contacts from the POLYMOD contact survey using a negative binomial regression model. Our analysis of contact counts differed from the one in Mossong *et al.* (2008b) in two ways. Professional contacts were not surveyed in the same way for all countries. Indeed, in the diary for BE, DE, FI and NL, participants were instructed not to list their professional contacts if their number would exceed

a certain threshold (see Table 3.4). Instead, participants had to provide an estimate for the average number of persons they encountered professionally each day. Whereas Mossong *et al.* (2008b) based their analysis on the contact counts only encompassing the fully recorded contacted persons, we took these 'extra' professional contacts into account and thus improved the comparability of the results between countries. The second difference is that apart from participant's age, gender, household size, day of contact recording, and country, we have added another covariate effect to the model, namely whether or not the sampled day was a (school or public) holiday.

Since for some of the surveys the number of possible contact entries was limited (see Table 3.4, max # contact entries), the number of contacts Y is right censored. For reasons of uniformity, the minimum of these limits i.e. 29 contacts for the survey in GB, is used for the censored negative binomial regression model. The loglikelihood function for all n respondents is then given by:

$$\sum_{i=1}^{n} w_i \left\{ \delta_i \log\big(P(Y = y_i|\boldsymbol{X}_i)\big) + (1-\delta_i) \log\left(1 - \sum_{j=0}^{28} P(Y = j|\boldsymbol{X}_i)\right)\right\}, \qquad (3.2)$$

where y_i is the observed number of contacts (including work contacts) for respondent i, w_i is the post-stratification weight as described in Section 3.2.2, and $\boldsymbol{X}_i$ is the vector of explanatory variables. Here, $\delta_i = 1$ if $y_i < 29$ and 0 otherwise, and P is the density function of the negative binomial distribution:

$$P(Y = y_i|\boldsymbol{X}_i) = \frac{\Gamma(y_i + \frac{1}{\alpha})}{\Gamma(y_i+1)\Gamma(\frac{1}{\alpha})} \left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}} \left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i}, \qquad (3.3)$$

where $\mu_i = \exp(\boldsymbol{X}_i\boldsymbol{\beta})$ is the mean which is linked to the covariates via a log-link function with $\boldsymbol{\beta}$ the vector of unknown coefficients, and $\alpha \geq 0$ is the overdispersion parameter. The variance is then given by $\mu_i + \alpha\mu_i^2$, thus when $\alpha = 0$, the negative binomial distribution simplifies to the Poisson distribution. Hens *et al.* (2009b) compared the performance of this model to a zero-inflated negative binomial counterpart to accommodate to excess zeros amongst the contact counts, and noted that zero-inflation was non-significant (p-value of 0.317). The more parsimonious regression model is therefore presented in Table 3.5.

The dispersion parameter estimate equals 0.41 (95% CI: [0.40, 0.43]), indicating the necessity of taking overdispersion into account. Participants in the 10-49 years age-category have the highest number of contacts, while participants above the age of 70 years have the lowest number of contacts followed by children younger than 5 years. There is no significant difference in the number of contacts made by males and females. Respondents living in larger households have a higher number of contacts.

Table 3.5: Weighted censored negative binomial regression model: sample mean and standard deviation (s.d.), and model-based relative number of contacts.

Covariate	Category	Sample Size	Mean (s.d.)	Relative # Reported Contacts [95% CI]
Age	< 5	660	10.21 (7.65)	1.00
	5-9	661	14.81 (10.09)	1.42 [1.27, 1.56]
	10-14	713	18.69 (13.40)	1.76 [1.58, 1.94]
	15-19	685	19.93 (21.14)	1.79 [1.61, 1.97]
	20-29	879	17.18 (25.72)	1.66 [1.51, 1.81]
	30-39	815	17.83 (21.68)	1.63 [1.49, 1.78]
	40-49	908	17.51 (23.29)	1.57 [1.43, 1.70]
	50-59	906	15.96 (20.84)	1.48 [1.35, 1.62]
	60-69	728	10.51 (14.47)	1.10 [1.00, 1.21]
	70+	270	7.71 (10.97)	0.81 [0.73, 0.89]
	missing*	65	10.40 (12.78)	0.94 [0.65, 1.23]
Gender	female	3808	16.13 (21.93)	1.00
	male	3429	15.14 (15.57)	0.97 [0.94, 1.01]
	missing**	53	10.92 (8.60)	1.60 [1.06, 2.14]
Household size	1	749	11.23 (18.26)	1.00
	2	1645	13.32 (17.89)	1.20 [1.13, 1.27]
	3	1683	14.67 (16.44)	1.23 [1.15, 1.31]
	4	2041	17.71 (17.67)	1.38 [1.29, 1.47]
	5	814	19.49 (29.12)	1.44 [1.34, 1.55]
	6+	358	19.30 (13.14)	1.63 [1.48, 1.79]
Day of the week	Sunday	862	11.98 (14.54)	1.00
	Monday	1032	16.36 (27.65)	1.35 [1.26, 1.45]
	Tuesday	1116	16.69 (20.16)	1.40 [1.31, 1.50]
	Wednesday	1017	16.93 (18.39)	1.40 [1.31, 1.50]
	Thursday	1069	16.86 (16.31)	1.41 [1.31, 1.51]
	Friday	1122	17.00 (18.25)	1.42 [1.33, 1.52]
	Saturday	936	12.85 (14.52)	1.19 [1.11, 1.28]
	missing***	136	12.85 (12.26)	1.44 [1.20, 1.68]
Country	BE	750	19.30 (24.31)	1.00
	DE	1341	7.95 (6.26)	0.49 [0.46, 0.53]
	FI	1006	18.46 (32.15)	0.86 [0.80, 0.93]
	GB	1012	11.74 (7.67)	0.72 [0.67, 0.77]
	IT	849	19.77 (12.27)	1.18 [1.08, 1.27]
	LU	1051	17.46 (12.81)	1.02 [0.94, 1.09]
	NL	269	24.92 (42.70)	1.41 [1.25, 1.56]
	PL	1012	16.31 (11.45)	0.97 [0.89, 1.04]
Period	regular	6106	16.15 (19.64)	1.00
	holiday	1048	12.93 (16.46)	0.91 [0.86, 0.96]
	missing***	136	12.85 (12.26)	1.09 [1.01, 1.16]
Overdispersion	α			0.41 [0.40, 0.43]

* Missing age was equally distributed over the other variables.

** Missing gender was associated with weekday, regular period and household size 1-4.

*** Missing day of the week/period was associated with DE, GB, LU and household size 2-4.

Participants have a larger number of contacts on weekdays compared to the weekend, and significantly fewer contacts on Sunday in comparison to Saturday. IT and NL have a relatively high number of contacts compared to BE, LU and PL whereas DE, FI and GB have a relatively low number of contacts. The mean number of contacts for DE, GB, IT, LU and PL are the same as published by Mossong *et al.* (2008b), however, there is a pronounced rise for BE, FI and NL by inclusion of contacts at work. The number of reported contacts is significantly lower during the holiday period compared to the regular period. The differences between the sample estimates (mean and s.d. in Table 3.5) and the model-based relative number of reported contacts indicate that it is important to control for the different participant characteristics.

3.2.4 Impact of School Closure on Disease Transmission

The POLYMOD contact survey allowed us to investigate the relative change in the basic reproduction number R_0 during the week versus weekends and during regular versus holiday periods, encompassing both school as well as public holidays (Hens *et al.*, 2009b). As schools are closed during weekends and holiday periods, the relative change in R_0 provides an indication of the impact collective school closures and prophylactic absenteeism may have during a pandemic (see e.g. Cauchemez *et al.*, 2008). Prophylactic absenteeism means 'healthy people avoiding social contact as a means of protection, including absence from work and school', and it reflects the role of public perception and confidence.

To this purpose, we relied on the social contact hypothesis introduced by Wallinga *et al.* (2006), which states that the age-specific transmission rates are directly proportional to the age-specific rates of making social contact (denoted by $c(a, a')$):

$$\beta(a, a') = q \cdot c(a, a'). \tag{3.4}$$

In subsequent Chapters, we will also refer to this assumption as the 'constant proportionality' assumption, since q represents a constant disease-specific proportionality factor. Assuming type I mortality (2.7), the relative change in R_0 i.e. the dominant eigenvalue of the next generation operator (2.13), was estimated from the smoothed contact rate matrices $\boldsymbol{C}$ as follows:

$$\frac{R_{0,1}}{R_{0,2}} = \frac{\text{max eigenvalue}\left(\frac{ND}{L}\, q\, \boldsymbol{C}_1\right)}{\text{max eigenvalue}\left(\frac{ND}{L}\, q\, \boldsymbol{C}_2\right)},$$

where indices 1 and 2 refer to the contacts recorded during the weekend (Saturday and Sunday) and the entire week (Monday to Sunday), or during the holiday and the

regular period, respectively. It is straightforward to see that the constants $\frac{ND}{L}\,q$ cancel out and that the ratio only relates to the contact matrices. For further methodological details we refer to Hens *et al.* (2009b).

Table 3.6 presents the resulting estimates for the weekend-week comparison together with 95% bootstrap-based percentile CIs (2.17). Extra professional contacts were not taken into account in the analysis due to the lack of covariate information. Based on all recorded contacts, a significant decrease in R_0 of at least 12% up to 26% during the weekend when compared to the entire week was shown in all countries, except for DE and FI. For close contacts, these differences were less pronounced and the significantly lower R_0 were again observed for BE, GB, IT, LU, NL and PL, ranging from 5% to 21% (Table 3.6).

Here, we do not present the results of the holiday versus regular period comparison, since it was only possible for half of the countries and somewhat compromised by regionally divergent holiday periods. For these results together with a visualization of the change in contact behavior via 'score matrices', we refer to Hens *et al.* (2009b). In short, the results revealed that social contact patterns differed substantially when comparing the week to the weekend and regular to holiday periods, and that this was mainly due to the reduction in work and/or school contacts. For most countries the basic reproduction number decreased by about 21% and 17%, respectively, although for some no significant decrease was observed.

School closure thus could have a substantial impact on the spread of a newly emerging infectious disease that is transmitted via non-sexual social contacts. On the other hand, House *et al.* (2010) showed for the UK that local, reactive school closures would require considerable coordination to achieve a substantial reduction in the number of hospitals that are over intensive care unit capacity at the peak of an influenza pandemic. Furthermore, there is an ethical tradeoff which needs to be made since school closures could result in severe economical costs due to childcare (Sadique *et al.*, 2008; Cauchemez *et al.*, 2009; Smith *et al.*, 2009; Keogh-Brown *et al.*, 2010).

Table 3.6: Relative change in R_0 for the weekend versus the week for all contacts and close contacts. '*' indicates a significant relative change in R_0.

Country	Sample Size	All Contacts		Close Contacts	
	weekend/week	rel. change R_0	95% CI	rel. change R_0	95% CI
BE	202/746	0.78*	[0.64, 0.94]	0.88*	[0.86, 0.93]
DE	266/1307	1.02	[0.83, 1.21]	1.03	[0.68, 1.39]
FI	283/999	0.78	[0.73, 1.16]	0.88	[0.85, 1.18]
GB	258/968	0.88*	[0.69, 0.90]	0.95*	[0.74, 0.97]
IT	226/840	0.80*	[0.63, 0.82]	0.79*	[0.68, 0.99]
LU	205/993	0.74*	[0.70, 0.74]	0.88*	[0.66, 0.89]
NL	68/257	0.78*	[0.59, 0.79]	0.79*	[0.62, 0.81]
PL	280/1002	0.77*	[0.66, 0.89]	0.84*	[0.71, 0.86]

Chapter 4

Mining the Belgian Contact Survey

The POLYMOD contact survey from Belgium differs from the surveys conducted in the other countries in several aspects. First, each participant recorded contacts during two randomly assigned days instead of one (random order): one weekday and one day in the weekend (Saturday or Sunday). The bivariate contact counts allow to estimate within subject contact correlation. Second, participants can be linked to their geographical location, enabling the comparison of contact behavior between different Belgian regions. Third, the sampling period included a school holiday period, which facilitates simulations of the impact school closure may have during an epidemic outbreak (cf. Section 3.2.4). Finally, to reduce reporting bias, participants with more than 20 professional contacts were requested not to record them in the diary (Table 3.4), but to separately give summary estimates in terms of number and age of the contacts instead. Mossong *et al.* (2008b) did not take these four aspects of the Belgian contact survey into account, and for reasons of comparability they analyzed a subsample containing one randomly selected day per participant.

In this chapter, we give a detailed description and thorough data analysis for the complete Belgian contact survey, and hereby closely follow Hens *et al.* (2009a). Section 4.1 completes the data collection description from Section 3.2.2 specifically for the Belgian situation, and introduces the imputation method we will use to augment the data with the extra professional contacts. In Section 4.2, we conduct data mining analyses using association rules and classification trees to discover patterns of contact characteristics. Generalized estimating equations are applied in Section 4.3 to regress the contact counts on participant-related information, taking into account the correlation between the bivariate contact counts. Finally, we make use of the next

generation matrix in Section 4.4 to mimic the spread of a newly emerging infection in Belgium. We finish the chapter with some concluding remarks.

4.1 Belgian Contact Survey

4.1.1 Data Collection

In a period from March until May 2006, 750 persons living in Belgium were recruited by random digit dialing on fixed telephone lines. The respondents were asked to anonymously complete a paper booklet (referred to as the 'diary') containing a questionnaire and a contact diary, without changing their usual behaviour. No persons were subject to interventions and no physical samples were collected as part of the study. The study protocol was approved by the ethical committee of the Antwerp University Hospital. Three different types of diaries were spread accommodating for the age of the respondent: children ($<$ 9y), teenagers (9-17y), and adults ($\geq$ 18y). In addition to a pilot study in Luxembourg, specific Belgian Dutch and French language versions of each type of diary were made and tested in Belgium (Beutels *et al.*, 2006; Mossong *et al.*, 2008b).

The diaries were sent and collected by mail. Each participant was reminded by phone that they had to fill in the diary, one day prior to each assigned day, and was followed up after the first day to check whether they had. If they had not filled in the diary, they were assigned to a new random day. If they had not returned the diary, they were reminded by a maximum of three follow up calls to send it in. If participants repeatedly failed to fill in the diary on their assigned day, they were excluded, and replaced by a new recruit. Each participant received a small token of appreciation for the amount of €5. All diaries were double entered in a computer database and checked manually.

In line with the total Belgian population (Chi-square test for equality of distributions, p-value 0.85), participants were recruited from the Flemish ($n = 441$), Walloon ($n = 239$) and Brussels geographic regions ($n = 70$). Sampling was undertaken in order to obtain the following age distribution: 10% in the 5-year age groups 0-4, 5-9, 10-14, 15-19 years, and 6% in age groups 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, $\geq$ 65 years. Most participants ($n = 480$) were adults (cf. age group definitions above) of whom 50% were male, which is representative for the Belgian population (p-value 0.70). There were also 130 teenagers (49% males, p-value 0.60) and 140 children (56% males, p-value 0.053).

Virtually all participants (98%) recorded their contacts in the four weeks from

March 18 to April 14, 2006, and 49% and 73% of the recorded weekdays and weekend days, respectively, were during the Eastern holiday period (April 03-17). Note that weekends at the start and the end of the holiday period were considered as holiday as well and represented 72% of all recorded weekend days. About half of the children aged 0-2 years attended childcare (55%), and almost all children aged 3-8 years attended childcare or school (93%), while school participation was 100% for participants aged 8-17 years. Adults aged 18 years or older were mostly employed (49%), unemployed (35%) or in further education (9%). Overall, 10% of the participants lived alone and these were all 18 years or older. Nearly 1 in 5 children aged 0-8 years lived in a single parent family, and 26%, 28% and 11% of the participants lived in a household of size 2, 3 and 4, respectively. Larger household sizes were only rarely observed (4%). These characteristics of the sample are all broadly in line with general Belgian population statistics (National Institute for Statistics, 2006, Belgium).

Although in order to keep the diary manageable, we intentionally did not ask to record the relationship with the contacted persons, we were able to extract household-like contact data from the survey. The persons in the household were identified as those contacts with exactly the same age as the registered ages of the household members, which took place at home. We performed a sensitivity analysis with respect to the selected contacts of the same age (at home).

4.1.2 Professional Contacts

In Section 3.2.3, we already raised the difference in professional contact recording between the different POLYMOD participating countries. Half of the countries (namely BE, DE, FI and NL) requested the respondents not to report their contacts at work in the diary, if their estimated number would exceed a predefined threshold value (Table 3.4). In the Belgian diaries, the participants were asked prior to filling in their dairy for the first time the following sequence of questions:

1. Do you have a profession through which you have a large number of contacts (e.g., clients, patients, students, etc)? ☐ YES ☐ NO;

2. If YES, please give an estimate of the average number of persons you contact professionally each day? ..... persons;

3. Tick in which of the following age categories these professional contacts mostly occur (multiple options possible):
   ☐ 0-5 years ☐ 6-11 years ☐ 12-17 years ☐ 18-60 years ☐ over 60 years;

4. If you estimated the number of these contacts to be more than 20, then please do not record these contacts in the diary, but only record the other (non-professional) contacts.

Although these questions were asked with the common intention to reduce reporting bias for people with many professional contacts (e.g. bus drivers), there are some important differences in how we approached this, in comparison to DE, FI and NL. First, based on information from the pilot studies, we set the threshold value at 20, whereas in DE, FI and NL, it was set at 10. Second, we asked first if the participants thought they had many professional contacts, and only if they did subjectively think so, how many they estimated these to be. Third, we only revealed in the last question what the consequence of their estimate was for the effort required to complete the diary. Fourth, we did not only ask about the number of professional contacts, but also about their usual age range.

In Section 3.2.3 we already presented a regression analysis from Hens *et al.* (2009b) for the total number of contacts, taking these 'extra' professional contacts into account. Hence, unlike Mossong *et al.* (2008b), underestimation of the average contact counts for BE, DE, FI and NL, was avoided. From now on, we make use of all available information about the professional contacts and perform imputation to complete the Belgian data. This too has its limitations, since imputation enables generating data from which reliable inferences can be made, but can not recreate the values that were not recorded (Little and Rubin, 1987). More specifically, the ticked age categories for the contacts at work provide a basis for the imputation procedure. Let n_i^w denote the estimated number of professional contacts, which needs to be imputed for participant i if $n_i^w > 20$, and let I_i^a denote the corresponding age range (e.g. $I_i^a = [6, 12)$ for a primary school teacher). We then sample n_i^w age values from I_i^a with sampling probabilities according to the population age distribution. Contrasting this method with contact data from GB, IT, LU and PL, indicates good performance.

In order to impute the other contact characteristics, plausible assumptions are made based on the available information from GB, IT, LU and PL. The distributions of the type of contact and the gender of the contacted person do not change substantially with the number n_i^w of contacts reported at work. Therefore, we impute the physical nature and the gender of the professional contact in the Belgian contact data set using the same distribution as when the number of professional contacts is $10 < n_i^w \leq 20$. The imputation of other contact characteristics like duration and frequency seemed more speculative. For instance, the higher the recorded n_i^w was, the shorter the contact durations were. Nonetheless, choosing a specific distribution for the duration

of contacts would be subjective, since the information for Belgium is clearly missing and the distributions vary substantially between countries. We consider it unlikely that a single professional contact in a large set of such contacts would last longer than 4 hours and reoccur daily. Therefore, we impute the two variables jointly by sampling from the bivariate distribution of duration and frequency for work contacts of participants for whom $10 < n_i^w \leq 20$. This method could also be applied more widely to all characteristics in an attempt to avoid disrupting dependencies, but could just the same also enforce dependencies.

For the remainder, we focus on the augmented Belgian data set, which in its POLYMOD version consists of 13 786 contacts recorded by 750 participants during one day (compared to the sample size of 8880 contacts presented in Table 3.4). Specifically in this chapter, we make use of all data available after imputing the professional contacts, which is 23 683 contacts recorded by 750 participants during two days. One female adult respondent recorded an estimate of 1000 contacts at work and was considered an outlier to the data set. Since this person is likely very influential e.g. when estimating the Belgian contact surface using bivariate smoothing, she is excluded from the analyses presented here and in subsequent chapters. Hence, the analyzed data comprise 749 participants who recorded a total of 12 775 contacts during one day, and 22 666 contacts during two days. Note that the results of the data analyses presented hereafter, are similar whenever the imputed professional contacts are left out, except when estimating the average number of contacts. The former was illustrated in Hens *et al.* (2009a) for the data mining analysis with classification trees, described in the next section.

4.2 Elucidating Highly Intimate Contacts

Person to person transmission of infectious diseases is generally more likely to occur during more intimate contacts, such as contacts involving skin-to-skin touching, contacts of long duration or contacts on a frequent basis. We use two data mining techniques, namely association rules and classification trees, to highlight interesting associations and relations between contact properties such as type of contact (close or non-close), location, frequency and duration. The aim is to characterize contacts with high risk of infectious disease transmission, when the main transmission route is through social contacts of the non-sexual type (e.g. droplet contact, airborne transmission, etc.).

4.2.1 Data Mining Methods

Association Rules

With the aim to discover meaningful patterns in large transactional databases, Agrawal *et al.* (1993) introduced the idea to mine association rules by means of finding frequently co-occurring items. An important example in this respect is the 'market basket analysis': identifying which supermarket products are frequently purchased together by a customer in a transaction. The event of a contact between two persons can also be interpreted as a transaction, and the contact characteristics (e.g. close contact) then constitute a set of binary items. Next, to introduce some basic concepts, we mainly follow the description from Hahsler *et al.* (2005) and Hahsler and Hornik (2007). An association rule is a rule of the form $X \Rightarrow Y$, where X and Y are two disjoint sets of items ($X \cap Y = \emptyset$). The rule means that if we find all items in X in a contact it is likely that the contact also has the items in Y. We focus on rules where Y is restricted to a single contact property, whereas X can consist of more than one property (e.g. a contact at home involving skin-to-skin touching). The length of a rule is the total number of items constituting that rule.

Association rules are selected from the set of all possible rules using measures of significance and interestingness. The support of a rule, the primary measure of significance, is the proportion of contacts in the data expressing all items in that rule:

$$\text{supp}(X \Rightarrow Y) = \text{supp}(X \cup Y) = \frac{c_{XY}}{n},$$

where c_{XY} represents the number of contacts which are characterized by all items in X and Y, and n is the total number of contacts in the data. A minimum support threshold value is chosen (often ad hoc) to select the most frequent - and hopefully important - item combinations called 'frequent itemsets'. This can be seen as a simplification of 'bump hunting' (Hastie *et al.*, 2001). From the frequent itemsets, one can further reduce the number of rules by selecting all rules which satisfy a threshold on a certain measure of interestingness, e.g. confidence and lift. The confidence of a rule is defined as:

$$\text{conf}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)},$$

and can be interpreted as an estimate of the conditional probability $P(E_Y|E_X)$, were E_X (E_Y) is the event that X (Y) characterizes a contact. One drawback of using confidence as a selection tool for association rules is that it always increases with the item in the right hand side of the rule (Y) getting more frequent. Although larger confidence values indicate stronger associations, these should not be confused with high correlation, neither with causality between X and Y.

Typically, rules mined using minimum support and confidence are ordered using their lift value:

$$\text{lift}(X \Rightarrow Y) = \frac{\text{conf}(X \Rightarrow Y)}{\text{supp}(Y)} = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)\,\text{supp}(Y)},$$

which is the deviation of the support of the whole rule from the support expected under independence. A lift value of 1 indicates that X and Y appear as frequently together as expected under independence, while a value > 1 indicates that the contact characteristics are (positively) associated. The lift of a rule directly captures correlation between the itemsets and is symmetric. Recently, alternative measures of interestingness have been developed such as hyper-lift and hyper-confidence (Hahsler and Hornik, 2007), though we do not consider them here. The `R`-package 'arules' (Hahsler *et al.*, 2006) is used to mine association rules for the Belgian contact data.

Classification Trees

Classification trees are used to predict membership of cases in the classes of a categorical dependent variable from their measurements on one or more predictor variables. To gain further insight in the factors determining contact intensity, we use the binary classification tree methodology as introduced by Breiman *et al.* (1984). This is a recursive partitioning method which is non-parametric in nature, simple and intuitively appealing. Table 4.1 lists the covariates used in the classification tree construction for the variables of interest: type of contact (close or non-close), location (home, work, school, transport, leisure, another place or multiple locations), frequency (daily, weekly, monthly, a few times a year, first time) and duration (0-5 min, 5-15 min, 15 min - 1 hour, 1-4 hours, > 4 hours).

At each step, the recursive partitioning algorithm determines an optimal cut off point based on some impurity measure, such that all contacts are split into two subpopulations (*binary* classification tree) to achieve high predictive classification with respect to the variable of interest. An example of such an impurity measure is the Gini index, which for node m with n_m observations in region R_m, is defined as:

$$\sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}), \quad \text{where} \quad \hat{p}_{mk} = \frac{1}{n_m} \sum_{x_i \in R_m} I(y_i = k),$$

is the proportion of class k observations in node m. When growing a tree, equal misclassification costs are assigned to the categories $k = 1, 2, \ldots, K$, of the response variable. The resulting subpopulations are split repeatedly until no additional partitioning is warranted, which results in a 'saturated tree': either a subpopulation

Table 4.1: Variables used in the classification trees analysis. 'Tree usage' refers to the use of the variables in the tree construction for tree 1: close contact; 2: contact location; 3: contact frequency and 4: contact duration. '*' indicates that the variable is used as the response variable.

Variable	Value	Code	Tree Usage
Type of contact	1='close';	touching	1*
	2='non-close'		
Location	1='home';	location	1,2*,3,4
	2='work';		
	3='school',		
	4='transport';		
	5='leisure';		
	6='other-multiple'		
Frequency	1='daily';	frequency	1,3*
	2='weekly';		
	3='monthly';		
	4='few times a year';		
	5='first time'		
Duration	1='0-5 min';	duration	1,4*
	2='5-15 min';		
	3='15 min - 1 hour';		
	4='1-4 hours';		
	5='>4 hours'		
Age contact	continuous	agecon	1,2,3,4
Gender contact	1='male';	gencon	1,2,3,4
	2='female'		
Age participant	continuous	agepar	1,2,3,4
Gender participant	1='male';	genpar	1,2,3,4
	2='female'		
Occupation participant	1='working';	occpar	1,2,3,4
	2='retired';		
	3='at home';		
	4='unemployed/job seeking';		
	5='in education';		
	6='other'		
Household size participant	household size	hhsize	1,2,3,4
	(including participant)		
Region	Brss='Brussels'	region	1,2,3,4
	WlsG='Wallonia'		
	VlmG='Flanders'		
Day of the week	0='Sunday';	dayofweek	1,2,3,4
	1='Monday';		
	2='Tuesday';		
	3='Wednesday';		
	4='Thursday';		
	5='Friday';		
	6='Saturday'		
Holiday period	0='no holiday period'	holiday	1,2,3,4
	1='holiday period';		

contains only one category of observed responses or its sample size is too small to divide any further.

To correct for overtraining, the saturated tree is then pruned to an optimal sized subtree, which is most predictive of the outcome and most robust against noise in the data. To select this optimal sized tree, the performance of the subtrees are evaluated by means of a 10-fold cross-validation (CV), using 10 randomly partitioned subsamples as test samples. The average proportion of misclassified observations then defines the CV cost of a tree. Imposing a maximal tree depth of three layers, we use the 1 SE rule described by Breiman *et al.* (1984) to select the three. The 1 SE rule states to chose the smallest-sized tree whose CV cost does not exceed the minimum CV cost plus 1 times the standard error of the minimum CV cost. We use the `R`-package 'mvpart' (De'ath, 2002), a modification of 'rpart', to generate the decision trees depicting the classification rules generated through recursive partitioning.

4.2.2 Application to the Data

Descriptive Analyses

Figure 4.1 presents histograms of the log-transformed number of contacts on weekdays and during the weekend, distinguishing between the regular and the holiday period. The log-transformation somewhat symmetrizes the contact distributions which are highly skewed on the original scale. The median contact counts on weekdays and during the weekend are 14 and 9 for the regular period, and 8 and 7 for the holiday period, respectively. Figure 4.2 shows contact intensity distributions inside and outside households, for type of contact, duration and frequency. The plots clearly indicate that contacts with household members are mostly intimate, which corresponds to darker shading, while contacts outside households display a fairly uniform distribution of contact characteristics. As a final descriptive statistic, the upper left panel of Figure 4.3 displays boxplots of the log-transformed number of (close) contacts inside and outside households. While there is a small variability in the number of household contacts, the contact counts outside households have a more pronounced variability. The three other panels in Figure 4.3 present boxplots of the log-transformed number of contacts for the different categories of location, duration and frequency. There is substantial heterogeneity in the number of contacts recorded at work/school and during leisure activities, with a high number of contacts mostly observed at work/school. Although there are no apparent differences, the lower left panel in Figure 4.3 indicates that there is a larger number of contacts of longer duration. Note that most contacts are frequent contacts whereas fewer yearly or first time contacts are recorded.

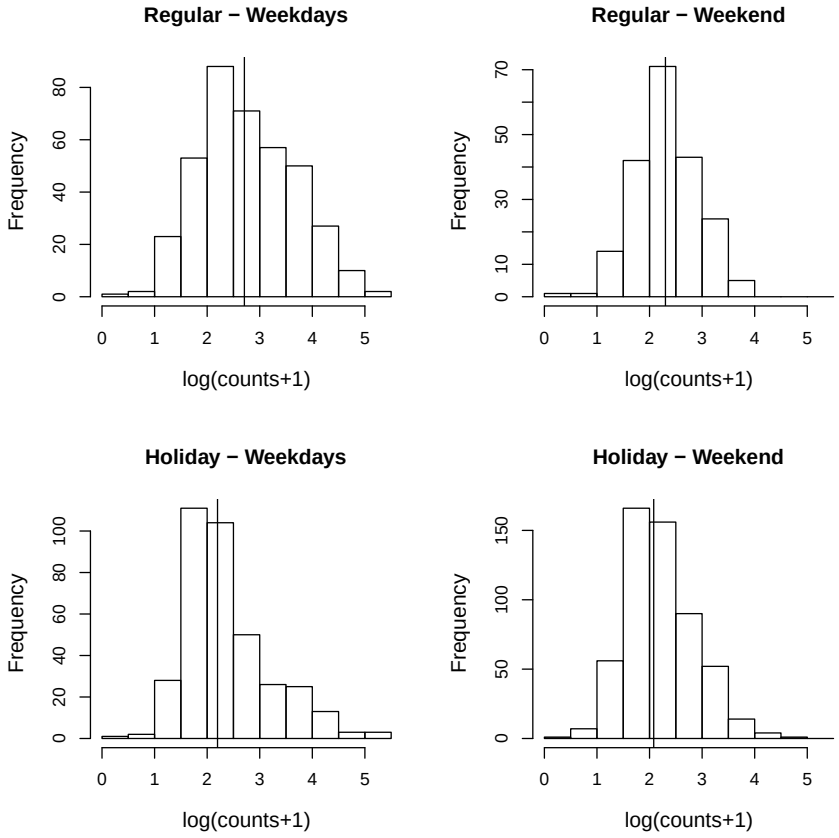

Figure 4.1: Histograms of log(number of contacts + 1) on weekdays (left) and during the weekend (right), distinguishing regular (top row) from holiday (bottom row) period. The vertical lines present the median values.

Association Rules

To discover properties of highly intimate contacts, we mine association rules with the following contact characteristics as right hand sides: close contact (supp=0.61), non-close contact (supp=0.39), 'long' contact i.e. lasting at least 4 hours (supp=0.26), and 'frequent' contact i.e. occurring on a daily basis (supp=0.31). All variables in Table 4.1 are used except for age of the contact and the participant, occupation of the participant and region. Day of the week is dichotomized into weekday and weekend. Association rules are selected using a minimum support threshold value of 0.02 (≈ 343 contacts) and a minimum confidence value of $\text{supp}(Y)$, such that $\text{lift}(X \Rightarrow Y) \geq 1$.

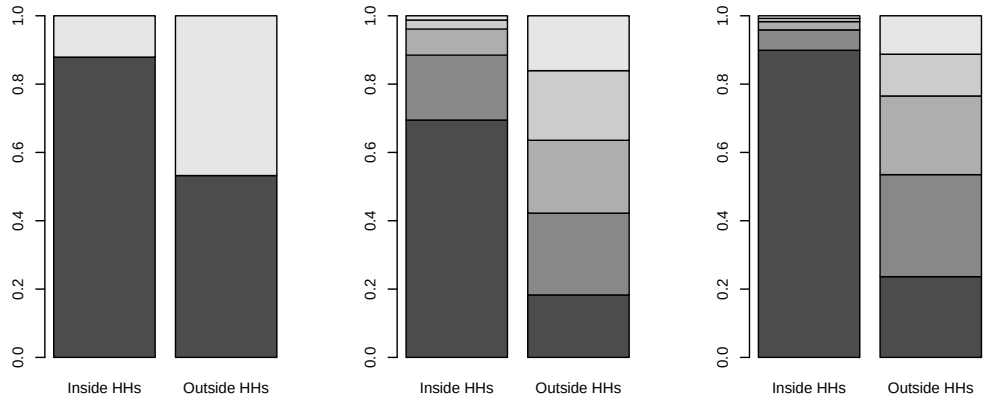

Figure 4.2: Contact intensity distributions in- and outside households: type of contact (left), duration (middle) and frequency (right). Darker colors correspond to close contact, longer duration and more frequent contacts, respectively.

For each contact characteristic of interest (Y), the rules of length 2 are ordered based on their lift value and Table 4.2 presents the three most interesting rules. More than 78% of the long duration, home and daily contacts, involve skin-to-skin touching (lift > 1.2). On the other hand, more than 64% of the first time contacts and short duration contacts (< 15 min) are conversations without physical touching (lift > 1.6). Frequent contacts, and contacts at home/school are in more than 38% of the cases, long contacts (lift > 1.4), and conversely long contacts, and contacts at home/school are in more than 52% of the cases, daily contacts (lift > 1.7). We now sum for each of the contact characteristics, the association rules of length 3 with the largest lift value. An association is found between long contacts at home and contacts involving skin-to-skin touching (conf=0.93, lift=1.53). Brief contacts (0-5 min) taking place during transport (car, bus, etc.) are mostly non-close (conf=0.93, lift=2.40). Similar to what we observe for rules of length 2, frequent contacts at home are associated with long contacts (conf=0.64, lift=2.43) and visa versa (conf=0.84, lift=2.76).

Classification Trees

Figures 4.4, 4.5, 4.6 and 4.7, depict the final classification trees for type of contact, location, frequency and duration, respectively. The length of a branch indicates its relative importance versus other branches. Together with the split variable, frequency plots for the terminal nodes are shown. All trees show an improvement with respect to misclassification compared to the null model, i.e. a tree with only a root node. In general, the misclassification rates are still considerably high due to several heterogeneous terminal nodes. The CV-based misclassification rates are close to or approximately

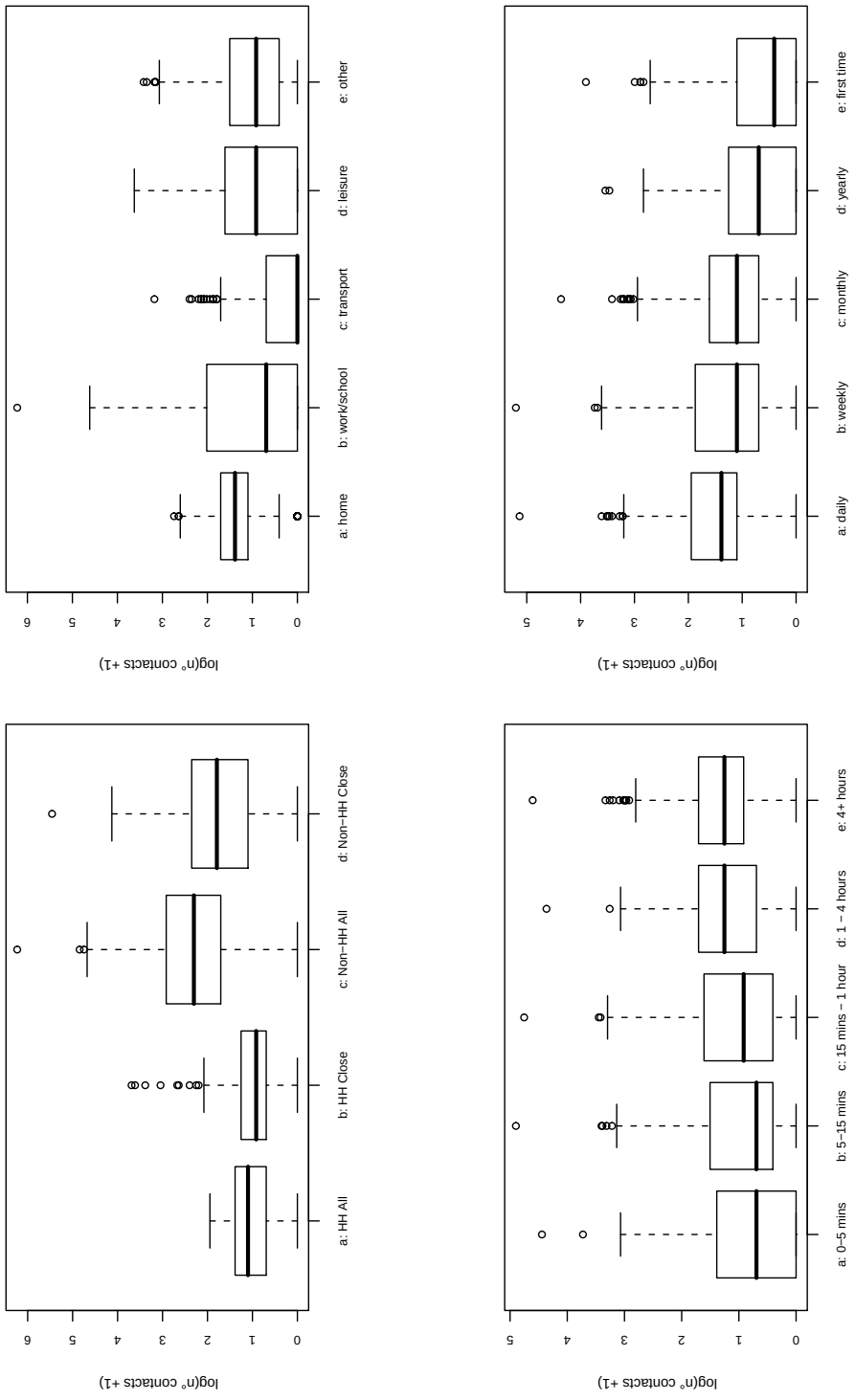

Figure 4.3: Boxplots of the number of (close) contacts in and outside households (left upper panel), the number of contacts per location (right upper panel), the number of contacts for different durations (left lower panel) and frequencies (right lower panel); all at the log-scale.

Table 4.2: Most interesting association rules of length 2 according to the lift value, for the following right hand side (rhs) characteristics: (non-)close contact, long contact and frequent contact; supp($X \Rightarrow Y$), conf($X \Rightarrow Y$) and lift($X \Rightarrow Y$) are provided for each rule.

Y (rhs)	X (lhs)	Supp	Conf	Lift
Close contact	duration: >4 hours	0.23	0.86	1.41
	location: home	0.21	0.81	1.33
	frequency: daily	0.24	0.78	1.28
Non-close contact	frequency: first time	0.08	0.76	1.96
	duration: 0-5 min	0.10	0.74	1.91
	duration: 5-15 min	0.10	0.64	1.65
Long contact (>4 hours)	frequency: daily	0.17	0.55	2.11
	location: home	0.12	0.47	1.78
	location: school	0.03	0.38	1.45
Frequent contact (daily)	duration: >4 hours	0.17	0.65	2.11
	location: home	0.16	0.62	2.02
	location: school	0.05	0.53	1.72

equal to the resubstitution misclassification rates. As a sensitivity analysis, the classification trees are grown for the non-augmented data as well, however, these are very similar (results can be found in Hens *et al.*, 2009a).

Duration, location and frequency of the contact, mainly determine whether or not a contact involves skin-to-skin touching (Figure 4.4). Two thirds of contacts of short duration ($<$ 15 min) are non-close, while longer contacts taking place at home, during transportation or leisure activities usually involve touching (78%). At work and school, there is a fairly even balance between close and non-close contacts (53% and 47%, respectively). As can be seen from Figure 4.5, on weekdays, contacts mainly occur at work (employed adults) or at school (children and teenagers), the latter except for the holiday period when students make more contacts at home and during leisure activities. In the weekend, contacts mainly take place during leisure activities and at home. Interestingly, a significantly smaller proportion of weekend contacts is spent as leisure activities in Wallonia (25%), compared to Flanders and Brussels (39%) (p-value < 0.001).

Figure 4.6 shows that contact frequency is mainly determined by location and the age of the contacted person. More frequent contacts are observed at home and school, and to a lesser extent at work. Contacts during transport, leisure or other activities, tend to be less frequent, especially when the contacted person is an adult. Contact

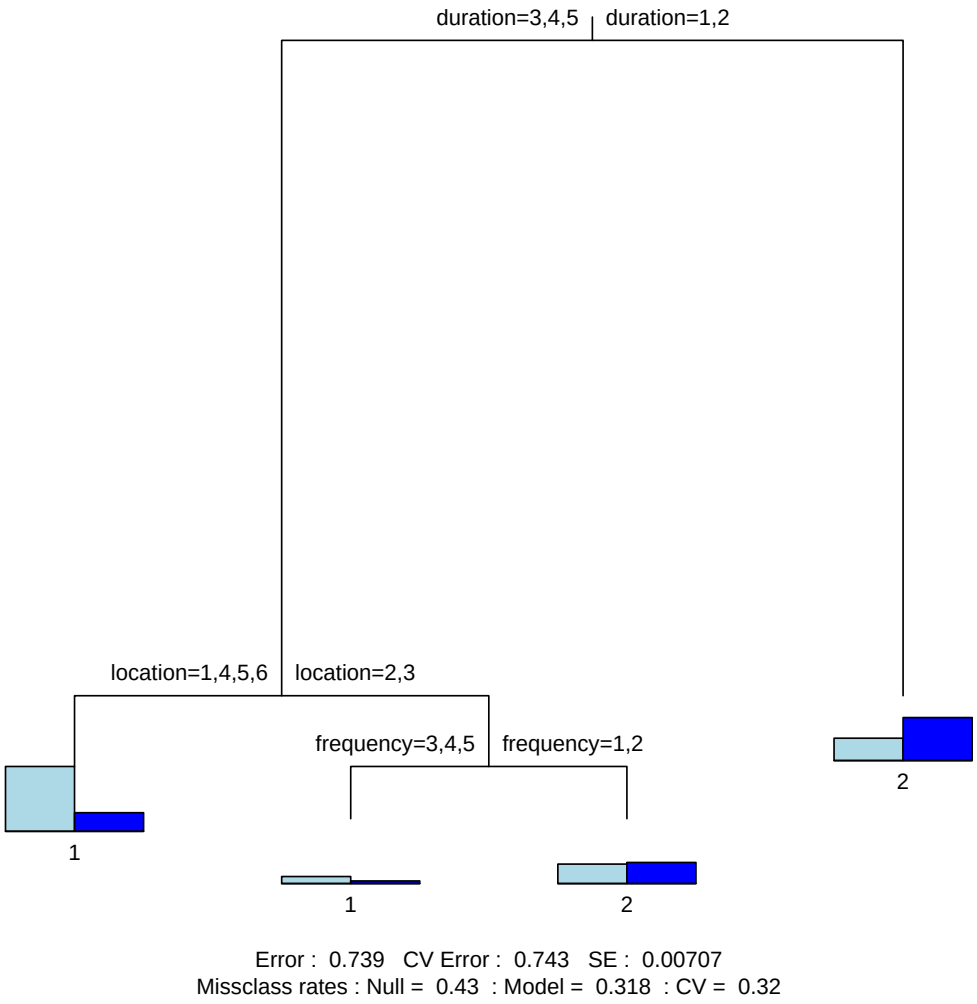

Figure 4.4: Classification tree for type of contact: close or non-close. Variable codes can be found in Table 4.1.

duration is highly dependent on contact location (Figure 4.7). At home and school, contacts are mostly of long duration, whereas contacts taking place during transport are generally more brief. Contacts at work typically constitute a mix of long and short duration contacts.

4.3 Modelling the Number of Contacts

In Section 3.2.3, the censored negative binomial regression analysis from Hens *et al.* (2009b) was presented, which aimed to relate the overall number of contacts to differ-

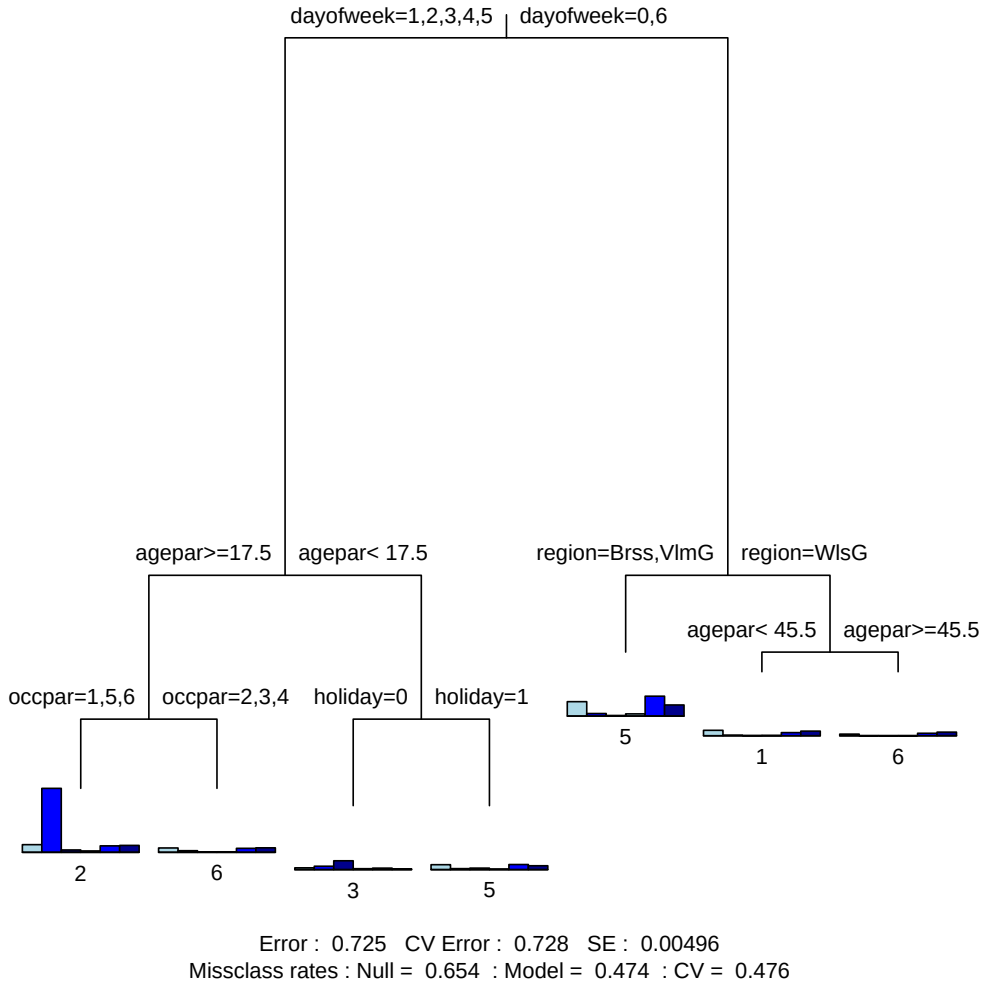

Figure 4.5: Classification tree for contact location. Variable codes can be found in Table 4.1.

ent participant characteristics for all POLYMOD countries. To analyze the contact counts for the complete Belgian data set, we will use generalized estimating equations (GEE) as introduced by Liang and Zeger (1986), since these can account for the correlation between the number of contacts recorded by the same individual on two different days.

4.3.1 Generalized Estimating Equations

GEE is a marginal modelling approach which only requires the specification of univariate marginal distributions and a working correlation matrix for the vector of clustered

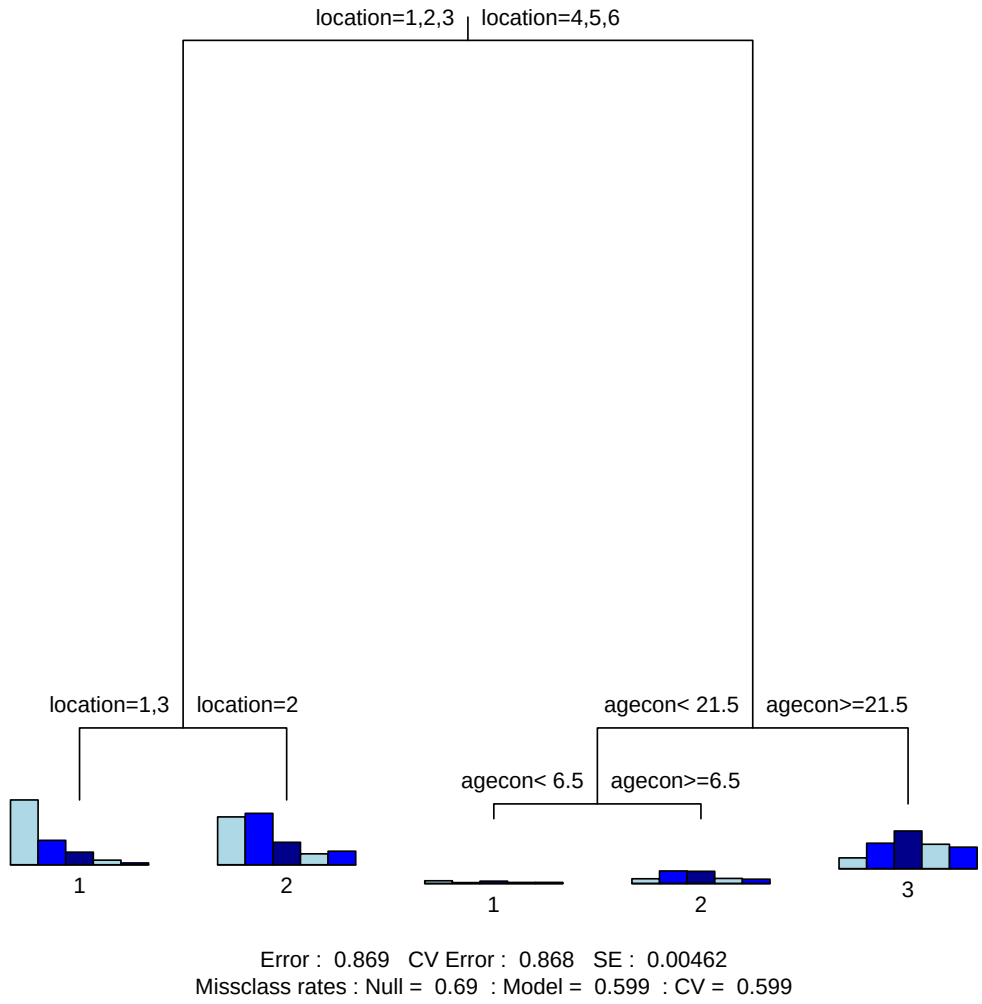

Figure 4.6: Classification tree for contact frequency. Variable codes can be found in Table 4.1.

observations per subject. We assume a Poisson distribution for the contact counts and incorporate an extra parameter to account for overdispersion. Post-stratification weights are included in the analyses, first, to adjust for the relative differences in age and household size representation as compared to Belgian demographic data (cf. Section 3.2.2), and second, to adjust for the differences in sampling proportions with respect to weekdays and holiday periods.

Model building is done a priori, using a non-parametric method called 'random forests' (Breiman, 2001). Random forests are constructed using binary partitioning *regression* trees (Breiman *et al.*, 1984) which are similar to classification trees, though

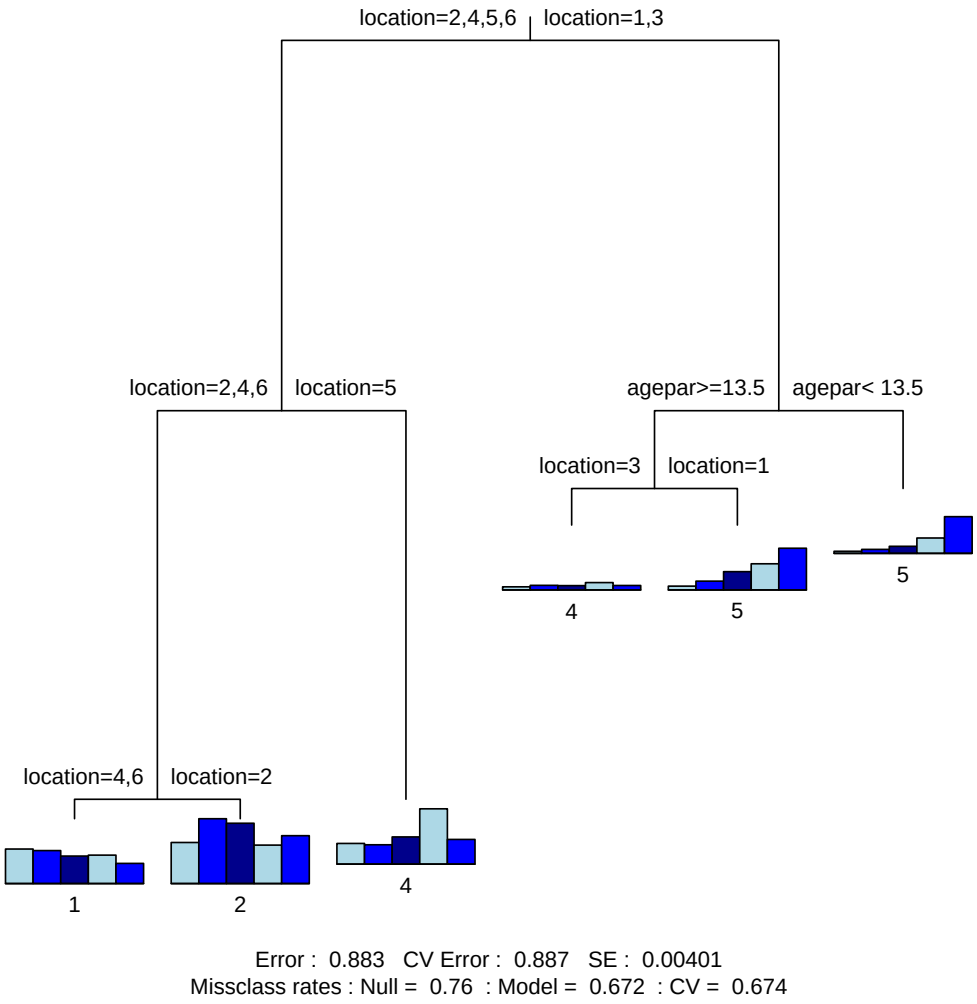

Figure 4.7: Classification tree for contact duration. Variable codes can be found in Table 4.1.

intended to relate a continuous response variable to explanatory variables. At each step, the partitioning algorithm splits the data into two parts such that the sum of the squared deviations from the mean in the separate parts, i.e. the mean squared error (MSE), is minimized. Random forests are constructed by joining several of these regression trees, each based on a random sample of the observations and the explanatory variables, to explicitly take into account the variability associated with the construction of a single tree. A by-product of this random forests methodology, is the so-called 'variable importance' list, which reflects how often a variable is used as a splitting criterion.

Table 4.3: Variables used as initial variables in the WGEE analysis for each of the diary types. 'X' indicates selection of the variable by random forests and inclusion in the WGEE analysis, and '*' indicates that the final models retain at maximum one of the two variables.

	Children ($<$ 9y)	Teenagers (9-17y)	Adults ($\geq$ 18y)
Age participant	X	X	X
Childcare	X		
Class size		X	
Day of the week*	X	X	X
Education participant			X
First or second day	X	X	X
Gender participant	X	X	X
Holiday period	X	X	X
Household size participant	X	X	X
Occupation participant			X
Region	X	X	X
Weekend*	X	X	X

Since the contact counts have a skew distribution, we construct a random forest for the log-transform of the number of contacts (Figure 4.1). The variables with the highest importance in the list, selected based on a threshold of a 5% increase in MSE, are retained for further model building (Table 4.3). One could chose to explore the effects of interactions between all pairs of selected variables. Since this inevitably leads to sparse cells, we only analyze the interaction between day of the week and holiday period, and for the remainder merely consider main effects. We then further reduce the model using a stepwise Poisson regression (with log-link) based on backward selection and the AIC-criterion (2.14), and finally apply the weighted GEE (WGEE) analysis to account for the correlation between two contact counts from the same individual. This procedure is applied separately for the three different types of diaries, considering 19 different contact counts as response variable (determined by type of contact, location, frequency and duration), which leads to a total of 57 analyses.

4.3.2 Application to the Data

Table 4.3 lists the variables used in the WGEE analysis of the contact counts for each of the diary types. The variables have been selected using random forests as described in Section 4.3.1. Most of these variables are explained in Table 4.1, except for childcare (number of children at the daycare center or class size for children attending kindergarten or school: < 10; 10-20; > 20; no childcare), class size (integer value for the number of persons in the classroom), education participant (no formal education; primary; secondary; higher), first or second day (two days of contact recording), and weekend (yes or no). Performing separate analyses for each of the diary types, allows us to incorporate diary-specific variables such as childcare, class size and education of the participant.

The WGEE results for the total number of contacts are presented in Table 4.4. We observe a significant decrease in the number of contacts made during the holiday period for children and teenagers, reflecting school closure. In the weekend, there is a significant decrease in the number of contacts made by teenagers and adults, but not by children. The number of children in childcare (and in the classroom for children attending school) is positively correlated with the number of contacts made, while there is no significant impact of the class size on the total number of contacts recorded by teenagers. Household size is a significant factor for all age groups, whereas gender, although retained in the Poisson regression backward selection, turns out to have no significant impact. Note that the latter two observations are in agreement with the findings from Hens *et al.* (2009b) for the 'all countries analysis' presented in Section 3.2.3. For adults, the number of contacts is highly influenced by their occupation, with employees or students making on average twice as many contacts as unemployed or retired adults, or adults staying at home (e.g. housewives). Overall, a significant level of overdispersion is observed in all models, however, the correlation between the two contact counts is not significantly different from zero.

Disentangling the total number of contacts in terms of type of contact, location, frequency and duration, yields further interesting insights in contact behavior (estimates can be found in Hens *et al.*, 2009a). For all types of diaries, the number of more intimate contacts increases significantly with household size. For children, the number of intimate contacts decreases significantly during the holiday period. Teenagers experience more intimate contacts with increasing class size and when living in Flanders. Work is the key factor for adults, where almost a two-fold increase in the number of intimate contacts for employed adults is observed. There is a marked increase in the number of intimate contacts for adults living in Flanders and Wallonia

Table 4.4: Results of the WGEE analysis using the total number of contacts as response variable, for each of the diary types. For each of the discrete variables with more than two realizations, the baseline (BL) category is shown. For each category, the sample size (Nr.) is given and the estimated rate is presented together with a 95% CI in square brackets. Boldfaced rates indicate a significant difference from the baseline.

Children		Nr.	Rate	95% CI	Teenagers		Nr.	Rate	95% CI	Adults		Nr.	Rate	95% CI
	(int)		5.27			(int)		11.09			(int)		6.37	
										Day of week	Tue	29	1.15	[0.81, 1.63]
										(BL: Mon)	Wed	37	0.95	[0.70, 1.31]
											Thu	73	0.81	[0.60, 1.10]
											Fri	108	0.93	[0.68, 1.27]
					Weekend	yes	19	**0.34**	[0.22, 0.53]		Sat	96	**0.54**	[0.42, 0.69]
											Sun	71	**0.46**	[0.35, 0.59]
Holiday period	yes	235	**0.73**	[0.58, 0.90]	Holiday period	yes	43	**0.38**	[0.25, 0.56]	Holiday period	yes	24	1.07	[0.87, 1.31]
					Holiday*weekend	y/y	106	**3.05**	[1.86, 5.02]					
Region	Fla	168	0.94	[0.69, 1.27]	Region	Fla	144	1.34	[0.91, 1.99]	Region	Fla	882	**1.84**	[1.29, 2.61]
(BL: Bru)	Wal	80	**0.72**	[0.52, 0.99]	(BL: Bru)	Wal	78	0.92	[0.60, 1.42]	(BL: Bru)	Wal	478	**1.70**	[1.21, 2.38]
Household size			**1.18**	[1.10, 1.26]	Household size			**1.25**	[1.11, 1.41]	Household size			**1.13**	[1.07, 1.20]
Childcare	11-20	34	**1.42**	[1.06, 1.90]	Class size			0.99	[0.97, 1.01]	Occupation	retired	172	**0.45**	[0.32, 0.62]
(BL: 1-10)	> 20	168	**1.50**	[1.19, 1.90]						(BL: working)	at home	88	**0.50**	[0.37, 0.68]
	none*	80	1.15	[0.92, 1.43]							unemployed	76	**0.53**	[0.40, 0.72]
											education	84	0.91	[0.64, 1.28]
											other	70	0.76	[0.52, 1.11]
										Age	25-44y	336	0.84	[0.66, 1.08]
										(BL: 18-24y)	45-64y	368	1.07	[0.80, 1.42]
											+65y	96	1.00	[0.63, 1.57]
										Education	primary	36	1.63	[0.78, 3.41]
										(BL: none)	secondary	498	**1.96**	[1.01, 3.81]
											higher	414	1.90	[0.97, 3.73]
Gender	male	140	1.13	[0.92, 1.37]						Gender	male	426	0.96	[0.81, 1.12]
					1st or 2nd day	2nd	125	**0.80**	[0.65, 0.97]	1st or 2nd day	2nd	479	0.98	[0.81, 1.18]
Total		280					260					958		
Overdispersion (s.e.)			**4.04**	(0.64)	Overdispersion (s.e.)			**7.02**	(1.11)	Overdispersion (s.e.)			**15.72**	(1.78)
Correlation (s.e.)			0.04	(0.12)	Correlation (s.e.)			0.18	(0.12)	Correlation (s.e.)			0.06	(0.05)

* Contact counts for children during the weekend are classified into the 'no childcare' category.

as opposed to Brussels. In general, there is a slight, mostly non-significant decrease in the number of contacts reported at day two, irrespective of whether this is a weekend or a weekday. This may indicate decreased compliance with contact recording over time due to survey fatigue, which has been confirmed by another contact survey in Australia involving three study days (McCaw *et al.*, 2010).

4.4 Mimicking the Spread of an Epidemic

We mimic the spread of a newly emerging infectious disease in a large population based on specific contact patterns. Note that, although the next generation matrix methodology provides a general idea of the age-specific dissemination during the initial phase of an epidemic, it does not aim to describe its stochastic nature.

4.4.1 Next Generation Methodology

In Section 2.1.2, it is described how the next generation operator (2.13) is related to the distribution of infectious cases during the initial exponential growth phase of an epidemic. Based on this concept, we illustrate the initial spread of a newly emerging airborne pathogen in Belgium. Considering J age intervals $[a_{[1]}, a_{[2]}), \ldots, [a_{[J]}, a_{[J+1]})$, the $J \times J$ next generation matrix has the following elements $(i, j = 1, \ldots, J)$:

$$\frac{ND}{L} \left\{ \int_{a_{[i]}}^{a_{[i+1]}} \exp\left(- \int_0^a \mu(u) du \right) da \right\} \beta_{ij}. \tag{4.1}$$

The leading right eigenvector of the next generation matrix is then proportional to the age-specific distribution of infected individuals in the early phase of an epidemic. The population size $N = 10\,547\,958$, the life expectancy at birth $L = 80$, and the age-specific mortality rates $\mu(a)$, specific for Belgium, are obtained from official government statistics (source: EUROSTAT). The mean infectious period D is set to five days, similar to the infectious period typically estimated for influenza (see e.g. Yang *et al.*, 2007).

We assume constant proportionality (3.4) for the transmission rates, which in the discrete age class framework translates into $(i, j = 1, \ldots, J)$:

$$\beta_{ij} = q \cdot c_{ij}, \tag{4.2}$$

where c_{ij} denotes the average per capita rate at which an individual of age class j makes contacts with a person of age class i, per unit time. Similar to the contact

matrix estimation approach from Mossong *et al.* (2008b) (see Figure 3.4), the contact rate matrices c_{ij} are estimated from the Belgian contact data using a bivariate smoothing approach (Wood, 2006). Further methodological details on the estimation of the contact rates are deferred to Chapter 5. Here, we specifically focus on contacts involving skin-to-skin touching, since these are the most determinant for the spread of influenza. During the estimation process, 1 year age intervals are considered and post-stratification weights are taken into account. Note that daily contacts are down weighted with a factor 1/5 to reflect the unlikely occurrence of transmission of infections during a second such contact. To approximate the emergence of pandemic influenza, we choose the proportionality factor such that the largest eigenvalue of (4.1) i.e. the basic reproduction number R_0, equals 2 (Mills *et al.*, 2004; Halloran *et al.*, 2008). We investigate the effect of school closure by estimating contact rates for the regular period, the holiday period and the weekend (excluding holidays), and by comparing the resulting relative impact on the leading eigenvector and R_0.

4.4.2 Application to the Data

The estimated close contact rate matrices are displayed on the left side of Figure 4.8. For the regular period (upper left panel), we observe a distinct assortative pattern in contact behavior, especially for children and teenagers, as well as a parent-child component. The assortativeness is less pronounced for adults, who make contact with individuals of a broader age range, particularly at work. These findings correspond to the observations made by Mossong *et al.* (2008b) as described in Section 3.2.2 (Figure 3.5). During holidays (Figure 4.8, second row) and weekends in the regular period (third row), similar though less assortative features arise from the close contact patterns. The upper right panel in Figure 4.8 illustrates the age distribution of infectious cases if a new infection, transmitted through close contacts, would emerge during a regular period. The disease-specific proportionality factor q is chosen such that $R_0 = 2.00$. The highest relative incidence is observed for the adolescent age group 14-20 years, while a second local maximum is observed for adults aged 35-45 years.

To investigate the impact of school closure as a control measure during the initial phase of an epidemic, we calculate the relative incidence and R_0 based on contact patterns for the holiday period and the weekend, respectively, while retaining the proportionality constant q. This results in $R_0 = 1.69$ and $R_0 = 1.33$, which is a relative change of 0.85 and 0.67, respectively. As depicted in Figure 4.8, both temporal conditions of holidays as well as weekends would decrease the relative incidence for

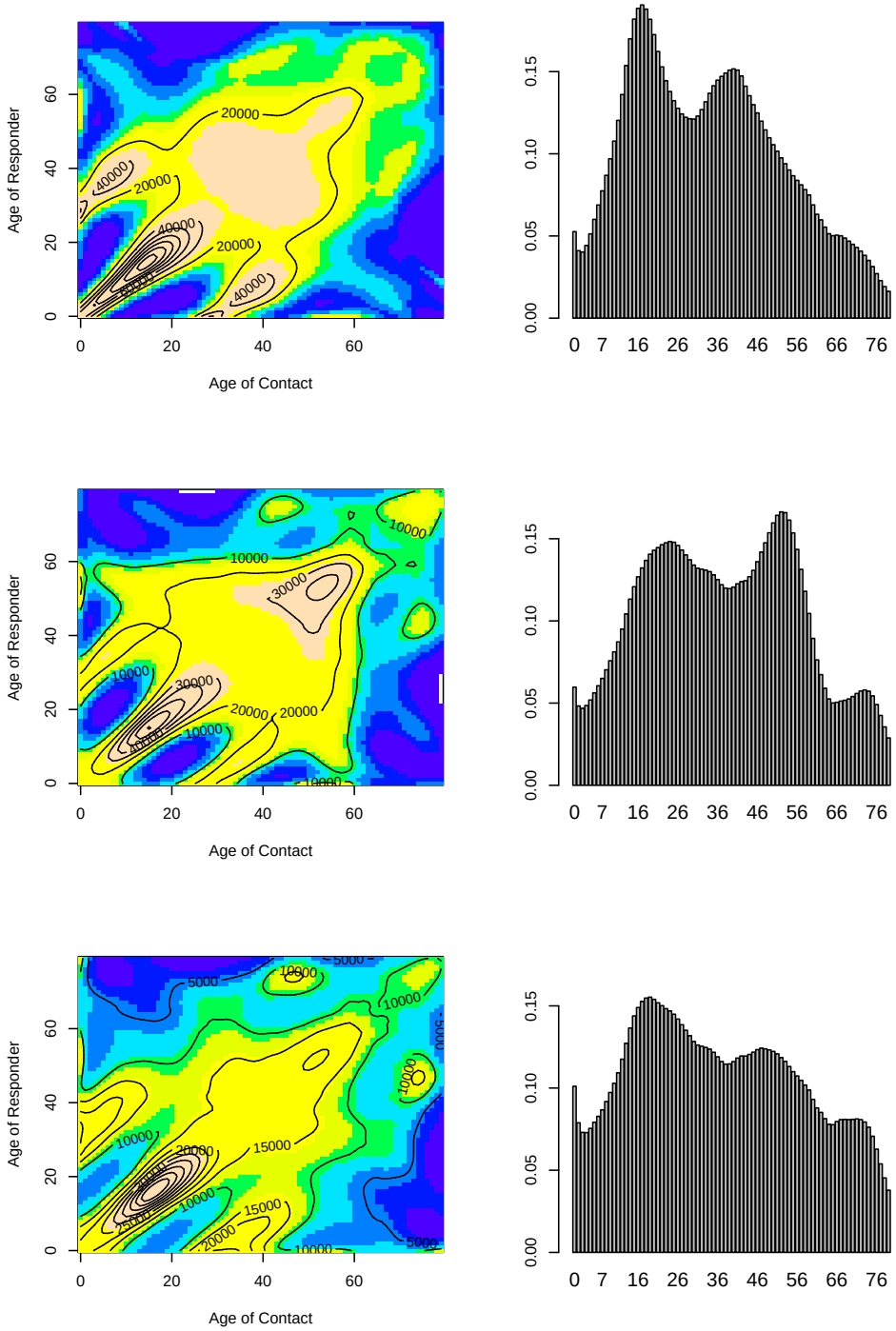

Figure 4.8: Close contact patterns (left) overlaid with contours at the population level, and the corresponding age distribution of infectious individuals (right) for a newly emerging infection in Belgium during the regular period (first row), the holiday period (second row) and the weekend (excluding holiday weekends) (third row).

teenagers considerably. Holidays seem to induce an age shift to the right for both peak incidences (weekends as well to a lesser extent), and elderly seem more involved in the transmission process when schools are closed. For the weekend, we observe a substantial decrease in relative incidence for adults due to the limited number of contacts at work. These findings motivated us to further study the relative influence of temporal conditions on R_0 using contact data from all POLYMOD countries (Hens *et al.*, 2009b). Some results were briefly described in Section 3.2.4, though they cannot be compared to the results here, since the study from Hens *et al.* (2009b) focused on the non-augmented, POLYMOD version of the Belgian contact data (one day) and their weekend-week comparison included both regular days as well as holidays.

4.5 Concluding Remarks

In this chapter, following Hens *et al.* (2009a), we extensively described the Belgian contact survey conducted in 2006. Respondents to the survey recorded their contacts during two randomly assigned days, one weekday and one day in the weekend. Quota sampling enhanced representativeness of the sample with respect to the Belgian population in terms of geographical spread (Flemish, Walloon and Brussels region), age and gender. On the other hand, representativeness of the sample may have been compromised by the sampling method of random digit dialing on fixed telephone lines (Pickery and Carton, 2005). The sampling period included a school holiday period, which was taken into account in the various analyses described in this chapter. In the diary, participants were instructed to estimate their professional contact count and to tick the age range of these contacts. If the estimated number of contacts at work would exceed 20, the participants were asked not to record them in the diary. We augmented the Belgian data set with these professional contacts by imputing the missing variables based on the ticked age categories and the data available for participants with a total number of recorded contacts at work between 10 and 20 (cf. Section 4.1.2). By explicitly imputing professional contacts at the individual level, the average number of contacts for the Belgian contact survey increases from 11.8 as reported by Mossong *et al.* (2008b), to 18.4.

In Sections 4.2, 4.3 and 4.4, several methods were presented to analyze various aspects of the Belgian contact survey. Most results were logical or intuitively expected, but nevertheless important to validate the contact survey and to distinguish between different levels of contact intimacy. Every piece of information available in the survey, and the patterns and associations detected between the variables, should be kept

in mind when using the contact data to inform mathematical models of infectious disease transmission. We found that contacts between household members are mostly intimate, i.e. close contacts taking longer than an hour and occurring on a daily basis. The association rules and classification trees analyses revealed that there are robust associations between general contact intimacy indicators, such as contacts taking place at home, lasting at least 4 hours, occurring on a daily basis, and involving skin-to-skin touching. This information may for instance be useful for contact tracing for tuberculosis, which is transmitted during more intimate contact as compared to the common cold or influenza.

The WGEE analysis showed that the number of reported contacts depends heavily on household size, size of the classroom for children (or the childcare center for infants) and daily occupation for adults. Differences in the average number of contacts are observed for the three regions in Belgium, showing the importance of considering regional differences within a country. The correlation between the two days of recording was generally found to be non-significant, though overdispersion of the contact counts again appeared to be essential. People mostly mix with people of similar age, or with their offspring, parents or grandparents. In the holiday period, the contact frequency for children and teenagers decreases considerably, while a similar observation is made for teenagers and adults during the weekend. In Section 4.4, we used the estimated contact rates to mimic the initial spread of a newly emerging airborne pathogen in Belgium, and observed an effect of temporal conditions on the age-specific relative incidence. The related effect of school closure on an emerging epidemic was further studied by Hens *et al.* (2009b) (cf. Section 3.2.4).

Linking the resulting contact patterns to data on disease prevalence or incidence, informs the estimation of crucial parameters for airborne infections such as the WAIFW matrix and R_0 (Wallinga *et al.*, 2006). Further, insights can be gained about which type of contact is most predictive for the dissemination of a certain infection in a large population. In the next chapter, we will illustrate these aspects for VZV in Belgium using both contact data as well as serological data (Goeyvaerts *et al.*, 2010a; Ogunjimi *et al.*, 2009).

Chapter 5

Estimating Varicella Zoster Virus Transmission from Data on Social Contacts

In Section 3.2, we already referred to the traditional approach in modelling transmission dynamics of infectious diseases, and more particularly in estimating age-dependent transmission rates, that was introduced by Anderson and May (1991). The idea is to, based on prior knowledge of age-related social mixing behavior, impose different mixing patterns on the WAIFW matrix, hereby constraining the number of distinct elements for identifiability reasons, and to estimate the parameters from serological data. Many authors have elaborated on this approach of Anderson and May (1991), however, the choice of the imposed mixing pattern is found to highly influence the outcome of the mathematical model and estimates of related parameters such as R_0 and the critical immunization level (Greenhalgh and Dietz, 1994). Both the traditional Anderson and May (1991) approach as well as the alternative method from Farrington and Whitaker (2005) to parameterize a continuous contact surface, involve a somewhat ad hoc choice. Alternatively, to estimate age-dependent transmission parameters, Wallinga *et al.* (2006) augmented seroprevalence data with auxiliary data on self-reported numbers of conversational contacts per person, whilst assuming that transmission rates are proportional to rates of conversational contact. The social contact surveys conducted as part of the POLYMOD project, which were extensively discussed in Chapters 3 and 4, allow us to elaborate on this methodology

presented by Wallinga *et al.* (2006).

The main parts of this chapter were published in Goeyvaerts *et al.* (2010a). In the first section, we illustrate the traditional approach of imposing mixing patterns to estimate the WAIFW matrix from serological data for VZV in Belgium, as introduced in Section 3.1.1. In Section 5.2, a transition is made to the novel approach of using social contact data to estimate R_0. The POLYMOD contact survey allows us to infer on age-specific mixing patterns in Belgium, which is of particular importance for varicella since social interactions between children determine the main routes of VZV spread. We show that a bivariate smoothing approach allows for a more flexible and better estimate of the contact surface compared to the maximum likelihood estimation method of Wallinga *et al.* (2006). Further, some refinements are proposed, among which an elicitation of contacts with high transmission potential and a non-parametric bootstrap approach, assessing sampling variability and accounting for age uncertainty, as suggested by Halloran (2006).

In general, however, contacts reported in social contact surveys are proxies of those events by which transmission may occur and there may exist age-specific characteristics related to susceptibility and infectiousness which are not captured by the contact rates. Therefore, our main result is the novel method of disentangling the WAIFW matrix into two age-specific components: the contact surface and an age-dependent proportionality factor, which entails an improvement of fit for the seroprevalence of VZV in Belgium. The proposed method, as described in Section 5.3, tackles two dimensions of uncertainty. First, by estimating the contact surface from data on social contacts, we overcome the problem of choosing a completely parametric model for the WAIFW matrix. Second, to deal with the problem of model selection uncertainty for the age-dependent proportionality factor, concepts of multimodel inference are applied and model averaged estimates for R_0 and the critical immunization level are calculated. Some concluding remarks are provided in the last section.

5.1 Estimation of R_0 by Imposing Mixing Patterns

5.1.1 Estimating Transmission Rates

To describe VZV transmission dynamics, we consider a compartmental MSIR model for a closed population of size N assuming demographic and endemic equilibrium, as described in Section 2.1. We thus explicitly take into account the fact that newborns are initially protected by maternal antibodies and do not take part in the transmission process. Type I mortality (2.7) with life expectancy L and type I maternal antibodies

(2.8) with A the age at which newborns lose their maternal immunity, are assumed. It is reasonable to assume type I mortality (see Figure 2.2) and to ignore mortality due to infection, when describing transmission dynamics for VZV in a developed country such as Belgium (see also Whitaker and Farrington, 2004b). Further, we do not explicitly model the potential reactivation of the virus as herpes zoster (Section 3.1.1), and therefore implicitly assume that infected individuals maintain lifelong immunity after recovery from chickenpox. Following Garnett and Grenfell (1992) and Whitaker and Farrington (2004b), we thus ignore varicella cases resulting from contact with persons suffering from shingles. Herpes zoster indeed has a limited impact on VZV transmission dynamics when considering large populations with no immunization program (Ferguson *et al.*, 1996), which is the case for Belgium.

Estimating transmission rates using seroprevalence data can not be done analytically since the integral equation (2.12) in general has no closed form solution. However, it is possible to solve this numerically by turning to a discrete age framework, assuming a constant force of infection in each age-class. Denote the first age interval $(a_{[1]}, a_{[2]})$ and the jth age interval $[a_{[j]}, a_{[j+1]})$, $j = 2, \ldots, J$, where $a_{[1]} = A$ and $a_{[J+1]} = L$. Making use of (2.9) and approximating $r(a)$ by $1 - s(a)$, $\forall a > A$, assuming $i(a)$ is small relative to $s(a)$, the prevalence of immune individuals of age a is given by:

$$r(a) = 1 - \exp\left(-\sum_{k=1}^{j-1} \lambda_k (a_{[k+1]} - a_{[k]}) - \lambda_j (a - a_{[j]})\right), \tag{5.1}$$

if a belongs to the jth age interval. Note that we allow the proportion of immune individuals to vary continuously with age and that we do not summarize the binary seroprevalence outcomes into a proportion per age class. Further, from (2.9) and (2.12) it follows that the force of infection for age class i equals ($i = 1, \ldots, J$):

$$\lambda_i = \frac{ND}{L} \sum_{j=1}^{J} \beta_{ij} \left[\exp\left(-\sum_{k=1}^{j-1} \lambda_k (a_{[k+1]} - a_{[k]})\right) - \exp\left(-\sum_{k=1}^{j} \lambda_k (a_{[k+1]} - a_{[k]})\right)\right], \tag{5.2}$$

where β_{ij} denotes the average per capita rate at which an individual of age class j makes effective contacts with a person of age class i, per year. Recall that the transmission rates β_{ij} make up the $J \times J$ WAIFW matrix.

Once the WAIFW matrix is estimated, the basic reproduction number R_0 can be calculated as the dominant eigenvalue of the $J \times J$ next generation matrix (Diekmann *et al.*, 1990) with elements defined by (4.1), which under type I mortality becomes ($i, j = 1, \ldots, J$): $\frac{ND}{L}\left(a_{[i+1]} - a_{[i]}\right)\beta_{ij}$. Recall that R_0 represents the average number

of secondary cases produced by one typical infected person during his or her entire period of infectiousness, when introduced into an entirely susceptible population (with the exception of newborns who are passively immune through maternal antibodies). In the next section, we illustrate the traditional approach of imposing mixing patterns to estimate the WAIFW matrix from seroprevalence data.

5.1.2 Imposing Mixing Patterns

The traditional approach of Anderson and May (1991) imposes different, somewhat ad hoc, mixing patterns on the WAIFW matrix. Note that, in the previous section, we ended up with a system of J equations with $J \times J$ unknown parameters (5.2) and thus restrictions on these patterns are necessary. Among the proposals in the literature, one distinguishes between several mixing assumptions such as homogeneous mixing ($\beta(a, a') = \beta$), proportional mixing ($\exists\ u : \beta(a, a') = u(a)u(a')$), separable mixing ($\exists\ u, v : \beta(a, a') = u(a)v(a')$) and symmetry ($\beta(a, a') = \beta(a', a)$). Note that the latter two mixing assumptions require additional restrictions to be made. As illustrated by Greenhalgh and Dietz (1994) and Van Effelterre *et al.* (2009), the structure imposed on the WAIFW matrix has a high impact on the estimate of R_0. In this section, we assume the transmission rates to be constant within six discrete age classes ($J = 6$). We follow Anderson and May (1991); Van Effelterre *et al.* (2009); Ogunjimi *et al.* (2009) and consider the following mixing patterns, based on prior knowledge of social mixing behavior, to model the WAIFW matrix for VZV:

$$W_1 = \begin{pmatrix} \beta_1 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 \\ \beta_6 & \beta_2 & \beta_6 & \beta_6 & \beta_6 & \beta_6 \\ \beta_6 & \beta_6 & \beta_3 & \beta_6 & \beta_6 & \beta_6 \\ \beta_6 & \beta_6 & \beta_6 & \beta_4 & \beta_6 & \beta_6 \\ \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_5 & \beta_6 \\ \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 \end{pmatrix}, \quad W_2 = \begin{pmatrix} \beta_1 & \beta_1 & \beta_3 & \beta_4 & \beta_5 & \beta_6 \\ \beta_1 & \beta_2 & \beta_3 & \beta_4 & \beta_5 & \beta_6 \\ \beta_3 & \beta_3 & \beta_3 & \beta_4 & \beta_5 & \beta_6 \\ \beta_4 & \beta_4 & \beta_4 & \beta_4 & \beta_5 & \beta_6 \\ \beta_5 & \beta_5 & \beta_5 & \beta_5 & \beta_5 & \beta_6 \\ \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 \end{pmatrix}$$

$$W_3 = \begin{pmatrix} \beta_1 & \beta_1 & \beta_1 & \beta_4 & \beta_5 & \beta_6 \\ \beta_1 & \beta_2 & \beta_3 & \beta_4 & \beta_5 & \beta_6 \\ \beta_1 & \beta_3 & \beta_3 & \beta_4 & \beta_5 & \beta_6 \\ \beta_4 & \beta_4 & \beta_4 & \beta_4 & \beta_5 & \beta_6 \\ \beta_5 & \beta_5 & \beta_5 & \beta_5 & \beta_5 & \beta_6 \\ \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 \end{pmatrix}, \quad W_4 = \begin{pmatrix} \beta_1 & \beta_1 & \beta_1 & \beta_1 & \beta_1 & \beta_1 \\ \beta_2 & \beta_2 & \beta_2 & \beta_2 & \beta_2 & \beta_2 \\ \beta_3 & \beta_3 & \beta_3 & \beta_3 & \beta_3 & \beta_3 \\ \beta_4 & \beta_4 & \beta_4 & \beta_4 & \beta_4 & \beta_4 \\ \beta_5 & \beta_5 & \beta_5 & \beta_5 & \beta_5 & \beta_5 \\ \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 \end{pmatrix} \tag{5.3}$$

$$W_5 = \begin{pmatrix} \beta_1 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 \\ \beta_6 & \beta_2 & \beta_6 & \beta_6 & \beta_6 & \beta_6 \\ \beta_6 & \beta_6 & \beta_3 & \beta_6 & \beta_6 & \beta_6 \\ \beta_6 & \beta_6 & \beta_6 & \beta_4 & \beta_6 & \beta_6 \\ \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_5 & \beta_6 \\ \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_6 & \beta_5 \end{pmatrix}, \quad W_6 = \begin{pmatrix} \beta_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \beta_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \beta_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & \beta_4 & 0 & 0 \\ 0 & 0 & 0 & 0 & \beta_5 & 0 \\ 0 & 0 & 0 & 0 & 0 & \beta_6 \end{pmatrix}$$

In order to estimate the transmission parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_6)^T$ from seroprevalence data, we follow an iterative procedure from Farrington *et al.* (2001) and Kanaan and Farrington (2005). First, one assumes plausible starting values for $\boldsymbol{\beta}$ and solves (5.2) iteratively for the piecewise constant force of infection $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_6)^T$, which in its turn can be contrasted to the serology. Second, this procedure is repeated under the constraint $\boldsymbol{\beta} \geq \mathbf{0}$, until the Bernoulli loglikelihood

$$\sum_{i=1}^{n} \big\{ y_i \log[r(a_i)] + (1 - y_i) \log[1 - r(a_i)] \big\},$$

has been maximized. Here, n denotes the size of the serological data set, y_i denotes a binary variable indicating whether subject i had experienced infection before age a_i and the prevalence $r(a_i)$ is obtained from (5.1).

5.1.3 Application to the Data

For the remainder, the following parameters specific for Belgium anno 2003 (Eurostat, 2007; FOD Economie Afdeling Statistiek, 2006), are kept constant when estimating the WAIFW matrix and R_0: size of the population aged 0 to 80 years, $N = 9\,943\,749$, and life expectancy at birth, $L = 80$. The mean duration of infectiousness for VZV is taken $D = 7/365$, and the age of losing maternal immunity is chosen $A = 0.5$ (Halloran *et al.*, 1994). By removing infants younger than 6 months, the size of the serological data set becomes $n = 2649$ (Section 3.1.1).

In this application, the population is divided into six age classes taking into account the schooling system in Belgium, following Van Effelterre *et al.* (2009): $(0.5, 2)$, $[2, 6)$, $[6, 12)$, $[12, 19)$, $[19, 31)$, $[31, 80)$. The last age class has a wide range because the serological data set only contains information for individuals up till 40 years. The following ML-estimate for $\boldsymbol{\lambda}$ is obtained assuming a piecewise constant force of infection and using constrained optimization to ensure monotonicity ($r'(a) \geq 0$): $\hat{\boldsymbol{\lambda}}^{\mathrm{ML}} = (0.313, 0.304, 0.246, 0.000, 0.082, 0.000)^T$. A graphical display of the fit is presented in Figure 5.1 and a dashed line is used to indicate the estimated prevalence and force of infection for the age interval $[40, 80)$, for which serological information is lacking.

During the estimation process, non-identifiability problems occur for mixing patterns W_1, W_5 and W_6, which is related to the fact that $\hat{\lambda}_4^{\mathrm{ML}} = \hat{\lambda}_6^{\mathrm{ML}} = 0$. Therefore, these mixing patterns are left from further consideration. For the remaining three, ML-estimates for $\boldsymbol{\beta}$ and R_0 are presented in Table 5.1. Note that mixing pattern W_4

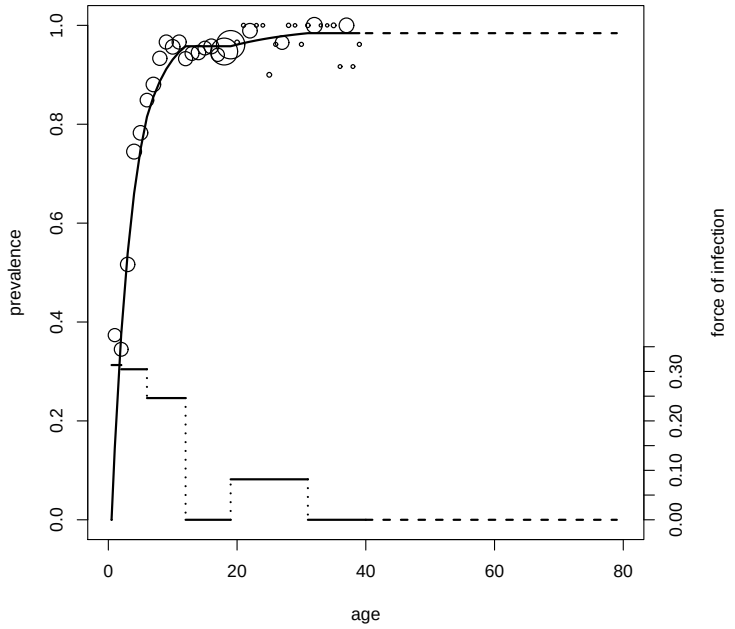

Figure 5.1: Estimated prevalence (upper curve) and force of infection (lower curve) for VZV assuming a piecewise constant force of infection. The dots represent the observed serological data with size proportional to the corresponding sample size. The dashed lines are used to indicate the estimated prevalence and force of infection for the age interval $[40, 80)$, for which serological information is lacking.

Table 5.1: Estimates for the transmission parameters (multiplied by 10^4) and for R_0, obtained by imposing mixing patterns W_2, W_3 and W_4 on the WAIFW matrix.

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\widehat{R}_0$	95% CI for R_0	AIC
W_2	1.413	1.335	1.064	0.000	0.343	0.000	3.51	[3.07, 13.42]	1372.819
W_3	1.362	1.441	0.873	0.000	0.343	0.000	3.37	[2.81, 13.38]	1372.819
W_4	1.334	1.298	1.049	0.000	0.349	0.000	4.21	[3.69, 13.13]	1372.756

has a regular configuration for the data, whereas W_2 and W_3 are non-regular since unconstrained ML-estimation induces negative estimates for β_4 (Farrington *et al.*, 2001). Both W_2 and W_3 entail the following ML-estimate for the force of infection: $\hat{\boldsymbol{\lambda}}^{\text{ML}} = (0.313, 0.305, 0.245, 0.002, 0.080, 0.000)^T$. The estimate of R_0 ranges from 3.37 to 4.21. 95% bootstrap-based percentile CIs (2.17) for R_0 are presented as well, applying a non-parametric bootstrap by taking $B = 1000$ samples with replacement from the serological data. The fit of the three mixing patterns can be compared using model selection criteria, such as AIC (2.14) and BIC (2.15). As can be seen from Table 5.1, the AIC-values (equivalent to BIC here) are virtually equal and do not provide any basis to guide the choice of a mixing pattern.

Note that these results differ somewhat from those obtained by Van Effelterre *et al.* (2009), where a different data set for VZV serology was used, collected from a large laboratory in the city of Antwerp between October 1999 and April 2000 (Thiry *et al.*, 2002).

5.2 Estimation of R_0 using Data on Social Contacts

5.2.1 Constant Proportionality of the Transmission Rates

In the previous section, we have illustrated some caveats involved in the traditional approach of imposing mixing patterns on the WAIFW matrix. In general, the choice of the structures as well as the choice of the age classes are somewhat ad hoc. Further, different regular mixing patterns may induce exactly the same fit to the serological data, though lead to considerably different estimates of β_{ij} and R_0. Since evidence for mixing patterns is thought to be found in social contact data, i.e. governing contacts with high transmission potential, an alternative approach to estimate transmission parameters has emerged: augmenting seroprevalence data with data on social contacts. Wallinga *et al.* (2006) assumed that the transmission rates β_{ij} are directly proportional to the contact rates c_{ij}, i.e. the constant proportionality (CP) assumption as

formulated in (4.2).

The proportionality factor q and the contact rates are not identifiable from serological data only. Therefore, in order to estimate the WAIFW matrix, one first needs to estimate the contact rates c_{ij} using social contact data. We illustrate this in the next section for the Belgian contact data, hereby following the sampling scheme of the POLYMOD project by only considering one assigned day for each participant (Section 4.1.2). Hence the analyzed data comprise 749 participants who recorded a total of 12 775 contacts, of which 3 contacts are omitted due to missing age values for the contacted person. Following the POLYMOD contact survey design, 'making contact with' is defined as a two-way conversation of at least three words in each others proximity and/or any sort of physical skin-to-skin touching (Section 3.2.2). In Section 5.2.5, we will refine on this definition and consider specific types of contact with high transmission potential. In a second step, keeping the estimated contact rates fixed, we estimate the proportionality factor from serological data using the estimation method described in Section 5.1.2.

5.2.2 Estimation Methods for the Contact Rates

Consider the random variable Y_{ij}, i.e. the number of contacts in age class j during one day as reported by a respondent in age class i ($i, j = 1, \ldots, J$), which has observed values $y_{ij,t}$, $t = 1, \ldots, T_i$, where T_i denotes the number of participants in the contact survey belonging to age class i. Now define $m_{ij} = E(Y_{ij})$, i.e. the mean number of contacts in age class j during one day as reported by a respondent in age class i. The elements m_{ij} make up a $J \times J$ matrix, which is called the 'social contact matrix'. Now, the contact rates c_{ij} are related to the social contact matrix as follows:

$$c_{ij} = 365 \cdot \frac{m_{ji}}{n_i},$$

where n_i denotes the population size in age class i, obtained from demographical data. When estimating the social contact matrix, the reciprocal nature of contacts needs to be taken into account (Wallinga *et al.*, 2006):

$$m_{ij} n_i = m_{ji} n_j, \tag{5.4}$$

which means that, on a population level, the total number of contacts from age class i to age class j must equal the total number of contacts from age class j to age class i.

Maximum Likelihood Approach of Wallinga *et al.* (2006)

Wallinga *et al.* (2006) took the following approach to estimate a contact matrix from social contact survey data. To allow for overdispersion, the Y_{ij} are assumed independently negative binomial distributed with mean m_{ij}, dispersion parameter k_{ij} and variance $m_{ij} + m_{ij}^2/k_{ij}$, inducing the following likelihood:

$$\prod_{i=1}^{J}\prod_{j=1}^{J}\prod_{t=1}^{T_i}\left[\text{NegBin}(y_{ij,t}; m_{ij}, k_{ij})\right]^{w_{it}},$$

where w_{it} is the post-stratification weight for the t^{th} participant in age class i, as described in Section 3.2.2. Therefore, an individual contribution to the loglikelihood function for Y_{ij} equals

$$w_{it} \log\left[\text{NegBin}(y_{ij,t}; \frac{c_{ji}n_j}{365}, k_{ij})\right], \tag{5.5}$$

where c_{ji} are the per capita contact rates which need to be estimated. Note that, as long as c_{ij} is modeled symmetrically: $c_{ij} = c_{ji}$, the reciprocal nature of contacts (5.4) is taken into account. In this case, the most saturated model for the contact rates has $J(J+1)/2$ parameters for the mean and J^2 dispersion parameters k_{ij}, which can be estimated by maximizing the summed loglikelihood contributions (5.5).

Bivariate Smoothing

We propose to estimate the elements m_{ij} of the social contact matrix using a bivariate smoothing approach as described by Wood (2006). In contrast to the maximum likelihood approach of Wallinga *et al.* (2006), the average number of contacts is modeled as a two-dimensional continuous function over the age of the respondent and the age of the contacted person, giving rise to a 'contact surface'. The basis is a tensor-product spline derived from two smooth functions of the respondent's and contact's age, ensuring flexibility:

$$Y_{ij} \sim \text{NegBin}(m_{ij}, k), \text{ where } g(m_{ij}) = \sum_{\ell=1}^{K}\sum_{p=1}^{K}\delta_{\ell p} b_\ell(a_{[i]}) d_p(a_{[j]}), \tag{5.6}$$

where g is some link function, $\delta_{\ell p}$ are unknown parameters, and b_ℓ and d_p are known basis functions for the marginal smoothers.

The basis dimension, K, should be chosen large enough in order to fit the data well, but small enough to maintain reasonable computational efficiency (Wood, 2006). For tensor-product smoothers, the upper limit of the degrees of freedom is given by

Table 5.2: Contact rate estimates for models W_1-W_3, W_5 and W_6, multiplied by 10^3, and corresponding AIC-values, obtained with maximum likelihood estimation.

Model	$\hat{c}_1$	$\hat{c}_2$	$\hat{c}_3$	$\hat{c}_4$	$\hat{c}_5$	$\hat{c}_6$	AIC
W_1	1.563	1.578	1.738	3.689	1.663	0.496	13788.48
W_2	0.747	1.578	1.004	2.679	0.938	0.519	13974.19
W_3	0.648	1.578	1.098	2.679	0.938	0.519	13971.12
W_5	1.563	1.578	1.738	3.689	0.769	0.416	13817.40
W_6	1.563	1.578	1.738	3.689	1.663	0.639	52652.23

the product of the K values provided for each marginal smooth, minus one, for the identifiability constraint. However, the actual effective degrees of freedom are also controlled by the degree of penalization selected during fitting. Thin plate regression splines are used to avoid the selection of knots and a log link is used in model (5.6). The post-stratification weights w_{it} are also taken into account in the smoothing process. By applying a smooth-then-constrain-approach as proposed by Mammen *et al.* (2001), the reciprocal nature of contacts (5.4) is allowed for.

5.2.3 Contact Rate Estimates for Belgium

Maximum Likelihood Approach of Wallinga *et al.* (2006)

The mixing patterns (5.3), previously used to estimate the WAIFW matrix, are now applied to model the contact rates c_{ij} in order to study the predictiveness of the social contact data for the transmission of VZV. Structure W_4 is not considered here, because it does not allow for the reciprocal nature of contacts. We focus on the same six age classes used in Section 5.1.3. By maximizing the negative binomial loglikelihood function, making use of (5.5), the estimates in Table 5.2 are obtained. The contact parameters are denoted by c_ℓ ($\ell = 1, \ldots, 6$) while the corresponding dispersion parameters (36 in total) were omitted from the table. Under the CP assumption (4.2), we expect the WAIFW matrix estimates obtained in Section 5.1.3 to be proportional to the contact rate estimates obtained here. However, comparing Tables 5.1 and 5.2 for W_2 and W_3, the β_ℓ estimates do not seem to be proportional to the c_ℓ estimates.

Further, a 'saturated model' as proposed by Wallinga *et al.* (2006), with 21 contact parameters and 36 dispersion parameters, is fitted to the Belgian contact data. The estimated contact rate matrix c_{ij} and corresponding dispersion parameters, obtained with maximum likelihood estimation, are displayed in Table 5.3 and the former is

Table 5.3: Contact rate estimates for the saturated model, multiplied by 10^3, and corresponding dispersion parameter estimates between brackets, obtained with maximum likelihood estimation ((*) indicates no overdispersion).

Age Class	$[0.5, 2)$	$[2, 6)$	$[6, 12)$	$[12, 19)$	$[19, 31)$	$[31, 101)$
$[0.5, 2)$	1.563 (0.33)	0.685 (0.53)	0.373 (0.11)	0.262 (0.42)	0.314 (1.49)	0.244 (3.04)
$[2, 6)$	0.685 (7.19)	1.578 (0.51)	0.606 (0.55)	0.121 (0.07)	0.274 (3.72)	0.266 (5.89)
$[6, 12)$	0.373 (0.27)	0.606 (1.74)	1.738 (0.90)	0.469 (0.40)	0.194 (1.45)	0.350 (3.38)
$[12, 19)$	0.262 (0.05)	0.121 (0.20)	0.469 (0.21)	3.689 (0.81)	0.679 (0.56)	0.349 (1.66)
$[19, 31)$	0.314 (*)	0.274 (0.38)	0.194 (0.18)	0.679 (0.14)	1.663 (1.39)	0.619 (0.89)
$[31, 101)$	0.244 (0.09)	0.266 (0.33)	0.350 (0.09)	0.349 (0.28)	0.619 (0.61)	0.639 (1.08)

depicted in Figure 5.2, on the left side. Contact rate estimates range from $0.121 \cdot 10^{-3}$ between age classes $[2, 6)$ and $[12, 19)$, to $3.689 \cdot 10^{-3}$ between individuals from age class $[12, 19)$. Note that the contact rate estimates on the diagonal are identical to the ones obtained for model W_6 in Table 5.2, as expected. The AIC-value for this saturated model is 13618.73, which is considerably smaller than the AIC-values for the mixing patterns considered above (Table 5.2).

Bivariate Smoothing

The smoothing is performed in `R` with the `gam` function from the `mgcv 1.3-30` package (Wood, 2006), considering one year age intervals, $[0, 1), [1, 2), \ldots, [100, 101)$. An informal check (by comparing the estimated degrees of freedom and the basis dimension) shows that $K = 11$ is a satisfactory basis dimension choice for the Belgian contact data. On the right hand side of Figure 5.2, the estimated contact surface obtained with the bivariate smoothing approach, is displayed. The smoothing approach seems better able to capture important features of human contacting behavior. Three components clearly arise in the smoothed contact surface. First of all, one can see a pronounced assortative structure on the diagonal, representing high contact rates between individuals of the same age. Second, an off-diagonal parent-child component comes forward, reflecting a very natural form of contact between parents and children, which might be important in modelling certain childhood infections such as parvovirus B19 (Mossong *et al.*, 2008a). Finally, there seems to be evidence for a grandparent-grandchild component. Except for the assortativeness, these features are not reflected by the contact rates c_{ij}, estimated from Wallinga *et al.* (2006)'s saturated model.

Further, we would like to compare the two estimation methods more formally us-

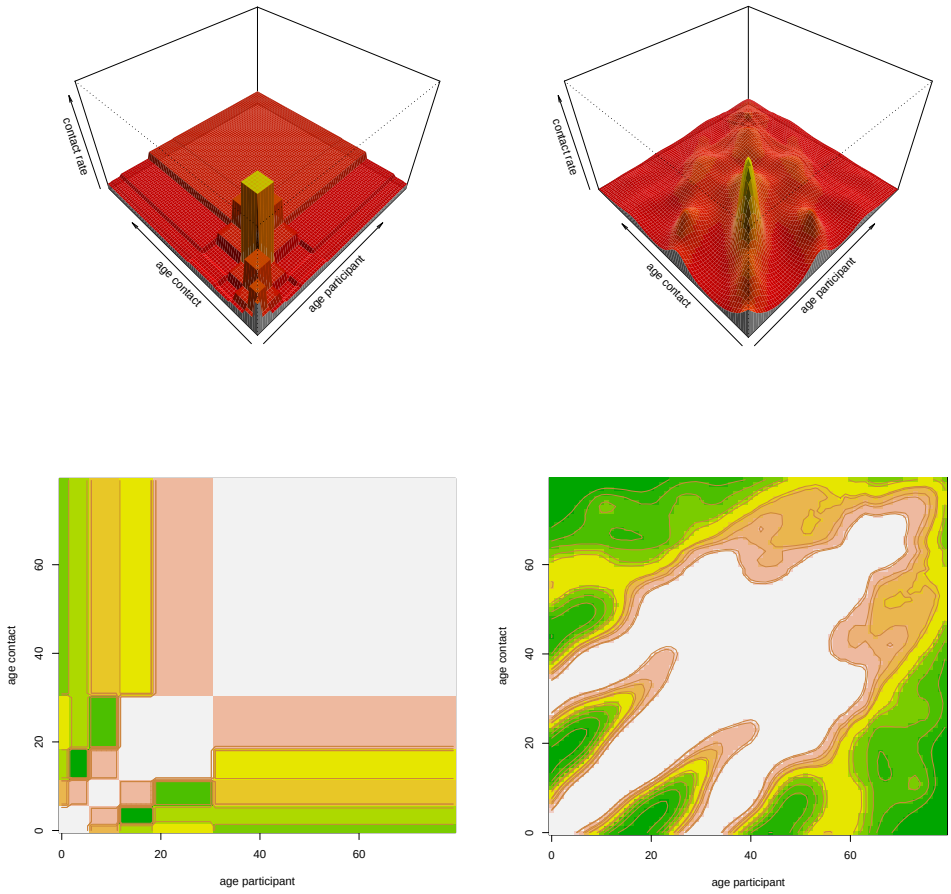

Figure 5.2: Perspective (above) and image (below) plot of the estimated contact rates c_{ij} obtained with maximum likelihood estimation for Wallinga *et al.* (2006)'s saturated model (left) and bivariate smoothing (right). The X- and Y-axis represent the age of the respondent and the age of the contact, respectively.

ing information criteria. The models are however not comparable, since they are fitted to different types of responses. Therefore, we refit the model proposed by Wallinga *et al.* (2006), considering one year age intervals and using the `gam` function for comparability. The AIC-value for the saturated model then equals 63009.60 while the AIC-value for the saturated model with merely one dispersion parameter is smaller, namely 62902.84, showing it is unnecessary to consider different overdispersion parameters. Finally, the AIC-value for the smoothing approach equals 58187.06, indicating improved estimation of the contact surface using non-parametric techniques. Note that the BIC-criterion also indicates the smoothing method to outperform both versions of Wallinga *et al.* (2006)'s saturated model (59130.66 for the smoothing method as opposed to 63244.49 and 63674.43 for both saturated models with constant and heterogeneous overdispersion, respectively).

5.2.4 Estimating Transmission Rates and R_0 for VZV

Under the CP assumption (4.2), we are now able to estimate the WAIFW matrix for VZV using serological data. Keeping the estimated contact rates $\hat{c}_{ij}$ fixed, we estimate the proportionality factor q using the estimation method described in Section 5.1.2. In Table 5.4, estimates for q and R_0 together with their corresponding 95% profile likelihood CIs (2.16), and AIC-values, are presented for each one of the contact rate models considered in Section 5.2.3. Note that the 95% CIs are implausibly narrow, resulting from the fact that the estimated contact rates are held constant. The estimates for R_0 range from 8.8 to 17.3. AIC-values vary widely between 1377 and 1585. The largest AIC-values are obtained for the contact rates estimated by maximizing the likelihood for models W_3 and W_2, which have the least assortative structure. This contrasts with the results in Section 5.1.3, where these mixing patterns actually performed well in describing VZV transmission, but is consistent with the discrepancy from the CP assumption we noted in Section 5.2.3. We elaborate on this issue in Section 5.3.

The smallest AIC-value is obtained for the saturated model as proposed by Wallinga *et al.* (2006). This is a rather odd result, since the bivariate smoothing approach allowed for a more flexible and better estimate of the contact surface (Section 5.2.3). If transmission rates are indeed proportional to rates of making conversational contact, one would expect the latter model to perform better. However, by comparing both model fits in Figure 5.3 this counterintuitive result can be clarified. Most infections with VZV occur early in life, leading to an initial, steep increase in the fraction of seropositives, which then plateaus after the age of ten. Therefore, contact

Table 5.4: ML-estimates for the proportionality factor and R_0, obtained from contact rates estimated by models W_1-W_3, W_5-W_6, Wallinga *et al.* (2006)'s saturated model, and bivariate smoothing, assuming CP.

Model for c_{ij}	$\hat{q}$	95% CI for q	$\widehat{R}_0$	95% CI for R_0	AIC
W_1	0.120	[0.113, 0.128]	13.21	[12.42, 14.04]	1419.858
W_2	0.065	[0.061, 0.070]	8.83	[8.23, 9.48]	1570.947
W_3	0.066	[0.061, 0.071]	8.93	[8.28, 9.64]	1585.177
W_5	0.130	[0.122, 0.138]	14.93	[14.06, 15.90]	1409.554
W_6	0.232	[0.221, 0.243]	17.30	[16.52, 18.16]	1411.112
Saturated	0.124	[0.117, 0.132]	14.08	[13.26, 14.94]	1377.146
Smoothing	0.132	[0.124, 0.140]	15.69	[14.74, 16.69]	1386.618

rate estimates between children will mainly determine the fit to the serological data, limiting the advantage of a better contact surface estimate. For infectious diseases which are less prevalent in the population such as PVB19, the smoothing approach is expected to yield a better fit to the seroprevalence data, providing more realistic estimates for the WAIFW matrix.

All models presented in Table 5.4, except for W_2 and W_3, induce rather large estimates of the basic reproduction number for VZV, compared to the R_0 range of 3-12 reported in the literature (e.g. Whitaker and Farrington, 2004b; Nardone *et al.*, 2007). This could be due to the fact that the transmission rates are assumed proportional to the total number of contacts recorded during one day, including e.g. short duration contacts at work. In the next section, we therefore investigate whether specific contact characteristics are more predictive of VZV spread.

5.2.5 Refinements to the Social Contact Data Approach

The aim is to clearly disentangle the WAIFW matrix into the contact process and the transmission potential. Therefore, in the following, contact rates are estimated using a bivariate smoothing approach, since this method outperforms the saturated model estimated using maximum likelihood as proposed by Wallinga *et al.* (2006) (Section 5.2.3). Following Ogunjimi *et al.* (2009) and Melegaro *et al.* (2010), contacts with high transmission potential are filtered from the social contact data. Further, we briefly describe the main results from a comparative analysis conducted by Ogunjimi *et al.* (2009) of two ML estimation methods for the transmission rates when making use of social contact data. Finally, to improve statistical inference, we present a non-

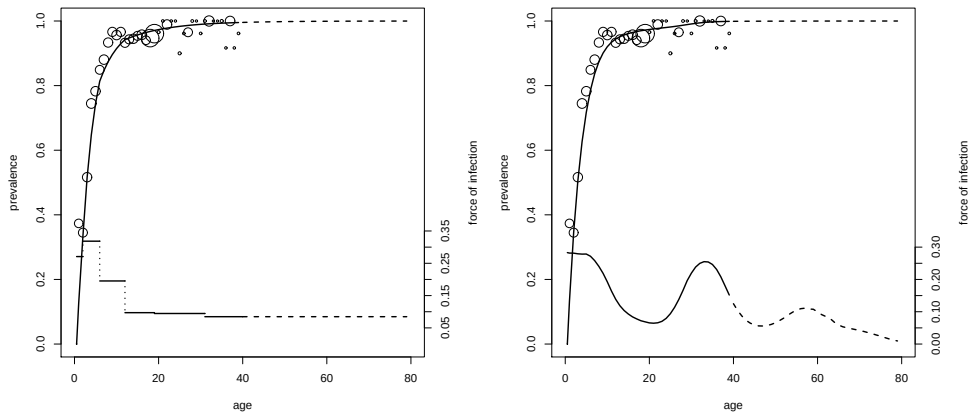

Figure 5.3: Estimated prevalence (upper curve) and force of infection (lower curve) obtained from contact rates estimated using maximum likelihood for Wallinga *et al.* (2006)'s saturated model (left panel) and using bivariate smoothing (right panel).

parametric bootstrap approach, explicitly accounting for all sources of variability.

Contacts with High Transmission Potential

The aim is to trace the kind of contact which is most likely to be responsible for VZV transmission, hereby exploiting the following details provided on each contact: duration and type of contact, which is either close or non-close. First, the contact rates $c(a, a')$ are estimated using all reported contacts (denoted 'C_1') as we did in Section 5.2.4 and further, as presented in Table 5.5, four specific contact types with high transmission potential for VZV are selected (Ogunjimi *et al.*, 2009; Melegaro *et al.*, 2010). From our analyses in Chapter 4, we know that these contact types very likely encompass all contacts within the household, between infants at the daycare center and between children at school. We will explore which contact type induces the best fit to the serological data.

Assuming CP, maximum likelihood estimates for the transmission parameters q_k ($k = 1, \ldots, 5$) and for the basic reproduction number R_0 together with their corresponding 95% profile likelihood CIs (first entry), are presented in Table 5.6. For each model C_k, the AIC-value, AIC difference Δ_k, Akaike weight w_k (2.18), and evidence ratio ER, are calculated following Burnham and Anderson (2002) (cf. Section 2.2.3). According to the AIC-criterion, although AIC differences are minor, the contact ma-

Table 5.5: Candidate models assuming various sorts of contact underlying VZV transmission.

Model	Parameter	Contact Type
C_1	q_1	all contacts
C_2	q_2	close contacts
C_3	q_3	close contacts $>$ 15 minutes
C_4	q_4	close contacts and non-close contacts $>$ 1 hour
C_5	q_5	close contacts $>$ 15 minutes and non-close contacts $>$ 1 hour

Table 5.6: ML-estimates for the proportionality factor and R_0, 95% profile likelihood CIs (first entry), 95% bootstrap-based percentile CIs (second entry) and several measures related to model selection, obtained from contact rates estimated using bivariate smoothing, considering different types of contact C_1-C_5, assuming CP.

Model	$\hat{q}_k$	95% CI for q_k	$\widehat{R}_0$	95% CI for R_0	AIC	Δ_k	w_k	ER
C_1	0.132	[0.124, 0.140]	15.69	[14.74, 16.69]	1386.618	11.660	0.002	340.4
		[0.103, 0.175]		[12.34, 21.41]				
C_2	0.160	[0.150, 0.169]	10.24	[9.65, 10.85]	1379.581	4.623	0.057	10.1
		[0.126, 0.208]		[8.21, 13.68]				
C_3	0.173	[0.163, 0.184]	8.68	[8.18, 9.20]	1374.958	0.000	0.574	1.0
		[0.133, 0.221]		[6.89, 11.34]				
C_4	0.145	[0.136, 0.154]	11.73	[11.05, 12.47]	1380.354	5.396	0.039	14.9
		[0.113, 0.188]		[9.41, 15.95]				
C_5	0.156	[0.147, 0.166]	10.40	[9.79, 11.04]	1376.068	1.110	0.329	1.7
		[0.119, 0.204]		[8.05, 14.10]				

trix consisting of close contacts longer than 15 minutes (model C_3) implies the best fit to the serological data. A graphical representation of the estimated prevalence and force of infection is omitted here, since the result is very close to the one obtained for model C_1 in Figure 5.3. Further, there is evidence for model C_5 as well, having an Akaike weight of 0.329 and an evidence ratio of 1.7. The latter model adds non-close contacts longer than one hour to model C_3, so the models are closely related.

SEIR-ODE Method

In this context, we note that Ogunjimi *et al.* (2009) compared two different ML-estimation methods for the proportionality factor q, considering the same serological

data set for VZV and the same contact matrices as listed in Table 5.5. The first estimation method corresponded to our approach in iteratively solving the equations determined by (5.2), implicitly assuming that the recovery rate is much larger than the force of infection. The second estimation method of Ogunjimi *et al.* (2009) was based on an age and time dependent MSEIR model. This extension of the MSIR model includes an extra E-compartment to account for the latent period (infected but not yet infectious to others) when infected with VZV. The set of PDEs (2.1) were thus extended with an extra transition stage between S and I and reformulated using a widely applied discretization of the continuous age variable, to transform the PDEs into approximating ODEs. The estimation method then involved solving these ODEs in order to obtain equilibrium values for the proportion of seropositives. The first, iterative method thus starts from assuming endemic equilibrium, while the MSEIR-ODE method only applies equilibrium at the end of the calculations in order to calculate the loglikelihood value and is computationally much more time consuming. The analyses from Ogunjimi *et al.* (2009) showed that the iterative method gave a smaller deviance and thus a better fit to the seroprevalence profile compared to the MSEIR-ODE method, though the difference was only minor and the estimated q and R_0 were broadly similar.

Non-Parametric Bootstrap

We explicitly acknowledge that up till now, by keeping the estimated contact rates fixed, we have ignored the variability originating from the contact data. In order to assess sampling variability for the social contact data and the serological data altogether, we will use a non-parametric bootstrap approach (cf. Section 2.2.2). Furthermore, building in a randomization process, uncertainty concerning age is accounted for. After all, in the social contact data, ages of respondents are rounded down, which is also the case for some individuals in the serological data set. Concerning the age of contacts, a lower and upper age limit is given by the respondents. Instead of using the mean value of these age limits, a random draw is now taken from the uniform distribution on the corresponding age interval. In summary, each bootstrap cycle consists of the following six steps:

1. randomize ages in the social contact data and the serological data set;
2. take a sample with replacement from the respondents in the social contact data;
3. recalculate post-stratification weights based on age and household size of the selected respondents;

4. estimate the social contact matrix (smooth-then-constrain approach);

5. take a sample with replacement from the serological data;

6. estimate the transmission parameters and R_0.

This bootstrap approach allows one to calculate bootstrap CIs for the transmission parameters and for the basic reproduction number, which take into account all sources of variability.

The impact on statistical inference is now illustrated for the models considered in the previous section. Nine hundred bootstrap samples are taken from the contact data and from the serological data simultaneously, while ages are being randomized. Merely $B = 587$ bootstrap samples lead to convergence in all five smoothing procedures, which might be induced by the sparse structure of the contact data. However, by individual monitoring of non-converging `gam` functions, convergence was reached after all and a comparison of the bootstrap results showed little difference whether or not these samples were included. 95% percentile CIs for q and R_0 are calculated based on the $B = 587$ bootstrap samples (see Table 5.6, second entry). Taking into account sampling variability for the social contact data has a noticeable impact, as can be seen from the wider 95% CIs.

5.3 Age-Dependent Proportionality of the Transmission Rates

The proportionality factor q might depend on several characteristics related to susceptibility and infectiousness, which could be e.g. ethnic-, climate-, disease- or age-specific. Examples of age-specific characteristics related to susceptibility and infectiousness include the mean infectious period, mucus secretion and hygiene. In the situation of seasonal and pandemic influenza this has been established and used in realistic simulation models (see e.g. Cauchemez *et al.* (2004) and Longini *et al.* (2005)). Furthermore, the conversational and physical contacts reported in the diaries serve as proxies of those events by which an infection can be transmitted. For example, sitting close to someone in a bus without actually touching each other, may also lead to transmission of infection. In light of these discrepancies, q can be considered as an age-specific adjustment factor which relates the true contact rates underlying infectious disease transmission to the social contact proxies.

In view of this, we will explore whether q varies with age, an assumption we will

refer to as 'age-dependent proportionality' (AP):

$$\beta(a, a') = q(a, a') \cdot c(a, a'), \tag{5.7}$$

which in the discrete framework turns into: $\beta_{ij} = q_{ij} \cdot c_{ij}$ $(i, j = 1, \ldots, J)$. In the previous section, it was observed that, under the CP assumption, close contacts longer than 15 minutes imply the best fit to the serological data for VZV. Therefore in the following, the contact rates are estimated from the recorded number of close contacts that last longer than 15 minutes and we will elaborate on model C_3 by assuming age dependence for q. Both discrete matrix structures as well as 'continuous' loglinear regression models are used to allow for an AP factor. Finally, we assess the level of model selection uncertainty and calculate a model averaged estimate of the basic reproduction number and the critical immunization level for VZV in Belgium.

5.3.1 Discrete Structures

The proportionality factor q_{ij} is now allowed to differ between age classes. Discrete matrix structures, involving two transmission parameters θ_1 and θ_2, are explored in modelling q_{ij}. Five models are considered, which fit the following structures for q_{ij} to the seroprevalence data:

$$M_1 = \begin{pmatrix} \theta_1 & \theta_2 \\ \theta_2 & \theta_2 \end{pmatrix}, \; M_2 = \begin{pmatrix} \theta_1 & \theta_1 \\ \theta_2 & \theta_2 \end{pmatrix}, \; M_3 = \begin{pmatrix} \theta_1 & \theta_2 \\ \theta_2 & \theta_1 \end{pmatrix}, \tag{5.8}$$
$$M_4 = \begin{pmatrix} \theta_1 & 0 \\ 0 & \theta_2 \end{pmatrix}, \; M_5 = \begin{pmatrix} \theta_1 & \theta_2 \\ \theta_1 & \theta_2 \end{pmatrix}.$$

The population is divided into two age classes, namely $[0.5, 12)$ and $[12, 80)$, a choice based on the dichotomy of the population according to the schooling system in Belgium (Section 5.1.3), yielding the smallest AIC-value. Note that higher order extensions, considering more parameters and/or number of age classes, were fitted to the serological data as well. The improvement in loglikelihood, however, does not outweigh the increase in the number of transmission parameters.

Notice that the structures of M_1-M_5 resemble the mixing patterns imposed on the WAIFW matrix in the traditional Anderson and May (1991) approach. We would like to emphasize that the method proposed here differs greatly from the latter, since the WAIFW matrix is now estimated using the estimated contact rates: $\beta_{ij} = q_{ij} \cdot \hat{c}_{ij}$. Hence, in contrast with the approach of Anderson and May (1991) who estimate β_{ij} by fixing the structure of the mixing pattern, in our approach we estimate the contact

pattern from the contact survey data and use several proportionality structures to select the best model from which the β_{ij} are estimated.

Table 5.7 displays ML-estimates for θ_1, θ_2 and the basic reproduction number R_0, together with their corresponding 95% percentile CIs ($B = 603$ bootstrap samples converged out of 700). For model M_4, θ_2 is non-identifiable, and unconstrained optimization of model M_5 would not lead to convergence. According to AIC, the remaining models fit equally well and are informative with respect to VZV transmission dynamics. Most likely, this is due to the fact that the main transmission routes for VZV are between children and from infectious children to susceptible adults, embodied by the first column $(\theta_1, \theta_2)^T$. The three models result in approximately the same estimates for θ_1 and θ_2 and consequently the differences in AIC are only minor.

It is clear from Table 5.7 that we estimate a difference (though non-significant according to the 95% CIs) in transmissibility between susceptibles younger and older than 12 years, which cannot be solely explained by the estimated contact rates. More specifically, in case of mixing with an infectious child, q is estimated to be about 2.5 times larger for susceptible children compared to susceptible adolescents or adults. A possible explanation is that when infectious children make close contact with susceptible children during a sufficient amount of time, the probability of effective VZV transmission is higher compared to the same situation with susceptible adolescents or adults. Another potential cause is underreporting of contacts between children. After all, up to the age of eight, the contact diaries were filled in by the parents, which may have induced some reporting bias (cf. Section 3.2.2).

5.3.2 Continuous Modelling

As opposed to the previous, the proportionality factor $q(a, a')$ is now allowed to vary continuously over age. Loglinear regression models are considered for $q(a, a')$, since we expect an exponential decline of q over a due to hygienic habits as well as an exponential decline of q over a' due to decreasing mucus secretion. The following loglinear models are fitted to the data:

$$
\begin{aligned}
M_6 : \quad & \log\{q(a)\} = && \gamma_0 + \gamma_1 a; \\
M_7 : \quad & \log\{q(a)\} = && \gamma_0 + \gamma_1 a + \gamma_2 a^2; \\
M_8 : \quad & \log\{q(a')\} = && \gamma_0 + \gamma_1 a'; \\
M_9 : \quad & \log\{q(a')\} = && \gamma_0 + \gamma_1 a' + \gamma_2 (a')^2; \\
M_{10} : \quad & \log\{q(a, a')\} = && \gamma_0 + \gamma_1 a + \gamma_2 a'.
\end{aligned}
$$

Table 5.7: Candidate models for the proportionality factor together with ML-estimates for the transmission parameters and R_0, 95% bootstrap-based percentile CIs, and several measures related to model selection.

Model	Parameter		95% CI	$\widehat{R}_0$	95% CI for R_0	K	AIC	Δ_k	w_k	ER
C_3	$\hat{q}$	0.173	[0.133, 0.221]	8.68	[6.89, 11.34]	1	1374.958	8.884	0.003	84.9
M_1	$\hat{\theta}_1$	0.185	[0.136, 0.244]	4.79	[4.15, 9.98]	2	1366.306	0.232	0.261	1.1
	$\hat{\theta}_2$	0.079	[0.006, 0.196]							
M_2	$\hat{\theta}_1$	0.183	[0.138, 0.240]	5.37	[4.47, 9.68]	2	1366.285	0.211	0.264	1.1
	$\hat{\theta}_2$	0.078	[0.006, 0.187]							
M_3	$\hat{\theta}_1$	0.185	[0.136, 0.244]	8.26	[6.82, 11.25]	2	1366.074	0.000	0.293	1.0
	$\hat{\theta}_2$	0.069	[0.006, 0.199]							
M_6	$\hat{\gamma}_0$	-1.622	[-2.028, -1.212]	5.79	[4.63, 12.60]	2	1368.709	2.635	0.079	3.7
	$\hat{\gamma}_1$	-0.023	[-0.067, 0.016]							
M_7	$\hat{\gamma}_0$	-1.720	[-2.441, -1.182]	5.03	[4.20, 1318.68]	3	1368.325	2.251	0.095	3.1
	$\hat{\gamma}_1$	0.014	[-0.086, 0.305]							
	$\hat{\gamma}_2$	-0.002	[-0.024, 0.001]							
M_8	$\hat{\gamma}_0$	-1.517	[-2.224, -0.446]	3.55	[1.76, 159.96]	2	1374.324	8.250	0.005	61.9
	$\hat{\gamma}_1$	-0.065	[-0.403, 0.064]							

Model M_6 models q as a first degree function of the age of the susceptible and model M_7 allows for an additional quadratic effect of age, a^2. Models M_8 and M_9 are the analogue of M_6 and M_7 for the age of the infectious person, a'. Finally, M_{10} models q as an exponential function of a and a' simultaneously. For model M_9, no convergence was obtained and model M_{10} gives rise to an estimated proportionality factor which is exponentially increasing over a', inducing unrealistically large estimates for q at older ages.

Maximum likelihood estimates for the model parameters and the basic reproduction number R_0 are presented in Table 5.7, together with the corresponding 95% percentile CIs ($B = 603$ bootstrap samples converged out of 700). According to the AIC-criterion, M_6 and M_7 fit equally well. Allowing the proportionality factor to vary by age of infectious persons, does not seem to substantially improve model fit, as can be seen by comparing the AIC-values of C_3 and M_8. Clearly for models M_7 and M_8, the upper limits of the CIs for R_0 are very large, as a consequence of estimated proportionality factors which are exponentially increasing over a and a', respectively. This result originates from two things: first, there is lack of serological information for individuals aged 40 and older, and second, VZV is highly prevalent in the population and most individuals become infected with VZV before the age of ten. Mathemat-

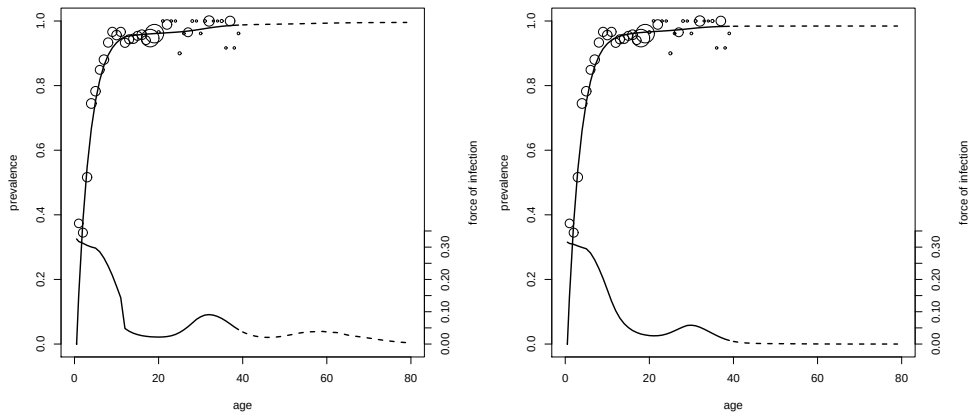

Figure 5.4: Estimated prevalence (upper curve) and force of infection (lower curve) for the discrete model M_3 (left panel) and the continuous model M_7 (right panel).

ically the latter means that from a certain age on, $r(a) \approx 1$ and $r'(a) \approx 0$, leading to an indeterminate force of infection $\lambda(a) = r'(a)/\{1 - r(a)\}$. In Section 5.3.4, we assess the sensitivity of the results to the former issue, repeating all analyses using simulated serological data for the age range $[40, 80)$.

Figure 5.4 displays the estimated prevalence function and force of infection for the discrete model M_3 (left) and the continuous model M_7 (right). The results are remarkably similar. The effect of making q age-dependent is visualized by comparing Figure 5.4 to the fit of model C_1, which was very close to model C_3, in Figure 5.3 (on the right). The models assuming AP estimate an initially higher force of infection and a steeper decrease from the age of ten, after which the force of infection is reduced by a factor two, compared to the CP model. While the latter model predicts total immunity for VZV at older ages, the AP models estimate a fraction of seropositives which is below one at all times.

5.3.3 Model Selection and Multimodel Inference

Table 5.7 presents all candidate models for the proportionality factor q we have collected up till now, among which the CP model C_3, the discrete AP models M_1, M_2 and M_3, and the continuous AP models M_6, M_7 and M_8. Further for each model, the number of parameters K, the AIC-value, the AIC difference Δ_k, the Akaike weight w_k and the evidence ratio (ER) are displayed.

Model M_3 with an assortative component θ_1 and a background component θ_2 is the 'best' model for q according to AIC. However, model selection uncertainty is likely to be high since the selected best model has an Akaike weight of only 0.293 (Burnham and Anderson, 2002). The evidence ratios for M_3 versus M_1 and M_2 are both 1.1, which means there is weak support for the best model. If many independent samples could be drawn, the three discrete age-dependent models would probably compete each other for the 'best' model position. The continuous models M_6 and M_7 have evidence ratios around 3.5, indicating that these models also contribute some information. Models C_3 and M_8 have the largest AIC difference Δ_k, a very small Akaike weight (≤ 0.005) and very large evidence ratios (84.9 and 61.9, respectively), which means there is little support for these two models.

Since there is no single model in the candidate set that is clearly superior to the others and since the estimate for the basic reproduction number R_0 varies noticeably over the candidate models, we are not inclined to base prediction only on M_3. Applying concepts of model averaging (Burnham and Anderson, 2002), a weighted estimate of R_0 is calculated based on the model estimates and their corresponding Akaike weights, as formulated in (2.19):

$$\widehat{\overline{R}}_0 = \sum_{k=1}^{7} w_k (\widehat{R}_0)_k = 6.07.$$

With the bootstrap procedure, we obtain a 95% percentile CI for this model averaged estimate $\widehat{\overline{R}}_0$, namely $[4.4, 351.6]$. Again, there is a large upper limit induced by the same issues reported in Section 5.3.2.

5.3.4 Sensitivity Analysis

In order to assess the lack-of-data-problem, we simulate serological data for the age range $[40, 80)$ assuming a constant prevalence of 0.983, which is estimated from a thin plate regression spline model for the original serological data. Sample sizes for one-year age groups are chosen according to the Belgian population distribution in 2003 and the total size of serological data now amounts to $n = 3856$. The seven candidate models for the proportionality factor q are now applied to the original serological data augmented with the simulated data. The results are presented in Table 5.8 and are, overall, quite similar to the results obtained before (Table 5.7). The 95% percentile CIs for R_0 ($B = 599$ bootstrap samples converged out of 700), however, are narrower since the simulated data for the age range $[40, 80)$ are 'forcing' the proportionality factor q to follow a natural pace. This is illustrated for model M_7

Table 5.8: Candidate models for the proportionality factor applied to the serological data set augmented with simulated data, together with ML-estimates for the transmission parameters and R_0, 95% bootstrap-based percentile CIs, and several measures related to model selection.

Model	Parameter		95% CI	$\widehat{R}_0$	95% CI for R_0	K	AIC	Δ_k	w_k	ER
C_3	$\hat{q}$	0.159	[0.126, 0.195]	7.98	[6.60, 10.19]	1	1618.747	70.774	$\ll 0.0001$	$\gg 10^3$
M_1	$\hat{\theta}_1$	0.189	[0.137, 0.250]	4.20	[3.88, 5.74]	2	1548.714	0.741	0.201	1.4
	$\hat{\theta}_2$	0.052	[0.021, 0.095]							
M_2	$\hat{\theta}_1$	0.186	[0.136, 0.247]	4.74	[4.36, 6.07]	2	1548.627	0.654	0.210	1.4
	$\hat{\theta}_2$	0.052	[0.020, 0.091]							
M_3	$\hat{\theta}_1$	0.189	[0.137, 0.250]	8.28	[6.43, 11.52]	2	1548.344	0.371	0.242	1.2
	$\hat{\theta}_2$	0.044	[0.016, 0.082]							
M_6	$\hat{\gamma}_0$	-1.561	[-1.934, -1.120]	4.96	[4.47, 6.54]	2	1551.321	3.348	0.055	5.3
	$\hat{\gamma}_1$	-0.035	[-0.067, -0.014]							
M_7	$\hat{\gamma}_0$	-1.793	[-2.247, -1.079]	5.22	[4.60, 7.51]	3	1547.973	0.000	0.292	1.0
	$\hat{\gamma}_1$	0.030	[-0.074, 0.126]							
	$\hat{\gamma}_2$	-0.002	[-0.006, 0.001]							
M_8	$\hat{\gamma}_0$	-1.458	[-2.061, -0.844]	2.69	[2.08, 12.97]	2	1610.113	62.140	$\ll 0.0001$	$\gg 10^3$
	$\hat{\gamma}_1$	-0.103	[-0.254, 0.016]							

in Figure 5.5, where the estimated function $q(a)$ is depicted for 100 randomly chosen bootstrap samples. Particularly, right CI limits for R_0 are smaller, whereas for most models the R_0 estimate seems to have decreased just a little bit. Note that the 95% CIs for θ_2 are narrower as well, and that the difference in transmissibility which we observed between susceptibles younger and older than 12 years, is now significant. Model selection uncertainty is illustrated quite nicely here, since four models, M_7, M_3, M_2 and M_1, have Akaike weights close to 0.24 and these models also had the most support for the original data set (Table 5.7). The model averaged estimate $\widehat{\overline{R}}_0$ now equals 5.64 and the 95% bootstrap-based percentile CI is $[4.7, 7.5]$.

5.3.5 Critical Immunization Level

Following Whitaker and Farrington (2004b)'s analysis of VZV, we also estimate the CIL (cf. Section 2.1.2), i.e. the minimal proportion of the population that must be immunized by vaccination to eliminate the infection from the population. We assume that a fraction v of individuals is immunized at a fixed age τ, such that maternal antibodies do not interfere with the vaccine ($\tau > A$). The reproduction number R_v is then defined as the dominant eigenvalue of the $J \times J$ matrix with elements $\{1 - v\ I(a_{[i]} \geq \tau)\}\ G_{ij}$, where $I(.)$ denotes the indicator function and G_{ij} are the

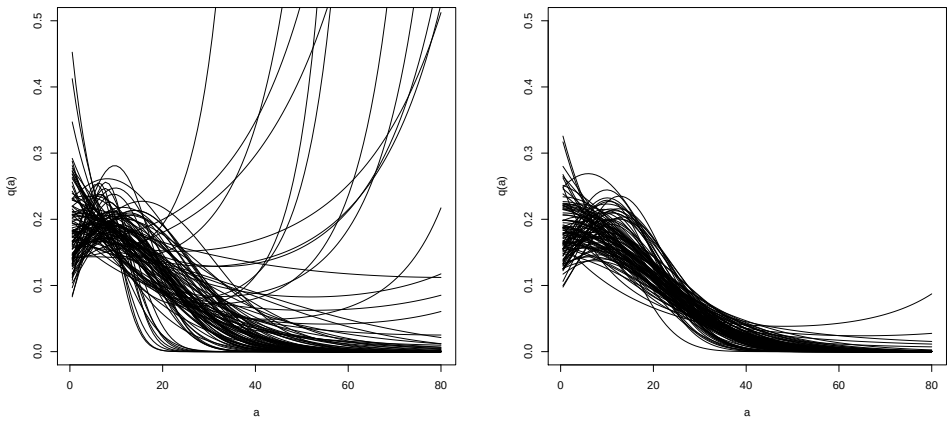

Figure 5.5: $q(a)$ estimates for model M_7, shown for 100 randomly chosen bootstrap samples from the original serological data (left panel) and from this data augmented with simulated data for $[40, 80)$ (right panel).

elements of the next generation matrix defined by (4.1). R_v represents the expected number of secondary cases produced when a typical infected individual is introduced into the population, assuming that all immunity (apart from passive immunity) is vaccine-derived. The CIL is then the smallest value v for which $R_v = 1$. We also consider an alternative CIL (denoted v_A) which is directly related to the basic reproduction number and is often applied in the literature: $v_A = 1 - 1/R_0$, assuming immunization takes place immediately after waning of maternal antibodies, at age A. Note that a 95% CI for v_A can be easily computed from the 95% CI for R_0 (monotone transformation).

In our application, we consider a vaccination strategy in which individuals are immunized at the age of 12 months, which is consistent with current MMR vaccination in Belgium. Results are presented for both the original serological data as well as the serological data augmented with simulated data for the age range $[40, 80)$, described in Section 5.3.4. Table 5.9 shows that $\hat{v}_{\text{MMR}}$ is consistently larger than $\hat{v}_A$, as expected, but still quite close since there is only a six months difference between the ages of immunization. 95% bootstrap-based percentile CIs for v_A are provided, as well as model averaged (MA) estimates for the CILs (range from 81% to 84%). Note that we assumed that immunization is with a single dose of a 100% effective vaccine, which is a rather simplistic though widely used setting (see e.g. Whitaker and Farrington, 2004b). For a more realistic assessment of the CIL for VZV, the values obtained

Table 5.9: Estimated critical immunization levels for vaccination at the age of 6 and 12 months, $\hat{v}_A$ and $\hat{v}_{\text{MMR}}$, respectively, using the original serological data (left) and using this data augmented with simulated data for $[40, 80)$ (right).

	Original Data			Simulated Data		
Model	$\hat{v}_A$	95% CI for v_A	$\hat{v}_{\text{MMR}}$	$\hat{v}_A$	95% CI for v_A	$\hat{v}_{\text{MMR}}$
C_3	0.885	[0.855, 0.912]	0.891	0.875	[0.848, 0.902]	0.880
M_1	0.791	[0.759, 0.900]	0.802	0.762	[0.742, 0.826]	0.775
M_2	0.814	[0.776, 0.897]	0.825	0.789	[0.771, 0.835]	0.802
M_3	0.879	[0.853, 0.911]	0.881	0.879	[0.844, 0.913]	0.880
M_6	0.827	[0.784, 0.921]	0.837	0.798	[0.776, 0.847]	0.810
M_7	0.801	[0.762, 0.999]	0.808	0.808	[0.783, 0.867]	0.814
M_8	0.718	[0.432, 0.994]	0.738	0.628	[0.519, 0.923]	0.659
MA	0.827	[0.77, 0.91]	0.835	0.811	[0.79, 0.87]	0.819

actually need to be corrected for other factors such as the vaccine efficacy (take and degree, see e.g. Hill and Longini, 2003), waning vaccine-induced immunity, and the circulation of herpes zoster. However, the goal was merely to illustrate the sensitivity of the estimated CIL with respect to different parametric models assumed for q when making use of the social contact data approach.

5.4 Concluding Remarks

In this chapter, an overview of different estimation methods for infectious disease parameters from data on social contacts and serological status, was given. The theoretical framework included a compartmental MSIR model, taking into account the presence of maternal antibodies, and the mass action principle, as presented by Anderson and May (1991). An important assumption made was the one of endemic equilibrium, which means that infection dynamics are in a steady state. The serological data set we used was collected over 17 months, averaging over potential epidemic cycles of VZV in Belgium during that period. In Section 5.1, we have illustrated the traditional Anderson and May (1991) approach of imposing mixing patterns on the WAIFW matrix to estimate transmission parameters from serological data. In contrast, the novel approach of using social contact data to estimate infectious disease parameters, avoids the choice of a parametric model for the entire WAIFW matrix.

The idea is fairly simple: transmission rates for infections that are transmitted

from person to person in a non-sexual way, such as VZV, are assumed to be proportional to rates of making conversational and/or physical contact, which can be estimated from contact surveys. Although more time consuming, the bivariate smoothing approach as proposed in Section 5.2, was better able to capture important features of human mixing behavior, compared to the maximum likelihood estimation method of Wallinga *et al.* (2006). However, when a non-parametric bootstrap approach was applied to take into account sampling variability, convergence problems arose, probably due to the large number of zeros in combination with the log-link. Therefore, a mixture of Poisson distributions or a zero-inflated negative binomial distribution could be more appropriate. Further, in Section 5.2, we dealt with a couple of challenges posed by Halloran (2006). The social contact survey contained useful additional information on the contact itself, which allowed us to target very specific contact types with high transmission potential for VZV. Furthermore, a non-parametric bootstrap approach was proposed to improve statistical inference.

The CP assumption was relaxed in Section 5.3 and we have shown that an improvement of fit could be obtained by disentangling the transmission rates into a product of two age-specific variables: the age-specific contact rate and an age-specific proportionality factor. The latter may reflect, for instance, differences in characteristics related to susceptibility and infectiousness or discrepancies between the social contact proxies measured in the contact survey and the true contact rates underlying infectious disease transmission. We would like to emphasize that there probably exist other models for $q(a, a')$ than the ones considered in Section 5.3, which fit the data even better. Our choice of a set of plausible candidate models was directed by parsimony on the one hand, limiting the total number of parameters to three, and prior knowledge on the other hand, considering loglinear models. Furthermore, we restricted analyses to close contacts lasting longer than 15 minutes, which means that close contacts of short duration and non-close contacts are assumed not to contribute to transmission of VZV.

It is important to note that different assumptions concerning the underlying type of contact as well as different parametric models for $q(a, a')$, are likely to entail different estimates of R_0, however, they may still induce a similar fit to the serological data. In order to deal with this problem of model selection uncertainty we have turned to multimodel inference in Section 5.3.3. In Figure 5.6, estimates of R_0 are presented for the main estimation methods considered in this chapter: the traditional method of imposing mixing patterns to the WAIFW matrix (W_4) and the method of using data on social contacts, assuming CP (the saturated model SA, C_1 and C_3) and AP (M_1, M_2 and M_3). There is a pronounced variability in the estimates of R_0, which is

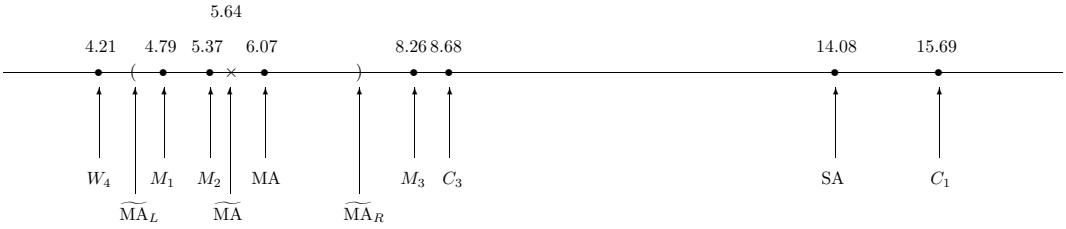

Figure 5.6: R_0 estimates for mixing pattern W_4, applied to the serological data in Section 5.1.3, and for the following models using social contact data: the saturated model (SA) as proposed by Wallinga *et al.* (2006), applied in Section 5.2.4 assuming CP, and further bivariate smoothing models: CP models C_1 and C_3 considering all and close contacts longer than 15 minutes, respectively (Section 5.2.5) and discrete AP models M_1, M_2 and M_3 (Section 5.3.1). The model averaged estimates for R_0 calculated from Table 5.7 (MA), based on the original serological data, and from Table 5.8 ($\widetilde{\text{MA}}$), based on the serological data set augmented with simulated data, are displayed, as well as 95% bootstrap-based percentile CI limits for the latter: $[\widetilde{\text{MA}}_L, \widetilde{\text{MA}}_R]$.

partially captured by the model averaged estimate MA, calculated from Table 5.7.

When estimating $q(a, a')$, we were actually faced with three problems of indeterminacy. First, there is lack of serological information for individuals aged 40 and older, second, prevalence of VZV rapidly stagnates at a high level, leading to an indeterminate force of infection, and third, serological surveys do not provide information related to infectiousness. The sensitivity analysis in Section 5.3.4 showed that lack of serological data had a large impact on CIs for R_0. We simulated data for the age range $[40, 80)$, giving rise to a model averaged estimate $\widetilde{\text{MA}}$ as displayed in Figure 5.6 with corresponding CI limits $[\widetilde{\text{MA}}_L, \widetilde{\text{MA}}_R]$. Nevertheless, parameter estimates were fairly close when comparing the fit to the original data with the fit to the augmented data. Furthermore, after finalizing the study, extra blood samples from individuals aged 40 years and older were tested for VZV, and we obtained similar estimates when the original VZV serology was complemented with these data.

Still, even for the complemented serological data set we encountered the issue of stagnating prevalence since the percentage of seropositives in older age groups was close to 100%. This problem might be controlled by combining information on the same infection over different countries or on different airborne infections, assuming there is a relation between the country- or disease-specific $q(a, a')$, respectively. This strategy already appeared beneficial when estimating R_0 directly from seroprevalence

data, without using social contact data (Farrington *et al.*, 2001). The third problem of indeterminacy manifested through models which only expressed age differences in q for infectious individuals, such as the discrete model M_5 (Section 5.3.1) and the continuous models M_8 and M_9 (Section 5.3.2). These models either did not lead to convergence or induced unrealistically large bootstrap estimates for q at older ages.

Further, the impact of intervention strategies such as school closures (cf. Section 3.2.4), might be investigated by incorporating transmission parameters, estimated from data on social contacts and serological status, in an age-time-dynamical setting. Baguelin *et al.* (2010), for example, developed such a real-time model for pandemic influenza in the UK, which was used by House *et al.* (2010) to estimate the effect of local, reactive school closure on intensive care provision. Finally, it is important to note that the models rely on the assumptions of type I mortality and type I maternal antibodies in order to facilitate calculations. Consequently, model improvements could be made through a more realistic approach of demographical dynamics.

Chapter 6

Model Structure Analysis to Estimate Basic Immunological Processes and Maternal Risk for Parvovirus B19

After being infected with PVB19, individuals acquire immunoglobulin G (IgG) antibodies against PVB19 and it is generally assumed that these antibodies persist for a lifetime (Young and Brown, 2004). Since the presence of IgG antibodies indicates past infection with PVB19 and the duration of exposure to infection increases with age, the proportion of seropositives should be monotone increasing with age, provided that there is time equilibrium at the disease (endemic) and population level (demographic), and that mortality attributable to PVB19 infection can be ignored. However, after a steep monotone rise with age, the seroprevalence profiles for PVB19 in each of five European countries (1995-2004) display a decrease or plateau between the ages of 20 and 40 years, after which the prevalence continues to monotonically increase with age (Figures 3.2 and 3.3). This phenomenon does not support the assumption of lifelong immunity. A cohort effect due to an epidemic or a demographical shift seems very unlikely since Nascimento *et al.* (1990) noted a similar decrease in adults for serological studies conducted in the 1980's in Rio De Janeiro (Brazil), England and Wales, Japan, and Germany (Nascimento *et al.*, 1990; Cohen and Buckley,

1988; Nunoue *et al.*, 1985; Schwarz *et al.*, 1987). Additionally, we find a decrease or plateau in the age class 20-40 years for PVB19 seroprevalence studies conducted in the 1990's in Japan, Australia and The Netherlands (Matsunaga *et al.*, 1995; Kelly *et al.*, 2000; Zaaijer *et al.*, 2004).

Furthermore, Schoub *et al.* (1993) used an avidity test to establish that most PVB19 infections in pregnancy are not primary infections but reinfections, and in 2007 a case report was published of a secondary symptomatic PVB19 infection in a healthy, immunocompetent adult two years after a positive PVB19 IgG antibody test during prenatal care (Kaufmann *et al.*, 2007). This may imply that reinfection with PVB19 remains possible after an adequate level of antibodies is produced upon primary infection. Hypotheses of waning of IgG antibodies, boosting by exposure to infectious individuals and reinfections, were suggested before (Schoub *et al.*, 1993; Kaufmann *et al.*, 2007; Vyse *et al.*, 2007; Huatuco *et al.*, 2008; Schneider *et al.*, 2008), however up till now these hypotheses have never been tested using empirical data. Gay (1996) used a mixture modelling approach to describe the distribution of continuous PVB19 IgG antibody titers and noted a significant increase with age in the left skewness of the seropositive population, particularly after age 20 years, suggesting a decay of antibody levels. Gaining insight in the processes underlying PVB19 transmission dynamics is of major public health interest, since the decrease or plateau in IgG seroprevalence is specifically observed in women of childbearing age.

In the absence of longitudinal antibody titer data for PVB19 which would enable us to study the evolution of IgG antibodies directly (e.g. Teunis *et al.*, 2002, for pertussis), we propose an alternative approach. We explore several immunological scenarios through mathematical modelling and infer on waning and boosting rates by augmenting the serological data with data on social contact patterns (Chapter 5), assessing whether the scenarios are able to explain the observed decrease in the seroprevalence profile for adults. Similar models were considered before to determine the effect of waning and boosting of immunity on vaccination schedules for measles (Rouderfer *et al.*, 1994) and to identify the causes of an epidemic outbreak of pertussis (van Boven *et al.*, 2000, 2001). However in these studies, values for waning and boosting rates were predefined and, in the absence of representative social contact surveys, proportionate mixing was assumed to specify the transmission rates. Inference on transmission dynamics of PVB19 is important for diagnosis, assessing the risk of prenatal infection and designing future vaccination policies. If a significant proportion of the population is infected twice or more with PVB19, it is likely that many secondary infections are asymptomatic or atypical and hence may not be noticed by traditional surveillance systems based on case reporting. The risk in pregnant women is then

likely underestimated and a larger proportion of undiagnosed fetal complications may therefore be attributable to PVB19 infection during pregnancy.

This chapter covers the study by Goeyvaerts *et al.* (2010b) and is organized as follows. In the first section, in addition to the serological data introduction in Section 3.1.2, we briefly describe the demographic data and social contact data, that inform our model structure analysis for PVB19 in five European countries. The compartmental dynamic transmission models we consider for PVB19 are introduced in Section 6.2. We divide the mathematical scenarios into three types of dynamics; the first type discerning between high and low 'waned' immunity (MSIRW), the second type allowing for multiple infections (MSIRS) and the third type being a mixture of the two previous ones (MSIRWS). For each scenario, exact formulas for the age-specific proportions of susceptible and seropositive individuals are derived. These are incorporated into the ML procedure to estimate the unknown parameters on PVB19 transmission, immunology and risk in pregnancy, which is described further on in Sections 6.2.2 and 6.2.3. Additionally, we propose some model extensions to assess age-specific heterogeneity in the immunity transition rates and in the proportionality factor q for the transmission rates.

In Section 6.3, we present the results of this model structure analysis and summarize the main findings using different inferential means. Our results show that for four countries, model selection criteria favor the scenarios allowing for waning immunity at an age-specific rate over the assumption of lifelong immunity, assuming that the transmission rates are directly proportional to the contact rates. Different views on the evolution of the immune response to PVB19 infection lead to altered estimates of the age-specific force of infection and the basic reproduction number. The scenarios which allow for multiple infections during one lifetime, predict a higher frequency of PVB19 infection in pregnant women and of associated fetal deaths. Some final conclusions and a discussion are provided in Section 6.5. When pre-vaccination serological data are available, the framework developed in this chapter could prove worthwhile to investigate these different scenarios for other infections as well, such as cytomegalovirus.

6.1 Introduction

6.1.1 Demographic Data

Some demographic figures for each country from the time of data collection will be used when modelling the serological data for PVB19 (cf. Section 3.1.2). First, to make

the data representative of the different populations, post-stratification weights w_i are calculated from demographic data on population sizes per age class, obtained from EUROSTAT (http://epp.eurostat.ec.europa.eu) and the Office for National Statistics, United Kingdom (http://www.statistics.gov.uk/popest). The reference years chosen for BE, EW, FI, IT and PL, are 2003, 1996, 1998, 2004 and 1999, respectively (Table 3.1). The weights w_i are truncated, applying a cut-off c, $\tilde{w}_i = \min(w_i, c)$, to reduce the influence of individuals with extreme weights and to avoid excessive variability. Based on the distributions of the post-stratification weights for all countries, we have chosen c equal to 7.

Further, we will consider a large population of fixed size N and assume demographic equilibrium with $N(a)$ the stationary age distribution for the population size and $\mu(a)$ the age-specific mortality rate, defined as in (2.3). The mortality rates $\mu(a)$ are estimated from the population sizes and additional data on age stratified numbers of deaths in the reference year, obtained from EUROSTAT. A Poisson generalized additive model with log link is used to model the number of deaths as a function of age with population size as an offset factor (Hens *et al.*, 2011). Thin plate regression splines are chosen via the `gam` function (`R`-package `mgcv 1.3-30`). Then, the life expectancy L is estimated from $\hat{\mu}(a)$ using (2.4), and presented in Table 3.1 together with the total population size N, obtained from the demographic data.

Finally, to estimate the frequency and burden of PVB19 infection during pregnancy, data on the number of live births in the reference year stratified by age of the mother at her last birthday, are retrieved from EUROSTAT. The maternal age distribution for live births is denoted by $B(a)$, thus the total number of live births equals $B = \int_0^\infty B(a)da$ (Table 3.1).

6.1.2 Social Mixing

Comparable to rubella, PVB19 is primarily spread by infected respiratory droplets and outbreaks tend to occur during winter and spring time. Close contacts, i.e. with physical skin-to-skin touching, are likely to play an important role in the transmission of PVB19, considering the reports of school outbreaks (Woolf *et al.*, 1989; Rice and Cohen, 1996; Gonçalves *et al.*, 2005), high attack rates in households (Chorba *et al.*, 1986) and outbreaks in hospital wards (Bell *et al.*, 1989; Pillay *et al.*, 1992). Furthermore in different studies, high risk estimates are reported for daycare and after-school clubs personnel, nursery and elementary school teachers (Valeur-Jensen *et al.*, 1999; Gillespie *et al.*, 1990; Cartter *et al.*, 1991), identifying young children as the main spreaders of PVB19. The social contact data approach, which we applied

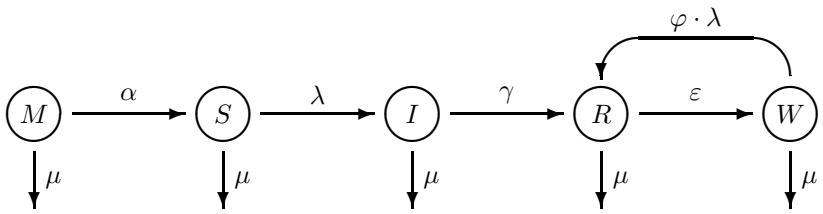

Figure 6.1: Illustration of the MSIRWb-ext compartmental model.

to VZV in Chapter 5, thus seems very convenient to estimate transmission rates for PVB19 as well. For each of the five countries under study for PVB19, POLYMOD contact survey data are available (Section 3.2.2): BE, GB (which encompasses EW), FI, IT and PL. The post-stratification weights are truncated to a maximum of 5; a value chosen based on the weight distributions. Recall that a short summary of the contact data collection and sample sizes for each country are provided in Table 3.4, and that for BE the augmented contact data set for one sampled day is considered (cf. Section 4.1.2).

6.2 Transmission Scenarios for PVB19

6.2.1 Mathematical Models

We will consider several compartmental scenarios to model the dynamics of PVB19 transmission assuming endemic equilibrium. The basic building block will be an MSIR structure as introduced in Section 2.1, assuming that after PVB19 infection, individuals recover and acquire immunity marked by a discernible IgG antibody level. Further, we assume type I maternal antibodies (2.8), and denote by A the age at which maternal antibodies are lost. Mortality due to infection is ignored, which is justifiable for PVB19. To investigate the assumption of lifelong immunity we fit the basic MSIR model to the serological data and compare its fit to specific mathematical scenarios described hereunder, comprising processes of waning, boosting and reinfection with PVB19.

MSIRW Models

Figure 6.1 shows a graphical representation of the 'MSIRWb-ext' model, which allows for waning of disease-acquired antibodies without loss of protective ('cellular')

immunity. Individuals then move at a rate $\varepsilon(a)$ from a high immunity state R to a low immunity state W, in which they are still protected from infection however categorized as being seronegative, i.e. with antibody levels (indicating 'humoral' immunity) falling below the serostatus threshold. We assume that low immunity can be boosted by exposure to infectious individuals. The boosting rate and the force of infection are then directly proportional with a proportionality constant φ, such that the rate at which individuals move back from W to R equals $\varphi \cdot \lambda(a)$. By solving the corresponding set of differential equations, one finds that the fraction in state S is given by (2.9) and that the proportion in state W equals

$$w(a) = \int_A^a \varepsilon(u) \exp\left(-\int_u^a \{\varphi\lambda(v) + \varepsilon(v)\}dv\right)\left\{1 - \exp\left(-\int_A^u \lambda(v)dv\right)\right\} du,$$

if $a > A$. Approximating $r(a)$ by $1 - s(a) - w(a)$, $\forall a > A$, assuming $i(a)$ is small relative to $s(a)$ and $w(a)$, we obtain the following expression for the proportion seropositives:

$$r(a) = \int_A^a \left\{(1-\varphi)\lambda(u) \exp\left(-\int_A^u \lambda(v)dv\right) + \varphi\lambda(u)\right\} \exp\left(-\int_u^a \{\varphi\lambda(v) + \varepsilon(v)\}dv\right) du,$$

if $a > A$. The two special cases in which there is no boosting of low immunity, $\varphi = 0$, and in which the boosting rate exactly equals the force of infection, $\varphi = 1$, as assumed by Rouderfer *et al.* (1994), are considered as well and denoted by 'MSIRW' and 'MSIRWb', respectively.

MSIRS Models

The MSIRS model, displayed in Figure 6.2, allows for loss of disease-acquired immunity and potential reinfection. Individuals are assumed to move from R back to the susceptible state S at a rate $\sigma(a)$. Again, by solving the corresponding set of differential equations and making use of $r(a) \approx 1 - s(a)$, $\forall a > A$, expressions for the proportion of susceptibles and seropositives can be obtained (for $a > A$):

$$\begin{aligned} s(a) &= \exp\left(-\int_A^a \{\lambda(u) + \sigma(u)\}\, du\right) + \int_A^a \sigma(u) \exp\left(-\int_u^a \{\lambda(v) + \sigma(v)\}dv\right) du, \\ r(a) &= \int_A^a \lambda(u) \exp\left(-\int_u^a \{\lambda(v) + \sigma(v)\}dv\right) du. \end{aligned} \tag{6.1}$$

The $MS_1I_1RS_2I_2RS_2$ model (denoted 'MSIRS-ext'), presented in Figure 6.3, is an extension of the MSIRS model and closely follows the model of van Boven *et al.* (2000, 2001) for pertussis. This scenario allows to distinguish between infection in immunologically naive individuals (I_1) and infection in individuals whose immune

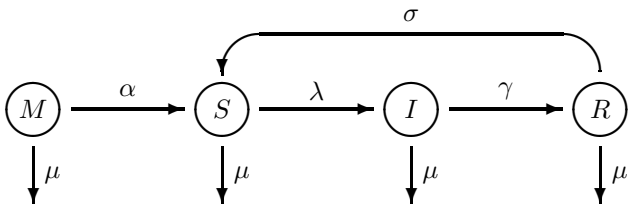

Figure 6.2: Illustration of the MSIRS compartmental model.

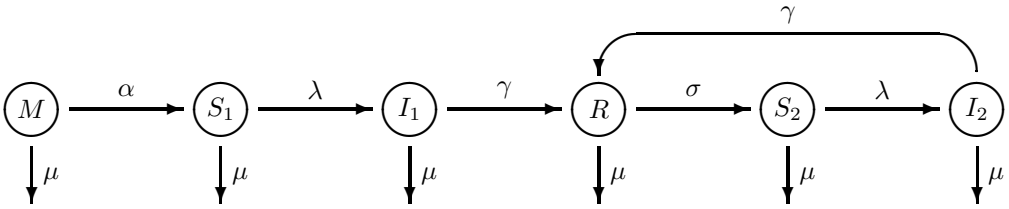

Figure 6.3: Illustration of the MSIRS-ext compartmental model.

system has been primed by infection before (I_2). The proportion of immunologically naive susceptibles $s_1(a)$ is given by equation (2.9) and the set of differential equations yields the following solutions, assuming γ is large and thus $r(a) \approx 1 - s_1(a) - s_2(a)$:

$$s_2(a) \quad = \quad \int_A^a \sigma(u) \left\{ 1 - \exp\left(- \int_A^u \lambda(v) dv \right) \right\} \exp\left(- \int_u^a \{\sigma(v) + \lambda(v)\} dv \right) du,$$

if $a > A$, and the fraction of seropositives is given by formula (6.1). In the MSIRS-ext framework, the mass action principle (2.12) is rewritten as

$$\lambda(a) = D \int_A^\infty \Big\{ \beta_1(a, a') \lambda(a') S_1(a') + \beta_2(a, a') \lambda(a') S_2(a') \Big\} da', \tag{6.2}$$

where $\beta_1(a, a')$ and $\beta_2(a, a')$ are the group-specific age-dependent transmission rates.

MSIRWS Model

The MSIRWS model (Figure 6.4) is an adaptation of the model by Rouderfer *et al.* (1994) for measles, and can be seen as a mixture of the MSIRWb-ext model and the MSIRS model. Individuals in the low immunity state W can either be boosted by exposure to infectious individuals and move back to the high immunity state R at a

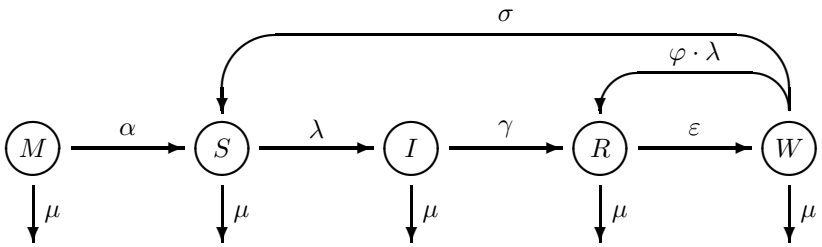

Figure 6.4: Illustration of the MSIRWS compartmental model.

rate $\varphi \cdot \lambda(a)$, or their immunity wanes to such an extent that they become susceptible again at a rate $\sigma(a)$. Approximating $r(a)$ by $1 - s(a) - w(a)$, assuming $i(a)$ is small, we obtain the following system of differential equations for $s(a)$ and $w(a)$:

$$\begin{cases} s'(a) &= \sigma(a)w(a) - \lambda(a)s(a), \\ w'(a) &= \varepsilon(a)\{1 - s(a)\} - \{\varphi\lambda(a) + \sigma(a) + \varepsilon(a)\}w(a). \end{cases}$$

This system of inhomogeneous linear differential equations of order 1 and dimension 2 cannot be solved explicitly for $s(a)$ and $w(a)$. However, by turning to discrete age classes, the solutions can be approximated recursively (cf. Appendix A).

6.2.2 Inference on PVB19 Immunology

In Chapter 5, we have shown that the method of estimating contact rates from social contact surveys and using them to inform transmission rates for infections transmitted predominantly through non-sexual social contacts, is more efficient than the traditional Anderson and May (1991) approach of imposing parametric mixing patterns on the WAIFW matrix (Wallinga *et al.*, 2006; Ogunjimi *et al.*, 2009; Goeyvaerts *et al.*, 2010a). Given the transmission routes and outbreak reports for PVB19 as summarized in Section 6.1.2, and given our findings for VZV in Chapter 5, the WAIFW matrix is assumed proportional to rates of making close contact, and particularly those for which the total contact time per day exceeds 15 minutes. The contact rates are estimated from the POLYMOD contact data by applying a smooth-then-constrain-approach as described in Section 5.2.2. In short, the mean contact surface is estimated using a bivariate smoothing approach with a thin plate regression spline basis (Wood, 2006), assuming a negative binomial distribution for the number of reported contacts over one year age intervals and taking into account post-stratification weights (`gam` function, `R`-package `mgcv 1.3-30`). Subsequently, the estimated con-

tact surface is constrained using age-specific population sizes (Section 6.1.1) to entail reciprocity (Wallinga *et al.*, 2006).

In a first application, we assume constant proportionality (CP, 3.4) for the transmission rates which requires estimation of an unknown proportionality parameter q. Note that for the MSIRS-ext scenario, we have two q parameters in the mass action principle (6.2), q_1 and q_2, to differentiate between infectivity of individuals with primary and secondary infection. Further, it is assumed that the immunity transition rates ε and σ are independent of age. Next, this assumption is relaxed by modelling the waning rate as a piecewise constant function with a cut-off point at a predetermined age H: $\varepsilon(a) = \varepsilon_1$, if $a \in (A, H)$, and $\varepsilon(a) = \varepsilon_2$, if $a \geq H$, and similar for $\sigma(a)$. This model is able to identify age differences in the rate at which antibody levels decay. Finally, we assess the sensitivity of our results with respect to the CP assumption by allowing for an age-dependent q (5.7). This age-specific proportionality factor $q(a, a')$ may reflect, for instance, discrepancies between the social contact proxies measured in the contact survey and the 'true' contact rates underlying infectious disease transmission, or differences in characteristics related to susceptibility or infectiousness; though the latter is not estimable from serological surveys. Similar to what we did in Chapter 5 for VZV, we consider the matrix structures $\boldsymbol{M_1}, \boldsymbol{M_2}, \boldsymbol{M_3}$ for $q(a, a')$ defined in (5.8), involving two transmission parameters θ_1 and θ_2 for the population dichotomized by a cut-off at a predetermined age G.

Since the integral equation (2.12) has no closed form solution, we turn to discrete age classes to estimate the scenario-specific parameters $\boldsymbol{\varepsilon}, \boldsymbol{\sigma}, \varphi$ and $\boldsymbol{q}$ (cf. Appendix A). Through an iterative procedure, the Bernoulli loglikelihood for the serological data is being maximized (cf. `Matlab` code in Appendix C):

$$\ell(\boldsymbol{\varepsilon}, \boldsymbol{\sigma}, \varphi, \boldsymbol{q} | y_1, \ldots, y_n) = \sum_{i=1}^{n} \tilde{w}_i \big\{ y_i \log[r(a_i | \boldsymbol{\varepsilon}, \boldsymbol{\sigma}, \varphi, \boldsymbol{q})] + (1 - y_i) \log[1 - r(a_i | \boldsymbol{\varepsilon}, \boldsymbol{\sigma}, \varphi, \boldsymbol{q})] \big\},$$

where n denotes the sample size of the serological data set, $\tilde{w}_i$ is the truncated post-stratification weight for subject i as defined in Section 6.1.1, and y_i is the binary variable indicating whether subject i of age a_i is classified as being seropositive. Once the ML-estimates for these parameters are obtained, the basic reproduction number R_0 can be computed as the dominant eigenvalue of the $J \times J$ next generation matrix (Diekmann *et al.*, 1990) with elements defined by (4.1).

6.2.3 Risk in Pregnancy

To assess the infection risk in pregnant women, we estimate the average maternal proportion of susceptibles ($\bar{s}_p$) and the average maternal force of infection ($\bar{\lambda}_p$),

$$\bar{s}_p = \frac{\int_0^\infty s(a)B(a)da}{\int_0^\infty B(a)da}, \qquad \bar{\lambda}_p = \frac{\int_0^\infty \lambda(a)s(a)B(a)da}{\int_0^\infty s(a)B(a)da},$$

where $B(a)$ represents the maternal age distribution of live births as introduced in Section 6.1.1. The annual number of PVB19 infections in pregnant women is calculated as follows (Gay *et al.*, 1994):

$$I_p = 0.77 \int_0^\infty \lambda(a)s(a)B(a)da,$$

where 0.77 years embodies the mean duration of pregnancy (40 weeks). To estimate the frequency of fetal deaths due to PVB19 infection during pregnancy, we calculate an average risk of fetal loss using data from the two largest prospective cohort studies of pregnant women with confirmed PVB19 infection reported in the literature: a study from EW (1985-1988 and 1992-1995) by Miller *et al.* (1998) and from DE (1993-1998) by Enders *et al.* (2004). We find an average excess fetal death rate during the first 20 weeks of gestation of 7.7%, when comparing the study populations to a control group of women in EW and DE who were followed up prospectively after varicella infection in pregnancy (Enders *et al.*, 1994). Pastuszak *et al.* (1994) showed that there is no significant difference in the rate of fetal loss between women with and women without primary varicella infection during pregnancy. Although Tolfvenstam *et al.* (2001) suggest that fetal death due to PVB19 infection in late second and third trimester of pregnancy could be more common than previously reported, a recent study by Riipinen *et al.* (2008) confirms the results of Miller *et al.* (1998) and Enders *et al.* (2004) that this is overall a very rare event.

6.3 Results

6.3.1 Constant Waning

For the remainder of the chapter, we assume that the mean duration of infectiousness for PVB19 is $D = 6/365$ years (Anderson and Cherry, 2004) and that maternally derived antibodies are lost at the age of $A = 0.5$ years, implying that neonates younger than 6 months are assumed not to take part in the PVB19 transmission process. The latter is in line with the serological findings of Cohen and Buckley (1988) in London

and Eis-Hübinger *et al.* (1998) in Germany, and the decay estimated by Huatuco *et al.* (2008) from a seroprevalence study in São Paulo, Brazil. Removing the few serological samples of neonates younger than 6 months, which are only covered by the sample for BE, the sample size for the latter becomes $n = 3069$ (Table 3.1).

Further, we consider integer age intervals for all countries: $(0.5, 1), [1, 2), [2, 3), \ldots, [79, 80)$. The different dynamical models are fitted to the serological data, assuming constant waning (CW) rates ε and σ, and CP with respect to close contacts > 15 minutes. Confidence intervals are obtained using the non-parametric bootstrap approach described in Section 5.2.5, taking into account all sources of sampling variability and age uncertainty. The ML-estimates for the scenario-specific parameters and R_0 are displayed in Table 6.1, together with 95% bootstrap-based percentile CIs (2.17) and information criteria AIC (2.14) and BIC (2.15). Figure 6.5 depicts the estimated seroprevalence and force of infection resulting from each compartmental scenario, for all five countries.

Likelihood ratio (LR) tests are performed to test the null hypotheses $H_0 : \varepsilon = 0$ and $H_0 : \sigma = 0$ for the MSIRW and MSIRS models, respectively. Since these null hypotheses are on the boundary of the parameter space $\mathbb{R}^+$, the asymptotic distribution of the LR-test statistic is a 50:50 mixture of χ^2_0 and χ^2_1 (Self and Liang, 1987). The p-values together with the information criteria in Table 6.1 indicate substantial evidence against the assumption of lifelong immunity for PVB19 in BE, EW and IT, and this is also clear from the model fit in Figure 6.5. Note that, except for MSIRW in IT and EW, the same conclusion can be made from the 95% CIs for ε and σ, which also take into account the variability originating from the contact data. For these countries, The MSIRW scenario with boosting comes out as the 'best' model, though AIC and BIC values for MSIRW and MSIRS models are fairly close. For FI and PL, however, the scenarios are not able to elicit any evidence of waning immunity from the PVB19 serology.

The results from MSIRWb-ext and MSIRS-ext for both FI and PL are omitted since ε and σ are estimated to be zero, making neither $\hat{\varphi}$ nor $\hat{q}_2$ estimable. The results from MSIRWb-ext are omitted for Italy as well since 90% of the bootstrap replicates of $\hat{\varepsilon}$ are larger than 10^3. The unboundedness of the parameters and the structure of the Italian serological data conduce to extremely large bootstrap estimates for both ε and φ, which is unrealistic and non-interpretable. For BE and EW, we additionally test whether the proportionality constant φ in the MSIRWb-ext model equals 0 or 1, corresponding to MSIRW and MSIRWb respectively. The former null hypothesis is on the boundary of the parameter space $\mathbb{R}^+$, while the latter hypothesis of $H_0 : \varphi = 1$ implies a classical LR-test. For BE and EW, there is a significant amount of boosting

Table 6.1: ML-estimates for the scenario-specific parameters $\boldsymbol{q}, \varepsilon, \sigma, \varphi$, and the basic reproduction number R_0, with 95% bootstrap-based percentile CIs in square brackets, information criteria AIC and BIC (minima indicated in boldface), and LR-test null hypotheses and p-values, obtained under the assumption of CW.

Country	Model				$\hat{R}_0$		AIC	BIC	LR-test	
									H_0	p-value
BE	MSIR	$\hat{q}$	0.056	[0.047, 0.062]	2.48	[2.27, 2.72]	3477.08	3483.10		
	MSIRW	$\hat{q}$	0.073	[0.058, 0.087]	3.21	[2.70, 3.93]	3390.20	3402.26		
		$\hat{\varepsilon}$	0.004	[0.002, 0.006]					$\varepsilon = 0$	< 0.001
	MSIRWb	$\hat{q}$	0.076	[0.060, 0.093]	3.35	[2.77, 4.17]	**3384.02**	**3396.07**		
		$\hat{\varepsilon}$	0.010	[0.005, 0.014]					$\varepsilon = 0$	< 0.001
	MSIRWb-ext	$\hat{q}$	0.076	[0.059, 0.093]	3.35	[2.77, 4.17]	3385.98	3404.06		
		$\hat{\varepsilon}$	0.009	[0.005, 0.021]					$\varphi = 0$	0.006
		$\hat{\varphi}$	0.91	[0.30, 2.56]					$\varphi = 1$	0.841
	MSIRS	$\hat{q}$	0.064	[0.054, 0.072]	2.84	[2.54, 3.22]	3387.51	3399.56		
		$\hat{\sigma}$	0.013	[0.006, 0.022]					$\sigma = 0$	< 0.001
	MSIRS-ext	$\hat{q}_1$	0.076	[0.000, 0.091]	3.35	[0.00, 4.10]	3386.02	3404.10		
		$\hat{q}_2$	0.000	[0.000, 0.132]					$q_1 = q_2$	0.062
		$\hat{\sigma}$	0.010	[0.005, 0.044]						
EW	MSIR	$\hat{q}$	0.053	[0.047, 0.057]	1.72	[1.64, 1.81]	3551.25	3557.20		
	MSIRW	$\hat{q}$	0.058	[0.050, 0.064]	1.87	[1.72, 2.04]	3533.53	3545.42		
		$\hat{\varepsilon}$	0.003	[0.000, 0.005]					$\varepsilon = 0$	< 0.001
	MSIRWb	$\hat{q}$	0.059	[0.051, 0.065]	1.90	[1.73, 2.07]	**3531.65**	**3543.54**		
		$\hat{\varepsilon}$	0.004	[0.001, 0.008]					$\varepsilon = 0$	< 0.001
	MSIRWb-ext	$\hat{q}$	0.059	[0.051, 0.065]	1.91	[1.74, 2.09]	3532.21	3550.05		
		$\hat{\varepsilon}$	0.008	[0.002, 0.025]					$\varphi = 0$	0.034
		$\hat{\varphi}$	3.16	[0.94, 12.5]					$\varphi = 1$	0.230
	MSIRS	$\hat{q}$	0.057	[0.050, 0.061]	1.83	[1.71, 1.96]	3531.88	3543.77		
		$\hat{\sigma}$	0.005	[0.001, 0.008]					$\sigma = 0$	< 0.001
	MSIRS-ext	$\hat{q}_1$	0.059	[0.026, 0.064]	1.90	[0.89, 2.06]	3533.65	3551.49		
		$\hat{q}_2$	0.000	[0.000, 0.364]					$q_1 = q_2$	0.632
		$\hat{\sigma}$	0.004	[0.001, 0.011]						
FI	MSIR	$\hat{q}$	0.052	[0.045, 0.057]	1.56	[1.52, 1.64]	**3055.50**	**3061.32**		
	MSIRW	$\hat{q}$	0.052	[0.045, 0.057]	1.56	[1.52, 1.65]	3057.50	3069.15		
		$\hat{\varepsilon}$	0.000	[0.000, 0.001]					$\varepsilon = 0$	1.000
	MSIRWb	$\hat{q}$	0.052	[0.045, 0.057]	1.56	[1.52, 1.65]	3057.50	3069.15		
		$\hat{\varepsilon}$	0.000	[0.000, 0.002]					$\varepsilon = 0$	1.000
	MSIRS	$\hat{q}$	0.052	[0.045, 0.057]	1.56	[1.52, 1.65]	3057.50	3069.15		
		$\hat{\sigma}$	0.000	[0.000, 0.002]					$\sigma = 0$	1.000
IT	MSIR	$\hat{q}$	0.025	[0.021, 0.027]	1.68	[1.60, 1.79]	3192.52	3198.35		
	MSIRW	$\hat{q}$	0.027	[0.023, 0.030]	1.86	[1.68, 2.04]	3176.16	3187.82		
		$\hat{\varepsilon}$	0.003	[0.000, 0.005]					$\varepsilon = 0$	< 0.001
	MSIRWb	$\hat{q}$	0.028	[0.023, 0.030]	1.89	[1.69, 2.08]	**3175.12**	**3186.78**		
		$\hat{\varepsilon}$	0.004	[0.001, 0.007]					$\varepsilon = 0$	< 0.001
	MSIRS	$\hat{q}$	0.027	[0.023, 0.029]	1.83	[1.68, 1.99]	3175.96	3187.62		
		$\hat{\sigma}$	0.005	[0.001, 0.008]					$\sigma = 0$	< 0.001
	MSIRS-ext	$\hat{q}_1$	0.028	[0.022, 0.030]	1.89	[1.58, 2.08]	3177.12	3194.61		
		$\hat{q}_2$	0.000	[0.000, 0.118]					$q_1 = q_2$	0.359
		$\hat{\sigma}$	0.004	[0.001, 0.008]						
PL	MSIR	$\hat{q}$	0.047	[0.041, 0.050]	2.16	[1.97, 2.31]	**2785.69**	**2791.51**		
	MSIRW	$\hat{q}$	0.047	[0.041, 0.051]	2.16	[1.97, 2.32]	2787.69	2799.33		
		$\hat{\varepsilon}$	0.000	[0.000, 0.000]					$\varepsilon = 0$	1.000
	MSIRWb	$\hat{q}$	0.047	[0.041, 0.051]	2.16	[1.97, 2.32]	2787.69	2799.33		
		$\hat{\varepsilon}$	0.000	[0.000, 0.001]					$\varepsilon = 0$	1.000
	MSIRS	$\hat{q}$	0.047	[0.041, 0.050]	2.16	[1.97, 2.31]	2787.69	2799.33		
		$\hat{\sigma}$	0.000	[0.000, 0.001]					$\sigma = 0$	1.000

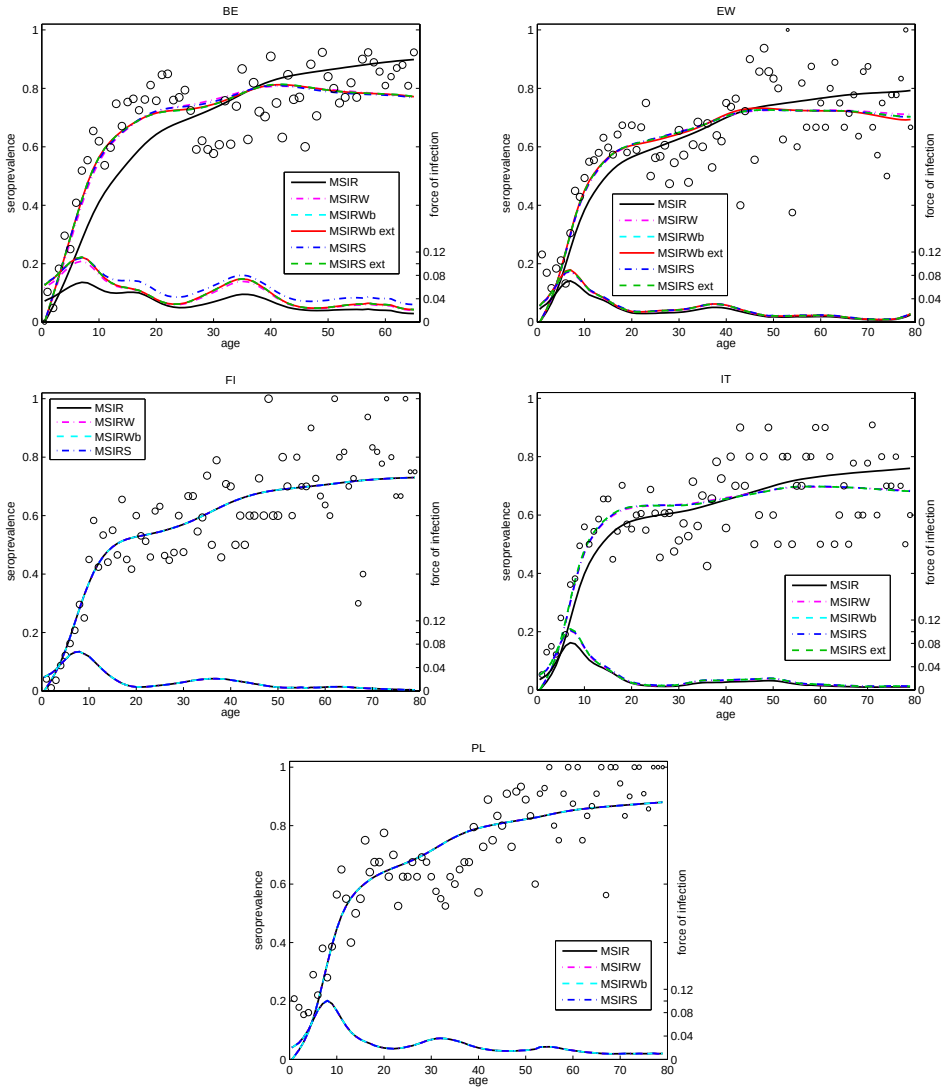

Figure 6.5: Estimated seroprevalence (upper curves) and corresponding force of infection (lower curves) obtained for each compartmental scenario for BE, EW, FI, IT and PL, assuming CW. The dots represent the observed serological data with size proportional to the population age distribution obtained from demographic data.

and the boosting rate is not significantly different from the force of infection.

To test the need of extending the MSIRS model, an LR-test of $H_0 : q_1 = q_2$ is performed for MSIRS-ext. The non-significant p-values together with AIC and BIC values demonstrate the limited impact on the fit to the data. Further, the bootstrap samples mainly give rise to two discrepant solutions: $\hat{q}_1 = 0$ or $\hat{q}_2 = 0$. We believe we cannot identify differences in transmission potential with respect to the immunological status of the infected individual, because the serological data only provide information related to susceptibility. Therefore, MSIRS-ext is not considered further when relaxing the CW assumption. Note that in the latter framework we have calculated the basic reproduction number based on a typical *primary* infected person, such that bootstrap replicates $\hat{q}_1 = 0$ correspond to $\hat{R}_0 = 0$, clarifying the lower CI-limits for BE and EW.

Solutions for the MSIRWS model are obtained using numerical approximation, however, these results are not presented here since they are not directly comparable with those for the other models in Table 6.1, which are obtained from analytical solutions. Nevertheless, for BE and EW, we are able to compare the fit of MSIRWS with MSIRWb-ext and MSIRS by constraining the parameters to the following values: $\sigma = 0$ and $(\varphi, \sigma) = (0, 200)$, respectively. The MSIRWS model is not better according to the BIC criterion and therefore omitted from further consideration.

Considering the best models in terms of AIC/BIC, the following estimates are obtained for the basic reproduction number R_0: 3.35 for BE, 1.90 for EW, 1.56 for FI, 1.89 for IT and 2.16 for PL. The estimated basic reproduction number for PVB19 is similar for EW, IT and PL. R_0 is significantly smaller for Finland and significantly larger for Belgium compared to the other countries, which may indicate an epidemiological difference. A visual inspection of the fit to the data for BE, EW and PL (Figure 6.5) reveals that the scenarios considered up till now, are not able to capture the decrease or plateau observed in the seroprofile for young adults. Therefore, in the next section, we further generalize these scenarios and relax the assumption that the waning rates are independent of age.

6.3.2 Age-specific Waning

We extend the CW models from the previous section to allow for age differences in the immunity transition rates ε and σ. Ten piecewise constant functions are fitted to the data with cut-off points H ranging from 5 to 50 years in 5 years steps. For FI, allowing for age-related heterogeneity in the rate at which antibody levels wane over time, has virtually no effect on the fit to the seroprevalence data and the resulting

parameter estimates. For BE, EW and PL, there is a large improvement in fit and the likelihood values for the four scenarios as functions of H clearly show maxima between the ages 20-50. For IT, the impact on the likelihood is rather limited and the curves for the four scenarios show distinct optimal values for H, ranging from 5 to 35 years. The tendency towards lower cut-off values for IT and EW seems to be driven by ill fitting points in infants, which is confirmed through a sensitivity analysis (cf. Section 6.5). A comparison of the overall likelihood, combined over all countries, for the different values of H and a visual inspection of the fit to the data lead us to the choice of $H = 35$ years.

Table 6.2 presents the ML-estimates and 95% bootstrap-based percentile CIs for the scenario-specific parameters and R_0, assuming a piecewise constant function for ε and σ with a cut-off at $H = 35$ years. Figure 6.6 displays the estimated seroprevalence and force of infection for the age-specific waning (AW) models. The results for Finland are omitted since these are the same as in the CW case (Table 6.1 and Figure 6.5). As described for the CW counterpart in Section 6.3.1, the bootstrap replicates for MSIRWb-ext are problematic for Italy and therefore the results are not considered here. For BE, EW and PL, the AW models perform markedly better than their constant counterparts according to AIC and BIC, with the single exception of the MSIRW scenario for PL. The fitted seroprofiles now clearly display a decrease or plateau in young adults (Figure 6.6). For IT, the AIC values are virtually equal while the CW models have smaller BIC values than the age-specific ones. For BE, EW and IT, the MSIRW and MSIRS scenarios are quite competitive when it comes to model selection and it is difficult to discern whether a dynamics involving waning and boosting of immunity or complete loss of immunity potentially leading to multiple infections, is more plausible for PVB19. The Polish results support the latter scenario in which protection acquired through PVB19 infection in childhood may be lost (≈ 24 years after infection), after which secondary infections with PVB19 could occur up till the age of 35 years.

Consistent for all countries and all scenarios, is the finding that the immunity transition rates ε and σ in individuals above 35 years of age are either estimated to be 2 to 7 times smaller than the corresponding rate in younger individuals, or that the transition from R to W and S for MSIRW and MSIRS respectively, even does not occur in individuals of age 35 years and older, which is the case for Poland. This may reflect the general observation that infection or boosting through exposure to individuals who are infectious with PVB19, elicits higher antibody responses in mature immune systems, which could prolong the process of antibody waning. Further, we obtain the following $\hat{R}_0$ ranges for the AW scenarios: 2.86-3.75 for BE, 1.96-2.19 for

Table 6.2: ML-estimates for the scenario-specific parameters $\boldsymbol{q}, \boldsymbol{\varepsilon}, \boldsymbol{\sigma}, \varphi$, and the basic reproduction number R_0, with 95% bootstrap-based percentile CIs in square brackets, information criteria AIC and BIC (minima indicated in boldface), obtained under the assumption of AW with cut-off $H = 35$ years.

Country	Model				$\hat{R}_0$		AIC	BIC
BE	MSIRW	$\hat{q}$	0.080	[0.063, 0.096]	3.53	[2.94, 4.30]	3359.11	**3377.19**
		$\hat{\varepsilon}_1$	0.007	[0.005, 0.009]				
		$\hat{\varepsilon}_2$	0.000	[0.000, 0.000]				
	MSIRWb	$\hat{q}$	0.084	[0.065, 0.102]	3.70	[3.05, 4.57]	3361.77	3379.86
		$\hat{\varepsilon}_1$	0.019	[0.012, 0.027]				
		$\hat{\varepsilon}_2$	0.005	[0.001, 0.010]				
	MSIRWb-ext	$\hat{q}$	0.085	[0.067, 0.103]	3.75	[3.08, 4.63]	**3353.63**	3377.74
		$\hat{\varepsilon}_1$	0.013	[0.008, 0.020]				
		$\hat{\varepsilon}_2$	0.000	[0.000, 0.005]				
		$\hat{\varphi}$	0.35	[0.05, 0.94]				
	MSIRS	$\hat{q}$	0.065	[0.056, 0.072]	2.86	[2.61, 3.20]	3359.25	3377.34
		$\hat{\sigma}_1$	0.030	[0.018, 0.049]				
		$\hat{\sigma}_2$	0.010	[0.004, 0.018]				
EW	MSIRW	$\hat{q}$	0.064	[0.054, 0.070]	2.05	[1.84, 2.27]	3521.81	3539.65
		$\hat{\varepsilon}_1$	0.008	[0.004, 0.011]				
		$\hat{\varepsilon}_2$	0.000	[0.000, 0.002]				
	MSIRWb	$\hat{q}$	0.068	[0.056, 0.076]	2.18	[1.92, 2.46]	3514.79	3532.63
		$\hat{\varepsilon}_1$	0.017	[0.008, 0.025]				
		$\hat{\varepsilon}_2$	0.003	[0.000, 0.006]				
	MSIRWb-ext	$\hat{q}$	0.068	[0.057, 0.076]	2.19	[1.93, 2.46]	3514.61	3538.39
		$\hat{\varepsilon}_1$	0.026	[0.010, 0.048]				
		$\hat{\varepsilon}_2$	0.007	[0.000, 0.017]				
		$\hat{\varphi}$	2.03	[0.30, 5.59]				
	MSIRS	$\hat{q}$	0.061	[0.053, 0.065]	1.96	[1.82, 2.10]	**3512.43**	**3530.27**
		$\hat{\sigma}_1$	0.021	[0.010, 0.032]				
		$\hat{\sigma}_2$	0.003	[0.000, 0.007]				
IT	MSIRW	$\hat{q}$	0.029	[0.024, 0.031]	1.96	[1.72, 2.17]	3176.10	3193.59
		$\hat{\varepsilon}_1$	0.006	[0.000, 0.009]				
		$\hat{\varepsilon}_2$	0.001	[0.000, 0.005]				
	MSIRWb	$\hat{q}$	0.029	[0.024, 0.032]	1.99	[1.74, 2.24]	**3174.87**	**3192.36**
		$\hat{\varepsilon}_1$	0.008	[0.000, 0.014]				
		$\hat{\varepsilon}_2$	0.004	[0.000, 0.007]				
	MSIRS	$\hat{q}$	0.028	[0.023, 0.030]	1.90	[1.72, 2.08]	3175.53	3193.02
		$\hat{\sigma}_1$	0.010	[0.000, 0.017]				
		$\hat{\sigma}_2$	0.004	[0.000, 0.008]				
PL	MSIRW	$\hat{q}$	0.049	[0.041, 0.054]	2.24	[2.00, 2.49]	2788.15	2805.62
		$\hat{\varepsilon}_1$	0.001	[0.000, 0.004]				
		$\hat{\varepsilon}_2$	0.000	[0.000, 0.000]				
	MSIRWb	$\hat{q}$	0.057	[0.044, 0.068]	2.64	[2.11, 3.16]	2770.01	2787.48
		$\hat{\varepsilon}_1$	0.013	[0.001, 0.022]				
		$\hat{\varepsilon}_2$	0.000	[0.000, 0.000]				
	MSIRWb-ext	$\hat{q}$	0.058	[0.047, 0.066]	2.67	[2.29, 3.04]	2752.17	2775.46
		$\hat{\varepsilon}_1$	0.030	[0.018, 0.048]				
		$\hat{\varepsilon}_2$	0.000	[0.000, 0.001]				
		$\hat{\varphi}$	2.45	[1.59, 5.18]				
	MSIRS	$\hat{q}$	0.053	[0.046, 0.056]	2.44	[2.23, 2.60]	**2740.45**	**2757.91**
		$\hat{\sigma}_1$	0.042	[0.014, 0.082]				
		$\hat{\sigma}_2$	0.000	[0.000, 0.001]				

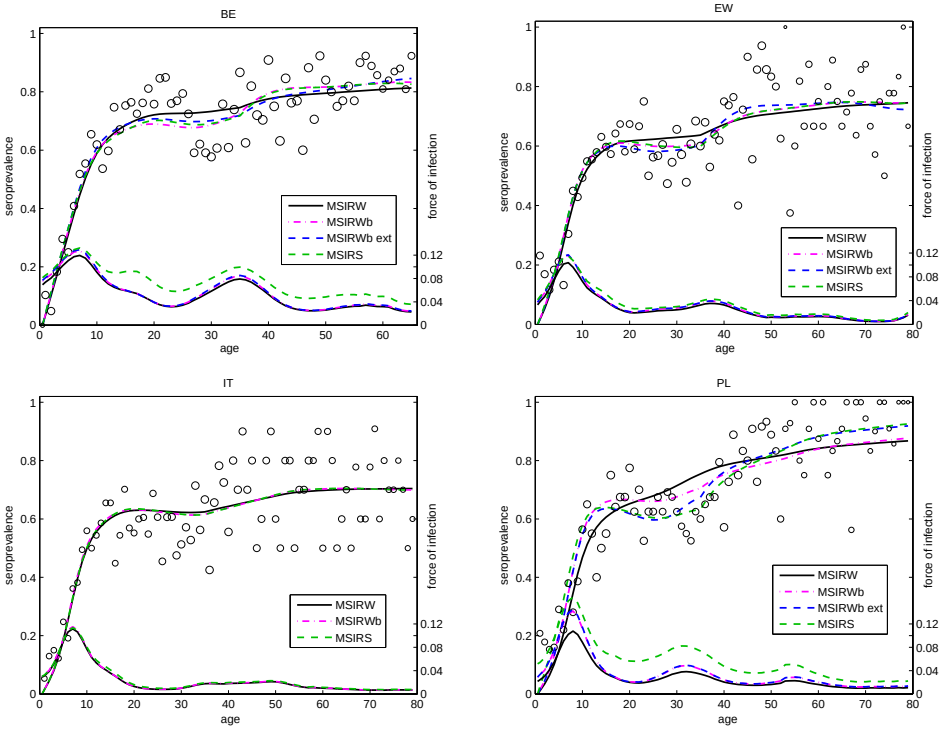

Figure 6.6: Estimated seroprevalence (upper curves) and corresponding force of infection (lower curves) obtained for each compartmental scenario for BE, EW, IT and PL, assuming AW. The dots represent the observed serological data with size proportional to the population age distribution obtained from demographic data.

EW, 1.90-1.99 for IT, and 2.24-2.67 for PL, which are all significantly larger than the basic reproduction number for Finland (Table 6.1).

6.3.3 Risk in Pregnancy

For each of the scenarios considered for each country in the two previous sections, Table 6.3 presents the ML-estimates for $\bar{s}_p$, $\bar{\lambda}_p$, I_p, the frequency of PVB19 infection in pregnancy, and the annual number of fetal deaths (FD) due to PVB19 infection, with corresponding 95% bootstrap-based percentile CIs. Our results for the MSIR model can be compared to the results of Mossong *et al.* (2008a) who analyzed the same serological surveys using local quadratic models based on the assumption of lifelong immunity. With the social contact data approach, we find similar estimates for the average maternal proportion of susceptibles $\bar{s}_p$ to PVB19: 27%, 38%, 43%, 38% and 31%, for BE, EW, FI, IT and PL, respectively. The largest difference is found for Poland, for which Mossong *et al.* (2008a) obtained an estimate of 37%. It should be noted that our MSIR scenario does not provide a good fit to the Polish serology since it is not flexible enough to capture the decrease in young adults. The maternal risk $\bar{\lambda}_p$ of acquiring PVB19 infection when still susceptible is estimated to be 0.034 (BE), 0.018 (EW), 0.014 (FI), 0.010 (IT) and 0.024 (PL), which in case of Belgium is significantly larger than the estimate of 0.006 obtained by Mossong *et al.* (2008a). Also for EW and PL, we estimate a larger maternal force of infection and in summary for BE and EW, we estimate a significantly higher frequency of PVB19 infection in pregnancy compared to Mossong *et al.* (2008a) and Vyse *et al.* (2007).

For the MSIRW scenarios presented in Table 6.3, we notice either no change or a slight decrease in the estimated frequency of PVB19 infection in pregnancy and the induced number of fetal deaths, when broadly comparing them to the MSIR model. In contrast for the MSIRS scenarios, the estimated frequency is much higher for BE, EW, IT and PL (AW) with a significant difference observed for the former and latter country. These two trends continue when comparing the CW models to their age-specific counterparts: for the MSIRW scenarios, allowing for AW induces a decrease in the estimated frequency of PVB19 infection in pregnancy, while it induces an increase for the MSIRS scenarios. The annual number of fetal deaths due to PVB19 infection in pregnancy, estimated from MSIR, MSIRWb AW and MSIRS AW, respectively, equals 31, 23, 77 for BE, 130, 122, 237 for EW, 10 for FI, 61, 61, 91 for IT, and 85, 71, 280 for PL. Our estimates for the average maternal force of infection for PVB19 are in line with the seroconversion rates reported in literature, which are estimated from prospective cohort studies in pregnant women (Alanen *et al.*, 2005; van Gessel

Table 6.3: ML-estimates for the average maternal proportion of susceptibles ($\bar{s}_p$), the average maternal force of infection ($\bar{\lambda}_p$) and the annual number of PVB19 infections (I_p) in pregnant women, the frequency of PVB19 infection in pregnancy (freq), and the annual number of fetal deaths (FD) due to PVB19 infection, together with 95% bootstrap-based percentile CIs in square brackets. First entry: constant waning (CW); second entry (if available): age-specific waning (AW) with cut-off $H = 35$ years.

Country	Model	Waning	$\hat{\bar{s}}_p$		$\hat{\bar{\lambda}}_p$		$\hat{I}_p$		Freq		FD	
BE	MSIR		0.27	[0.23, 0.30]	0.034	[0.028, 0.039]	797	[656, 905]	1 in 143	[126, 174]	31	[25, 35]
	MSIRW	CW	0.17	[0.12, 0.21]	0.046	[0.036, 0.057]	677	[552, 777]	1 in 168	[147, 207]	26	[21, 30]
		AW	0.14	[0.10, 0.17]	0.051	[0.040, 0.062]	622	[509, 721]	1 in 183	[158, 224]	24	[20, 28]
	MSIRWb	CW	0.15	[0.11, 0.19]	0.048	[0.038, 0.060]	652	[526, 760]	1 in 175	[150, 217]	25	[20, 29]
		AW	0.13	[0.09, 0.16]	0.054	[0.042, 0.066]	595	[479, 701]	1 in 192	[163, 238]	23	[18, 27]
	MSIRWb-ext	CW	0.15	[0.11, 0.20]	0.048	[0.038, 0.060]	653	[525, 762]	1 in 175	[150, 217]	25	[20, 29]
		AW	0.12	[0.09, 0.16]	0.054	[0.042, 0.067]	587	[473, 684]	1 in 194	[167, 241]	23	[18, 26]
	MSIRS	CW	0.24	[0.21, 0.27]	0.059	[0.042, 0.082]	1256	[915, 1703]	1 in 91	[67, 125]	48	[35, 66]
		AW	0.29	[0.26, 0.33]	0.077	[0.057, 0.107]	1990	[1431, 2825]	1 in 57	[40, 80]	77	[55, 109]
	MSIRS-ext	CW	0.25	[0.21, 0.27]	0.050	[0.038, 0.145]	1091	[839, 2889]	1 in 105	[39, 136]	42	[32, 111]
EW	MSIR		0.38	[0.35, 0.41]	0.018	[0.015, 0.020]	3373	[2874, 3659]	1 in 192	[177, 226]	130	[111, 141]
	MSIRW	CW	0.32	[0.28, 0.36]	0.021	[0.017, 0.024]	3365	[2867, 3655]	1 in 193	[178, 226]	130	[110, 141]
		AW	0.27	[0.23, 0.32]	0.025	[0.019, 0.028]	3277	[2814, 3580]	1 in 198	[181, 231]	126	[108, 138]
	MSIRWb	CW	0.31	[0.27, 0.36]	0.022	[0.017, 0.025]	3358	[2861, 3649]	1 in 193	[178, 227]	129	[110, 140]
		AW	0.23	[0.19, 0.29]	0.027	[0.020, 0.032]	3181	[2743, 3499]	1 in 204	[186, 237]	122	[106, 135]
	MSIRWb-ext	CW	0.31	[0.27, 0.36]	0.022	[0.017, 0.025]	3354	[2865, 3647]	1 in 194	[178, 227]	129	[110, 140]
		AW	0.23	[0.19, 0.28]	0.028	[0.021, 0.032]	3172	[2741, 3488]	1 in 205	[186, 237]	122	[106, 134]
	MSIRS	CW	0.35	[0.33, 0.38]	0.022	[0.017, 0.026]	3919	[3113, 4469]	1 in 166	[145, 208]	151	[120, 172]
		AW	0.39	[0.36, 0.43]	0.031	[0.023, 0.038]	6156	[4187, 7603]	1 in 105	[85, 155]	237	[161, 293]
	MSIRS-ext	CW	0.36	[0.33, 0.38]	0.022	[0.017, 0.031]	3845	[3140, 5333]	1 in 169	[122, 207]	148	[121, 205]

Table 6.3: (continued) ML-estimates for the average maternal proportion of susceptibles ($\bar{s}_p$), the average maternal force of infection ($\bar{\lambda}_p$) and the annual number of PVB19 infections (I_p) in pregnant women, the frequency of PVB19 infection in pregnancy (freq), and the annual number of fetal deaths (FD) due to PVB19 infection, together with 95% bootstrap-based percentile CIs in square brackets. First entry: constant waning (CW); second entry (if available): age-specific waning (AW) with cut-off $H = 35$ years.

Country	Model	Waning	$\hat{\bar{s}}_p$		$\hat{\bar{\lambda}}_p$		$\hat{I}_p$		Freq		FD	
FI	MSIR		0.43	[0.39, 0.45]	0.014	[0.011, 0.016]	260	[200, 289]	1 in 220	[197, 285]	10	[8, 11]
	MSIRW(b)	CW	0.43	[0.39, 0.45]	0.014	[0.011, 0.016]	260	[200, 290]	1 in 220	[197, 285]	10	[8, 11]
	MSIRS	CW	0.43	[0.39, 0.45]	0.014	[0.011, 0.016]	260	[201, 294]	1 in 220	[194, 284]	10	[8, 11]
IT	MSIR		0.38	[0.36, 0.41]	0.010	[0.007, 0.011]	1590	[1277, 1764]	1 in 354	[319, 440]	61	[49, 68]
	MSIRW	CW	0.32	[0.28, 0.37]	0.012	[0.009, 0.014]	1594	[1282, 1782]	1 in 353	[316, 439]	61	[49, 69]
		AW	0.29	[0.25, 0.36]	0.013	[0.009, 0.014]	1583	[1278, 1778]	1 in 355	[316, 440]	61	[49, 68]
	MSIRWb	CW	0.31	[0.27, 0.37]	0.012	[0.009, 0.014]	1591	[1282, 1781]	1 in 354	[316, 439]	61	[49, 69]
		AW	0.28	[0.24, 0.36]	0.013	[0.009, 0.015]	1577	[1269, 1776]	1 in 357	[317, 443]	61	[49, 68]
	MSIRS	CW	0.36	[0.34, 0.39]	0.013	[0.009, 0.015]	1957	[1452, 2294]	1 in 288	[245, 387]	75	[56, 88]
		AW	0.38	[0.34, 0.42]	0.014	[0.010, 0.018]	2356	[1509, 3019]	1 in 238	[186, 373]	91	[58, 116]
	MSIRS-ext	CW	0.36	[0.34, 0.39]	0.012	[0.009, 0.015]	1888	[1443, 2387]	1 in 298	[236, 390]	73	[56, 92]
PL	MSIR		0.31	[0.28, 0.33]	0.024	[0.019, 0.028]	2208	[1753, 2577]	1 in 173	[148, 218]	85	[67, 99]
	MSIRW(b)	CW	0.31	[0.28, 0.33]	0.024	[0.019, 0.028]	2208	[1753, 2569]	1 in 173	[149, 218]	85	[67, 99]
	MSIRW	AW	0.29	[0.25, 0.32]	0.025	[0.019, 0.030]	2147	[1726, 2469]	1 in 178	[155, 221]	83	[66, 95]
	MSIRWb	AW	0.21	[0.16, 0.29]	0.030	[0.021, 0.037]	1854	[1470, 2225]	1 in 206	[172, 260]	71	[57, 86]
	MSIRWb-ext	AW	0.21	[0.17, 0.26]	0.030	[0.023, 0.036]	1832	[1477, 2192]	1 in 208	[174, 259]	71	[57, 84]
	MSIRS	CW	0.31	[0.28, 0.33]	0.024	[0.019, 0.028]	2208	[1753, 2586]	1 in 173	[148, 218]	85	[67, 100]
		AW	0.38	[0.32, 0.42]	0.065	[0.036, 0.100]	7277	[3386, 12305]	1 in 52	[31, 113]	280	[130, 474]

et al., 2006; Valeur-Jensen *et al.*, 1999).

6.3.4 Age-Dependent Proportionality

We assess the sensitivity of the results from our model structure analysis for PVB19 with respect to the CP assumption of the transmission rates, hereby restricting attention to MSIR, MSIRW(b) and MSIRS to ensure estimability. We choose the same dichotomy of the population, namely with a cut-off point at age $G = 12$ years, which performed well in our application of the MSIR model to VZV serology described in Section 5.3. By parameterizing $q(a, a')$ according to the matrix structures $\boldsymbol{M_1}, \boldsymbol{M_2}, \boldsymbol{M_3}$ as defined in (5.8), the evidence of waning immunity arising from the CW models for BE, EW and IT, is almost completely absorbed, which is expressed by the very small estimates for the waning rates. For these countries, under the assumption of lifelong immunity, the AP model is always selected according to AIC/BIC. For Belgium, the AP constant waning models fit the seroprofile much better than the CP models and the estimates for R_0 vary around the estimates obtained previously (Table 6.1), with a pronounced dependence on the configuration type $\boldsymbol{M}$ which is similar to what we observed for VZV (Section 5.3 and Figure 5.6). When making pairwise comparisons of the CP versus AP constant waning models for EW, the AIC values are always in favor of the AP scenarios, while the selection based on BIC depends on the waning scenario and the parametric model considered for $q(a, a')$. For IT, however, the BIC values always select the CP models over their AP counterpart. For BE, EW and IT, the force of infection is now estimated to be smaller in adults which reduces the estimated maternal frequency of PVB19 infection (Table 6.3). Finally for FI and PL, allowing q to be age-dependent, does not substantially affect the fit of the CW scenarios to the serological data and nearly preserves the estimates obtained previously (Tables 6.1 and 6.3). The CP-models are better in terms of BIC and AIC, the latter with the exception of $\boldsymbol{M_2}$ for Poland.

For the AW models, however, the evidence in favor of waning immunity is sustained for BE, EW and PL, under the AP assumption for the transmission rates. Furthermore, the ranking of the different waning scenarios according to AIC/BIC remains approximately the same for each country compared to the results in Table 6.2. Under AP, the estimates for the waning rates $\boldsymbol{\varepsilon}, \boldsymbol{\sigma}$ slightly decrease for BE and slightly increase for EW and FI, while for IT and PL these fluctuate around the estimates obtained before depending on the parametric structure considered for $q(a, a')$. Further, the estimates for R_0 are generally close to the estimates obtained before (Table 6.2), though we observe somewhat larger deviations in case of $\boldsymbol{M_1}$ and $\boldsymbol{M_2}$ for the MSIRWb

scenario for BE, EW and PL. For these three countries, information criteria based pairwise selection of the CP versus AP counterparts differs depending on the waning scenario and the configuration type $\boldsymbol{M}$ considered, but overall the smallest AIC/BIC values are obtained for the AP MSIRS scenarios based on $\boldsymbol{M_3}$ for BE and $\boldsymbol{M_1}$ for EW and PL. A visual inspection of the fit to the serological data shows that this model more pronouncedly captures the shoulder effect in teenagers and 20 year olds for BE and EW, and that the fit to the initial prevalence rise in children is improved for PL, compared to the scenarios depicted in Figure 6.6. For FI, the MSIR CP model is still the best one according to information criteria based selection and for IT, the CP models are again selected over their AP counterpart. The frequency of PVB19 infection in pregnancy is now estimated to be lower in BE, slightly higher for FI, and for EW, IT and PL it fluctuates around the estimates displayed in Table 6.3 depending on the AP matrix for $q(a, a')$. For all countries, the annual number of PVB19-induced fetal deaths estimated from the MSIRS scenario seems to be the most sensitive with respect to the proportionality assumption.

6.4 Simulation Study

We conduct a simulation study to assess the performance of the different mathematical scenarios and the ability of the model selection criteria AIC and BIC to select the true underlying disease dynamics. Without loss of generality, we simulate $n_s = 200$ serological data sets of size $n = 3075$, taking the ages of the individuals from the Belgian seroprevalence data as a basis (cf. Table 3.1). The binary responses are simulated by considering each one of the ten scenarios studied in Section 6.3 as the 'true' model, and by using the ML-estimates for Belgium as parameter values. All compartmental models are then fitted to the n_s simulated serological data sets. For each model, the bias of the estimator $\bar{\hat{q}} = \frac{1}{n_s}\sum_{i=1}^{n_s}\hat{q}_i$, for the proportionality factor q is calculated as follows: $\text{bias}(\bar{\hat{q}}) = \bar{\hat{q}} - q$. The mean squared error (MSE) is computed as:

$$\text{MSE}\left(\bar{\hat{q}}\right) = \text{bias}^2(\bar{\hat{q}}) + \widehat{\text{Var}}(\bar{\hat{q}}), \quad \text{with} \quad \widehat{\text{Var}}(\bar{\hat{q}}) = \frac{1}{n_s - 1}\sum_{i=1}^{n_s}(\hat{q}_i - \bar{\hat{q}})^2,$$

where the latter formula is the estimated sample variance of $\bar{\hat{q}}$. Further, we calculate the same figures for a few other 'global' parameters: the basic reproduction number R_0, the average maternal proportion of susceptibles $\bar{s}_p$, the average maternal force of infection $\bar{\lambda}_p$, and the annual number of fetal deaths (FD). Note that for the MSIRS-ext model, we use the proportionality factor q_1 for infected individuals with a primary infection, as a surrogate for q.

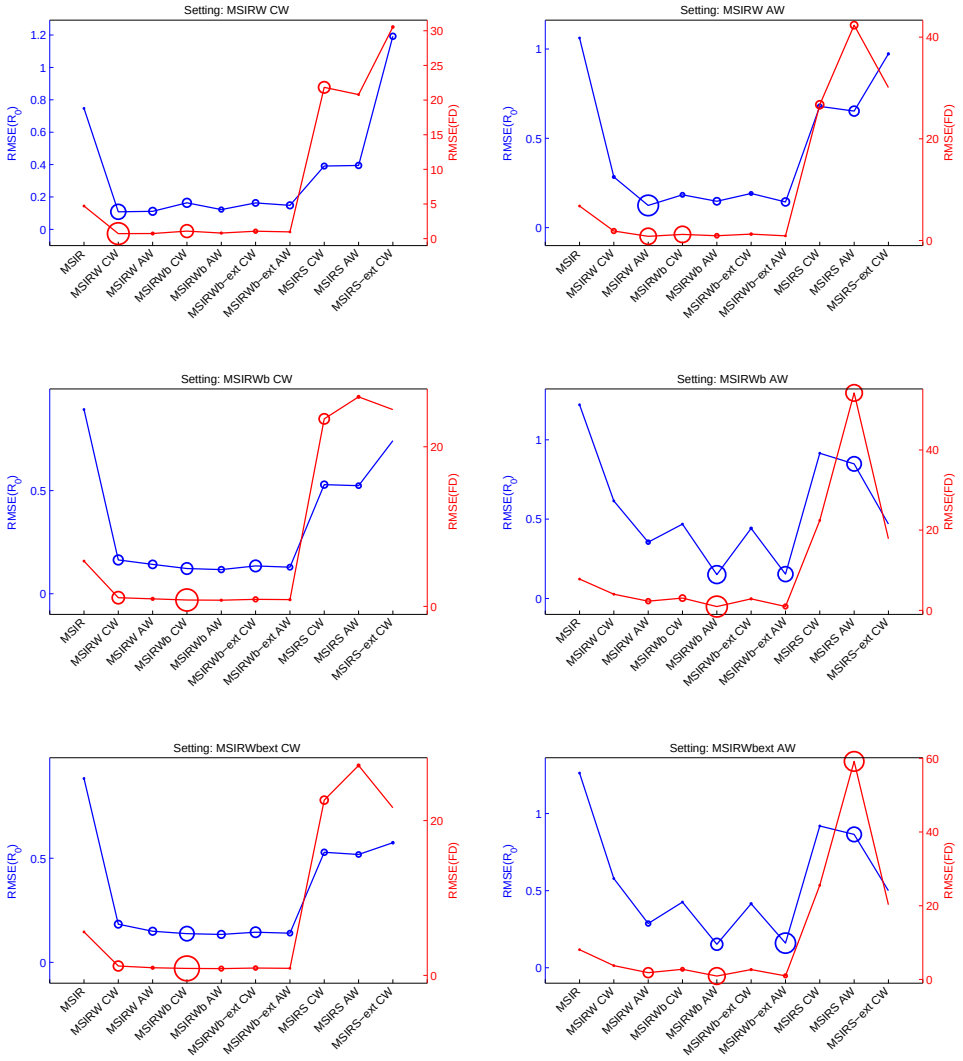

Figure 6.7: The root mean squared error (RMSE) of $\hat{\bar{R}}_0$ and $\widehat{\overline{\text{FD}}}$, where FD denotes the annual number of fetal deaths, obtained for each scenario. The sizes of the dots are proportional to the AIC (blue) and BIC (red) model selection probabilities. Each panel corresponds to a fixed simulation setting.

Tables D.1-D.10 in Appendix D, present the simulation results in full, while Figures 6.7 and 6.8 provide a summary of the main findings. Each panel corresponds to a specific simulation setting and the dots represent the square root of MSE, the so-called root mean squared error (RMSE), of $\bar{\hat{R}}_0$ (left Y-axis) and $\overline{\widehat{\text{FD}}}$ (right Y-axis) for each mathematical scenario. The RMSE values have the same unit as the quantity being estimated. The sizes of the dots are proportional to the AIC (blue) and BIC (red) model selection percentages: $\pi_{\text{sel,AIC}}$ and $\pi_{\text{sel,BIC}}$, respectively. The results when simulating serology under the lifelong immunity hypothesis (MSIR), are presented in Table D.1. As expected, all mathematical models perform well and have low MSE values, while AIC and BIC are well able to detect MSIR as the true underlying scenario. It is less interesting to depict these results as we did for the MSIRW models in Figure 6.7. The panels on the left side of this figure reveal that, under a CW setting, the MSIRW scenarios entail low RMSE values, whereas the MSIR and MSIRS models produce higher RMSEs due to fairly large biases (see Tables D.2, D.4, D.6). Under AW (Figure 6.7 on the right), we see a similar pattern, though the MSIRW AW scenarios now clearly outperform the CW counterparts in terms of RMSE.

Figure 6.8 displays the RMSE values in case the simulation setting is of the MSIRS type. For MSIRS CW, the MSIRS CW/AW models entail low RMSE values, whereas for MSIRS AW only the true model performs well in general. Below in Figure 6.8, the results for the MSIRS-ext setting are depicted. As opposed to the previous, the peculiar result arises that the RMSE value of $\bar{\hat{R}}_0$ for the true underlying dynamical model is rather large compared to the other scenarios. Table D.10 shows that this is due to a large estimated variance for $\bar{\hat{q}}$, and consequently for $\bar{\hat{R}}_0$ as well, when fitting MSIRS-ext to the simulated data sets.

In most settings, the BIC model selection probability for the true model is larger than the corresponding AIC model selection probability. Nevertheless, it seems important to calculate and compare both criteria in practice, since for MSIRW AW and the MSIRWb-ext models, the BIC selection percentage for the actual underlying dynamics is lower than the AIC selection percentage. Furthermore, the simulation study reveals that the model selection criteria have difficulties to identify certain waning scenarios: MSIRS CW and the MSIRWb-ext and MSIRS-ext models. For the latter settings, the information criteria tend to select more parsimonious models.

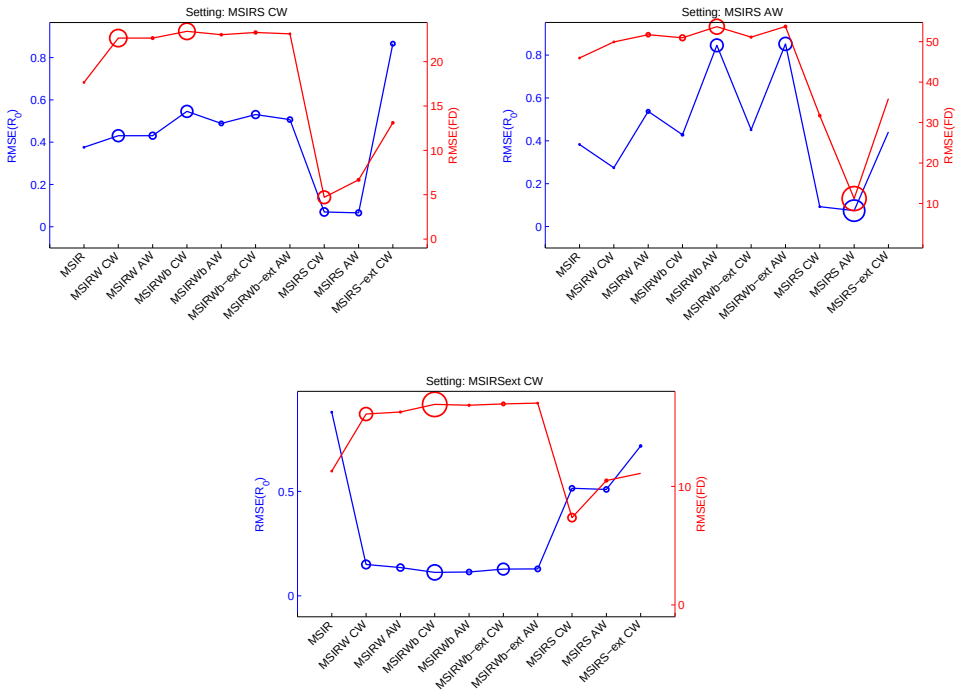

Figure 6.8: The root mean squared error (RMSE) of $\bar{\hat{R}}_0$ and $\overline{\widehat{\mathrm{FD}}}$, where FD denotes the annual number of fetal deaths, obtained for each scenario. The sizes of the dots are proportional to the AIC (blue) and BIC (red) model selection probabilities. Each panel corresponds to a fixed simulation setting.

6.5 Concluding Remarks

The results in Sections 6.3.1 and 6.3.2 for BE, EW and PL, indicate substantial evidence towards processes of age-specific waning immunity for PVB19. Furthermore, this finding is preserved when we relax the constant proportionality assumption of the transmission rates (Section 6.3.4). Figure 6.6 shows that the age-specific MSIRWb and MSIRS scenarios are able to explain at least partly the observed decrease in the seroprofile for adults. The waning rates $\boldsymbol{\varepsilon}$ and $\boldsymbol{\sigma}$ are consistently estimated to be smaller in individuals above 35 years of age, which may reflect a stronger antibody response in more mature immune systems when exposed to PVB19, prolonging the subsequent waning process of IgG antibodies. It is however difficult to discern from the data whether a scenario involving waning and boosting of low immunity or a scenario allowing for reinfections, is more plausible for PVB19. The simulation study presented in Section 6.4 and Appendix D, illustrates that this finding may also hold in general when inferring on disease dynamics using serological data. Elucidating the underlying immunological process for PVB19 is nevertheless important with respect to maternal-fetal risk assessment as we have shown in Section 6.3.3, in which the MSIRS scenarios predict a higher risk of PVB19 infection in pregnancy and a larger associated number of fetal deaths.

For IT, the evidence against lifelong immunity for PVB19 is merely sustained under the assumption of constant proportionality and is less pronounced than for BE, EW and PL. From the Finnish serological data we cannot infer any evidence of waning immunity for PVB19, which relates to the shape of the seroprofile. The Finnish seroprofile plateaus between the ages of 20 and 40 years and does not display a decrease as for the other countries. For both FI and IT we obtain smaller estimates for the basic reproduction number R_0 and it could be hypothesized that the reduced potential of spread for PVB19 in these countries makes it more difficult to observe long-term waning processes at the population level. There is a limit to what can be inferred from serological surveys and we have reached the boundary of what is estimable by considering models such as MSIRWb-ext, MSIRS-ext and MSIRWS. In Appendix B, we provide, in addition to the results here, estimates for the average number of transitions from one stage to the other per person during their lifetime and the average age at which these transitions occur.

In our model structure analysis, we have assumed endemic equilibrium for PVB19 which means that disease incidence fluctuates around a stationary average over time. The few reports in the literature suggest that PVB19 has 3-5 year epidemic cycles in European countries with a seasonal peak in the first half of each year (Bosman *et al.*,

2002; Riipinen *et al.*, 2008; Vyse *et al.*, 2007), comparable to rubella. Using auxiliary data on case reports, Whitaker and Farrington (2004a) show that cyclic epidemics have only a marginal effect on estimates obtained under endemic equilibrium from serological surveys for immunizing infections with short latent and infectious periods. Whether these findings can be extended towards non-immunizing infections has not been investigated yet and is beyond the scope of this thesis.

It was noted that the serological data reveal a rather high proportion of seropositive 1-year old infants (Mossong *et al.*, 2008a), which decreases until the second or third year of life and then starts to increase gradually, except for IT where a similar pattern is detected from three years of age. One would however expect that the proportion of seropositive infants immediately starts to build up after the loss of maternal antibodies. Mossong *et al.* (2008a) suggest that this could be due to a lack of assay specificity for these age groups exposed to many other viral agents. On the other hand, cyclic PVB19 epidemics in relation to the timing of the data collection could perhaps also explain these observations. The proportion of seropositive neonates born in the period after an epidemic will be lower than expected whereas the number of congenital infections during an epidemic, and thus the proportion of seropositive newborns, will be larger. Yet, we are not able to verify this hypothesis due to lack of data on the epidemic patterns for the countries involved in this study.

Given these seropositivity 'deviations' in infants, we performed a sensitivity analysis by omitting the serological samples for infants aged 0.5-3 years and re-fitting all models. The same scenarios and cut-off points for the AW models are selected according to AIC and BIC, except for EW and IT where in case of the MSIRW scenarios the cut-off point is not anymore selected at young ages. Overall, the ML-estimates of the model parameters, R_0 and the risk in pregnancy are approximately the same. Only for the MSIRS scenario in EW, IT and PL we observe a slight decrease in $\hat{q}$ and $\hat{\boldsymbol{\sigma}}$, inducing smaller estimates for the number of fetal deaths.

There is a need for additional large prospective cohort studies in pregnant women in order to obtain more precise estimates of the risk of fetal death and hydrops fetalis due to PVB19 infection. In Miller *et al.* (1998) and Enders *et al.* (2004), only pregnant women who were reported because they had rash, arthropathy or other symptoms, and/or contact with a suspected case of erythema infectiosum, were included in the analysis (at the point when maternal PVB19 infection was serologically confirmed). This selection, with a reduced probability of asymptomatic PVB19 cases to be reported, may compromise the generalization of the estimated risk to the entire population of pregnant women.

Chapter 7

Estimating Measles-Mumps-Rubella Vaccination Coverage from Trivariate Current Status Data

Universal vaccination of infants is an important tool to control or even eliminate vaccine preventable infectious diseases that are potential causes of illness and death. In order for vaccination strategies to be effective, it is crucial to achieve and maintain a high vaccination coverage. Although trivalent measles, mumps and rubella (MMR) immunization has been widely implemented in the whole of Europe, still several small-scale measles epidemics occur, as was recently the case in Northern Ireland (Smithson *et al.*, 2010). The suboptimal vaccination coverage raises serious doubts that the World Health Organization's stated goal of eliminating measles from Europe by 2010 can be attained (Muscat *et al.*, 2009). In this chapter, we explore methods to estimate MMR vaccination coverage from trivariate current status data using the serological data sets from Belgium and Ireland, which were introduced in Section 3.1.4.

Up till now in Belgium, age-cohort-specific MMR vaccination coverage has been estimated from surveys based on the expanded program on immunization (EPI) cluster sampling technique (see e.g. Theeten *et al.*, 2007). Vaccine uptake is documented retrospectively by conducting face-to-face interviews with the selected families at home. These EPI-surveys, however, suffer from incomplete documentation of vaccinations after infancy and possible participation bias, which may induce biased estimates of the vaccination coverage. In Flanders, a web-based system called 'Vaccinnet' for the

ordering and registration of vaccines has been recently developed, but it is still to be assessed whether this system is able to tackle the aforementioned issues. In Ireland, each Health Service Executive Area maintains a childhood immunization register in which uptake statistics are compiled from vaccine return forms completed by general practitioners. The database is limited to the uptake of the first MMR dose measured at 2 years of age only, and does not allow for linkage between the different areas to ensure uniqueness (O'Flanagan and others as the 'Measles and Rubella Elimination Committee of the Department of Health and Children', 2007).

As an alternative to these methods relying on parental recall or registered evidence, Gay (2000) proposed to exploit the information contained in multivariate antibody prevalence data obtained from large national serum banks. Gay (2000) argued that the higher the multivalent vaccination coverage, the greater the extent to which seropositivity to each infection within individuals coincides. From the age-specific data on serological status with respect to MMR, Gay (2000) obtained ML-estimates for the vaccination coverage, the seroconversion rates and the proportions of unvaccinated who are seropositive as a result of natural infection ('exposure probabilities'). From Gay (2000)'s modelling equations, Altmann and Altmann (2000) derived exact solutions for the various parameters, using computer aided elimination theory of variables. These exact formulas, however, do not take into account the uncertainty originating from the data nor the interdependency between the different age cohorts, and may yield biologically implausible solutions. Using the Belgian data set, we illustrate the methods of Gay (2000) and Altmann and Altmann (2000) in Section 7.1.

In Section 7.2, we elaborate on the estimation method presented by Gay (2000) by taking into account the dependency in acquisition of measles, mumps and rubella, which are transmitted via the same route. In a likelihood-based marginal model framework, we use the Bahadur model to describe the association between the exposure probabilities. Further, instead of considering a saturated, age-specific structure for the exposure probabilities, we propose a semiparametric approach with restricted cubic splines to model the probability of acquiring natural infection as a smooth function of age. Allowing for dependence between the exposure probabilities, has a clear effect on the MMR vaccination coverage estimates and the corresponding uncertainty, as is shown for the Belgian and Irish data sets. Further, the restricted cubic splines allow for a more parsimonious model, which is flexible enough to capture the main trends of the exposure probability profiles. An extensive discussion on the results and some further research prospects are provided in Section 7.3.

7.1 Existing Methods

From now on, the index $d = 1, 2, 3$ refers to the three diseases: measles, mumps and rubella, respectively, and the index j refers to the right-open age interval $[j, j + 1)$ where $j = 1, \ldots, m$. We use the same parameter definitions as introduced by Gay (2000). The main parameter of interest is the vaccination coverage ν_j, i.e. the proportion of individuals aged j years who have received the trivalent MMR vaccine. Here, the implicit assumption is made that each individual receives no more than one dose of the vaccine. Thus, the effect of the second dose of MMR, which mainly influences the older age cohorts, is ignored. Further, let ζ_{jd} denote the 'seroconversion rate', i.e. the proportion of individuals of age j who acquire detectable antibodies against disease d when being vaccinated, and let η_{jd} denote the proportion of unvaccinated individuals aged j years who are seropositive for disease d as a result of naturally acquired infection ('exposure probability').

7.1.1 Gay's Estimation Approach

Gay (2000) developed a model to estimate the trivalent vaccination coverage from trivariate serology under the following assumptions:

1. Vaccinated individuals who did not seroconvert as a result of vaccination, have the same probability of being seropositive as an unvaccinated individual of the same age (i.e. η_{jd});

2. Within an individual, seroconversion to each vaccine component is independent;

3. The risk of exposure to infection is homogeneous within each age cohort and infection with each disease is independent.

Using the first assumption, the probability for a vaccinated individual in age class $[j, j + 1)$ of being seropositive for disease d equals:

$$\pi_{jd} = \zeta_{jd} + (1 - \zeta_{jd})\eta_{jd}.$$

Based on the trivariate binary response variable $\boldsymbol{Y}_{ij} = (Y_{ij1}, Y_{ij2}, Y_{ij3})$ indicating whether or not individual i 'nested' in age class j tested seropositive for measles, mumps and rubella (1 = seropositive, 0 = seronegative), he/she can be categorized into one of eight different immunity states $(\pm, \pm, \pm)$, where $+$ indicates seropositive and $-$ seronegative. The probability that a person of age j is classified into category k is denoted by p_{jk} $(k = 1, \ldots, 8)$. Using the second and third assumption, Gay

(2000) determined the classification probabilities for each age group, which in case of $[j, j+1)$ led to:

$$\begin{aligned}
p_{j1} = f_j(+,+,+) = &\quad \nu_j\pi_{j1}\pi_{j2}\pi_{j3} &&+ (1-\nu_j)\eta_{j1}\eta_{j2}\eta_{j3}\\
p_{j2} = f_j(+,+,-) = &\quad \nu_j\pi_{j1}\pi_{j2}(1-\pi_{j3}) &&+ (1-\nu_j)\eta_{j1}\eta_{j2}(1-\eta_{j3})\\
p_{j3} = f_j(+,-,+) = &\quad \nu_j\pi_{j1}(1-\pi_{j2})\pi_{j3} &&+ (1-\nu_j)\eta_{j1}(1-\eta_{j2})\eta_{j3}\\
p_{j4} = f_j(+,-,-) = &\quad \nu_j\pi_{j1}(1-\pi_{j2})(1-\pi_{j3}) &&+ (1-\nu_j)\eta_{j1}(1-\eta_{j2})(1-\eta_{j3})\\
p_{j5} = f_j(-,+,+) = &\quad \nu_j(1-\pi_{j1})\pi_{j2}\pi_{j3} &&+ (1-\nu_j)(1-\eta_{j1})\eta_{j2}\eta_{j3}\\
p_{j6} = f_j(-,+,-) = &\quad \nu_j(1-\pi_{j1})\pi_{j2}(1-\pi_{j3}) &&+ (1-\nu_j)(1-\eta_{j1})\eta_{j2}(1-\eta_{j3})\\
p_{j7} = f_j(-,-,+) = &\quad \nu_j(1-\pi_{j1})(1-\pi_{j2})\pi_{j3} &&+ (1-\nu_j)(1-\eta_{j1})(1-\eta_{j2})\eta_{j3}\\
p_{j8} = f_j(-,-,-) = &\quad \nu_j(1-\pi_{j1})(1-\pi_{j2})(1-\pi_{j3}) + &&(1-\nu_j)(1-\eta_{j1})(1-\eta_{j2})(1-\eta_{j3}),
\end{aligned} \tag{7.1}$$

where $f_j(\boldsymbol{y}_{ij})$ indicates the joint probability density function of $\boldsymbol{Y}_{ij}$.

Gay (2000) assumed that the disease-specific seroconversion rates are independent of age (ζ_d), which is plausible from a biological point of view, unless there is a cohort effect due to changing vaccine compounds. As a consequence, the sets of equations (7.1) for all age classes are interlinked because of the mutual parameters $\zeta_1, \zeta_2, \zeta_3$. Let n_{jk} denote the number of individuals of age j classified into category k. Gay (2000) maximized the likelihood for the observed n_{jk} to estimate the $4m+3$ unknown parameters: $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_m), \boldsymbol{\zeta} = (\zeta_1, \zeta_2, \zeta_3), \boldsymbol{\eta}_1 = (\eta_{11}, \ldots, \eta_{m1})$ and $\boldsymbol{\eta}_2, \boldsymbol{\eta}_3$ defined analogously. During the optimization process, Gay (2000) put some additional constraints in order to obtain biologically relevant estimates: $0 \leq \nu_j \leq 1$, $0 \leq \zeta_d \leq 1$ and $0 \leq \eta_{jd} \leq \eta_{j+1,d} \leq 1$. Note that the latter constraint implies monotonicity of the exposure probabilities with respect to age. Alternatively, we use link functions to transform the probabilities to the real number scale:

$$\text{logit}(\nu_j) = \log\left(\frac{\nu_j}{1-\nu_j}\right) = \alpha_j, \quad \text{logit}(\zeta_d) = \delta_d, \quad \text{logit}(\eta_{jd}) = \beta_{jd}, \tag{7.2}$$

and we do not require the exposure probabilities to monotonically increase over age. We come back to the motivation for the latter in Section 7.2.2. Since the response variables $(n_{j1}, \ldots, n_{j8})$ follow a multinomial distribution with $n_j = \sum_k n_{jk}$ the total number of individuals aged j years, and p_{jk} the event probabilities, the loglikelihood is given by

$$\ell(\boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \boldsymbol{\beta_3} | n_{11}, \ldots, n_{18}, n_{21}, \ldots, n_{m8}) = \sum_{j=1}^{m}\sum_{k=1}^{8} n_{jk}\log(p_{jk}) + C, \tag{7.3}$$

where C is a constant.

7.1.2 Exact Solutions by Altmann & Altmann

Using computer aided elimination theory of variables, Altmann and Altmann (2000) derived exact formulas for ν_j, ζ_{jd} and η_{jd} from Gay's set of equations (7.1). The theory of Altmann and Altmann (2000) only works if Gay's assumption that the seroconversion rates are independent of age, is eliminated. If the so-called primary feasibility conditions, $g_{4j} > 0$ and $g_{2jd} > 0$ ($d = 1, 2, 3$), hold, there is a unique, feasible solution for the vaccination coverage ν_j:

$$\nu_j = \frac{g_{1j}\sqrt{g_{4j}} + g_{3j}}{2g_{1j}\sqrt{g_{4j}}},$$

where $g_{1j} = n_j$,

$$\begin{aligned} g_{2j1} &= (n_{j8} + n_{j4})(n_{j1} + n_{j5}) - (n_{j7} + n_{j3})(n_{j2} + n_{j6}) \\ g_{2j2} &= (n_{j8} + n_{j6})(n_{j1} + n_{j3}) - (n_{j4} + n_{j2})(n_{j5} + n_{j7}) \\ g_{2j3} &= (n_{j8} + n_{j7})(n_{j1} + n_{j2}) - (n_{j6} + n_{j5})(n_{j3} + n_{j4}), \end{aligned}$$

and

$$\begin{aligned} g_{3j} = {} & \{(n_{j8} + n_{j5} + n_{j3} + n_{j2}) - (n_{j1} + n_{j4} + n_{j6} + n_{j7})\}(n_{j8}n_{j1} + n_{j5}n_{j4} \\ & + n_{j3}n_{j6} + n_{j2}n_{j7}) - 2(n_{j8}n_{j1}(n_{j8} - n_{j1}) + n_{j5}n_{j4}(n_{j5} - n_{j4}) + n_{j3}n_{j6} \\ & \cdot(n_{j3} - n_{j6}) + n_{j2}n_{j7}(n_{j2} - n_{j7})) + 2\{(n_{j5}n_{j3}n_{j2} + n_{j8}n_{j3}n_{j2} + n_{j8}n_{j5}n_{j2} \\ & + n_{j8}n_{j5}n_{j3}) - (n_{j4}n_{j6}n_{j7} + n_{j1}n_{j6}n_{j7} + n_{j1}n_{j4}n_{j7} + n_{j1}n_{j4}n_{j6})\} \end{aligned}$$

$$\begin{aligned} g_{4j} = {} & (n_{j8}^2 n_{j1}^2 + n_{j5}^2 n_{j4}^2 + n_{j3}^2 n_{j6}^2 + n_{j2}^2 n_{j7}^2) + 4(n_{j8}n_{j5}n_{j3}n_{j2} + n_{j1}n_{j4}n_{j6}n_{j7}) \\ & -2(n_{j8}n_{j1}n_{j5}n_{j4} + n_{j8}n_{j1}n_{j3}n_{j6} + n_{j8}n_{j1}n_{j2}n_{j7} + n_{j5}n_{j4}n_{j3}n_{j6} \\ & + n_{j5}n_{j4}n_{j2}n_{j7} + n_{j3}n_{j6}n_{j2}n_{j7}). \end{aligned}$$

By 'feasible' it is meant that the solutions are real numbers, contained in the interval $[0, 1]$. Moreover, if $g_{2jd} \geq \sqrt{g_{4j}} + |h_{2jd}|$, Altmann and Altmann (2000) showed that feasible solutions for the disease-specific seroconversion rates and exposure probabilities can be obtained as well ($d = 1, 2, 3$):

$$\zeta_{jd} = \frac{2\sqrt{g_{4j}}}{g_{2jd} - h_{2jd} + \sqrt{g_{4j}}}, \qquad \eta_{jd} = \frac{g_{2jd} + h_{2jd} - \sqrt{g_{4j}}}{2g_{2jd}},$$

where

$$\begin{aligned} h_{2j1} &= -n_{j8}n_{j5} + n_{j7}n_{j6} + n_{j4}n_{j1} - n_{j3}n_{j2} \\ h_{2j2} &= -n_{j8}n_{j3} + n_{j7}n_{j4} + n_{j6}n_{j1} - n_{j5}n_{j2} \\ h_{2j3} &= -n_{j8}n_{j2} + n_{j7}n_{j1} + n_{j6}n_{j4} - n_{j5}n_{j3}. \end{aligned}$$

Altmann and Altmann (2000) revealed that the set of equations (7.1) has a special geometrical interpretation, which implies that the vaccination coverage ν_j is more robust against data noise than the seroconversion rates or the exposure probabilities. This is reflected by the fact that one only needs the weaker primary feasibility conditions to obtain an appropriate result for the vaccination coverage.

7.1.3 Illustration of the Independence Models

We now illustrate both methods to assess MMR coverage using the Belgian 2002 data set presented in Table 3.2. Table 7.1 shows the age-specific estimates (indicated with a tilde) obtained using Gay's maximum likelihood approach, and the exact solutions of Altmann and Altmann (2000). The ML-estimates for the vaccination coverage and the disease-specific exposure probabilities are also depicted together with 95% pointwise CIs in the left upper panels of Figure 7.1 and 7.2, respectively. The vaccination coverage is estimated to be 47% in 1 year olds, which is expected since the first MMR dose is administered during the second year of life. For the other age groups, the coverage estimates fluctuate between 68% and 92%. There seems to be a decreasing trend of MMR coverage in the younger age cohorts with a minimum at 9 years of age, while the coverage broadly increases again afterwards. This effect could be due to the second MMR dose recommended at 10-13 years. Figure 7.2 indicates that the estimated variability of the exposure probabilities is remarkably large, especially in the younger age groups where estimates are at the boundary of zero. Although the exposure estimates are very wiggly, there seems to be some increasing trend over age for all three diseases.

The Altmann and Altmann (2000) approach entails feasible solutions for the vaccination coverage, but the other parameters sometimes fall outside the biologically plausible parameter range, as indicated in bold in Table 7.1. Some solutions for the seroconversion rates ζ_{jd} are larger than 1 and some negative values are obtained for the exposure probabilities η_{jd}. Recall that in Gay's estimation approach, the parameter estimates were constrained to lie within $[0, 1]$. Overall, the vaccination coverage estimates from Gay's model are quite close to the exact Altmann and Altmann (2000) solutions in younger age groups. In the older age cohorts, there is some difference

Table 7.1: Comparison of the exact solutions as derived by Altmann & Altmann and the estimates obtained using Gay's maximum likelihood approach (indicated with a tilde) for the Belgian 2002 data.

Age	ν	$\tilde{\nu}$	ζ_1	ζ_2	ζ_3	η_1	$\tilde{\eta}_1$	η_2	$\tilde{\eta}_2$	η_3	$\tilde{\eta}_3$
$[1,2)$	0.47	0.47	0.96	0.87	**1.01**	0.10	0.10	0.04	0.04	0.10	0.11
$[2,3)$	0.87	0.88	0.98	0.84	1.00	0.22	0.21	0.00	0.00	0.10	0.10
$[3,4)$	0.92	0.92	1.00	0.86	0.98	**-0.03**	0.00	0.00	0.00	0.17	0.16
$[4,5)$	0.87	0.86	0.97	0.86	0.98	**-0.02**	0.00	0.00	0.00	0.06	0.08
$[5,6)$	0.83	0.84	**1.01**	0.84	1.00	0.00	0.00	0.20	0.17	0.21	0.16
$[6,7)$	0.81	0.81	1.00	0.81	0.97	0.16	0.19	0.00	0.00	0.18	0.18
$[7,8)$	0.84	0.81	1.00	0.86	0.92	0.29	0.44	0.00	0.11	0.50	0.46
$[8,9)$	0.88	0.88	0.98	0.77	1.00	0.12	0.12	0.00	0.00	0.09	0.13
$[9,10)$	0.67	0.68	**1.03**	0.77	**1.04**	0.51	0.53	0.14	0.06	0.34	0.38
$[10,11)$	0.88	0.87	1.00	0.83	1.00	0.57	0.63	0.00	0.00	0.25	0.33
$[11,12)$	0.88	0.86	0.95	0.90	0.99	0.40	0.37	0.25	0.40	0.22	0.33
$[12,13)$	0.74	0.73	0.75	**1.04**	**1.05**	0.67	0.48	0.36	0.59	0.51	0.61
$[13,14)$	0.80	0.78	0.93	0.95	0.99	0.40	0.34	0.25	0.43	0.49	0.53
$[14,15)$	0.89	0.84	0.95	0.97	0.95	**-0.01**	0.13	0.32	0.65	0.60	0.65
$[15,16)$	0.90	0.79	0.86	0.96	0.93	0.49	0.49	**-0.09**	0.55	0.74	0.80
$[16,17)$	0.96	0.89	0.92	0.88	1.00	0.00	0.00	0.50	0.68	0.39	0.76
$[17,18)$	0.88	0.84	1.00	0.90	0.91	0.29	0.53	0.33	0.51	0.50	0.44
$[18,19)$	0.89	0.86	0.95	0.89	0.98	0.27	0.29	**-0.01**	0.21	0.52	0.60

with respect to vaccination coverage, especially in the 15 and 16 year olds. To compare the seroconversion rates predicted by the two models, we calculate a weighted average over age of the exact Altmann and Altmann (2000) values: 0.96 for measles, 0.88 for mumps and 0.98 for rubella. These are fairly similar to the seroconversion rates estimated using Gay's model, which are presented in the first row of Table 7.2. As expected, the lowest vaccine efficacy is obtained for mumps.

The deviance -2ℓ and the AIC and BIC values for Gay's model are presented in Table 7.3. To assess the goodness-of-fit of Gay's model, we also fit the most saturated biologically relevant model, thus allowing for the seroconversion rates to depend on age. The following fit statistics are obtained for this saturated model: $-2\ell = 2590.11$, # par = 126, AIC = 2842.11, and BIC = 3499.51. The χ^2 goodness-of-fit test leads to an approximate p-value of 0.666, indicating a good fit of Gay's model.

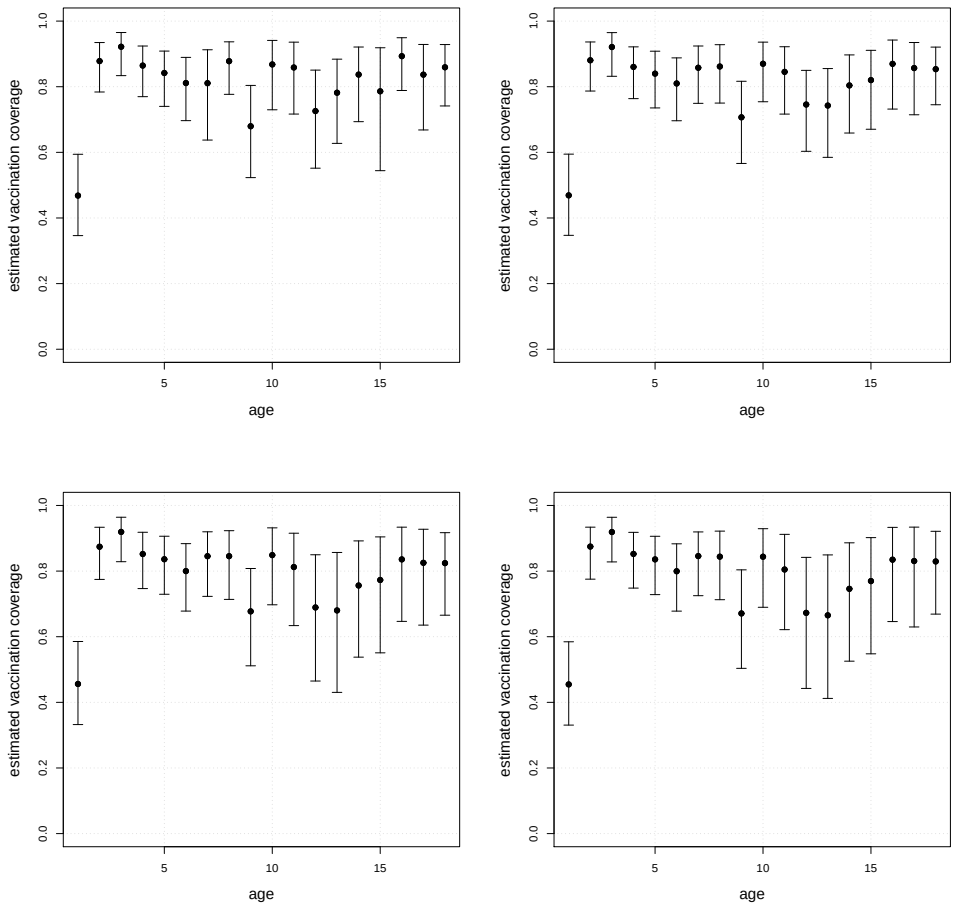

Figure 7.1: Estimated vaccination coverage with 95% pointwise CI for each age class of the Belgian 2002 data, obtained for Gay's model with 'saturated' exposure probabilities (left upper panel), and for Gay's model (right upper panel), BAH I (left lower panel) and BAH II (right lower panel) with RCS structure (4 knots) for the exposure probabilities.

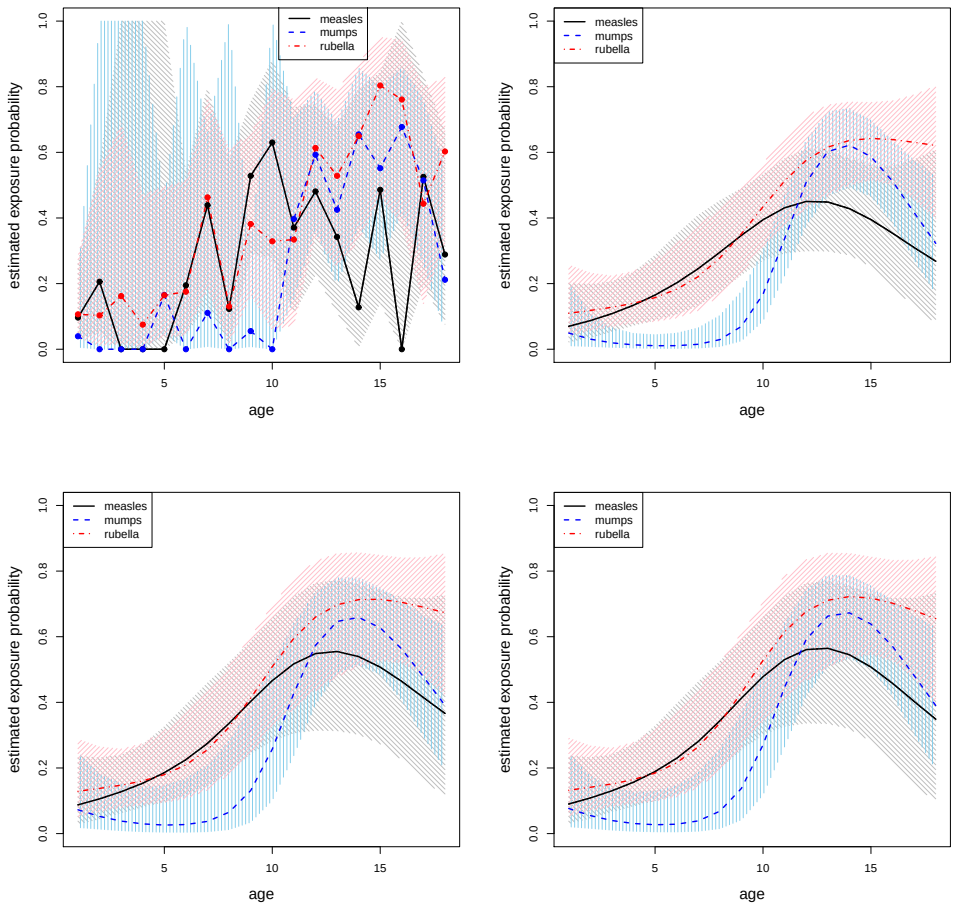

Figure 7.2: Estimated exposure probabilities with 95% pointwise CIs for each age class of the Belgian 2002 data, obtained for Gay's model with 'saturated' exposure probabilities (left upper panel), and for Gay's model (right upper panel), BAH I (left lower panel) and BAH II (right lower panel) with RCS structure (4 knots) for the exposure probabilities.

Table 7.2: Estimates of the seroconversion rates to each of the MMR vaccine components with 95% CIs in square brackets, obtained for a selected set of models fitted to the Belgian (BE) 2002 and the Irish (IE) 2003 data.

Data	Model	Exposure	Measles (ζ_1)	Mumps (ζ_2)	Rubella (ζ_3)
BE 2002	Gay	Saturated	0.99 [0.96, 0.99]	0.85 [0.82, 0.88]	0.99 [0.97, 1.00]
	Gay	RCS $K = 4$	0.99 [0.96, 0.99]	0.85 [0.82, 0.87]	0.99 [0.97, 1.00]
	BAH I	RCS $K = 4$	0.99 [0.96, 1.00]	0.84 [0.81, 0.87]	0.99 [0.97, 1.00]
	BAH II	RCS $K = 4$	0.99 [0.96, 1.00]	0.84 [0.81, 0.87]	0.99 [0.96, 1.00]
IE 2003	Gay	Saturated	0.99 [0.97, 1.00]	0.83 [0.79, 0.87]	0.97 [0.95, 0.98]
	Gay	RCS $K = 3$	0.99 [0.95, 1.00]	0.83 [0.79, 0.86]	0.97 [0.94, 0.98]
	BAH I	RCS $K = 3$	1.00 [0.00, 1.00]	0.82 [0.78, 0.86]	0.98 [0.93, 0.99]
	BAH III	RCS $K = 3$	1.00 [0.00, 1.00]	0.82 [0.72, 0.89]	0.98 [0.57, 1.00]

7.2 Likelihood-Based Marginal Modelling

The approaches of Gay (2000) and Altmann and Altmann (2000) both rely on the three assumptions formulated by Gay as listed in Section 7.1.1. A violation of the first assumption, e.g. in the sense that vaccinated individuals who did not seroconvert as a result of vaccination, may have a smaller probability of acquiring natural infection than an unvaccinated individual of the same age, would not affect our results too much because of the high immunogenicity of the MMR vaccine. The second assumption is supported by a study from England (Pebody *et al.*, 2002), where no evidence for correlation of seropositivity to each antigen was found after MMR vaccination other than that produced by a small excess of children seronegative to all three diseases. The third assumption ignores the natural heterogeneity in the population in that some individuals are more prone than others to acquire infection with all three diseases. Measles, mumps and rubella are transmitted through the same route, thus the individual's social contact behavior underlies the probability of exposure to MMR. The exposure probabilities are therefore not independent, and we assess the impact of relaxing the third assumption by extending Gay's model to take into account the dependency in acquisition of the three diseases. Note that the exact approach of Altmann and Altmann (2000) cannot be extended in a similar way, because it relies on the dimension of the probability space and the parameter space being both equal to 7, which results from the complete separation of the system of equations (7.1) per age class. It is convenient to model the dependency at the marginal level, since subject-specific models are computationally very intensive. In Section 7.2.1, we introduce

Table 7.3: Estimates of the correlation parameters with 95% CIs, the deviance and information criteria obtained for Gay's model and specific Bahadur models with different mean structures for the exposure probabilities fitted to the Belgian 2002 data.

Model	Par	Est	95% CI	-2ℓ	#par	AIC	BIC
Saturated exposure probabilities							
Gay (indep.)				2636.28	75	2786.28	3177.59
RCS exposure probabilities $K = 3$							
Gay (indep.)				2708.29	30	2768.29	2924.82
BAH I	$\rho_{(2)}$	0.16	[-0.03, 0.34]	2706.21	31	2768.21	2929.95
BAH II	$\rho_{(2)}$	0.15	[-0.08, 0.36]	2705.99	32	2769.99	2936.95
	$\rho_{(3)}$	-0.04	[-0.22, 0.14]				
RCS exposure probabilities $K = 4$							
Gay (indep.)				2686.58	33	2752.58	2924.75
BAH I	$\rho_{(2)}$	0.11	[-0.07, 0.28]	2684.99	34	2752.99	2930.38
BAH II	$\rho_{(2)}$	0.13	[-0.03, 0.29]	2684.45	35	2754.45	2937.06
	$\rho_{(3)}$	0.08	[-0.13, 0.28]				
RCS exposure probabilities $K = 5$							
Gay (indep.)				2675.60	36	2747.60	2935.43

the Bahadur model to describe the association between the exposure probabilities in terms of correlation coefficients. But before doing so, we derive expressions for the joint probabilities in the vaccinated population.

In general, the joint probabilities (7.1) for age class $[j, j+1)$ can be written as:

$$f_j(\boldsymbol{y}_{ij}) = \nu_j f_j(\boldsymbol{y}_{ij}|\text{V}) + (1-\nu_j) f_j(\boldsymbol{y}_{ij}|\text{NV}), \tag{7.4}$$

where V indicates 'vaccinated' and NV indicates 'non-vaccinated'. Given the marginal model for the joint exposure probabilities $f_j(\boldsymbol{y}_{ij}|\text{NV})$, we are able to derive the joint density for $\boldsymbol{Y}_{ij}$ in the vaccinated population, and by (7.4) also the full joint distribution. In the vaccinated population, the processes of seroconversion and natural infection are interdependent and we need the law of total probability, hereby conditioning on the seroconversion status, to obtain the joint probability for each of the

eight different immunity states:

$$
\begin{aligned}
f_j(+,+,+|\mathrm{V}) &= \zeta_1\zeta_2\zeta_3 + \zeta_1\zeta_2(1-\zeta_3)f_j(.,.,+|\mathrm{NV}) + \zeta_1(1-\zeta_2)\zeta_3 f_j(.,+,.|\mathrm{NV}) \\
&\quad +\zeta_1(1-\zeta_2)(1-\zeta_3)f_j(.,+,+|\mathrm{NV}) + (1-\zeta_1)\zeta_2\zeta_3 f_j(+,.,.|\mathrm{NV}) \\
&\quad +(1-\zeta_1)\zeta_2(1-\zeta_3)f_j(+,.,+|\mathrm{NV}) + (1-\zeta_1)(1-\zeta_2)\zeta_3 f_j(+,+,.|\mathrm{NV}) \\
&\quad +(1-\zeta_1)(1-\zeta_2)(1-\zeta_3)f_j(+,+,+|\mathrm{NV}) \\
f_j(+,+,-|\mathrm{V}) &= \zeta_1\zeta_2(1-\zeta_3)f_j(.,.,-|\mathrm{NV}) + \zeta_1(1-\zeta_2)(1-\zeta_3)f_j(.,+,-|\mathrm{NV}) + (1-\zeta_1) \\
&\quad \cdot\zeta_2(1-\zeta_3)f_j(+,.,-|\mathrm{NV}) + (1-\zeta_1)(1-\zeta_2)(1-\zeta_3)f_j(+,+,-|\mathrm{NV}) \\
f_j(+,-,+|\mathrm{V}) &= \zeta_1(1-\zeta_2)\zeta_3 f_j(.,-,.|\mathrm{NV}) + \zeta_1(1-\zeta_2)(1-\zeta_3)f_j(.,-,+|\mathrm{NV}) + (1-\zeta_1) \\
&\quad \cdot(1-\zeta_2)\zeta_3 f_j(+,-,.|\mathrm{NV}) + (1-\zeta_1)(1-\zeta_2)(1-\zeta_3)f_j(+,-,+|\mathrm{NV}) \\
f_j(+,-,-|\mathrm{V}) &= \zeta_1(1-\zeta_2)(1-\zeta_3)f_j(.,-,-|\mathrm{NV}) + (1-\zeta_1)(1-\zeta_2)(1-\zeta_3) \\
&\quad \cdot f_j(+,-,-|\mathrm{NV}) \\
f_j(-,+,+|\mathrm{V}) &= (1-\zeta_1)\zeta_2\zeta_3 f_j(-,.,.|\mathrm{NV}) + (1-\zeta_1)\zeta_2(1-\zeta_3)f_j(-,.,+|\mathrm{NV}) + (1-\zeta_1) \\
&\quad \cdot(1-\zeta_2)\zeta_3 f_j(-,+,.|\mathrm{NV}) + (1-\zeta_1)(1-\zeta_2)(1-\zeta_3)f_j(-,+,+|\mathrm{NV}) \\
f_j(-,+,-|\mathrm{V}) &= (1-\zeta_1)\zeta_2(1-\zeta_3)f_j(-,.,-|\mathrm{NV}) + (1-\zeta_1)(1-\zeta_2)(1-\zeta_3) \\
&\quad \cdot f_j(-,+,-|\mathrm{NV}) \\
f_j(-,-,+|\mathrm{V}) &= (1-\zeta_1)(1-\zeta_2)\zeta_3 f_j(-,-,.|\mathrm{NV}) + (1-\zeta_1)(1-\zeta_2)(1-\zeta_3) \\
&\quad \cdot f_j(-,-,+|\mathrm{NV}) \\
f_j(-,-,-|\mathrm{V}) &= (1-\zeta_1)(1-\zeta_2)(1-\zeta_3)f_j(-,-,-|\mathrm{NV}).
\end{aligned}
$$

The joint densities for immunity states 4, 6 and 7 in the vaccinated population can be simplified further by working out the marginal densities:

$$
\begin{aligned}
f_j(+,-,-|\mathrm{V}) &= \zeta_1(1-\zeta_2)(1-\zeta_3)f_j(-,-,-|\mathrm{NV}) + (1-\zeta_2)(1-\zeta_3)f_j(+,-,-|\mathrm{NV}) \\
f_j(-,+,-|\mathrm{V}) &= (1-\zeta_1)\zeta_2(1-\zeta_3)f_j(-,-,-|\mathrm{NV}) + (1-\zeta_1)(1-\zeta_3)f_j(-,+,-|\mathrm{NV}) \\
f_j(-,-,+|\mathrm{V}) &= (1-\zeta_1)(1-\zeta_2)\zeta_3 f_j(-,-,-|\mathrm{NV}) + (1-\zeta_1)(1-\zeta_2)f_j(-,-,+|\mathrm{NV}).
\end{aligned}
$$

7.2.1 The Bahadur Model for Trivariate Binary Data

We use the Bahadur model (Bahadur, 1961), a marginal model conceived for binary data, to describe the joint density of the test outcomes $\boldsymbol{Y}_{ij}$ in the non-vaccinated population. For the model description, we mainly follow Molenberghs and Verbeke (2005). The Bahadur model (abbreviated 'BAH') can be represented as the product

of two components:

$$f_j(\boldsymbol{y}_{ij}|\text{NV}) = g_j(\boldsymbol{y}_{ij}|\text{NV}) \cdot c_j(\boldsymbol{y}_{ij}|\text{NV}).$$

In our trivariate context, the first factor represents the independence model (7.1) used by Gay (2000):

$$g_j(\boldsymbol{y}_{ij}|\text{NV}) = \prod_{d=1}^{3} \eta_{jd}^{y_{ijd}}(1-\eta_{jd})^{(1-y_{ijd})},$$

and the second component,

$$c_j(\boldsymbol{y}_{ij}|\text{NV}) = 1 + \rho_{j12}y^*_{ij1}y^*_{ij2} + \rho_{j13}y^*_{ij1}y^*_{ij3} + \rho_{j23}y^*_{ij2}y^*_{ij3} + \rho_{j123}y^*_{ij1}y^*_{ij2}y^*_{ij3},$$

embodies the correction factor. In these formulas, the y^*_{ijd} equal $\frac{y_{ijd}-\eta_{jd}}{\sqrt{\eta_{jd}(1-\eta_{jd})}}$ and can be interpreted as standardized response values, while the

$$\rho_{jd_1d_2} = E(Y^*_{ijd_1}Y^*_{ijd_2}) = \text{Corr}(Y_{ijd_1}, Y_{ijd_2}),$$

are defined as the marginal two-way correlation coefficients and $\rho_{j123} = E(Y^*_{ij1}Y^*_{ij2}Y^*_{ij3})$ as the third order correlation coefficient. We use Fisher's z-transform for the correlation coefficients:

$$z(\rho) = \log\left(\frac{1+\rho}{1-\rho}\right) = \gamma \Leftrightarrow \rho = \frac{\exp(\gamma)-1}{\exp(\gamma)+1},$$

and the γ parameters are estimated by maximizing the loglikelihood (7.3) using the same link functions for the other parameters as before (7.2).

Given the estimates obtained for the exposure probabilities and the two-way correlation coefficients, we can also estimate the pairwise probability that an unvaccinated individual of age j had a past infection with both disease d_1 and d_2:

$$\eta_{jd_1d_2} = P(Y_{ijd_1} = 1, Y_{ijd_2} = 1|\text{NV}) = \eta_{jd_1}\eta_{jd_2} + \rho_{jd_1d_2}[\eta_{jd_1}(1-\eta_{jd_1})\eta_{jd_2}(1-\eta_{jd_2})]^{1/2},$$

or the joint probability that an unvaccinated individual of age j had a past infection with all three diseases:

$$\begin{aligned}\eta_{j123} = P(\boldsymbol{Y}_{ij} = (1,1,1)|\text{NV}) &= \eta_{j1}\eta_{j2}\eta_{j3} + \rho_{j12}\eta_{j3}[\eta_{j1}(1-\eta_{j1})\eta_{j2}(1-\eta_{j2})]^{1/2} + \rho_{j13} \\ &\cdot \eta_{j2}[\eta_{j1}(1-\eta_{j1})\eta_{j3}(1-\eta_{j3})]^{1/2} + \rho_{j23}\eta_{j1}[\eta_{j2}(1-\eta_{j2})\eta_{j3}(1-\eta_{j3})]^{1/2} + \rho_{j123} \\ &\cdot [\eta_{j1}(1-\eta_{j1})\eta_{j2}(1-\eta_{j2})\eta_{j3}(1-\eta_{j3})]^{1/2}.\end{aligned}$$

During the analyses, we make some simplifying assumptions with respect to the association structure. First, we assume that the correlation coefficients are the same for all age groups, which reduces the total number of correlation parameters from $4m$ to

Table 7.4: Summary of the different Bahadur models considered.

Name	#par	Second order	Third order
BAH I	1	$\rho_{(2)} = \rho_{12} = \rho_{13} = \rho_{23}$	none
BAH II	2	$\rho_{(2)} = \rho_{12} = \rho_{13} = \rho_{23}$	$\rho_{(3)}$
BAH III	3	ρ_{12}, ρ_{13}, ρ_{23}	none
BAH IV	4	ρ_{12}, ρ_{13}, ρ_{23}	$\rho_{(3)}$

4. Further, we explore the assumption of 'exchangeability' or 'equicorrelation' which states that the associations do not depend on the disease: $\rho_{12} = \rho_{13} = \rho_{23} = \rho_{(2)}$ and $\rho_{123} = \rho_{(3)}$, thus leading to 2 correlation parameters. Finally, we investigate whether the third order correlation coefficient $\rho_{(3)}$ can be ignored. The four resulting Bahadur models that are considered in this chapter, are summarized in Table 7.4.

Bahadur (1961) indicates that the sum of the probabilities of all possible outcomes is one, i.e. $\sum_{k=1}^{8} p_{jk|\mathrm{NV}}$, even when higher order correlations are set equal to zero. However, the requirement of having non-negative probabilities for all possible outcomes results in restrictions on the parameter space of the Bahadur model. These restrictions have been studied in the specific context of exchangeably clustered data, i.e. each subject within a cluster has the same response probability and the associations of a particular order are assumed constant within a cluster (see e.g. Bahadur, 1961; Declerck *et al.*, 1998). Bahadur (1961) discusses the restrictions on the second order correlation when all higher order associations are left out. The bounds on $\rho_{(2)}$, required to obtain a valid probability mass function, depend on the response probability. When the size of the clusters equals three, the lower bounds vary between -1/3 and 0, while the upper bounds vary between 0.5 and 1. Within the context of developmental toxicity studies, Declerck *et al.* (1998) studied the lower and upper bound of $\rho_{(2)}$ in the presence of higher order correlations. The inclusion of a third order correlation in the Bahadur model for clusters of size three, somewhat relaxes the bounds on $\rho_{(2)}$. In our setting of (unexchangeable) trivariate data, the analytical calculations for these bounds would be much more complex and are beyond the scope of this study.

7.2.2 Semiparametric Model for the Exposure Probabilities

Up till now, the exposure probabilities η_{jd} were modelled in a 'saturated' manner (7.2) with one parameter β_{jd} for each disease d and each age class j, leading to a total of $3m$ parameters. Gay (2000) put the additional constraint that the exposure probabilities

are monotonically increasing with age, which is to be expected for an immunizing infection in a population with no vaccination programme. However, mass vaccination against MMR has been implemented in most of Europe since the 1980s, and the effect of herd immunity has altered MMR disease dynamics. Therefore, the exposure probabilities estimated from 21st century current status data do not anymore reflect the monotone age-specific seroprevalence profile expected under pre-vaccination equilibrium. Since we do not know the functional relationship between the marginal probability for a non-vaccinated individual of acquiring infection with measles, mumps or rubella, and age, we propose to use a semiparametric model of cubic regression splines (3.1). Cubic regression splines are an easy way of including an explanatory variable in a smooth non-linear way in a wide variety of models, however, they may behave poorly in the tails.

'Restricted cubic splines' (abbreviated 'RCS') also known as 'natural cubic splines' are cubic splines with the constraint that they are linear in their tails beyond the boundary knots. This is enforced by putting $s''_d = s'''_d = 0$, where $s_d()$ represents the cubic splines function (3.1). Restricted cubic splines thus allow for a more parsimonious model. The RCS model with K knots $\kappa_{1d}, \ldots, \kappa_{Kd}$ for the independent variable 'age' is given by (Devlin and Weeks, 1986):

$$\text{logit}(\eta_{jd}) = s_d(j) = \beta_{0d} + \beta_{1d} j_{1d} + \beta_{2d} j_{2d} + \ldots + \beta_{K-1,d} j_{K-1,d},$$

where $j_{1d} = j$ and for $q = 1, \ldots, K-2$:

$$j^*_{q+1,d} = (j - \kappa_{qd})^3_+ - \frac{(j - \kappa_{K-1,d})^3_+ (\kappa_{Kd} - \kappa_{qd})}{(\kappa_{Kd} - \kappa_{K-1,d})} + \frac{(j - \kappa_{Kd})^3_+ (\kappa_{K-1,d} - \kappa_{qd})}{(\kappa_{Kd} - \kappa_{K-1,d})},$$

where $(j - \kappa_{qd})_+ = j - \kappa_{qd}$ if $j > \kappa_{qd}$ and $(j - \kappa_{qd})_+ = 0$ if $j \leq \kappa_{qd}$. The variables $j_{q+1,d}$ are normalized as follows: $j_{q+1,d} = j^*_{q+1,d}/(\kappa_{Kd} - \kappa_{1d})^2$, such that the RCS components are on the original age scale. This RCS model has a total of K parameters to describe the exposure probability for disease d. The `SAS` macro `%rcspline` from Harrell is used to generate the RCS components (Harrell, 2001).

7.2.3 Application to the Data

We now fit Gay's model and the four Bahadur models (see Table 7.4) to the Belgian 2002 and the Irish 2003 serology, hereby considering different structures for the exposure probabilities: saturated as in the original framework of Gay (2000), and semiparametric RCS models with 3 to 5 knots placed at predefined quantiles of the age range, hereby following Harrell (2001). Due to the sparseness of the data (Tables 3.2

and 3.3) and the structure of the joint density (7.4), the optimization procedure for some of the Bahadur models leads to parameter estimates which do not yield a valid multinomial probability mass function in the non-vaccinated population for all age categories. Building in parameter restrictions to avoid negative probabilities, would require complex analytical derivations of bounds for all correlation parameters (for each Bahadur model considered), which is beyond the scope of this study. Alternatively, penalized likelihood methods to constrain the probabilities could be explored, but we will not consider these in this thesis. Therefore, we here present the results of the Bahadur models for which the ML-estimates entail a valid probability mass function for all age categories, as well as a positive definite hessian matrix.

The correlation parameter estimates with 95% CIs, the deviance and the AIC and BIC values are presented in Tables 7.3 and 7.5 for Belgium and Ireland, respectively. For the Belgian serology, both LR-tests as well as AIC and BIC values indicate that the correlation parameters in BAH I and BAH II for the RCS models with 3 or 4 knots, are not significantly different from zero. For Ireland, however, AIC selects BAH I over Gay's model and the LR-tests for $H_0 : \rho_{(2)} = 0$ entail a significant p-value of 0.044 for $K = 3$ and a non-significant p-value of 0.060 for $K = 4$. For the Belgian data, when we further increase the number of knots in the RCS model to 6, the AIC and BIC value for Gay's model are both larger than when $K = 5$ (2752.39 and 2955.87, respectively). Furthermore, the corresponding Bahadur models are not valid. The RCS models with $K = 4$ and $K = 3$ were broadly the best in terms of AIC and BIC for Belgium and Ireland, respectively. For these models, the ML-estimates of the seroconversion rates are shown in Table 7.2, while the ML-estimates of the vaccination coverage and the exposure probabilities are displayed together with 95% pointwise CIs in Figures 7.1 and 7.2 (Belgium) and Figures 7.3 and 7.4 (Ireland), respectively.

From the upper panels in Figures 7.1 and 7.3, it is observed that replacing the saturated exposure probabilities by a more parsimonious semiparametric model, has a moderate effect on the estimated MMR vaccination coverage: the estimates are pressed towards one another, inducing a more stable pattern with less fluctuations. By taking into account the dependency in acquisition of the three pathogens (Figures 7.1 and 7.3, lower panels), the vaccination coverage estimates decrease everywhere to some extent, especially in the older age cohorts. Further, the estimated uncertainty increases markedly which is embodied by the wider 95% confidence intervals. There is no substantial difference between the different Bahadur models with regard to the estimated vaccination coverage, except that the variability is larger for BAH III than for BAH I when considering the Irish serology in Figure 7.3. The MMR seroconversion

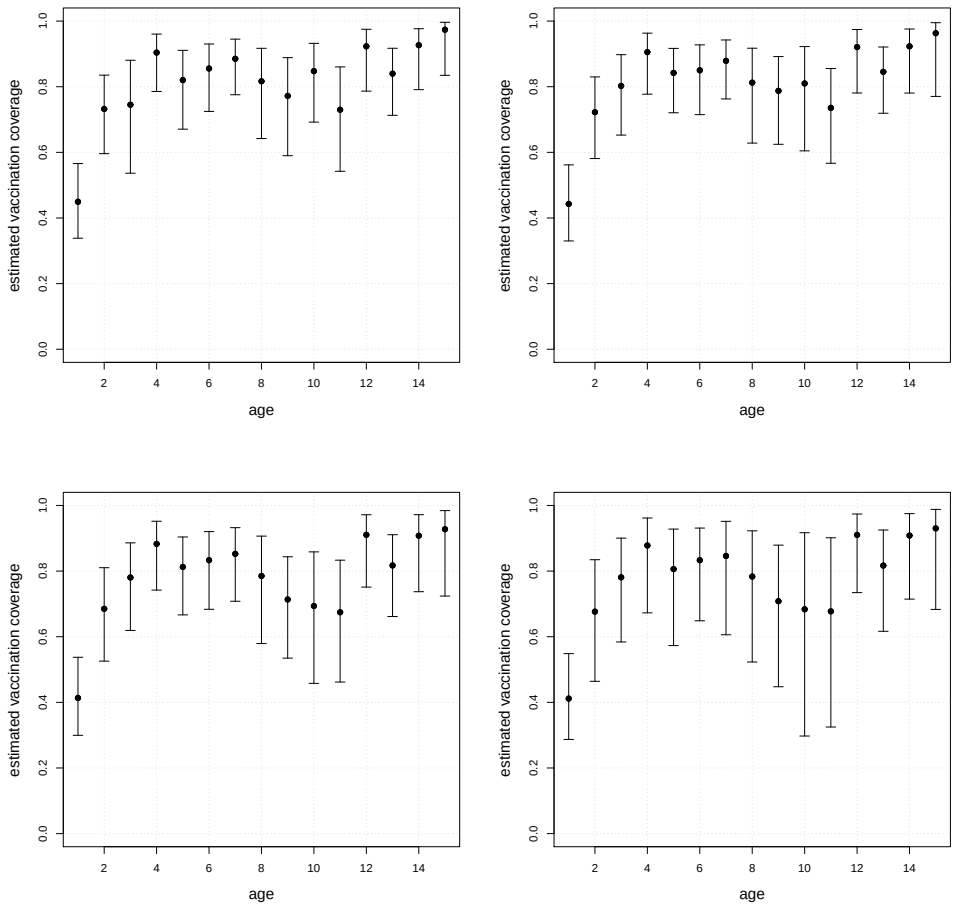

Figure 7.3: Estimated vaccination coverage with 95% pointwise CI for each age class of the Irish 2003 data, obtained for Gay's model with 'saturated' exposure probabilities (left upper panel), and for Gay's model (right upper panel), BAH I (left lower panel) and BAH III (right lower panel) with RCS structure (3 knots) for the exposure probabilities.

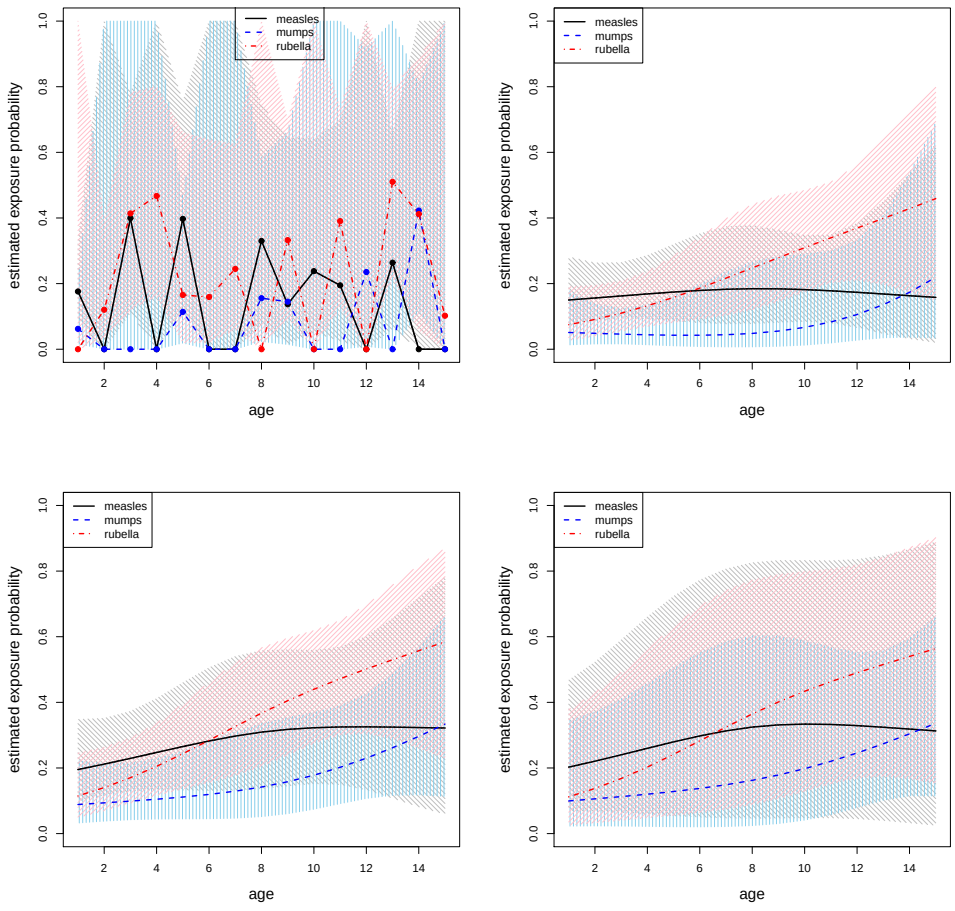

Figure 7.4: Estimated exposure probabilities with 95% pointwise CIs for each age class of the Irish 2003 data, obtained for Gay's model with 'saturated' exposure probabilities (left upper panel), and for Gay's model (right upper panel), BAH I (left lower panel) and BAH III (right lower panel) with RCS structure (3 knots) for the exposure probabilities.

Table 7.5: Estimates of the correlation parameters with 95% CIs, the deviance and information criteria obtained for Gay's model and specific Bahadur models with different mean structures for the exposure probabilities fitted to the Irish 2003 data.

Model	Par	Est	95% CI	-2ℓ	#par	AIC	BIC
Saturated exposure probabilities							
Gay (indep.)				1541.67	63	1667.67	1956.59
RCS exposure probabilities $K = 3$							
Gay (indep.)				1588.68	27	1642.68	1766.51
BAH I	$\rho_{(2)}$	0.21	[0.03, 0.37]	1584.64	28	1640.64	1769.05
BAH III	ρ_{12}	0.27	[-0.32, 0.70]	1584.63	30	1644.63	1782.21
	ρ_{13}	0.18	[-0.56, 0.77]				
	ρ_{23}	0.20	[0.00, 0.38]				
RCS exposure probabilities $K = 4$							
Gay (indep.)				1585.07	30	1645.07	1782.65
BAH I	$\rho_{(2)}$	0.21	[0.03, 0.39]	1581.52	31	1643.52	1785.70
RCS exposure probabilities $K = 5$							
Gay (indep.)				1581.10	33	1647.10	1798.45

rate estimates in Table 7.2 are, except for the estimated uncertainty, nearly unaffected by the assumptions regarding the mean and association structure for the exposure probabilities. Further, it is quite agreeable to observe such a strong consistency between the disease-specific estimates for Belgium and Ireland.

The upper panels in Figures 7.2 and 7.4 clearly show the difference between the saturated and the RCS model for the exposure probabilities. The fitted RCS profiles are very smooth and the 95% pointwise CIs are more narrow compared to the saturated model. As can be seen from the two lower panels, the profiles are estimated to be higher and the corresponding 95% CIs are wider when the association between the probabilities of acquiring infection with MMR is incorporated into the model. There is virtually no effect of allowing for a third order correlation for the Belgian data, or by relaxing the assumption of exchangeability for the two-way correlations for the Irish serology. Overall, the estimated MMR exposure probabilities seem to be somewhat larger in older age cohorts which could be due to the accumulating risk of

exposure or to the higher circulation of MMR at the introduction of universal vaccination (herd immunity effect). Further, Figures 7.2 and 7.4 indicate that unvaccinated teenagers seem more likely to have acquired past infection with rubella compared with measles or mumps, which is a strange result since the pre-vaccination prevalence of immune teenagers is estimated to be smaller for rubella in comparison to measles and mumps (Farrington, 1990). This could however reflect the delay of replacement of the monovalent rubella vaccine by the second MMR-dose in pre-adolescent girls.

7.2.4 Sensitivity Analysis

One could argue that the second and third order correlations between the exposure probabilities for MMR should be positive because of the common, behavioral social contact aspect which drives exposure to all three diseases. As a sensitivity analysis, we re-fit the Bahadur models considered in the previous section under the constraint that $0 \leq \rho \leq 1$, where ρ represents any correlation parameter present in the models. When allowing for a saturated structure for the exposure probabilities, as before, none of the Bahadur models entail a valid fit for the two data sets. The results of the RCS models for the Irish data set are not presented here, since the hessian obtained for BAH III ($K = 3$) is not positive definite anymore, and the results for the BAH I models are exactly the same as before (Table 7.5), except for a slight shift of the 95% CIs for $\rho_{(2)}$ to the right.

The sensitivity analysis results for Belgium are more interesting (see Table 7.6), since more Bahadur models now lead to a valid solution. While BAH III with an RCS structure of 5 knots is the best model in terms of AIC, Gay's model with an RCS structure of 4 knots is the best model in terms of BIC. For the RCS models with $K = 4$, a LR-test for Gay's model versus BAH III entails a non-significant p-value of 0.097. When the number of knots is increased to 5, the three pairwise LR-tests always reject the null hypothesis of no correlation (p-values range from 0.027 to 0.043). LR-tests for the nested Bahadur models indicate that extending to BAH IV is not informative. Figure 7.5 shows the ML-estimates of the vaccination coverage and the exposure probabilities together with 95% pointwise CIs for BAH III with an RCS structure of 5 knots. The vaccination coverage estimates are broadly somewhere in between Gay's model and BAH I with an RCS structure of 4 knots (Figure 7.1), whereas the estimated exposure probability curve for measles in the right panel has quite a different shape with a more pronounced peak around the age of 10 years. In conclusion, the sensitivity analysis makes clear that it might be important for the Belgian data as well to take into account the association between the MMR acquisition

Table 7.6: Estimates of the correlation parameters with 95% CIs, the deviance and information criteria obtained for specific Bahadur models with different mean structures for the exposure probabilities fitted to the Belgian 2002 data under the constraint of positive correlation (Gay's model added as a reference).

Model	Par	Est	95% CI	-2ℓ	#par	AIC	BIC
RCS exposure probabilities $K = 3$							
Gay (indep.)				2708.29	30	2768.29	2924.82
BAH I	$\rho_{(2)}$	0.16	[0.05, 0.43]	2706.20	31	2768.20	2929.95
BAH II	$\rho_{(2)}$	0.17	[0.06, 0.41]	2706.29	32	2770.29	2937.25
	$\rho_{(3)}$	0.01	[0.00, 0.99]				
RCS exposure probabilities $K = 4$							
Gay (indep.)				2686.58	33	2752.58	2924.75
BAH I	$\rho_{(2)}$	0.11	[0.02, 0.43]	2684.99	34	2752.99	2930.38
BAH II	$\rho_{(2)}$	0.14	[0.04, 0.38]	2684.45	35	2754.45	2937.06
	$\rho_{(3)}$	0.07	[0.00, 0.65]				
BAH III	ρ_{12}	0.02	[0.00, 1.00]	2680.27	36	2752.27	2940.09
	ρ_{13}	0.01	[0.00, 0.98]				
	ρ_{23}	0.25	[0.11, 0.47]				
RCS exposure probabilities $K = 5$							
Gay (indep.)				2675.60	36	2747.60	2935.43
BAH II	$\rho_{(2)}$	0.24	[0.12, 0.43]	2669.28	38	2745.28	2943.55
	$\rho_{(3)}$	0.00	[0.00, 1.00]				
BAH III	ρ_{12}	0.09	[0.01, 0.63]	2666.44	39	2744.44	2947.92
	ρ_{13}	0.00	[0.00, 1.00]				
	ρ_{23}	0.28	[0.16, 0.46]				
BAH IV	ρ_{12}	0.14	[0.02, 0.55]	2665.74	40	2745.74	2954.44
	ρ_{13}	0.00	[0.00, 1.00]				
	ρ_{23}	0.33	[0.17, 0.54]				
	$\rho_{(3)}$	0.11	[0.01, 0.68]				

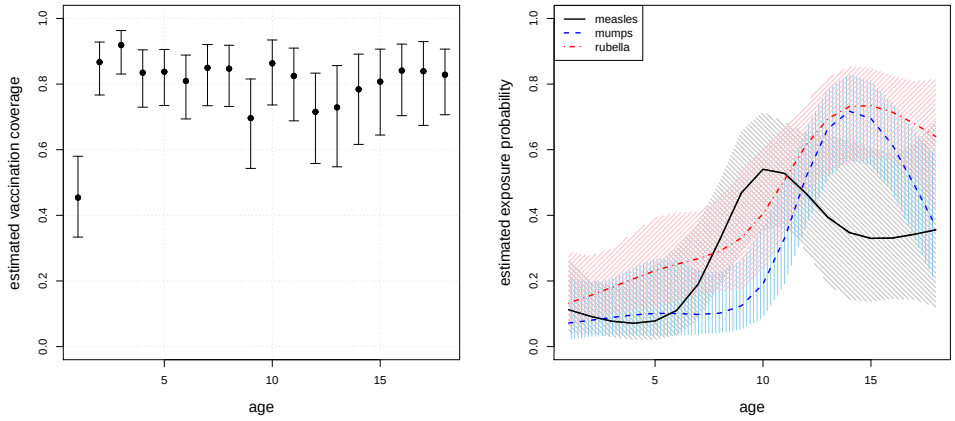

Figure 7.5: Estimated vaccination coverage (left panel) and exposure probabilities (right panel) with 95% pointwise CIs for each age class of the Belgian 2002 data, obtained for BAH III with RCS structure (5 knots) for the exposure probabilities under the constraint of positive correlation.

probabilities.

7.3 Concluding Remarks

In Section 7.1, we described and illustrated the two existing approaches of Gay (2000) and Altmann and Altmann (2000) to, respectively, estimate and theoretically calculate trivalent vaccination coverage from trivariate serological data. Although the method of Altmann and Altmann (2000) was founded on Gay's modelling equations, it is less interesting from a statistical point of view since it ignores the variability in the data and attributes biologically implausible values to insufficient data quality rather than incorrect modelling assumptions. As a by-product of Gay's method, estimates for the vaccine seroconversion rates and the natural exposure probabilities are obtained. In Section 7.2, we relaxed one of the assumptions made by Gay (2000) and explicitly modelled the association between the probabilities of acquiring infection with each of the three pathogens, by means of the Bahadur model for trivariate binary data (Bahadur, 1961). When fitting several configurations of the Bahadur model to the Belgian and Irish MMR serology, we were confronted with the problem of negative probabilities for the multinomial distribution. Bahadur (1961) and Declerck *et al.* (1998), amongst others, have studied the correlation parameter restrictions needed to

Table 7.7: Estimates of MMR vaccine coverage in infants and adolescents obtained from retrospective EPI-surveys in Flanders (95% CIs in square brackets). Note that in the EPI-survey of 1999, infants with missing vaccination documents were omitted from the coverage estimates, while for the other surveys the doses were considered not given.

Survey year	Age	Birth cohort	Coverage dose 1	Coverage dose 2	Source
1999	18-24m	1997	83.4% [80.3, 86.5]		Swennen *et al.* (2002)
2005	18-24m	2003	94.0% [92.6, 95.3]		Theeten *et al.* (2007)
2005	7-8y	1997	88.0% [85.6, 90.4]		Theeten *et al.* (2009)
2005	13-14y	1991	80.6% [78.2, 83.0]	83.6% [81.4, 85.8]	Vandermeulen *et al.* (2008)
2008	18-24m	2006	96.6% [95.2, 97.6]		Boonen *et al.* (2009)
2008	13-14y	1994	88.1% [86.1, 90.0]	90.6% [89.0, 92.2]	Boonen *et al.* (2009)

obtain a valid density function, but only in the less complex case of exchangeably clustered data. Since analytically deriving the parameter bounds was beyond the scope of this study, we used an ad hoc approach by merely considering those models for which a valid fit to the data was obtained. Further, in Section 7.2.2, as an alternative to Gay's saturated model, we proposed a more parsimonious structure based on restricted cubic splines to model the exposure probabilities as a function of age. In our MMR applications, these RCS models clearly outperformed the saturated one.

Taking into account the dependency between the exposure probabilities somewhat decreases the MMR vaccination coverage estimates while it increases the associated variability (see Figures 7.1 and 7.3). For example, considering the Irish data with an RCS structure of 3 knots, MMR coverage estimates vary between 72% in the 2 years age cohort and 96% in the 15 years age cohort for Gay's model, whereas they vary between 67% in the 11 years age cohort and 93% in the 15 years age cohort for BAH I. In general, the estimated vaccination coverage profiles for Belgium and Ireland are quite different, with seemingly low MMR coverage in Irish infants in the early 2000's. While the Irish coverage shows an abrupt increase from 11 to 12 year olds, the Belgian coverage shows decreases in 9 and 12-13 year olds. Yet, we do not have an explanation for this. In Table 7.7, a brief overview is presented of the MMR vaccination coverage estimates obtained from retrospective EPI-surveys in Flanders. The MMR coverage estimates for Belgium are 84%, 86% and 85% according to Gay's model with $K = 4$ and 84%, 85% and 82% according to BAH III with $K = 5$, for the 1997 (dose 1), 1994

(dose 1) and 1991 (between dose 1 and 2) birth cohorts, respectively. These estimates match quite well with the EPI-survey results in Table 7.7.

One could constrain the RCS model for the exposure probabilities to take into account the protective period induced by maternally-derived immunity, for instance by imposing that $\eta_d(A) = \text{expit}(s_d(A)) = 0$, where A is the average age at which maternal immunity is lost. Further, we have additional covariate information about the serum samples such as gender, and specifically for Belgium: the region of residence (Flanders, Wallonia and Brussels). Stratifying the outcomes according to these variables, however, would enlarge the problem of sparseness of the data. It would be interesting to investigate whether the modelling equations can be extended to incorporate the coverage of the second MMR dose and the process of waning vaccine-induced immunity. The more realistic the model created to account for these two aspects, the more difficult it will be to identify the parameters from the serological data and to separate the effects from the first MMR dose coverage and the exposure probabilities. If the loss of vaccine-induced immunity would be studied, the effect of boosting of immunity by exposure to infection has to be considered as well (Rouderfer *et al.*, 1994).

The ad hoc method we used to deal with the parameter space restrictions of the Bahadur model, inhibited us to fully explore the dependency between the three exposure probabilities. The sensitivity analysis in Section 7.2.4, where we put the additional constraint of positive correlation, indicated good performance of BAH III for the Belgian data, which we had not inferred before since no valid solution was obtained under the 'unconstrained' optimization. For the Irish serology, the constraint of positive correlation did not enhance convergence towards valid ML-estimates. In the absence of analytical parameter bounds or a penalized likelihood approach, the construction of profile likelihood CIs, which would be a better alternative for the inverted Wald CIs (Agresti, 2002), is also precluded because of the problems with the Bahadur model. Therefore, we are currently exploring the use of the trivariate Dale model as a complement to the Bahadur model to describe the association between the natural acquisition probabilities for the three diseases.

Chapter 8

Discussion and Further Research

8.1 Summary of the Thesis

In this thesis, we explored diverse modelling methods for current status data and social contact data to enhance our understanding of the transmission of endemic or actively immunized infectious diseases which spread from person to person. In Chapter 4, we thoroughly studied the Belgian contact survey, collected as part of the POLYMOD project. The data mining analyses revealed that there are robust associations between general contact intimacy indicators, such as contacts taking place at home, lasting at least 4 hours, occurring on a daily basis, and involving skin-to-skin touching. The total number of reported contacts in the survey increased significantly with increasing household size and class size for children, and for adults who were employed or in further education, whereas it decreased significantly for children and teenagers during a school holiday period.

We proposed a semiparametric, bivariate smoothing approach to estimate contact rates from social contact survey data in Chapter 5, and found this method to outperform Wallinga *et al.* (2006)'s low dimensional, fully parametric maximum likelihood approach. Furthermore, the bivariate smoothing method revealed a common pattern in the contact surfaces for all countries in the POLYMOD project: individuals mostly mix assortatively i.e. with people of similar age, which also includes contact with a person's partner and siblings, and non-assortatively with (grand)children or (grand)parents, i.e. first-degree and second-degree relatives. However, there is still room for improvement as our generalized additive model did not directly take into account zero-inflation, digit preference or clustering of the contact counts, though the

latter two aspects were accounted for in the non-parametric bootstrap approach.

Further in Chapter 5, we estimated age-specific transmission rates for VZV in Belgium by augmenting the serological data with the estimated contact rates, hereby extending the work of Wallinga *et al.* (2006). An improvement of fit to the seroprevalence was obtained by modeling transmission as the product of two age-specific variables: the age-specific contact rate and an age-specific proportionality factor $q(a, a')$. Despite the fact that the social contact data approach tackles the main disadvantages of the traditional Anderson and May (1991) method, it still involves two dimensions of uncertainty: the choice of the type of contact underlying actual transmission of disease, and the choice of a parametric model relating the contact rates to the transmission rates. Focussing on close contacts lasting longer than 15 minutes, which induced the best fit to the VZV data under constant proportionality, different models for $q(a, a')$ resulted in a similar fit, while entailing different estimates of the basic reproduction number R_0. To overcome this problem of model selection uncertainty, we turned to multimodel inference and computed a model averaged estimate of R_0.

We conducted a compartmental model structure analysis in Chapter 6, to estimate basic immunological processes for PVB19, such as waning immunity, natural boosting of immunity and secondary infections, and to assess the impact on the inferred maternal risk. The social contact data approach revealed evidence towards long term processes of waning immunity for PVB19, however, it was difficult to discern from the current status data whether individuals with low immunity remain protected and can be boosted, or become susceptible again and potentially get reinfected. Our results showed that for four of the five European countries studied, model selection criteria favor the scenarios allowing for waning immunity at an age-specific rate over the assumption of lifelong immunity, assuming that the transmission rates are directly proportional to the contact rates. Different views on the evolution of the immune response to PVB19 infection led to altered estimates of the age-specific force of infection and R_0. The scenarios which allowed for multiple infections during one lifetime predicted a higher frequency of PVB19 infection in pregnant women and of associated fetal deaths.

Finally, in Chapter 7, we reviewed the work of Gay (2000) and Altmann and Altmann (2000) on the estimation of trivalent vaccination coverage from trivariate serological data. While the exact, algebraic method of Altmann and Altmann (2000) was found less interesting from a statistical point of view, we elaborated on Gay (2000)'s maximum likelihood approach by explicitly modelling the association between the probabilities of exposure to each of the three diseases for a non-vaccinated individual. To this purpose, the Bahadur model for trivariate binary data was used,

which produced a decrease in the estimated MMR vaccination coverage and an increase in the corresponding estimated variability when applied to the serology for Belgium and Ireland. Because of the restrictive Bahadur parameter space, we are currently exploring the trivariate Dale model as well.

8.2 Current Status Data

The use of current status data in infectious disease modelling relies on the correct classification of the antibody titers by the manufacturer's cut-off range for the ELISA test, with respect to the immunity status of the individual. However, not for all viruses we have studied in this thesis, a serological correlate of protection is agreed upon. For example, rubella antibody titers > 10 IU/ml are considered seroprotective according to the current consensus, whereas there is no International Standard for mumps sera (Tischer *et al.*, 2007). For measles, the antibody level which gives protection against infection is still under debate. Further, by dichotomizing the antibody results and excluding the equivocal samples, obviously some information is lost. Recently, new methods have emerged to interpret continuous antibody titers using mixture models (see e.g. Vyse *et al.*, 2006; Rota *et al.*, 2008), however, this should be done with caution since the fitted mixture components do not necessarily reflect prior known group structures in the data, such as different immunity states (e.g. susceptible, vaccinated or naturally infected). Furthermore, from a modelling point of view, the mixture-model methods for continuous antibody levels developed up till now are confined to estimating the seroprevalence and the force of infection, hereby implicitly assuming SIR (Bollaerts *et al.*, 2011). Whether the methods can be extended to establish a link with the mass action principle, to estimate age-dependent transmission rates from continuous titers, or with compartmental models for waning immunity as we considered for PVB19 in Chapter 6, needs yet to be explored.

For an endemic infectious disease with an SIR dynamics, one expects the seroprevalence to be monotonically increasing with age. However, in practice, the seroprofile may display distortions with respect to monotonicity, which can be visualized via a semi- or non-parametric estimate (cf. Section 3.1.3). According to Hens *et al.* (2010a), these distortions could be due to the presence of maternally derived immunity, a violation of time homogeneity, waning antibodies or plain randomness. Hens *et al.* (2010a) presented a practical flow chart with different remedial techniques to obtain a 'regularized' estimate of the force of infection. In all analyses presented here, we took into account the fact that newborns are initially protected by assuming 'type I maternal

antibodies'. For VZV and PVB19, we explicitly assumed endemic equilibrium, even though for both diseases there exist regular epidemic cycles. This violation very likely does not influence our results for VZV (Whitaker and Farrington, 2004a), however if immunity for PVB19 would truly wane over time, the same is not guaranteed there. The PVB19 seroprofile distortions in adults was our motivation in the first place to investigate the hypothesis of waning antibodies. That these distortions would be due to time heterogeneity seemed very unlikely, since the same pattern was observed for several countries in Europe, America, Asia and Australia, over at least two decades. There were also some distortions in infants, which seemed more likely due to PVB19 epidemics, but a sensitivity analysis showed that these were not influential.

Throughout this thesis, we experienced that there is a limit to what can be inferred from current status data. Serological surveys do not provide information related to infectiousness, which was one of the main problems with the traditional Anderson and May (1991) approach to estimate the WAIFW matrix. Thanks to the social contact data approach, the transmission rates can now be disentangled into two components and the (symmetric) age related contact heterogeneities can be estimated directly from social contact surveys. Consequently, the problem of indeterminacy is shifted to the age-specific proportionality factor $q(a, a')$. Therefore, it is important to assess the sensitivity of the results with regards to different assumptions for $q(a, a')$. Other factors which complicate inference from serological data are lack of data for certain age groups or extremely prevalent diseases such as VZV (cf. Chapter 5). The fact that we were not able to show improved performance of more complex immunological scenarios for PVB19, such as MSIRWb-ext, MSIRS-ext and MSIRWS, does not necessarily imply that the other, simpler models are closer to reality. We believe that these scenarios cannot be identified without the use of auxiliary data, which was supported by the bootstrap and simulation study results in Chapter 6. We suspect that something similar happens for the MMR vaccination coverage and the third order correlation between the MMR exposure probabilities, which are both informed by the occurrence of samples seropositive to all three infections (cf. Chapter 7).

Ideally, for the purpose of validation, the results obtained from serological data could be compared to results from other data sources such as incidence data or EPI-surveys. Auxiliary sources of data could also help to sustain certain modelling assumptions. Good quality age-specific incidence data, however, were not available to us.

8.3 Social Contact Data Approach

Although there is a positive trend in the sense that more time and thinking is spent on the survey design in recent social contact surveys with the aim to collect data to inform infectious disease transmission models (e.g. POLYMOD project), there are still some points which need continuous attention in the future. Apart from the fact that a deliberate decision needs to be made with respect to the timing of the survey, reporting of contacts by children and of professional contacts by adults need to be carefully considered. Up till now, parents were asked to fill in the contact diary for their young child, which is the most practicable method though it might lead to underreporting of contacts. Adults were either not given any instructions related to professional contacts or they were explicitly instructed not to record them in the diary in case the number would exceed a fixed threshold. We find the latter approach the most efficient provided that first, the ordering of the questions is similar to the one from the Belgian contact survey (cf. Section 4.1.2) and second, that additional age related information is required from the participant. In this way, one anticipates the problem of contact underreporting for adults who encounter many individuals professionally each day, and one can still estimate the age-specific contact rate surface by sampling at random from the recorded age distribution.

The per capita contact rates estimated from social contact surveys are constrained to be symmetric to take into account the reciprocal nature of contact events. Although we have not considered this, it might be interesting to measure the sample deviation from symmetry, like Wallinga *et al.* (2006), to assess underreporting of contacts or oversampling for specific age groups. For a contact survey from Utrecht, Wallinga *et al.* (2006) found that different age classes overall agree on the number of contacts occurring between them. Since the year of serological data collection will not always (perfectly) correspond to the timing of social contact surveys available, assumptions are needed to impose the symmetry constraint. By using age-specific population sizes obtained from demographical data corresponding to the year of serological data collection, we implicitly assumed that the mean number of contacts m_{ij} remained constant between the two data collection periods. In fact, because of the cohort interpretation of the serological data, the transmission models we considered in this thesis implicitly assume that the contact rates have been constant over several decades. We return to this issue in the next paragraph. Further, more research is needed about measures of (dis)assortativeness for contact patterns and the relation with their predictiveness for serological data. Farrington *et al.* (2009) proposed a summary measure of disassortativeness, i.e. the extent to which contacts occur between individuals from different

age groups, and found that this index was remarkably constant across the contact patterns of countries involved in the POLYMOD survey.

Recently, another large social contact survey was conducted in Indonesia, Taiwan, Thailand and Vietnam as part of the WHO-project on 'Influenza illness and vaccination in Asia: data collection of social contacts and mixing patterns' (SMILI). Not all countries, however, have the means or interest to conduct social contact surveys. In addition, contact data are cross-sectional i.e. a snapshot of social mixing patterns, which are likely to change over time due to changing relationships between humans and their social and physical environments. It is therefore of interest to parameterize the contact surface (e.g. using POLYMOD data) and to develop a sort of generic model for contact rates, which can be informed by longitudinal, demographic data for a specific country. The aim is to derive a flexible, though parsimonious parametric structure which captures the most important features of the contact surface, by incorporating variables such as the average age of becoming a parent, the duration of compulsory education, the average age at retirement, etc. To the same purpose, Italian colleagues are currently working on so-called 'synthetic contact matrices', which are reconstructed from individual-based simulation models for social-demographic dynamics (Iozzi *et al.*, 2009).

Finally, we would like to recall some of the issues inherent to the mass action principle formulated in (2.12), which were also discussed by Ogunjimi *et al.* (2009). The mass action principle allows that individuals make contact with each other at different rates depending on their age, but the contacts can in principle occur with whomever in the population (randomly distributed). Hence, the mass action concept assumes that there does not exist a predefined relationship or a causal circumstance by which a certain pair of individuals is more likely to interact with each other than another pair of individuals having the same age. It would be interesting to investigate whether the mass action principle (2.12) can be extended to incorporate clustering of contacts, perhaps by borrowing concepts from network theory. On the other hand, social contact data may also be useful to inform network or individual-based models, which are used to study the epidemic spread of infectious diseases and the impact of control measures. Further, the mass action principle (2.12) implicitly assumes that an infected individual's social mixing behavior does not change during the infectious period. However, when people are seriously ill they are likely to adapt their normal activities and have fewer contacts which mostly occur within the household (Eames *et al.*, 2010). Whether this is relevant for the estimation of the WAIFW matrix and R_0, depends on the latency and incubation period of the infectious disease considered, i.e. whether the clinical symptoms occur before, during or after the infectious period.

Bibliography

Agrawal, R., Imielinski, T. and Swami, A. (1993) Mining association rules between sets of items in large databases. In: *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, 207216. ACM Press, New York, NY, USA.

Agresti, A. (2002) *Categorical Data Analysis, Second Edition.* Wiley Series in Probability and Statistics.

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In: *Proceedings of the Second International Symposium on Information Theory* (Eds. B. N. Petrov and F. Csaki), 267–281. Akademiai Kiado, Budapest.

Alanen, A., Kahala, K., Vahlberg, T., Koskela, P. and Vainionpää, R. (2005) Seroprevalence, incidence of prenatal infections and reliability of maternal history of varicella zoster virus, cytomegalovirus, herpes simplex virus and parvovirus B19 infection in South-Western Finland. *BJOG: an International Journal of Obstetrics and Gynaecology*, **112**, 50–56.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P. (2002) *Molecular Biology of the Cell, 4th edition.* New York: Garland Science.

Altmann, D. and Altmann, K. (2000) Estimating vaccine coverage by using computer algebra. *IMA Journal of Mathematics Applied in Medicine and Biology*, **17**, 137–146.

Anderson, M. J. and Cherry, J. D. (2004) *Textbook of Pediatric Infectious Diseases*, chap. 17, 1796. Saunders, Philadelphia, Pa.

Anderson, R. and May, R. (1990) Modern vaccines: Immunisation and herd immunity. *Lancet*, **335**, 641–645.

Anderson, R. M. and May, R. M. (1991) *Infectious Diseases of Humans: Dynamics and Control*. Oxford: Oxford University Press.

Andersson, H. and Britton, T. (2000) *Stochastic Epidemic Models and their Statistical Analysis*. Springer: New York.

Baguelin, M., Hoek, A. J., Jit, M., Flasche, S., White, P. J. and Edmunds, W. J. (2010) Vaccination against pandemic influenza A/H1N1v in England: a real-time economic evaluation. *Vaccine*, **28**, 2370–2384.

Bahadur, R. (1961) *Studies in Item Analysis and Prediction*, chap. A representation of the joint distribution of responses to n dichotomous items. Stanford University Press.

Bailey, N. J. T. (1975) *The Mathematical Theory of Infectious Diseases and its Application*. Griffin, London.

Bansal, S., Grenfell, B. T. and Meyers, L. A. (2007) When individual behaviour matters: homogeneous and network models in epidemiology. *Journal of the Royal Society Interface*, **4**, 879–891.

Becker, N. G. (1989) *Analysis of Infectious Disease Data*. Chapman and Hall/CRC.

Bell, L. M., Naides, S. J., Stoffman, P., Hodinka, R. L. and Plotkin, S. A. (1989) Human parvovirus B19 infection among hospital staff members after contact with infected patients. *The New England Journal of Medicine*, **321**, 485–491.

Bernoulli, D. (1760) Essai dune nouvelle analyse de la mortalité causée par la petite vérole et des avantages de l'inoculation pour la prévenir. In: *Mémoires de l'Académie Royale des Sciences, Paris (1766)*.

Beutels, P., Shkedy, Z., Aerts, M. and Van Damme, P. (2006) Social mixing patterns for transmission models of close contact infections: exploring self-evaluation and diary-based data collection through a web-based interface. *Epidemiology and Infection*, **134**, 1158–1166.

Bollaerts, K., Aerts, M., Hens, N., Shkedy, Z., Faes, C., Van der Stede, Y. and Beutels, P. (2011) Estimating the population prevalence and force of infection directly from antibody titers. *Tech. rep.*, Hasselt University.

Boonen, M., Theeten, H., Vandermeulen, C., Roelants, M., Depoorter, A., Van Damme, P. and Hoppenbrouwers, K. (2009) Vaccinatiegraad bij jonge kinderen en adolescenten in Vlaanderen in 2008. *Vlaams Infectieziektebulletin*, **68**, 9–14.

Bosman, A., Wallinga, J. and Kroes, A. C. M. (2002) Elke vier jaar de vijfde ziekte: parvovirus B19. *Infectieziekten Bulletin*, **6**, 215–219.

van Boven, M., de Melker, H., Schellekens, J. and Kretzschmar, M. (2000) Waning immunity and sub-clinical infection in an epidemic model: implications for pertussis in The Netherlands. *Mathematical Biosciences*, **164**, 161–182.

van Boven, M., de Melker, H. E., Schellekens, J. F. P. and Kretzschmar, M. (2001) A model based evaluation of the 1996-7 pertussis epidemic in the Netherlands. *Epidemiology and Infection*, **127**, 73–85.

Breiman, L. (2001) Random forests. *Machine Learning*, **45**, 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. (1984) *Classification and Regression Trees*. Wadsworth International Group.

Briss, P. A., Fehrs, L. J., Parker, R. A., Wright, P. F., Sannella, E. C., Hutcheson, R. H. and Schaffner, W. (1994) Sustained transmission of mumps in a highly vaccinated population - assessment of primary vaccine failure and waning vaccined-induced immunity. *The Journal of Infectious Diseases*, **169**, 77–82.

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

Brisson, M., Gay, N., Edmunds, W. J. and Andrews, N. (2002) Exposure to varicella boosts immunity to herpes-zoster: implications for mass vaccination against chickenpox. *Vaccine*, **20**, 2500–2507.

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

Cartter, M. L., Farley, T., Rosengren, S., Quinn, D. L., Gillespie, S. M., Gary, G. W. and Hadler, J. L. (1991) Occupational risk factors for infection with parvovirus B19 among pregnant women. *The Journal of Infectious Diseases*, **163**, 282–285.

Cauchemez, S., Carrat, F., Viboud, C., Valleron, A. J. and Boëlle, P. Y. (2004) A Bayesian MCMC approach to study transmission of influenza: application to household longitudinal data. *Statistics in Medicine*, **23**, 3469–3487.

Cauchemez, S., Ferguson, N. M., Wachtel, C., Tegnell, A., Saour, G., Duncan, B. and Nicoll, A. (2009) Closure of schools during an influenza pandemic. *The Lancet Infectious Diseases*, **9**, 473–481.

Cauchemez, S., Valleron, A. J., Boëlle, P., Flahault, A. and Ferguson, N. M. (2008) Estimating the impact of school closure on influenza transmission from sentinel data. *Nature*, **452**, 750–754.

Chorba, T., Coccia, P., Holman, R. C., Tattersall, P., Anderson, L. J., Sudman, J., Young, N. S., Kurczynski, E., Saarinen, U. M., Moir, R. *et al.* (1986) The role of parvovirus B19 in aplastic crisis and erythema infectiosum (fifth disease). *The Journal of Infectious Diseases*, **154**, 383–393.

Cochi, S. L., Wharton, M. and Plotkin, S. A. (1994) *Vaccines*, chap. Mumps Vaccine, 277–301. Saunders, Philadelphia, Pa.

Cohen, B. (1995) Parvovirus B19: an expanding spectrum of disease. *British Medical Journal*, **311**, 1549–1552.

Cohen, B. J. and Buckley, M. M. (1988) The prevalence of antibody to human parvovirus B19 in England and Wales. *Journal of Medical Microbiology*, **25**, 151–153.

Cox, D. R. (1970) *Analysis of Binary Data*. Methuen, London.

Daley, D. J. and Gani, J. (1999) *Epidemic Modelling: An Introduction*. Cambridge University Press.

Davidkin, I., Jokinen, S., Broman, M., Leinikki, P. and Peltola, H. (2008) Persistence of measles, mumps, and rubella antibodies in an MMR-vaccinated cohort: a 20-year follow-up. *The Journal of Infectious Diseases*, **197**, 950–956.

De'ath, G. (2002) Multivariate regression trees : A new technique for modeling species-environment relationships. *Ecology*, **83**, 1105–1117.

Declerck, L., Aerts, M. and Molenberghs, G. (1998) Behaviour of the likelihood ratio test statistic under a Bahadur model for exchangeable binary data. *Journal of Statistical Computation and Simulation*, **61**, 15–38.

Del Valle, S. Y., Hyman, J. M., Hethcote, H. W. and Eubank, S. G. (2007) Mixing patterns between age groups in social networks. *Social Networks*, **29**, 539–554.

Devlin, T. and Weeks, B. (1986) Spline functions for logistic regression modeling. In: *Proceedings of the Eleventh Annual SAS Users Group International* (Ed. C. N. S. Institute), 646–651.

Diekmann, O., Heesterbeek, J. A. P. and Metz, J. A. J. (1990) On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations. *Journal of Mathematical Biology*, **28**, 365–382.

Dietz, K. (1975) Transmission and control of arbovirus diseases. In: *Epidemiology, SIMS Utah Conference Proceedings* (Eds. D. Ludwig and K. L. Cooke), 104–121. SIAM, Philadelphia.

Dietz, K. (1993) The estimation of the basic reproduction number for infectious diseases. *Statistical Methods in Medical Research*, **2**, 23–41.

Eames, K., Tilston, N., White, P., Adams, E. and Edmunds, W. J. (2010) The impact of illness and the impact of school closure on social contact patterns. *Health Technology Assessment*, **14**, 267–312.

Edmunds, W. J., Kafatos, G., Wallinga, J. and Mossong, J. R. (2006) Mixing patterns and the spread of close-contact infectious diseases. *Emerging Themes in Epidemiology*, **3**.

Edmunds, W. J., O'Callaghan, C. J. and Nokes, D. J. (1997) Who mixes with whom? A method to determine the contact patterns of adults that may lead to the spread of airborne infections. *Proceedings of the Royal Society B: Biological Sciences*, **264**, 949–957.

Efron, B. (1979) Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, **7**, 1–26.

Efron, B. and Tibshirani, R. (1993) *An introduction to the Bootstrap*. Chapman and Hall/CRC.

Eis-Hübinger, A. M., Dieck, D., Schild, R., Hansmann, M. and Schneweis, K. E. (1998) Parvovirus B19 infection in pregnancy. *Intervirology*, **41**, 178–184.

Enders, G., Miller, E., Cradock-Watson, J., Bolley, I. and Ridehalgh, M. (1994) Consequences of varicella and herpes zoster in pregnancy: prospective study of 1739 cases. *Lancet*, **343**, 1548–1551.

Enders, M., Weidner, A., Zoellner, I., Searle, K. and Enders, G. (2004) Fetal morbidity and mortality after acute human parvovirus B19 infection in pregnancy: prospective evaluation of 1018 cases. *Prenatal Diagnosis*, **24**, 513–518.

Eurostat (2007) Population table for Belgium, 2003. http://epp.eurostat.ec.europa.eu/.

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

Farrington, C., Whitaker, H., Wallinga, J. and Manfredi, P. (2009) Measures of disassortativeness and their application to directly transmitted infections. *Biometrical Journal*, **51**, 387–407.

Farrington, C. P. (2003) *Modelling Epidemics*. The Open University.

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. *Applied Statistics*, **50**, 251–292.

Farrington, C. P. and Whitaker, H. J. (2005) Contact surface models for infectious diseases: estimation from serologic survey data. *Journal of the American Statistical Association*, **100**, 370–379.

Ferguson, N. M., Anderson, R. M. and Garnett, G. P. (1996) Mass vaccination to control chickenpox: the influence of zoster. *Proceedings of the National Academy of Science USA*, **93**, 7231–7235.

FOD Economie Afdeling Statistiek (2006) Levensverwachting bij de geboorte, per gewest en internationale vergelijking. http://mineco.fgov.be/.

Garnett, G. P. and Grenfell, B. T. (1992) The epidemiology of varicella-zoster virus infections: a mathematical model. *Epidemiology and Infection*, **108**, 495–511.

Gay, N. (2000) A method for estimating coverage of a multivalent vaccine from antibody prevalence data: application to MMR vaccine in 3 European countries. Unpublished manuscript.

Gay, N. J. (1996) Analysis of serological surveys using mixture models: application to a survey of parvovirus B19. *Statistics in Medicine*, **15**, 1567–1573.

Gay, N. J., Hesketh, L. M., Cohan, B. J., Rush, M., Bates, C., Morgan-Capner, P. and Miller, E. (1994) Age specific antibody prevalence to parvovirus B19: how many women are infected in pregnancy? *Communicable Disease Report*, **4**, R104–107.

van Gessel, P. H., Gaytant, M. A., Vossen, A. C. T. M., Galama, J. M. D., Ursem, N. T. C., Steegers, E. A. P. and Wildschut, H. I. J. (2006) Incidence of parvovirus B19 infection among an unselected population of pregnant women in the Netherlands: A prospective study. *European Journal of Obstetrics & Gynecology and Reproductive Biology*, **128**, 46–49.

Gillespie, S. M., Cartter, M. L., Asch, S., Rokos, J. B., Gary, G. W., Tsou, C. J., Hall, D. B., Anderson, L. J. and Hurwitz, E. S. (1990) Occupational risk of human parvovirus B19 infection for school and day-care personnel during an outbreak of erythema infectiosum. *Journal of the American Medical Association*, **263**, 2061–2065.

Goeyvaerts, N., Hens, N., Aerts, M. and Beutels, P. (2010b) Model structure analysis to estimate basic immunological processes and maternal risk for parvovirus B19. *Biostatistics*, doi: 10.1093/biostatistics/kxq059.

Goeyvaerts, N., Hens, N., Ogunjimi, B., Aerts, M., Shkedy, Z., Van Damme, P. and Beutels, P. (2010a) Estimating infectious disease parameters from data on social contacts and serological status. *Applied Statistics*, **59**, 255–277.

Goeyvaerts, N., Hens, N., Theeten, H., Aerts, M., Van Damme, P. and Beutels, P. (2011) Estimating vaccination coverage for the trivalent measles-mumps-rubella vaccine from trivariate current status data. *Tech. rep.*, Hasselt University.

Gonçalves, G., Correia, A. M., Palminha, P., Rebelo de Andrade, H. and Alves, A. (2005) Outbreaks caused by parvovirus B19 in three Portuguese schools. *Eurosurveillance*, **10**, pii=549.

Greenhalgh, D. and Dietz, K. (1994) Some bounds on estimates for reproductive ratios derived from the age-specific force of infection. *Mathematical Biosciences*, **124**, 9–57.

Grenfell, B. T. and Anderson, R. M. (1985) The estimation of age-related rates of infection from case notifications and serological data. *Journal of Hygiene*, **95**, 19–36.

Hahsler, M., Grün, B. and Hornik, K. (2005) arules - a computational environment for mining association rules and frequent item sets. *Journal of Statistical Software*, **14**, 1–25.

Hahsler, M., Grün, B. and Hornik, K. U. (2006) arules: Mining association rules and frequent itemsets. URL `http://cran.r-project.org/`. R package version 0.4-3.

Hahsler, M. and Hornik, K. (2007) New probabilistic interest measures for association rules. *Intelligent Data Analysis*, **11**, 437–455.

Halloran, M. E. (2006) Invited commentary: Challenges of using contact data to understand acute respiratory disease transmission. *American Journal of Epidemiology*, **164**, 945–946.

Halloran, M. E., Cochi, S. L., Lieu, T. A., Wharton, M. and Fehrs, L. (1994) Theoretical epidemiologic and morbidity effects of routine varicella immunization of preschool children in the United States. *American Journal of Epidemiology*, **140**, 81–104.

Halloran, M. E., Ferguson, N. M., Eubank, S., Longini, Jr., I. M., Cummings, D. A. T., Lewis, B., Xu, S., Fraser, C., Vullikanti, A., Germann, T. C., Wagener, D., Beckman, R., Kadau, K., Barrett, C., Macken, C. A., Burke, D. S. and Cooley, P. (2008) Modeling targeted layered containment of an influenza pandemic in the United States. *PNAS*, **105**, 4629–4644.

Harrell, F. (2001) *Regression modelling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer Series in Statistics, New York, N.Y.

Hastie, T., Tibshirani, R. and Friedman, J. (2001) *The Elements of Statistical Learning*. Springer Series in Statistics, New York, N.Y.

Hastie, T. J. and Tibshirani, R. J. (1990) *Generalized Additive Models*. Chapman and Hall, London.

Heegaard, E. D. and Brown, K. E. (2002) Human parvovirus B19. *Clinical Microbiology Reviews*, **15**, 485–505.

Heesterbeek, H. (2002) A brief history of R_0 and a recipe for its calculation. *Acta Biotheoretica*, **50**, 189–204.

Hens, N., Aerts, M., Faes, C., Shkedy, Z., Lejeune, O., Van Damme, P. and Beutels, P. (2010a) Seventy-five years of estimating the force of infection from current status data. *Epidemiology and Infection*, **138**, 802–812.

Hens, N., Aerts, M., Shkedy, Z., Theeten, H., Van Damme, P. and Beutels, P. (2008) Modelling multisera data: the estimation of new joint and conditional epidemiological parameters. *Statistics in Medicine*, **27**, 2651–2664.

Hens, N., Goeyvaerts, N., Aerts, M., Shkedy, Z., Van Damme, P. and Beutels, P. (2009a) Mining social mixing patterns for infectious disease models based on a two-day population survey in Belgium. *BMC Infectious Diseases*, **9**, 5.

Hens, N., Kvitkovicova, A., Aerts, M., Hlubinka, D. and Beutels, P. (2010b) Modelling distortions in seroprevalence data using change-point fractional polynomials. *Statistical Modelling*, **10**, 159–175.

Hens, N., Minalu Ayele, G., Goeyvaerts, N., Aerts, M., Mossong, J., Edmunds, J. W. and Beutels, P. (2009b) Estimating the impact of school closure on social mixing behaviour and the transmission of close contact infections in eight European countries. *BMC Infectious Diseases*, **9**, 187.

Hens, N., Shkedy, Z., Aerts, M., Faes, C., Van Damme, P. and Beutels, P. (2011) *Modeling Infectious Disease Parameters Based on Serological and Social Contact Data: a Modern Statistical Perspective*. Springer-Verlag New York Inc. Forthcoming.

Hill, A. N. and Longini, Jr., I. M. (2003) The critical vaccination fraction for heterogeneous epidemic models. *Mathematical Biosciences*, **181**, 85–106.

House, T., Baguelin, M., Van Hoek, A., White, P. J., Sadique, Z., Eames, K., Read, J. M., Hens, N., Melegaro, A., Edmunds, W. J. and Keeling, M. (2010) Modelling the impact of local, reactive school closures on critical care provision during an influenza pandemic. *Proceedings of the Royal Society B: Biological Sciences*, submitted.

Huatuco, E. M. M., Durigon, E. L., Lebrun, F. L. A. S., Passos, S. D., Gazeta, R. E., Azevedo Neto, R. S. and Massad, E. (2008) Seroprevalence of human parvovirus B19 in a suburban population in Sao Paulo, Brazil. *Revista de Saúde Pública*, **42**, 443–449.

Iozzi, F., Chinazzi, M., Trusiano, F., Manfredi, P., Billari, F. and Zagheni, E. (2009) Little-Italy: An agent-based approach to the estimation of contact data.

In: *Epidemics*^{2}*, Second International Conference on Infectious Disease Dynamics, Athens Greece, Abstract Book*.

Kafatos, G., Andrews, N. and Nardone, A. (2005) Model selection methodology for inter-laboratory standardisation of antibody titres. *Vaccine*, **23**, 5022–5027.

Kanaan, M. N. and Farrington, C. P. (2005) Matrix models for childhood infections: a Bayesian approach with applications to rubella and mumps. *Epidemiology and Infection*, **133**, 1009–1021.

Kaufmann, J., Buccola, J. M., Stead, W., Rowley, C., Wong, M. and Bates, C. K. (2007) Secondary symptomatic parvovirus B19 infection in a healthy adult. *Journal of General Internal Medicine*, **22**, 877–878.

Keeling, M. J. and Eames, K. T. D. (2005) Networks and epidemic models. *Journal of the Royal Society Interface*, **2**, 295–307.

Kelly, H. A., Siebert, R., Hammond, R., Leydon, J. and Maskill, W. (2000) The age-specific prevalence of human parvovirus immunity in Victoria, Australia compared with other parts of the world. *Epidemiology and Infection*, **124**, 449–457.

Keogh-Brown, M. R., Smith, R. D., Edmunds, W. J. and Beutels, P. (2010) The macroeconomic impact of pandemic influenza: estimates from models of the United Kingdom, France, Belgium and The Netherlands. *The European Journal of Health Economics*, **11**, 543–554.

Kermack, W. O. and McKendrick, A. G. (1927) A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society London A*, **115**, 700–721.

Liang, K. and Zeger, S. (1986) Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.

Little, R. and Rubin, D. (1987) *Statistical Analysis with Missing Data*. New York: John Wiley & Sons, Inc.

Longini, Jr., I. M., Nizam, A., Xu, S., Ungchusak, K., Hanshaoworakul, W., Cummings, D. A. T. and Halloran, M. E. (2005) Containing pandemic influenza at the source. *Science*, **309**, 1083–1087.

Mammen, E., Marron, J. S., Turlach, B. A. and Wand, M. P. (2001) A general projection framework for constrained smoothing. *Statistical Science*, **16**, 232–248.

Matsunaga, Y., Takeda, N., Yamazaki, S., Kamata, K. and Kurosawa, D. (1995) Seroepidemiology of human parvovirus B19 using recombinant VP1+VP2 particle antigen. *Kansenshogaku Zasshi*, **69**, 1371–1375.

McCaw, J. M., Forbes, K., Nathan, P. M., Pattison, P. E., Robins, G. L., Nolan, T. M. and McVernon, J. (2010) Comparison of three methods for ascertainment of contact information relevant to respiratory pathogen transmission in encounter networks. *BMC Infectious Diseases*, **10**, 166.

Melegaro, A., Jit, M., Gay, N., Zagheni, E. and Edmunds, W. J. (2010) What types of contacts are important for the spread of infections? Using contact survey data to explore European mixing patterns. *Tech. rep.*, Health Protection Agency, Centre for Infections, London, UK.

Mikolajczyk, R. T., Akmatov, M. K., Rastin, S. and Kretzschmar, M. (2008) Social contacts of school children and the transmission of respiratory-spread pathogens. *Epidemiology and Infection*, **136**, 813–822.

Miller, E., Fairley, C. K., Cohen, B. J. and Seng, C. (1998) Immediate and long term outcome of human parvovirus B19 infection in pregnancy. *British Journal of Obstetrics and Gynaecology*, **105**, 174–178.

Mills, C. E., Robins, J. M. and Lipsitch, M. (2004) Transmissibility of 1918 pandemic influenza. *Nature*, **432**, 904–906.

Molenberghs, G. and Verbeke, G. (2005) *Models for Discrete Longitudinal Data*. Springer Series in Statistics, New York, N.Y.

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. (2008a) Parvovirus B19 infection in five European countries: seroepidemiology, force of infection and maternal risk of infection. *Epidemiology and Infection*, **136**, 1059–1068.

Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K. *et al.* (2008b) Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Medicine*, **5(3)**, 381–391.

Muench, H. (1934) Derivation of rates from summation data by the catalytic curve. *Journal of the American Statistical Association*, **29**, 25–38.

Muscat, M., Bang, H., Wohlfahrt, J., Glismann, S. and Mølbak, K. (2009) Measles in Europe: an epidemiological assessment. *Lancet*, **373**, 383–389.

Namata, H., Shkedy, Z., Faes, C., Aerts, M., Molenberghs, G. and Theeten, H. (2007) Estimation of the force of infection from currents status data using generalized linear mixed models. *Journal of Applied Statistics*, **34**, 923–939.

Nardone, A., de Ory, F., Carton, M., Cohen, D., van Damme, P., Davidkin, I., Rota, M. C. *et al.* (2007) The comparative sero-epidemiology of varicella zoster virus in 11 countries in the European region. *Vaccine*, **25**, 7866–7872.

Nascimento, J. P., Buckley, M. M., Brown, K. E. and Cohen, B. J. (1990) The prevalence of antibody to human parvovirus B19 in Rio De Janeiro, Brazil. *Revista do Instituto de Medicina Tropical de Sao Paulo*, **32**, 41–45.

Nunoue, T., Okochi, K., Mortimer, P. P. and Cohen, B. J. (1985) Human parvovirus (B19) and erythema infectiosum. *The Journal of Pediatrics*, **107**, 38–40.

O'Flanagan, D. and others as the 'Measles and Rubella Elimination Committee of the Department of Health and Children' (2007) Eliminating measles and rubella and preventing congenital rubella infection. Health Protection Surveillance Centre, Dublin, Ireland. URL `http://www.hpsc.ie/`.

Ogunjimi, B., Hens, N., Goeyvaerts, N., Aerts, M., Van Damme, P. and Beutels, P. (2009) Using empirical social contact data to model person to person infectious disease transmission: an illustration for varicella. *Mathematical Biosciences*, **218**, 80–87.

O'Neill, P. D. (2010) Introduction and snapshot review: Relating infectious disease transmission models to data. *Statistics in Medicine*, **29**, 2069–2077.

de Ory, F., Echevarria, J. M., Kafatos, G., Anastassopoulou, C., Andrews, N., Backhouse, J. *et al.* (2006) European seroepidemiology network 2: standardisation of assays for seroepidemiology of varicella zoster virus. *Journal of Clinical Virology*, **36**, 111–118.

Pastuszak, A. L., Levy, M., Schick, B., Zuber, C., Feldkamp, M., Gladstone, J., Bar-Levy, F., Jackson, E., Donnenfeld, A., Meschino, W. and Koren, G. (1994) Outcome after maternal varicella infection in the first 20 weeks of pregnancy. *The New England Journal of Medicine*, **330**, 901–905.

Pebody, R. G., Gay, N. J., Hesketh, L. M., Vyse, A., Morgan-Capner, P., Brown, D., Litton, P. and Miller, E. (2002) Immunogenicity of second dose measles-mumps-rubella (MMR) vaccine and implications for serosurveillance. *Vaccine*, **20**, 1134–1140.

Pickery, J. and Carton, A. (2005) Hoe representatief zijn telefonische surveys in Vlaanderen? (how representative are telephone surveys in Flanders?). URL `http://www4.vlaanderen.be/dar/svr/publicaties/Publicaties/nota/2005-03_telefoon.pdf`.

Pillay, D., Patou, G., Hurt, S., Kibbler, C. C. and Griffiths, P. D. (1992) Parvovirus B19 outbreak in a children's ward. *Lancet*, **339**, 107–109.

Potter, G. E., Handcock, M. S., Longini, Jr., I. M. and Halloran, M. E. (2011) Estimating within-household contact networks from egocentric data. *submitted*.

Rice, P. S. and Cohen, B. J. (1996) A school outbreak of parvovirus B19 infection investigated using salivary antibody assays. *Epidemiology and Infection*, **116**, 331–338.

Riipinen, A., Väisänen, E., Nuutila, M., Sallmen, M., Karikoski, R., Lindbohm, M.-L., Hedman, K., Taskinen, H. and Söderlund-Venermo, M. (2008) Parvovirus B19 infection in fetal deaths. *Clinical Infectious Diseases*, **47**, 1519–1525.

Rota, M. C., Massari, M., Gabutti, G., Guido, M., De Donno, A. and Ciofi degli Atti, M. L. (2008) Measles serological survey in the Italian population: Interpretation of results using mixture model. *Vaccine*, **26**, 4403–4409.

Rouderfer, V., Becker, N. and Hethcote, H. (1994) Waning immunity and its effects on vaccination schedules. *Mathematical Biosciences*, **124**, 59–82.

Sadique, M. Z., Adams, E. J. and Edmunds, W. J. (2008) Estimating the costs of school closure for mitigating an influenza pandemic. *BMC Public Health*, **8**, 135.

Schneider, B., Höne, A., Tolba, R. H., Fischer, H. P., Blümel, J. and Eis-Hübinger, A. M. (2008) Simultaneous persistence of multiple genome variants of human parvovirus B19. *Journal of General Virology*, **89**, 164–176.

Schoub, B. D., Blackburn, N. K., Johnson, S. and McAnerney, J. M. (1993) Primary and secondary infection with human parvovirus B19 in pregnant women in South Africa. *South African Medical Journal*, **83**, 505–506.

Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Schwarz, T. F., Roggendorf, M. and Deinhardt, F. (1987) Hufigkeit der Parvoviras-B19-Infektionen: Seroepidemiologische Untersuchungen. *Deutsche Medicinische Wochenschrift*, **112**, 1526–1531.

Self, S. G. and Liang, K.-Y. (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, **82**, 605–610.

Shkedy, Z., Aerts, M., Molenberghs, G., Beutels, P. and Van Damme, P. (2003) Modelling forces of infection by using monotone local polynomials. *Applied Statistics*, **52**, 469–485.

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

Smith, R. D., Keogh-Brown, M. R., Barnett, T. and Tait, J. (2009) The economy-wide impact of pandemic influenza on the UK: a computable general equilibrium modelling experiment. *BMJ*, **339**, b4571.

Smithson, R., Irvine, N., Hutton, C., Doherty, L. and Watt, A. (2010) Spotlight on measles 2010: Ongoing measles outbreak in Northern Ireland following an imported case, September-October 2010. *Eurosurveillance*, **15**, pii=19698.

Swennen, B., Van Damme, P., Vellinga, A., Coppieters, Y. and Depoorter, A. (2002) Analysis of factors influencing vaccine uptake: perspectives from Belgium. *Vaccine*, **20**, S5–S7.

Teunis, P. F. M., Van der Heijden, O. G., de Melker, H. E., Schellekens, J. F. P., Versteegh, F. G. A. and Kretzschmar, M. E. E. (2002) Kinetics of the IgG antibody response to pertussis toxin after infection with B. pertussis. *Epidemiology and Infection*, **129**, 479–489.

Theeten, H., Hens, N., Vandermeulen, C., Depoorter, A., Roelants, M., Aerts, M., Hoppenbrouwers, K. and Van Damme, P. (2007) Infant vaccination coverage in 2005 and predictive factors for complete or valid vaccination in Flanders, Belgium: an EPI-survey. *Vaccine*, **25**, 4940–4948.

Theeten, H., Vandermeulen, C., Roelants, M., Hoppenbrouwers, K., Depoorter, A. and Van Damme, P. (2009) Coverage of recommended vaccines in children at 7-8 years of age in Flanders, Belgium. *Acta Paediatrica*, **98**, 1307–1312.

Thiry, N., Beutels, P., Shkedy, Z., Vranckx, R., Vandermeulen, C., Wielen, M. V. and Van Damme, P. (2002) The seroepidemiology of primary varicella-zoster virus infection in Flanders (Belgium). *European Journal of Pediatrics*, **161**, 588–593.

Thomas, S. L., Wheeler, J. G. and Hall, A. J. (2002) Contacts with varicella or with children and protection against herpes zoster in adults: a case-control study. *Lancet*, **360**, 678–682.

Tischer, A., Andrews, N., Kafatos, G., Nardone, A., Berbers, G., Davidkin, I., Aboudy, Y., Backhouse, J., Barbara, C., Bartha, K., Bruckova, B., Duks, A., Griskevicius, A., Hesketh, L., Johansen, K., Jones, L., Kuersteiner, O., Lupulescu, E., Mihneva, Z., Mrazova, M., De Ory, F., Prosenc, K., Schneider, F., Tsakris, A., Smelhausova, M., Vranckx, R., Zarvou, M. and Miller, E. (2007) Standardization of measles, mumps and rubella assays to enable comparisons of seroprevalence data across 21 European countries and Australia. *Epidemiology and Infection*, **135**, 787–797.

Tolfvenstam, T., Papadogiannakis, N., Norbeck, O., Petersson, K. and Broliden, K. (2001) Frequency of human parvovirus B19 in intrauterine fetal death. *Lancet*, **357**, 1494–1497.

Valeur-Jensen, A. K., Pedersen, C. B., Westergaard, T., Jensen, I. P., Lebech, M., Andersen, P. K., Aaby, P., Pedersen, B. N. and Melbye, M. (1999) Risk factors for parvovirus B19 infection in pregnancy. *Journal of the American Medical Association*, **281**, 1099–1105.

Van Effelterre, T., Shkedy, Z., Aerts, M., Molenberghs, G., Van Damme, P. and Beutels, P. (2009) Contact patterns and their implied basic reproductive numbers: an illustration for varicella-zoster virus. *Epidemiology and Infection*, **137**, 48–57.

Vandermeulen, C., Roelants, M., Theeten, H., Depoorter, A., Van Damme, P. and Hoppenbrouwers, K. (2008) Vaccination coverage in 14-year-old adolescents: Documentation, timeliness, and sociodemographic determinants. *Pediatrics*, **121**, e428–e434.

Vandermeulen, C., Roelants, M., Vermoere, M., Roseeuw, K., Goubau, P. and Hoppenbrouwers, K. (2004) Outbreak of mumps in a vaccinated child population: a question of vaccine failure? *Vaccine*, **22**, 2713–2716.

Vyse, A. J., Andrews, N. J., Hesketh, L. M. and Pebody, R. (2007) The burden of parvovirus B19 infection in women of childbearing age in England and Wales. *Epidemiology and Infection*, **135**, 1354–1362.

Vyse, A. J., Gay, N. J., Hesketh, L. M., Pebody, R., Morgan-Capner, P. and Miller, E. (2006) Interpreting serological surveys using mixture models: the seroepidemiology of measles, mumps and rubella in England and Wales at the beginning of the 21st century. *Epidemiology and Infection*, **134**, 1303–1312.

Wallinga, J., Teunis, P. and Kretzschmar, M. (2006) Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents. *American Journal of Epidemiology*, **164**, 936–944.

Whitaker, H. J. and Farrington, C. P. (2004a) Estimation of infectious disease parameters from serological survey data: the impact of regular epidemics. *Statistics in Medicine*, **23**, 2429–2443.

Whitaker, H. J. and Farrington, C. P. (2004b) Infections with varying contact rates: application to varicella. *Biometrics*, **60**, 615–623.

Wood, S. N. (2006) *Generalized Additive Models: an Introduction with R*. Chapman and Hall/CRC Press.

Woolf, A. D., Campion, G. V., Chishick, A., Wise, S., Cohen, B. J., Klouda, P. T., Caul, O. and Dieppe, P. A. (1989) Clinical manifestations of human parvovirus B19 in adults. *Archives of Internal Medicine*, **149(5)**, 1153–1156.

Yang, Y., Halloran, M. E., Sugimoto, J. D. and Longini, Jr., I. M. (2007) Detecting human-to-human transmission of avian influenza A (H5N1). *Emerging Infectious Diseases*, **13**, 1348–1353.

Young, N. S. and Brown, K. E. (2004) Parvovirus B19. *The New England Journal of Medicine*, **350**, 586–597.

Zaaijer, H. L., Koppelman, M. H. G. M. and Farrington, C. P. (2004) Parvovirus B19 viraemia in Dutch blood donors. *Epidemiology and Infection*, **132**, 1161–1166.

Zagheni, E., Billari, F. C., Manfredi, P., Melegaro, A., Mossong, J. and Edmunds, W. J. (2008) Using time-use data to parameterize models for the spread of close-contact infectious diseases. *American Journal of Epidemiology*, **168**, 1082–1090.

Appendix A

Discretized Formulas

The integral equation (2.12) has no closed form solution and therefore, we solve the system numerically by turning to a discrete age framework, assuming a constant force of infection in each age class. For this purpose, denote the first age interval $(a_{[1]}, a_{[2]})$ and the jth age interval $[a_{[j]}, a_{[j+1]})$, where $a_{[1]} = A$. For the MSIRW models introduced in Chapter 6, the proportion of susceptibles of age a (2.9), with $a \in [a_{[j]}, a_{[j+1]})$, reduces to

$$s(a) = \exp\left(-\sum_{k=1}^{j-1} \lambda_k (a_{[k+1]} - a_{[k]}) - \lambda_j (a - a_{[j]})\right).$$

Making use of this formula, the force of infection for age class i is approximated by:

$$\begin{aligned}\lambda_i =& \frac{ND}{L} \exp(-\mu_1 A) \sum_j \beta_{ij} \frac{\lambda_j}{\lambda_j + \mu_j} \left[\exp\left(-\sum_{k=1}^{j-1} (\lambda_k + \mu_k)(a_{[k+1]} - a_{[k]})\right)\right. \\ & \left. - \exp\left(-\sum_{k=1}^{j} (\lambda_k + \mu_k)(a_{[k+1]} - a_{[k]})\right)\right],\end{aligned}$$

β_{ij} denoting the per capita rate at which an individual of age class j makes effective contacts with a person of age class i, per year. The fraction of seropositives for the

MSIRWb-ext model is approximated by

$$
\begin{aligned}
r(a) =&(1-\varphi)\sum_{\ell=1}^{j-1}\frac{\lambda_\ell}{(1-\varphi)\lambda_\ell-\varepsilon_\ell}\exp\Bigg(-\sum_{k=1}^{\ell-1}\lambda_k(a_{[k+1]}-a_{[k]})-\sum_{m=\ell+1}^{j-1}(\varphi\lambda_m+\varepsilon_m)\\
&\cdot(a_{[m+1]}-a_{[m]})-(\varphi\lambda_j+\varepsilon_j)(a-a_{[j]})\Bigg)\cdot\Big[\exp\big(-(\varphi\lambda_\ell+\varepsilon_\ell)(a_{[\ell+1]}-a_{[\ell]})\big)\\
&-\exp\big(-\lambda_\ell(a_{[\ell+1]}-a_{[\ell]})\big)\Big]+\frac{(1-\varphi)\lambda_j}{(1-\varphi)\lambda_j-\varepsilon_j}\exp\left(-\sum_{k=1}^{j-1}\lambda_k(a_{[k+1]}-a_{[k]})\right)\\
&\cdot\big[\exp\big(-(\varphi\lambda_j+\varepsilon_j)(a-a_{[j]})\big)-\exp\big(-\lambda_j(a-a_{[j]})\big)\big]+\varphi\sum_{\ell=1}^{j-1}\frac{\lambda_\ell}{\varphi\lambda_\ell+\varepsilon_\ell}\\
&\cdot\exp\Bigg(-\sum_{m=\ell+1}^{j-1}(\varphi\lambda_m+\varepsilon_m)(a_{[m+1]}-a_{[m]})-(\varphi\lambda_j+\varepsilon_j)(a-a_{[j]})\Bigg)\big[1-\exp\big(\\
&-(\varphi\lambda_\ell+\varepsilon_\ell)(a_{[\ell+1]}-a_{[\ell]})\big)\big]+\frac{\varphi\lambda_j}{\varphi\lambda_j+\varepsilon_j}\cdot\big[1-\exp\big(-(\varphi\lambda_j+\varepsilon_j)(a-a_{[j]})\big)\big],
\end{aligned}
\tag{A.1}
$$

if a belongs to the jth age interval. The variants for the more parsimonious, nested MSIRW models are obtained by substituting φ with 0 and 1, respectively. From the integral equation (6.2) it follows that in case of an MSIRS-ext model, the force of infection for age class i can be approximated by:

$$
\begin{aligned}
\lambda_i =&\frac{ND}{L}\exp(-\mu_1 A)\sum_j\Bigg[\beta_{1ij}\frac{\lambda_j}{\lambda_j+\mu_j}\Bigg\{\exp\left(-\sum_{\ell=1}^{j-1}(\lambda_\ell+\mu_\ell)(a_{[\ell+1]}-a_{[\ell]})\right)\\
&-\exp\left(-\sum_{\ell=1}^{j}(\lambda_\ell+\mu_\ell)(a_{[\ell+1]}-a_{[\ell]})\right)\Bigg\}+\beta_{2ij}\frac{\lambda_j}{\lambda_j+\sigma_j+\mu_j}\exp\Big(-\sum_{\ell=1}^{j-1}\mu_\ell(a_{[\ell+1]}\\
&-a_{[\ell]})\Big)\Big\{1-\exp\big(-(\lambda_j+\sigma_j+\mu_j)(a_{[j+1]}-a_{[j]})\big)\Big\}\Bigg\{\sum_{\ell=1}^{j-1}\exp\Big(-\sum_{m=\ell+1}^{j-1}(\lambda_m+\sigma_m)\\
&\cdot(a_{[m+1]}-a_{[m]})\Big)\Bigg[\frac{\sigma_\ell}{\lambda_\ell+\sigma_\ell}\Big\{1-\exp\big(-(\lambda_\ell+\sigma_\ell)(a_{[\ell+1]}-a_{[\ell]})\big)\Big\}-\exp\Big(-\sum_{k=1}^{\ell}\lambda_k\\
&\cdot(a_{[k+1]}-a_{[k]})\Big)\Big\{1-\exp\big(-\sigma_\ell(a_{[\ell+1]}-a_{[\ell]})\big)\Big\}\Bigg]+\frac{\sigma_j}{\lambda_j+\sigma_j}\Bigg[\frac{\lambda_j+\sigma_j+\mu_j}{\mu_j}\\
&\cdot\frac{1-\exp\big(-\mu_j(a_{[j+1]}-a_{[j]})\big)}{1-\exp\big(-(\lambda_j+\sigma_j+\mu_j)(a_{[j+1]}-a_{[j]})\big)}-1\Bigg]-\exp\Big(-\sum_{\ell=1}^{j-1}\lambda_\ell(a_{[\ell+1]}-a_{[\ell]})\Big)\\
&\cdot\left[\frac{\lambda_j+\sigma_j+\mu_j}{\lambda_j+\mu_j}\cdot\frac{1-\exp\big(-(\lambda_j+\mu_j)(a_{[j+1]}-a_{[j]})\big)}{1-\exp\big(-(\lambda_j+\sigma_j+\mu_j)(a_{[j+1]}-a_{[j]})\big)}-1\right]\Bigg\}\Bigg].
\end{aligned}
$$

In case $\beta_{1ij} = \beta_{2ij}, \forall i, j$, it can be shown analytically that this expression reduces to the corresponding formula for the MSIRS model. The formula for the fraction of seropositives for both the MSIRS and MSIRS-ext model is identical to the one derived for the MSIRWb-ext model (A.1) with $\varphi = 1$, where the waning rate ε is replaced by the rate of re-entering the susceptible state, σ.

For the MSIRWS model, we partition the age classes $[a_{[j]}, a_{[j+1]})$ into smaller intervals of length δ and approximate the system of differential equations by a set of difference equations. The proportion of susceptibles and the proportion of individuals in the low immunity state, are then calculated as follows:

$$\begin{cases} s_{i+1} &= s_i + \sigma_i \delta w_i - \lambda_i \delta s_i, \\ w_{i+1} &= w_i + \varepsilon_i \delta \{1 - s_i\} - \{\varphi \lambda_i + \sigma_i + \varepsilon_i\} \delta w_i, \end{cases}$$

where $s_1 = 1$ and $w_1 = 0$. The force of infection for age class i is approximated by:

$$\lambda_i = \frac{N D e^{-\mu_1 A}}{L} \sum_j \frac{\beta_{ij} \lambda_j s_j}{\mu_j} \left[\exp\left(-\sum_{k=1}^{j-1} \mu_k (a_{[k+1]} - a_{[k]}) \right) - \exp\left(-\sum_{k=1}^{j} \mu_k (a_{[k+1]} - a_{[k]}) \right) \right]$$

and the fraction of seropositives $r(a)$ is approximated by $1 - s_i - w_i$, where the index i is chosen such that age a is located in the i^{th} age interval of length δ.

Appendix B

Immunity Transitions

Following Rouderfer *et al.* (1994), we estimate the number of certain PVB19 immunity transitions per person during their lifetime and the average age at which these transitions occur (Table B.1) for all scenarios considered in Chapter 6, hereby using the ML-estimates for the scenario-specific parameters. Note that for the MSIRS-ext model, the total fraction of susceptibles equals $s(a) = s_1(a) + s_2(a)$. For each country and each transmission scenario considered, the resulting estimates are presented in Table B.2.

Table B.1: The average number of transitions per person during their lifetime (*) and the average age at which these transitions occur.

Notation	Formula	Interpretation (all averages)
$\bar{n}_{SI}$	$\int_0^\infty \lambda(a)s(a)N(a)/N(0)da$	*number of infections
$\bar{A}_{SI}$	$\{\int_0^\infty a\lambda(a)s(a)N(a)/N(0)da\}/\bar{n}_{SI}$	age at infection
$\bar{n}_{RW}$	$\int_0^\infty \varepsilon(a)r(a)N(a)/N(0)da$	*number of transitions from high to low immunity
$\bar{A}_{RW}$	$\{\int_0^\infty a\varepsilon(a)r(a)N(a)/N(0)da\}/\bar{n}_{RW}$	age at waning from high to low immunity
$\bar{n}_{WR}$	$\int_0^\infty \varphi\lambda(a)w(a)N(a)/N(0)da$	*number of boosts from low to high immunity
$\bar{A}_{WR}$	$\{\int_0^\infty a\varphi\lambda(a)w(a)N(a)/N(0)da\}/\bar{n}_{WR}$	age at boosting from low to high immunity
$\bar{n}_{RS}$	$\int_0^\infty \sigma(a)r(a)N(a)/N(0)da$	*number of losses of disease-acquired immunity
$\bar{A}_{RS}$	$\{\int_0^\infty a\sigma(a)r(a)N(a)/N(0)da\}/\bar{n}_{RS}$	age at loss of disease-acquired immunity

Table B.2: ML-estimates for the average number of transitions per person during their lifetime and the average age at which these transitions occur (see Table B.1 for notations), together with 95% bootstrap-based percentile confidence intervals in square brackets. First entry: constant waning (CW); second entry (if available): age-specific waning (AW) with $H = 35$ years.

Country	Model	Waning	$\hat{\bar{n}}_{SI}$		$\hat{\bar{A}}_{SI}$		$\hat{\bar{n}}_{RW}*$ $\hat{\bar{n}}_{RS}+$		$\hat{\bar{A}}_{RW}*$ $\hat{\bar{A}}_{RS}+$		$\hat{\bar{n}}_{WR}$		$\hat{\bar{A}}_{WR}$	
BE	MSIR		0.90	[0.88, 0.92]	16.8	[15.6, 17.9]								
	MSIRW	CW	0.95	[0.93, 0.97]	13.3	[12.2, 14.2]	0.21 $*$	[0.13, 0.28]	41.8 $*$	[41.0, 42.5]				
		AW	0.96	[0.95, 0.98]	12.1	[11.2, 13.0]	0.15 $*$	[0.11, 0.19]	20.4 $*$	[20.2, 20.8]				
	MSIRWb	CW	0.96	[0.93, 0.98]	12.7	[11.6, 13.8]	0.50 $*$	[0.28, 0.72]	41.7 $*$	[41.0, 42.4]	0.27	[0.13, 0.43]	43.1	[41.6, 45.0]
		AW	0.97	[0.95, 0.98]	11.5	[10.6, 12.5]	0.55 $*$	[0.34, 0.76]	31.3 $*$	[22.3, 37.2]	0.39	[0.24, 0.56]	36.7	[33.6, 40.6]
	MSIRWb-ext	CW	0.96	[0.93, 0.98]	12.7	[11.6, 13.7]	0.47 $*$	[0.24, 1.06]	41.7 $*$	[40.9, 42.4]	0.24	[0.06, 0.80]	43.2	[41.4, 45.3]
		AW	0.97	[0.95, 0.98]	11.4	[10.4, 12.4]	0.26 $*$	[0.16, 0.48]	20.1 $*$	[19.9, 31.3]	0.14	[0.02, 0.35]	37.3	[34.3, 40.6]
	MSIRS	CW	1.38	[1.12, 1.78]	21.5	[18.2, 25.0]	0.68 +	[0.34, 1.16]	40.3 +	[39.0, 41.5]				
		AW	1.70	[1.38, 2.22]	21.2	[18.7, 23.9]	0.95 +	[0.57, 1.55]	30.6 +	[24.7, 35.2]				
	MSIRS-ext	CW	1.23	[1.07, 2.86]	19.3	[17.1, 30.7]	0.51 +	[0.29, 2.44]	40.6 +	[37.2, 41.6]				
EW	MSIR		0.78	[0.75, 0.80]	16.9	[15.9, 17.8]								
	MSIRW	CW	0.83	[0.78, 0.86]	15.6	[14.8, 16.6]	0.12 $*$	[0.02, 0.20]	42.8 $*$	[42.0, 43.8]				
		AW	0.87	[0.82, 0.90]	14.4	[13.6, 15.3]	0.14 $*$	[0.07, 0.20]	20.8 $*$	[20.6, 35.7]				
	MSIRWb	CW	0.84	[0.79, 0.87]	15.5	[14.6, 16.5]	0.20 $*$	[0.03, 0.34]	42.7 $*$	[41.9, 43.8]	0.06	[0.01, 0.12]	43.8	[41.9, 45.6]
		AW	0.89	[0.85, 0.92]	13.5	[12.7, 14.7]	0.37 $*$	[0.18, 0.52]	27.2 $*$	[20.2, 35.4]	0.21	[0.09, 0.32]	36.4	[34.1, 39.8]
	MSIRWb-ext	CW	0.84	[0.79, 0.87]	15.4	[14.5, 16.4]	0.36 $*$	[0.08, 1.09]	42.6 $*$	[41.9, 43.7]	0.22	[0.03, 0.91]	42.8	[40.8, 44.5]
		AW	0.90	[0.85, 0.92]	13.5	[12.7, 14.7]	0.62 $*$	[0.19, 1.22]	31.1 $*$	[20.3, 36.5]	0.44	[0.05, 1.04]	35.4	[32.2, 38.9]
	MSIRS	CW	0.91	[0.80, 1.01]	17.8	[16.2, 19.2]	0.21 +	[0.03, 0.37]	42.1 +	[41.2, 43.3]				
		AW	1.20	[0.98, 1.38]	18.8	[17.0, 20.0]	0.47 +	[0.23, 0.68]	26.9 +	[19.6, 34.2]				
	MSIRS-ext	CW	0.90	[0.81, 1.14]	17.5	[16.3, 22.3]	0.20 +	[0.04, 0.48]	42.1 +	[41.1, 43.3]				

Table B.2: (continued) ML-estimates for the average number of transitions per person during their lifetime and the average age at which these transitions occur (see Table B.1 for notations), together with 95% bootstrap-based percentile confidence intervals in square brackets. First entry: constant waning (CW); second entry (if available): age-specific waning (AW) with $H = 35$ years.

Country	Model	Waning	$\hat{\bar{n}}_{SI}$		$\hat{\bar{A}}_{SI}$		$\hat{\bar{n}}_{RW}*$ / $\hat{\bar{n}}_{RS}+$		$\hat{\bar{A}}_{RW}*$ / $\hat{\bar{A}}_{RS}+$		$\hat{\bar{n}}_{WR}$		$\hat{\bar{A}}_{WR}$	
FI	MSIR		0.72	[0.69, 0.75]	16.5	[15.6, 17.5]								
	MSIRW(b)	CW	0.72	[0.69, 0.75]	16.5	[15.6, 17.3]								
	MSIRS	CW	0.72	[0.69, 0.76]	16.5	[15.6, 17.6]								
IT	MSIR		0.75	[0.71, 0.77]	16.6	[15.2, 17.8]								
	MSIRW	CW	0.81	[0.75, 0.84]	15.2	[14.3, 16.4]	0.13 $*$	[0.02, 0.20]	43.0 $*$	[42.3, 44.0]				
		AW	0.83	[0.76, 0.86]	14.6	[13.9, 15.9]	0.13 $*$	[0.04, 0.20]	31.0 $*$	[20.7, 56.4]				
	MSIRWb	CW	0.82	[0.75, 0.85]	15.0	[14.2, 16.2]	0.20 $*$	[0.03, 0.32]	42.9 $*$	[42.2, 44.0]	0.06	[0.01, 0.11]	46.4	[45.1, 48.1]
		AW	0.84	[0.77, 0.87]	14.4	[13.5, 15.7]	0.24 $*$	[0.07, 0.36]	35.4 $*$	[20.8, 56.3]	0.10	[0.01, 0.16]	42.8	[38.4, 58.7]
	MSIRS	CW	0.89	[0.76, 0.98]	17.5	[15.5, 19.2]	0.21 $+$	[0.03, 0.34]	42.4 $+$	[41.7, 43.6]				
		AW	0.97	[0.79, 1.10]	17.9	[15.7, 19.6]	0.28 $+$	[0.07, 0.44]	34.6 $+$	[20.5, 55.9]				
	MSIRS-ext	CW	0.88	[0.77, 0.99]	17.2	[15.7, 19.2]	0.20 $+$	[0.05, 0.35]	42.4 $+$	[41.7, 43.6]				
PL	MSIR(W)(b)	CW	0.86	[0.83, 0.87]	16.0	[15.1, 17.1]								
	MSIRW	AW	0.87	[0.83, 0.89]	15.5	[14.8, 16.5]	0.02 $*$	[0.00, 0.07]	21.5 $*$	[21.3, 55.4]				
	MSIRWb	AW	0.92	[0.85, 0.94]	13.4	[12.1, 15.3]	0.23 $*$	[0.02, 0.41]	20.8 $*$	[20.4, 21.6]	0.16	[0.01, 0.32]	35.2	[33.8, 37.1]
	MSIRWb-ext	AW	0.92	[0.88, 0.94]	13.2	[12.5, 14.5]	0.52 $*$	[0.32, 0.81]	20.6 $*$	[20.4, 21.8]	0.48	[0.28, 0.79]	30.0	[26.5, 32.8]
	MSIRS	CW	0.86	[0.83, 0.88]	16.0	[15.1, 17.1]								
		AW	1.65	[1.11, 2.33]	19.1	[16.9, 20.2]	0.81 $+$	[0.26, 1.57]	18.9 $+$	[18.0, 20.5]				

Appendix C

Matlab Code

The MSIRW and MSIRS scenarios considered for PVB19 in Chapter 6 are implemented in `Matlab` and ML-estimates are obtained using `fminsearch`. We provide the `Matlab` functions below for the two most extensive models `MSIRWboostext` and `MSIRSext`, since all other scenarios are special cases. Both functions make use of the function `read` (not displayed here) to import the country-specific data: the estimated daily contact rates matrix `rij(:,:)`, the vectors containing the serological data i.e. the individuals' age `age(:,)`, serological status `resp(:,)` and post-stratification weight `weight(:,)`, the life expectancy `L`, the total population size `N`, the age-specific mortality rates `mu(:,)`, and the maternal age distribution `bi(:,)` for live births. Further, both functions make use of the function `Rfrac` displayed below to calculate the age-specific fraction of seropositives according to formula (A.1). Finally, both functions require the following input parameters: the `country{''}` specification, the cut-off point `H` for the age-specific waning scenario, the starting values `init(,:)` for the optimization procedure, and the `model{''}` specification for the waning rates.

```
function r = Rfrac(age,epsilon,phi,C,B1,Cb,foi,k)
alow = floor(age);
if length(epsilon)>1
    theta = exp(-(phi*foi(alow+1)+epsilon(alow+1)).*(age-max(0.5,alow)));
    r2 = sum(Cb(:,alow+1).*(theta*ones(1,k))')'+(foi(alow+1)./(phi*foi(alow+1)
    +epsilon(alow+1))).*(1-theta);
else
    theta = exp(-(phi*foi(alow+1)+epsilon).*(age-max(0.5,alow)));
    r2 = sum(Cb(:,alow+1).*(theta*ones(1,k))')'+(foi(alow+1)./(phi*foi(alow+1)
    +epsilon)).*(1-theta);
end
r1 = sum(C(:,alow+1).*(theta*ones(1,k))')'+(B1(alow+1).*(theta-exp(-foi(alow+1)
    .*(age-max(0.5,alow)))));
r1 = max(0,r1);
```

```
r2 = max(0,r2);
r = (1-phi)*r1 + phi*r2;
end
```

C.1 MSIRWb-ext Model

```
function [parhat,R0,risk,trans,aic,bic,exitflag,output] = MSIRWboostext(country,H,init,
                                                                        model)
[rij,age,resp,weight,L,N,mu,bi] = readd(country);

% age of maternal antibody waning (0<=A<1)
A = 0.5;

% mean duration of infectiousness
D = 6/365;

% k right-open age-intervals are considered: (A,1), [1,2),..., [k-1,k)
k = 80;
step = [1-A, ones(1,k-1)]';
ageint = [A+[0;cumsum(step(1:end-1))] A+cumsum(step)];
rij = rij(1:k,1:k);
mu = mu(1:k);
bi = bi(1:k);

% ages <= A and >= k are removed from the serological data
resp = resp(age>A & age<k);
age = age(age>A & age<k);

% Function "qestim" to calculate the FOI and likelihood
% conditional on the parameter values
%*******************************************************
function dev = qestim(par)
    q = exp(-par(1));
    if strcmp(model,'constant')
        epsilon = exp(-par(2));
    end
    if strcmp(model,'discrete')
        % piecewise constant function
        epsilon = [exp(-par(2))*ones(H,1) ; exp(-par(3))*ones(k-H,1)];
    end
    phi = exp(-par(end));
    bij = 365*q.*rij;
    foi = 0.1*ones(k,1);
    tol = 1;
    it = 0;
```

```
    while (tol>1D-15) && (it<2000)
        S = (N/L)*exp(-mu(1)*A)*exp(-cumsum([0;(foi+mu).*step]));
        I = foi./(foi+mu).*(S(1:end-1)-S(2:end));
        foinext = D*bij*I;
        tol = sum((foinext-foi).^2);
        it = it+1;
        foi = foinext;
    end
    if it==2000
        error('Maximum number of iterations exceeded')
    end
    % input from MSIRW framework
    s = exp(-cumsum([0;foi.*step]));
    if length(epsilon)>1
        f = @(i,j) exp(-sum((phi*foi(i+1:j-1)+epsilon(i+1:j-1)).*step(i+1:j-1)));
    else
        f = @(i,j) exp(-sum((phi*foi(i+1:j-1)+epsilon).*step(i+1:j-1)));
    end
    F = zeros(k);
    for j = 1:k
        for i = 1:j-1
            F(i,j) = f(i,j);
        end
    end
    E = [f(0,2) diag(F,2)' f(k-1,k+1)];
    B1 = foi./((1-phi)*foi-epsilon).*s(1:end-1);
    B2 = B1.*(E'-(s(2:end)./s(1:end-1)));
    C = (B2*ones(1,k)).*F;
    % input from MSIRWboost framework
    B = (foi./(phi*foi+epsilon)).*(1-E');
    Cb = (B*ones(1,k)).*F;
    % fraction of seropositives
    r = Rfrac(age,epsilon,phi,C,B1,Cb,foi,k);
    ll = resp.*log(r)+(1-resp).*log(1-r);
    dev = -2*sum(weight.*ll);
end

% Non-linear optimization of the function "qestim"
%************************************************
[parhat,dev,exitflag,output] = fminsearch(@qestim,init,optimset('FunValCheck',
                                'on','Display','final','MaxFunEvals',1500));
parhat = exp(-parhat);

% Next generation matrix and R0
%******************************
Na = (N/L)*exp(-mu(1)*A)*exp(-cumsum([0;mu.*step]));
```

```
M = (Na(1:end-1)-Na(2:end))./mu;
G = D*diag(M)*bij;
R0 = max(real(eig(G)));

% Risk in pregnancy
%******************
Iy = sum(bi.*(s(1:end-1)-s(2:end)));
slb = sum(bi./foi.*(s(1:end-1)-s(2:end)));
sp = slb/sum(bi);
foip = Iy/slb;
Ip = 0.77*Iy;
freqp = sum(bi)/Ip;
fetaldeath = Ip*0.077*(20/40);
risk = [sp foip Ip freqp fetaldeath];

% Transitions
%************
U1 = (1-exp(-(phi*foi+epsilon+mu).*step))./(phi*foi+epsilon+mu);
U2 = (1-exp(-(foi+mu).*step))./(foi+mu);
U3 = (1-exp(-mu.*step))./mu;
T1 = (ageint(:,1)-ageint(:,2).*exp(-(phi*foi+epsilon+mu).*step))./(phi*foi+epsilon+mu)
    +U1./(phi*foi+epsilon+mu);
T2 = T1-((ageint(:,1)-ageint(:,2).*exp(-(foi+mu).*step))./(foi+mu)+U2./(foi+mu));
T3 = ((ageint(:,1)-ageint(:,2).*exp(-mu.*step))./mu+U3./mu)-T1;
nSI = sum((L/N)*I);
ASI = sum((L/N)*foi./(foi+mu).*(S(1:end-1).*(ageint(:,1)+U2)-S(2:end)
    .*(ageint(:,2))))/nSI;

r1 = sum(C)'+(B1.*(1-(U2./U1)));
r2 = sum(Cb)'+(foi./(phi*foi+epsilon).*((U3./U1)-1));
radapt = (1-phi)*r1+phi*r2;
r1A = T1.*sum(C)'+T2.*B1;
r2A = T1.*sum(Cb)'+T3.*foi./(phi*foi+epsilon);
radaptA = (1-phi)*r1A+phi*r2A;
nRW = sum((L/N)*epsilon.*Na(1:end-1).*U1.*(radapt));
ARW = sum((L/N)*epsilon.*Na(1:end-1).*(radaptA))/nRW;

B1 = (epsilon./(phi*foi+epsilon)).*(1-E');
C1 = (B1*ones(1,k)).*F;
B2 = (epsilon./((1-phi)*foi-epsilon)).*s(1:end-1).*(E'-s(2:end)./s(1:end-1));
C2 = (B2*ones(1,k)).*F;
wadapt = sum(C1)'-sum(C2)'+epsilon./(phi*foi+epsilon).*((U3./U1)-1)-epsilon
        ./((1-phi)*foi-epsilon).*s(1:end-1).*(1-(U2./U1));
wadaptA = T1.*(sum(C1)'-sum(C2)')+epsilon./(phi*foi+epsilon).*T3-epsilon
        ./((1-phi)*foi-epsilon).*s(1:end-1).*T2;
nWR = sum((L/N)*phi*foi.*Na(1:end-1).*U1.*(wadapt));
```

```
AWR = sum((L/N)*phi*foi.*Na(1:end-1).*(wadaptA))/nWR;

trans = [nSI ASI nRW ARW nWR AWR];

% Information criteria
%*********************
aic = dev+2*length(parhat);
bic = dev+log(length(resp))*length(parhat);

end
```

C.2 MSIRS-ext Model

```
function [parhat,R0,risk,trans,aic,bic,exitflag,output] = MSIRSext(country,H,init,
                                                                   model)
[rij,age,resp,weight,L,N,mu,bi] = readd(country);

% age of maternal antibody waning (0<=A<1)
A = 0.5;

% mean duration of infectiousness
D = 6/365;

% k right-open age-intervals are considered: (A,1), [1,2),..., [k-1,k)
k = 80;
step = [1-A, ones(1,k-1)]';
ageint = [A+[0;cumsum(step(1:end-1))] A+cumsum(step)];
rij = rij(1:k,1:k);
mu = mu(1:k);
bi = bi(1:k);

% ages <= A and >= k are removed from the serological data
resp = resp(age>A & age<k);
age = age(age>A & age<k);

% Function "qestim" to calculate the FOI and likelihood
% conditional on the parameter values
%*****************************************************
function dev = qestim(par)
    q1 = exp(-par(1));
    q2 = exp(-par(2));
    if strcmp(model,'constant')
        sig = exp(-par(end));
    end
    if strcmp(model,'discrete')
```

```
    % piecewise constant function
    sig = [exp(-par(end-1))*ones(H,1) ; exp(-par(end))*ones(k-H,1)];
end
b1ij = 365*q1*rij;
b2ij = 365*q2*rij;
V2 = exp(-sig.*step);
V3 = exp(-mu.*step);
CV3 = cumprod([1;V3]);
foi = 0.1*ones(k,1);
tol = 1;
it = 0;
while (tol>1D-15) && (it<2000)
    % foi = term*(I1+I2)
    V1 = exp(-foi.*step);
    CV1 = cumprod([1;V1]);
    V12 = V1.*V2;
    CV13 = CV1.*CV3;
    % constructing the number of primary infectious individuals I1
    I1 = (N/L)*exp(-mu(1)*A)*foi./(foi+mu).*(CV13(1:end-1)-CV13(2:end));
    % constructing the number of secondary infectious individuals
    % I2 = term*(Q1+Q2-Q3)
    % constructing Q1
    f = @(l,j) prod(V12(l+1:j-1));
    g = @(l,j) prod(V2(l+1:j-1));
    F = zeros(k);
    G = zeros(k);
    for j = 1:k
        for l = 1:j-1
            F(l,j) = f(l,j);
            G(l,j) = g(l,j);
        end
    end
    B1 = (sig./(foi+sig)).*(1-V12);
    T1 = (B1*ones(1,k)).*F;
    B2 = CV1(1:end-1)*ones(1,k);
    T2 = ((1-V2)*ones(1,k)).*G.*(B2');
    Q1 = sum(T1)'-sum(T2)';
    % constructing Q2 en Q3
    Q2 = sig./(sig+foi).*((foi+sig+mu)./mu.*((1-V3)./(1-V1.*V2.*V3))-1);
    Q3 = CV1(1:end-1).*((foi+sig+mu)./(foi+mu).*((1-V1.*V3)./(1-V1.*V2.*V3))-1);
    I2 = (N/L)*exp(-mu(1)*A)*foi./(foi+sig+mu).*CV3(1:end-1).*(1-V1.*V2.*V3)
        .*(Q1+Q2-Q3);
    foinext = D*(b1ij*I1+b2ij*I2);
    tol = sum((foinext-foi).^2);
    it = it+1;
    foi = foinext;
```

```
    end
    if it==2000
        error('Maximum number of iterations exceeded')
    end
    V1 = exp(-foi.*step);
    V12 = V1.*V2;
    f = @(l,j) prod(V12(l+1:j-1));
    F = zeros(k);
    for j = 1:k
        for l = 1:j-1
            F(l,j) = f(l,j);
        end
    end
    E = [f(0,2) diag(F,2)' f(k-1,k+1)];
    B = (foi./(foi+sig)).*(1-E');
    C = (B*ones(1,k)).*F;
    r = Rfrac(age,sig,1,zeros(k),zeros(k,1),C,foi,k);
    ll = resp.*log(r)+(1-resp).*log(1-r);
    dev = -2*sum(weight.*ll);
end

% Non-linear optimization of the function "qestim"
%*************************************************
[parhat,dev,exitflag,output] = fminsearch(@qestim,init,optimset('FunValCheck',
                                'on','Display','final','MaxFunEvals',1500));
parhat = exp(-parhat);

% Next generation matrix and R0
%******************************
Na = (N/L)*exp(-mu(1)*A)*exp(-cumsum([0;mu.*step]));
M = (Na(1:end-1)-Na(2:end))./mu;
G = D*diag(M)*b1ij;
R0 = max(real(eig(G)));

% Risk in pregnancy
%******************
Tp = exp(-cumsum([0;(foi+sig).*step]));
% constructing Q1p
Q1p = (1-Tp(2:end)./Tp(1:end-1)).*sum(C)';
% constructing Q3p
Q3p = foi./(foi+sig).*(((foi+sig).*step)-(1-Tp(2:end)./Tp(1:end-1)));
Iy = sum(foi.*bi.*step)-sum(foi.*bi./(foi+sig).*(Q1p+Q3p));
slb = sum(bi.*step)-sum(bi./(foi+sig).*(Q1p+Q3p));
sp = slb/sum(bi);
foip = Iy/slb;
Ip = 0.77*Iy;
```

```
freqp = sum(bi)/Ip;
fetaldeath = Ip*0.077*(20/40);
risk = [sp foip Ip freqp fetaldeath];

% Transitions
%************
% constructing P
T = exp(-cumsum([0;(foi+sig+mu).*step]));
P = T(1:end-1)-T(2:end);
% constructing Q1
B1 = (sig./(foi+sig)).*(1-E');
C1 = (B1*ones(1,k)).*F;
Q1 = (1-T(2:end)./T(1:end-1)).*sum(C1)';
% constructing Q2
Q2 = sig./(foi+sig).*(((foi+sig+mu)./mu.*(1-exp(-mu.*step)))-(1-T(2:end)./T(1:end-1)));
% constructing Q
Q = exp(-cumsum([0;mu(1:end-1).*step(1:end-1)])).*(Q1+Q2);
nSI = sum(exp(-mu(1)*A)*foi./(foi+sig+mu).*(P+Q));
U1 = (1-exp(-mu.*step))./mu;
U2 = (ageint(:,1)-ageint(:,2).*exp(-mu.*step))./mu;
U3 = (1-T(2:end)./T(1:end-1))./(foi+sig+mu);
U4 = (ageint(:,1)-ageint(:,2).*T(2:end)./T(1:end-1))./(foi+sig+mu);
V1 = (L/N)*(foi./mu).*(Na(1:end-1).*(ageint(:,1)+U1)-Na(2:end).*(ageint(:,2)));
V2 = (L/N)*Na(1:end-1).*foi.*((U3./(foi+sig+mu)+U4).*sum(C)'+foi./(foi+sig)
    .*((U1./mu)+U2-(U3./(foi+sig+mu))-U4));
ASI = sum(V1-V2)/nSI;

T1 = U4+U3./(foi+sig+mu);
T3 = U2+U1./mu-T1;
radapt = sum(C)'+(foi./(foi+sig).*((U1./U3)-1));
radaptA = T1.*sum(C)'+T3.*foi./(foi+sig);
nRS = sum((L/N)*sig.*Na(1:end-1).*U1.*(radapt));
ARS = sum((L/N)*sig.*Na(1:end-1).*(radaptA))/nRS;

trans = [nSI ASI nRS ARS];

% Information criteria
%*********************
aic = dev+2*length(parhat);
bic = dev+log(length(resp))*length(parhat);

end
```

Appendix D

Simulation Results

Tables D.1-D.10 present the results of the simulation study for PVB19 described in Section 6.4. For each simulation setting considered (see table legend), the sample estimate, estimated sample standard deviation ($\widehat{\text{s.d.}}$), and mean squared error (MSE) are given for the proportionality factor q, the basic reproduction number R_0, the average maternal proportion of susceptibles $\bar{s}_p$, and the average maternal force of infection $\bar{\lambda}_p$, that are estimated by fitting each of the mathematical scenarios considered in Chapter 6 to the simulated serological data sets. Further, the AIC and BIC model selection percentages, $\pi_{\text{sel,AIC}}$ and $\pi_{\text{sel,BIC}}$ respectively, are provided as well, and the largest value for each is displayed in bold. Note that for the MSIRS-ext model, we use the proportionality factor q_1 for infectious individuals with a primary infection, as a surrogate for q. When using MSIR as the true underlying dynamics, identifiability problems arise for MSIRWb-ext (both CW and AW) and MSIRS-ext since the waning rate estimates are close to zero, which makes it difficult to estimate ϕ and q_2, respectively. Therefore, the results for these models are omitted from Table D.1, and for this reason the $\pi_{\text{sel,AIC}}$ do not sum up to 100%.

Table D.1: Simulation study results for PVB19 considering MSIR as the 'true' model and using ML-estimates for BE as parameter values: $q = 0.056$ ($R_0 = 2.48$, $\bar{s}_p = 0.27$, $\bar{\lambda}_p = 0.034$, $n_s = 188$).

model	waning	$\bar{\hat{q}}$	$\widehat{\text{s.d.}}\left(\bar{\hat{q}}\right)$	$\text{MSE}\left(\bar{\hat{q}}\right)$ $\cdot 10^3$	$\bar{\hat{R}}_0$	$\widehat{\text{s.d.}}\left(\bar{\hat{R}}_0\right)$	$\text{MSE}\left(\bar{\hat{R}}_0\right)$	$\pi_{\text{sel,AIC}}$
MSIR		0.056	0.001	0.002	2.49	0.06	0.00	**74%**
MSIRW	CW	0.057	0.001	0.003	2.52	0.06	0.00	2%
	AW	0.057	0.001	0.002	2.52	0.06	0.00	2%
MSIRWb	CW	0.057	0.001	0.003	2.52	0.06	0.01	5%
	AW	0.057	0.001	0.003	2.54	0.06	0.01	0%
MSIRS	CW	0.057	0.001	0.002	2.50	0.06	0.00	3%
	AW	0.057	0.001	0.002	2.51	0.05	0.00	4%
model	waning	$\overline{\hat{\bar{s}}_p}$	$\widehat{\text{s.d.}}\left(\overline{\hat{\bar{s}}_p}\right)$	$\text{MSE}\left(\overline{\hat{\bar{s}}_p}\right)$ $\cdot 10^2$	$\overline{\hat{\bar{\lambda}}_p}$	$\widehat{\text{s.d.}}\left(\overline{\hat{\bar{\lambda}}_p}\right)$	$\text{MSE}\left(\overline{\hat{\bar{\lambda}}_p}\right)$ $\cdot 10^3$	$\pi_{\text{sel,BIC}}$
MSIR		0.27	0.01	0.01	0.034	0.001	0.001	**97%**
MSIRW	CW	0.26	0.01	0.02	0.035	0.001	0.002	1%
	AW	0.26	0.01	0.02	0.035	0.001	0.001	0%
MSIRWb	CW	0.26	0.01	0.02	0.035	0.001	0.002	1%
	AW	0.26	0.01	0.02	0.035	0.001	0.002	0%
MSIRS	CW	0.26	0.01	0.01	0.035	0.001	0.003	2%
	AW	0.27	0.01	0.02	0.036	0.002	0.005	0%

Table D.2: Simulation study results for PVB19 considering MSIRW CW ($\varphi = 0$) as the 'true' model and using ML-estimates for BE as parameter values: $q = 0.073$ and $\varepsilon = 0.004$ ($R_0 = 3.21$, $\bar{s}_p = 0.17$, $\bar{\lambda}_p = 0.046$, $n_s = 185$).

model	waning	$\hat{\bar{q}}$	$\widehat{\text{s.d.}}\left(\hat{\bar{q}}\right)$	$\text{MSE}\left(\hat{\bar{q}}\right)$ $\cdot 10^3$	$\hat{\bar{R}}_0$	$\widehat{\text{s.d.}}\left(\hat{\bar{R}}_0\right)$	$\text{MSE}\left(\hat{\bar{R}}_0\right)$	$\pi_{\text{sel,AIC}}$
MSIR		0.056	0.001	0.286	2.46	0.06	0.56	0%
MSIRW	CW	0.073	0.002	0.006	3.22	0.11	0.01	**29%**
	AW	0.073	0.003	0.006	3.21	0.11	0.01	11%
MSIRWb	CW	0.075	0.003	0.014	3.32	0.12	0.03	14%
	AW	0.074	0.003	0.008	3.25	0.12	0.01	5%
MSIRWb-ext	CW	0.075	0.003	0.014	3.29	0.14	0.03	9%
	AW	0.074	0.003	0.011	3.26	0.14	0.02	9%
MSIRS	CW	0.064	0.002	0.079	2.82	0.07	0.15	7%
	AW	0.064	0.001	0.080	2.82	0.07	0.16	8%
MSIRS-ext	CW	0.054	0.019	0.730	2.38	0.86	1.42	8%
model	waning	$\overline{\hat{\bar{s}}_p}$	$\widehat{\text{s.d.}}\left(\overline{\hat{\bar{s}}_p}\right)$	$\text{MSE}\left(\overline{\hat{\bar{s}}_p}\right)$ $\cdot 10^2$	$\overline{\hat{\bar{\lambda}}_p}$	$\widehat{\text{s.d.}}\left(\overline{\hat{\bar{\lambda}}_p}\right)$	$\text{MSE}\left(\overline{\hat{\bar{\lambda}}_p}\right)$ $\cdot 10^3$	$\pi_{\text{sel,BIC}}$
MSIR		0.27	0.01	1.08	0.034	0.001	0.155	0%
MSIRW	CW	0.17	0.01	0.01	0.046	0.002	0.003	**49%**
	AW	0.17	0.01	0.01	0.046	0.002	0.003	2%
MSIRWb	CW	0.16	0.01	0.02	0.048	0.002	0.006	24%
	AW	0.16	0.01	0.01	0.047	0.002	0.004	0%
MSIRWb-ext	CW	0.16	0.01	0.02	0.047	0.002	0.006	4%
	AW	0.16	0.01	0.02	0.047	0.002	0.005	0%
MSIRS	CW	0.24	0.01	0.57	0.058	0.004	0.159	21%
	AW	0.24	0.02	0.54	0.057	0.004	0.135	0%
MSIRS-ext	CW	0.24	0.01	0.52	0.067	0.019	0.800	1%

Table D.3: Simulation study results for PVB19 considering MSIRW AW ($\varphi = 0$) as the 'true' model and using ML-estimates for BE as parameter values: $q = 0.080$, $\varepsilon_1 = 0.007$, and $\varepsilon_2 = 0.000$ ($R_0 = 3.53$, $\bar{s}_p = 0.14$, $\bar{\lambda}_p = 0.051$, $n_s = 181$).

model	waning	$\bar{\hat{q}}$	$\widehat{\text{s.d.}}\left(\bar{\hat{q}}\right)$	$\text{MSE}\left(\bar{\hat{q}}\right)$ $\cdot 10^3$	$\bar{\hat{R}}_0$	$\widehat{\text{s.d.}}\left(\bar{\hat{R}}_0\right)$	$\text{MSE}\left(\bar{\hat{R}}_0\right)$	$\pi_{\text{sel,AIC}}$
MSIR		0.056	0.001	0.579	2.47	0.06	1.13	0%
MSIRW	CW	0.074	0.003	0.041	3.27	0.12	0.08	2%
	AW	0.079	0.003	0.008	3.50	0.12	0.02	**45%**
MSIRWb	CW	0.077	0.003	0.017	3.42	0.14	0.03	5%
	AW	0.082	0.003	0.011	3.60	0.13	0.02	11%
MSIRWb-ext	CW	0.077	0.003	0.019	3.41	0.15	0.04	3%
	AW	0.081	0.003	0.011	3.59	0.13	0.02	12%
MSIRS	CW	0.065	0.002	0.237	2.86	0.07	0.46	5%
	AW	0.065	0.002	0.219	2.88	0.07	0.43	17%
MSIRS-ext	CW	0.068	0.018	0.486	2.99	0.81	0.95	1%
model	waning	$\overline{\hat{\bar{s}}_p}$	$\widehat{\text{s.d.}}\left(\overline{\hat{\bar{s}}_p}\right)$	$\text{MSE}\left(\overline{\hat{\bar{s}}_p}\right)$ $\cdot 10^2$	$\overline{\hat{\bar{\lambda}}_p}$	$\widehat{\text{s.d.}}\left(\overline{\hat{\bar{\lambda}}_p}\right)$	$\text{MSE}\left(\overline{\hat{\bar{\lambda}}_p}\right)$ $\cdot 10^3$	$\pi_{\text{sel,BIC}}$
MSIR		0.27	0.01	1.72	0.034	0.001	0.300	0%
MSIRW	CW	0.16	0.01	0.06	0.047	0.002	0.019	6%
	AW	0.14	0.01	0.01	0.051	0.002	0.004	**33%**
MSIRWb	CW	0.15	0.01	0.03	0.049	0.002	0.008	**33%**
	AW	0.13	0.01	0.01	0.052	0.002	0.005	4%
MSIRWb-ext	CW	0.15	0.01	0.03	0.049	0.002	0.009	1%
	AW	0.13	0.01	0.01	0.052	0.002	0.005	0%
MSIRS	CW	0.24	0.01	1.07	0.061	0.006	0.135	12%
	AW	0.27	0.02	1.74	0.072	0.006	0.459	11%
MSIRS-ext	CW	0.25	0.01	1.15	0.061	0.021	0.528	0%

Table D.4: Simulation study results for PVB19 considering MSIRWb CW ($\varphi = 1$) as the 'true' model and using ML-estimates for BE as parameter values: $q = 0.076$ and $\varepsilon = 0.010$ ($R_0 = 3.35$, $\bar{s}_p = 0.15$, $\bar{\lambda}_p = 0.048$, $n_s = 196$).

model	waning	$\bar{\hat{q}}$	$\widehat{\text{s.d.}}\left(\bar{\hat{q}}\right)$	$\text{MSE}\left(\bar{\hat{q}}\right)$ $\cdot 10^3$	$\bar{\hat{R}}_0$	$\widehat{\text{s.d.}}\left(\bar{\hat{R}}_0\right)$	$\text{MSE}\left(\bar{\hat{R}}_0\right)$	$\pi_{\text{sel,AIC}}$
MSIR		0.056	0.001	0.410	2.46	0.06	0.80	0%
MSIRW	CW	0.073	0.002	0.014	3.23	0.11	0.03	16%
	AW	0.074	0.003	0.010	3.26	0.11	0.02	13%
MSIRWb	CW	0.076	0.003	0.008	3.35	0.12	0.01	**20%**
	AW	0.076	0.003	0.007	3.34	0.12	0.01	8%
MSIRWb-ext	CW	0.076	0.003	0.009	3.35	0.13	0.02	**20%**
	AW	0.076	0.003	0.008	3.36	0.13	0.02	7%
MSIRS	CW	0.064	0.001	0.144	2.83	0.07	0.28	10%
	AW	0.064	0.001	0.141	2.83	0.06	0.27	6%
MSIRS-ext	CW	0.066	0.014	0.283	2.93	0.61	0.55	0%

model	waning	$\overline{\hat{\bar{s}}_p}$	$\widehat{\text{s.d.}}\left(\overline{\hat{\bar{s}}_p}\right)$	$\text{MSE}\left(\overline{\hat{\bar{s}}_p}\right)$ $\cdot 10^2$	$\overline{\hat{\bar{\lambda}}_p}$	$\widehat{\text{s.d.}}\left(\overline{\hat{\bar{\lambda}}_p}\right)$	$\text{MSE}\left(\overline{\hat{\bar{\lambda}}_p}\right)$ $\cdot 10^3$	$\pi_{\text{sel,BIC}}$
MSIR		0.27	0.01	1.39	0.034	0.001	0.218	0%
MSIRW	CW	0.17	0.01	0.03	0.046	0.002	0.007	22%
	AW	0.16	0.01	0.02	0.047	0.002	0.005	2%
MSIRWb	CW	0.15	0.01	0.01	0.048	0.002	0.004	**50%**
	AW	0.16	0.01	0.01	0.048	0.002	0.003	1%
MSIRWb-ext	CW	0.15	0.01	0.01	0.048	0.002	0.004	6%
	AW	0.15	0.01	0.01	0.049	0.002	0.004	1%
MSIRS	CW	0.24	0.01	0.80	0.059	0.004	0.126	17%
	AW	0.25	0.02	0.90	0.060	0.005	0.165	2%
MSIRS-ext	CW	0.24	0.01	0.85	0.058	0.014	0.278	0%

Table D.5: Simulation study results for PVB19 considering MSIRWb AW ($\varphi = 1$) as the 'true' model and using ML-estimates for BE as parameter values: $q = 0.084$, $\varepsilon_1 = 0.019$, and $\varepsilon_2 = 0.005$ ($R_0 = 3.70$, $\bar{s}_p = 0.13$, $\bar{\lambda}_p = 0.054$, $n_s = 200$).

model	waning	$\bar{\hat{q}}$	$\widehat{\text{s.d.}}\left(\bar{\hat{q}}\right)$	$\text{MSE}\left(\bar{\hat{q}}\right)$ $\cdot 10^3$	$\bar{\hat{R}}_0$	$\widehat{\text{s.d.}}\left(\bar{\hat{R}}_0\right)$	$\text{MSE}\left(\bar{\hat{R}}_0\right)$	$\pi_{\text{sel,AIC}}$
MSIR		0.056	0.001	0.765	2.48	0.06	1.49	0%
MSIRW	CW	0.070	0.002	0.194	3.09	0.11	0.38	0%
	AW	0.076	0.003	0.065	3.36	0.12	0.13	4%
MSIRWb	CW	0.074	0.003	0.113	3.25	0.13	0.22	0%
	AW	0.084	0.003	0.012	3.69	0.15	0.02	**38%**
MSIRWb-ext	CW	0.074	0.003	0.100	3.28	0.14	0.19	1%
	AW	0.084	0.003	0.012	3.70	0.15	0.02	30%
MSIRS	CW	0.063	0.001	0.430	2.79	0.07	0.84	0%
	AW	0.065	0.001	0.369	2.85	0.07	0.72	28%
MSIRS-ext	CW	0.074	0.003	0.113	3.25	0.13	0.22	0%
model	waning	$\overline{\hat{\bar{s}}_p}$	$\widehat{\text{s.d.}}\left(\overline{\hat{\bar{s}}_p}\right)$	$\text{MSE}\left(\overline{\hat{\bar{s}}_p}\right)$ $\cdot 10^2$	$\overline{\hat{\bar{\lambda}}_p}$	$\widehat{\text{s.d.}}\left(\overline{\hat{\bar{\lambda}}_p}\right)$	$\text{MSE}\left(\overline{\hat{\bar{\lambda}}_p}\right)$ $\cdot 10^3$	$\pi_{\text{sel,BIC}}$
MSIR		0.27	0.01	2.01	0.034	0.001	0.388	0%
MSIRW	CW	0.18	0.01	0.30	0.044	0.002	0.091	0%
	AW	0.15	0.01	0.08	0.049	0.002	0.029	6%
MSIRWb	CW	0.16	0.01	0.16	0.047	0.002	0.052	9%
	AW	0.13	0.01	0.01	0.053	0.002	0.005	**47%**
MSIRWb-ext	CW	0.16	0.01	0.14	0.047	0.002	0.046	0%
	AW	0.13	0.01	0.01	0.054	0.002	0.005	5%
MSIRS	CW	0.24	0.01	1.38	0.055	0.005	0.022	0%
	AW	0.29	0.02	2.86	0.076	0.007	0.558	34%
MSIRS-ext	CW	0.25	0.01	1.53	0.048	0.002	0.037	0%

Table D.6: Simulation study results for PVB19 considering MSIRWb-ext CW as the 'true' model and using ML-estimates for BE as parameter values: $q = 0.076$, $\varepsilon = 0.009$, and $\varphi = 0.91$ ($R_0 = 3.35$, $\bar{s}_p = 0.15$, $\bar{\lambda}_p = 0.048$, $n_s = 194$).

model	waning	$\bar{\hat{q}}$	$\widehat{\text{s.d.}}\left(\bar{\hat{q}}\right)$	$\text{MSE}\left(\bar{\hat{q}}\right)$ $\cdot 10^3$	$\bar{\hat{R}}_0$	$\widehat{\text{s.d.}}\left(\bar{\hat{R}}_0\right)$	$\text{MSE}\left(\bar{\hat{R}}_0\right)$	$\pi_{\text{sel,AIC}}$
MSIR		0.056	0.001	0.400	2.47	0.06	0.78	0%
MSIRW	CW	0.073	0.003	0.017	3.21	0.12	0.03	11%
	AW	0.074	0.003	0.011	3.27	0.12	0.02	11%
MSIRWb	CW	0.076	0.003	0.010	3.34	0.14	0.02	**28%**
	AW	0.076	0.003	0.009	3.36	0.13	0.02	11%
MSIRWb-ext	CW	0.076	0.003	0.011	3.35	0.14	0.02	18%
	AW	0.077	0.003	0.010	3.38	0.14	0.02	6%
MSIRS	CW	0.064	0.002	0.143	2.83	0.07	0.28	8%
	AW	0.064	0.002	0.138	2.84	0.07	0.27	6%
MSIRS-ext	CW	0.069	0.011	0.169	3.06	0.50	0.33	1%
model	waning	$\overline{\hat{\bar{s}}_p}$	$\widehat{\text{s.d.}}\left(\overline{\hat{\bar{s}}_p}\right)$	$\text{MSE}\left(\overline{\hat{\bar{s}}_p}\right)$ $\cdot 10^2$	$\overline{\hat{\bar{\lambda}}_p}$	$\widehat{\text{s.d.}}\left(\overline{\hat{\bar{\lambda}}_p}\right)$	$\text{MSE}\left(\overline{\hat{\bar{\lambda}}_p}\right)$ $\cdot 10^3$	$\pi_{\text{sel,BIC}}$
MSIR		0.27	0.01	1.34	0.034	0.001	0.212	0%
MSIRW	CW	0.17	0.01	0.03	0.046	0.002	0.008	16%
	AW	0.16	0.01	0.02	0.047	0.002	0.005	2%
MSIRWb	CW	0.16	0.01	0.02	0.048	0.002	0.005	**58%**
	AW	0.15	0.01	0.01	0.048	0.002	0.004	5%
MSIRWb-ext	CW	0.15	0.01	0.02	0.048	0.002	0.005	4%
	AW	0.15	0.01	0.01	0.049	0.002	0.005	0%
MSIRS	CW	0.24	0.01	0.78	0.058	0.005	0.114	12%
	AW	0.25	0.02	0.94	0.061	0.005	0.182	2%
MSIRS-ext	CW	0.25	0.01	0.85	0.055	0.011	0.162	0%

Table D.7: Simulation study results for PVB19 considering MSIRWb-ext AW as the 'true' model and using ML-estimates for BE as parameter values: $q = 0.085$, $\varepsilon_1 = 0.013$, $\varepsilon_2 = 0.000$, and $\varphi = 0.35$ ($R_0 = 3.75$, $\bar{s}_p = 0.12$, $\bar{\lambda}_p = 0.054$, $n_s = 198$).

model	waning	$\bar{\hat{q}}$	$\widehat{\text{s.d.}}\left(\bar{\hat{q}}\right)$	$\text{MSE}\left(\bar{\hat{q}}\right)$ $\cdot 10^3$	$\bar{\hat{R}}_0$	$\widehat{\text{s.d.}}\left(\bar{\hat{R}}_0\right)$	$\text{MSE}\left(\bar{\hat{R}}_0\right)$	$\pi_{\text{sel,AIC}}$
MSIR		0.056	0.001	0.818	2.49	0.06	1.59	0%
MSIRW	CW	0.072	0.003	0.172	3.18	0.13	0.34	0%
	AW	0.079	0.003	0.042	3.49	0.14	0.08	6%
MSIRWb	CW	0.076	0.003	0.093	3.35	0.15	0.18	0%
	AW	0.086	0.003	0.012	3.79	0.14	0.02	21%
MSIRWb-ext	CW	0.076	0.004	0.088	3.36	0.16	0.17	1%
	AW	0.086	0.003	0.013	3.79	0.15	0.03	**44%**
MSIRS	CW	0.064	0.002	0.434	2.83	0.08	0.84	0%
	AW	0.065	0.002	0.384	2.88	0.08	0.75	28%
MSIRS-ext	CW	0.076	0.006	0.128	3.33	0.28	0.25	0%
model	waning	$\overline{\hat{\bar{s}}_p}$	$\widehat{\text{s.d.}}\left(\overline{\hat{\bar{s}}_p}\right)$	$\text{MSE}\left(\overline{\hat{\bar{s}}_p}\right)$ $\cdot 10^2$	$\overline{\hat{\bar{\lambda}}_p}$	$\widehat{\text{s.d.}}\left(\overline{\hat{\bar{\lambda}}_p}\right)$	$\text{MSE}\left(\overline{\hat{\bar{\lambda}}_p}\right)$ $\cdot 10^3$	$\pi_{\text{sel,BIC}}$
MSIR		0.27	0.01	2.07	0.034	0.001	0.412	0%
MSIRW	CW	0.17	0.01	0.24	0.046	0.002	0.079	0%
	AW	0.14	0.01	0.05	0.051	0.002	0.019	17%
MSIRWb	CW	0.15	0.01	0.12	0.048	0.002	0.042	3%
	AW	0.12	0.01	0.01	0.055	0.002	0.005	34%
MSIRWb-ext	CW	0.15	0.01	0.11	0.049	0.002	0.040	1%
	AW	0.12	0.01	0.01	0.055	0.002	0.006	4%
MSIRS	CW	0.24	0.01	1.41	0.058	0.006	0.049	0%
	AW	0.30	0.02	3.03	0.081	0.007	0.743	**42%**
MSIRS-ext	CW	0.25	0.01	1.59	0.050	0.007	0.068	0%

Table D.8: Simulation study results for PVB19 considering MSIRS CW as the 'true' model and using ML-estimates for BE as parameter values: $q = 0.064$ and $\sigma = 0.013$ ($R_0 = 2.84$, $\bar{s}_p = 0.24$, $\bar{\lambda}_p = 0.059$, $n_s = 188$).

model	waning	$\bar{\hat{q}}$	$\widehat{\text{s.d.}}\left(\bar{\hat{q}}\right)$	$\text{MSE}\left(\bar{\hat{q}}\right)$ $\cdot 10^3$	$\bar{\hat{R}}_0$	$\widehat{\text{s.d.}}\left(\bar{\hat{R}}_0\right)$	$\text{MSE}\left(\bar{\hat{R}}_0\right)$	$\pi_{\text{sel,AIC}}$
MSIR		0.056	0.002	0.072	2.47	0.08	0.14	0%
MSIRW	CW	0.074	0.002	0.095	3.26	0.10	0.19	**21%**
	AW	0.074	0.002	0.095	3.26	0.10	0.19	10%
MSIRWb	CW	0.077	0.003	0.153	3.38	0.11	0.30	**21%**
	AW	0.075	0.002	0.123	3.32	0.11	0.24	5%
MSIRWb-ext	CW	0.076	0.003	0.144	3.36	0.13	0.28	12%
	AW	0.076	0.003	0.132	3.33	0.12	0.26	6%
MSIRS	CW	0.065	0.002	0.003	2.85	0.07	0.00	13%
	AW	0.065	0.001	0.002	2.85	0.07	0.00	7%
MSIRS-ext	CW	0.06	0.019	0.385	2.66	0.85	0.75	5%
model	waning	$\overline{\hat{\bar{s}}_p}$	$\widehat{\text{s.d.}}\left(\overline{\hat{\bar{s}}_p}\right)$	$\text{MSE}\left(\overline{\hat{\bar{s}}_p}\right)$ $\cdot 10^2$	$\overline{\hat{\bar{\lambda}}_p}$	$\widehat{\text{s.d.}}\left(\overline{\hat{\bar{\lambda}}_p}\right)$	$\text{MSE}\left(\overline{\hat{\bar{\lambda}}_p}\right)$ $\cdot 10^3$	$\pi_{\text{sel,BIC}}$
MSIR		0.27	0.01	0.10	0.034	0.001	0.651	0%
MSIRW	CW	0.16	0.01	0.63	0.047	0.002	0.156	**36%**
	AW	0.16	0.01	0.63	0.047	0.002	0.157	2%
MSIRWb	CW	0.15	0.01	0.80	0.049	0.002	0.116	34%
	AW	0.16	0.01	0.71	0.048	0.002	0.136	1%
MSIRWb-ext	CW	0.15	0.01	0.77	0.048	0.002	0.123	2%
	AW	0.16	0.01	0.74	0.048	0.002	0.131	0%
MSIRS	CW	0.24	0.01	0.01	0.06	0.004	0.017	23%
	AW	0.24	0.02	0.03	0.059	0.005	0.021	2%
MSIRS-ext	CW	0.24	0.02	0.02	0.064	0.018	0.351	1%

Table D.9: Simulation study results for PVB19 considering MSIRS AW as the 'true' model and using ML-estimates for BE as parameter values: $q = 0.065$, $\sigma_1 = 0.030$, and $\sigma_2 = 0.010$ ($R_0 = 2.86$, $\bar{s}_p = 0.29$, $\bar{\lambda}_p = 0.077$, $n_s = 199$).

model	waning	$\bar{\hat{q}}$	$\widehat{\text{s.d.}}\left(\bar{\hat{q}}\right)$	$\text{MSE}\left(\bar{\hat{q}}\right)$ $\cdot 10^3$	$\bar{\hat{R}}_0$	$\widehat{\text{s.d.}}\left(\bar{\hat{R}}_0\right)$	$\text{MSE}\left(\bar{\hat{R}}_0\right)$	$\pi_{\text{sel,AIC}}$
MSIR		0.056	0.002	0.075	2.49	0.07	0.15	0%
MSIRW	CW	0.071	0.002	0.039	3.12	0.10	0.08	0%
	AW	0.077	0.002	0.148	3.39	0.11	0.29	4%
MSIRWb	CW	0.074	0.003	0.094	3.27	0.13	0.18	2%
	AW	0.084	0.003	0.367	3.70	0.14	0.71	23%
MSIRWb-ext	CW	0.075	0.003	0.105	3.30	0.13	0.20	0%
	AW	0.084	0.003	0.372	3.70	0.14	0.73	24%
MSIRS	CW	0.063	0.002	0.004	2.80	0.07	0.01	0%
	AW	0.065	0.002	0.003	2.87	0.07	0.01	**48%**
MSIRS-ext	CW	0.074	0.005	0.100	3.25	0.21	0.19	0%
model	waning	$\overline{\hat{\bar{s}}_p}$	$\widehat{\text{s.d.}}\left(\overline{\hat{\bar{s}}_p}\right)$	$\text{MSE}\left(\overline{\hat{\bar{s}}_p}\right)$ $\cdot 10^2$	$\overline{\hat{\bar{\lambda}}_p}$	$\widehat{\text{s.d.}}\left(\overline{\hat{\bar{\lambda}}_p}\right)$	$\text{MSE}\left(\overline{\hat{\bar{\lambda}}_p}\right)$ $\cdot 10^3$	$\pi_{\text{sel,BIC}}$
MSIR		0.27	0.01	0.10	0.034	0.001	1.840	0%
MSIRW	CW	0.18	0.01	1.39	0.045	0.002	1.048	0%
	AW	0.15	0.01	2.07	0.049	0.002	0.790	5%
MSIRWb	CW	0.16	0.01	1.78	0.047	0.002	0.897	8%
	AW	0.13	0.01	2.81	0.054	0.002	0.552	29%
MSIRWb-ext	CW	0.16	0.01	1.84	0.047	0.002	0.876	0%
	AW	0.13	0.01	2.83	0.054	0.002	0.548	2%
MSIRS	CW	0.24	0.01	0.29	0.055	0.005	0.493	1%
	AW	0.29	0.02	0.04	0.077	0.007	0.054	**56%**
MSIRS-ext	CW	0.25	0.01	0.23	0.049	0.005	0.819	0%

Table D.10: Simulation study results for PVB19 considering MSIRS-ext CW as the 'true' model and using ML-estimates for BE as parameter values: $q_1 = 0.076$, $q_2 = 0.000$, and $\sigma = 0.010$ ($R_0 = 3.35$, $\bar{s}_p = 0.25$, $\bar{\lambda}_p = 0.050$, $n_s = 196$). Note that the proportionality factor q_1 for infectious individuals with a primary infection, is used as a surrogate for q.

model	waning	$\bar{\hat{q}}$	$\widehat{\text{s.d.}}\left(\bar{\hat{q}}\right)$	$\text{MSE}(\bar{\hat{q}})$ $\cdot 10^3$	$\bar{\hat{R}}_0$	$\widehat{\text{s.d.}}\left(\bar{\hat{R}}_0\right)$	$\text{MSE}\left(\bar{\hat{R}}_0\right)$	$\pi_{\text{sel,AIC}}$
MSIR		0.056	0.002	0.397	2.48	0.08	0.77	0%
MSIRW	CW	0.073	0.002	0.012	3.24	0.10	0.02	13%
	AW	0.074	0.002	0.009	3.26	0.10	0.02	10%
MSIRWb	CW	0.076	0.003	0.006	3.36	0.11	0.01	**30%**
	AW	0.076	0.003	0.007	3.35	0.11	0.01	6%
MSIRWb-ext	CW	0.076	0.003	0.008	3.37	0.13	0.02	21%
	AW	0.076	0.003	0.009	3.38	0.13	0.02	7%
MSIRS	CW	0.064	0.002	0.136	2.84	0.07	0.26	6%
	AW	0.065	0.001	0.133	2.85	0.07	0.26	6%
MSIRS-ext	CW	0.068	0.014	0.264	3.00	0.62	0.51	1%
model	waning	$\overline{\hat{\bar{s}}_p}$	$\widehat{\text{s.d.}}\left(\overline{\hat{\bar{s}}_p}\right)$	$\text{MSE}\left(\overline{\hat{\bar{s}}_p}\right)$ $\cdot 10^2$	$\overline{\hat{\bar{\lambda}}_p}$	$\widehat{\text{s.d.}}\left(\overline{\hat{\bar{\lambda}}_p}\right)$	$\text{MSE}\left(\overline{\hat{\bar{\lambda}}_p}\right)$ $\cdot 10^3$	$\pi_{\text{sel,BIC}}$
MSIR		0.27	0.01	0.06	0.034	0.001	0.257	0%
MSIRW	CW	0.16	0.01	0.72	0.047	0.002	0.013	24%
	AW	0.16	0.01	0.76	0.047	0.002	0.011	1%
MSIRWb	CW	0.15	0.01	0.93	0.049	0.002	0.005	**57%**
	AW	0.15	0.01	0.90	0.048	0.002	0.006	1%
MSIRWb-ext	CW	0.15	0.01	0.93	0.049	0.002	0.006	3%
	AW	0.15	0.01	0.95	0.049	0.002	0.005	0%
MSIRS	CW	0.24	0.01	0.02	0.059	0.004	0.095	13%
	AW	0.24	0.02	0.04	0.060	0.005	0.129	2%
MSIRS-ext	CW	0.24	0.01	0.02	0.057	0.014	0.236	0%

Samenvatting

Infectieziekten zijn ziektes in mensen, dieren of planten die veroorzaakt worden door ziektekiemen zoals bijvoorbeeld virussen, bacteriën of parasieten. Er bestaan verschillende wegen waarlangs deze ziektekiemen overgedragen kunnen worden van de ene 'gastheer' op de andere, bijvoorbeeld: via de lucht (*airborne*), druppelcontact (bijv. door te hoesten), direct of indirect fysiek contact, fecaal-orale overdracht (bijv. via besmet drinkwater of voedsel), seksueel contact of vectoroverdracht (bijv. via muggen). In deze thesis ligt de nadruk op modellen voor virale infectieziekten in mensen, die hoofdzakelijk via sociale contacten van een niet-seksuele aard overgedragen worden, bijvoorbeeld via de lucht, druppelcontact of direct fysiek contact.

In het algemeen, wanneer een persoon geïnfecteerd wordt met een virale infectieziekte, gaat het adaptief immuunsysteem complexe mechanismen activeren om de gastheer te beschermen. Het adaptief immuunsysteem bestaat uit twee soorten verdedigingsmechanismen: de cel-gemedieerde en de humorale afweer. Het is deze laatste soort die verantwoordelijk is voor de productie van virusspecifieke antilichamen die zorgen voor langetermijnbescherming. Wanneer er geen vaccinatie bestaat, wijst de aanwezigheid van virusspecifieke IgG antilichamen in het bloed op een historische infectie met het virus of op maternele antilichamen bij een pasgeborene. De voornaamste gegevensbron die gebruikt wordt in deze thesis, zijn cross-sectionele databanken bestaande uit bloedstalen. De bloedstalen worden getest met een virusspecifieke ELISA-kit (*Enzyme-Linked Immuno Sorbent Assay*). De resultaten hiervan worden serologische gegevens genoemd en geven informatie met betrekking tot de immuniteitsstatus van de individuen. Wij focussen hier op de gedichotomiseerde uitkomst van de ELISA-test die weergeeft of een persoon seropositief of seronegatief is voor het virus (*current status data*).

Deze thesis werd gemaakt in een interdisciplinair Belgisch onderzoeksconsortium om simulatiemodellen te ontwikkelen voor infectieziektenoverdracht en controleprocessen, gesteund door het Strategisch BasisOnderzoek (SBO) van het Agentschap voor Innovatie door Wetenschap en Technologie (IWT) in Vlaanderen (project 'SIMID', 060081). Het doel van de thesis was om statistische modellen te ontwikkelen gebaseerd op wiskundige modelvergelijkingen, om specifieke parameters te schatten over de persoon-tot-persoon overdracht van infectieziekten, die ofwel endemisch zijn of waarvoor actief gevaccineerd wordt, gebruik makend van gedichotomiseerde serologische gegevens. Een 'endemische' infectieziekte is een ziekte die over een langere tijd in een constante frequentie in een bevolking voorkomt. De incidentie van een endemische infectieziekte kan cyclische epidemieën ondergaan over de tijd, maar fluctueert steeds rond een stationair gemiddelde. Het schatten van zulke parameters is belangrijk, omdat het helpt om leeftijdsspecifieke patronen van ziekteverspreiding op populatieniveau af te leiden en te begrijpen. Verder worden deze parameters ook gebruikt in modellen om universele vaccinatieprogramma's te plannen en op te volgen, en om controlemaatregelen (schoolsluiting, vaccinatie, antivirale middelen, enzovoort) te evalueren wanneer een epidemie uitbreekt.

Een van die belangrijke parameters is de 'infectiedruk', de snelheid waarmee een vatbaar persoon geïnfecteerd wordt met een infectieziekte. Een ander basisconcept is de 'wie verkrijgt infectie van wie'- matrix (*'Who Acquires Infection From Whom'* of WAIFW matrix). Deze matrix geeft de leeftijdsspecifieke overdrachtsintensiteiten weer over twee dimensies, namelijk de leeftijd van diegene die vatbaar is voor de infectieziekte en de leeftijd van diegene die geïnfecteerd is. Hoe groter de overdrachtsintensiteit tussen twee leeftijdsgroepen, i.e. de frequentie van doeltreffende contacten tussen twee individuen uit deze leeftijdsgroepen, des te groter de kans dat het virus overgedragen wordt, gegeven dat één van de twee betrokkenen besmettelijk is en de andere vatbaar. In het verleden was het zeer moeilijk om de WAIFW matrix te kwantificeren omdat er geen gegevens waren over contactpatronen. Toen werd de WAIFW matrix voornamelijk geschat met behulp van de methode die gepopulariseerd werd door het boek van Anderson and May (1991). Deze methode veronderstelt dat de WAIFW matrix een bepaalde structuur heeft (*mixing pattern*) die geparametriseerd wordt onder een aantal beperkingen, zodat alle parameters identificeerbaar zijn. Gebruik makend van de wet van massa-actie voor de leeftijdsspecifieke infectiedruk, worden de parameters vervolgens geschat op basis van serologische gegevens.

Hoewel deze Anderson and May (1991) methode een realistischer alternatief aanbiedt voor de sterke veronderstelling van *homogeneous mixing*, die overeenkomt met een constante WAIFW matrix, zijn er ook nadelen aan verbonden. De keuze

van de structuur en de verdeling in leeftijdsgroepen is eerder subjectief en berust op een 'prior' idee dat de onderzoeker heeft over sociale contactpatronen of vatbaarheid/besmettelijkheid. Verder houdt de methode sterke parametrische veronderstellingen in die in de praktijk kunnen leiden tot onrealistische discontinuïteiten. Tenslotte kunnen verschillende matrixstructuren in gelijke mate ondersteund worden door de data, terwijl ze verschillende schattingen opleveren voor gerelateerde parameters zoals het basisreproductiegetal (Greenhalgh and Dietz, 1994). Het basisreproductiegetal R_0 stelt het gemiddeld aantal secundaire gevallen voor, voortgebracht door één typisch geïnfecteerd individu in een totaal vatbare populatie.

Wallinga *et al.* (2006) argumenteerden dat enquêtes over sociale contacten een nuttige bron van informatie zouden zijn om de persoon-tot-persoon overdracht van infectieziekten te modelleren en ze stelden een alternatieve schattingsmethode voor. Ze initieerden de 'sociale-contact-hypothese': leeftijdsspecifieke overdrachtsintensiteiten zijn recht evenredig aan frequenties van verbale contacten die geschat kunnen worden vanuit contactbevragingen. Door de geschatte contactfrequenties te integreren in een wiskundig transmissiemodel en dit te contrasteren tot een serologische dataset, kan de WAIFW matrix voor een bepaalde infectieziekte geschat worden. In navolging van dit onderzoek werd in het POLYMOD project een grootschalige enquête uitgevoerd over contactgedrag in acht Europese landen (Mossong *et al.*, 2008b). Deelnemers aan de enquête dienden gedurende één dag al hun contacten te rapporteren in een dagboekje. Een contact tussen twee personen werd gedefinieerd als een uitwisseling van tenminste drie woorden in elkaars nabijheid en/of een fysieke aanraking (bijv. een hand of kus geven). Het dagboekje bevat informatie over de deelnemer zelf maar ook details over zijn/haar contacten zoals de leeftijd en het geslacht van de betrokkene en de plaats, duur, frequentie en al dan niet fysieke aard van het contact.

In Hoofdstuk 4 hebben we een grondige analyse gemaakt van de Belgische contactenquête. In tegenstelling tot de andere Europese landen dienden de deelnemers hun contacten gedurende twee dagen te rapporteren, namelijk tijdens een weekdag en een dag in het weekend. Twee *data mining* technieken, namelijk associatieregels en classificatiebomen, toonden aan dat er robuuste associaties bestaan tussen verschillende indicatoren van 'intieme' contacten i.e. met een hoger risico op infectieziekteoverdracht. Deze indicatoren zijn bijvoorbeeld contacten die thuis plaatsvinden, langer dan vier uur duren, dagelijks gebeuren of gepaard gaan met fysieke aanrakingen. Het effect van verschillende factoren op het totaal aantal gerapporteerde contacten werd onderzocht, gebruik makend van *weighted generalized estimating equations* zodat de correlatie tussen de twee dagen in rekening gebracht kon worden. We stelden vast dat het aantal gerapporteerde contacten stijgt wanneer de huishoud-

grootte toeneemt. Hetzelfde effect werd geobserveerd bij kinderen wanneer het aantal leerlingen in de klas toeneemt en bij volwassenen wanneer ze tewerkgesteld zijn of voortgezet onderwijs volgen. Anderzijds is er tijdens de schoolvakantie een significante daling van de dagelijkse contactfrequentie voor kinderen en adolescenten.

In Hoofdstuk 5 hebben we de methodologie van Wallinga *et al.* (2006) om de WAIFW matrix te schatten gebruik makend van sociale contactgegevens, verder verfijnd in een toepassing voor waterpokken in België. We hebben een flexibel alternatief voorgesteld voor de laagdimensionale, parametrische schattingsmethode voor contactfrequenties: een semiparametrisch, bivariaat *smoothing* model dat toelaat om een continu 3D contactoppervlak te schatten. Dit levert een betere fit op voor de sociale contactgegevens. Via deze schattingsmethode vinden we voor alle Europese landen in het POLYMOD project een gemeenschappelijk patroon terug: mensen maken voornamelijk contact met leeftijdsgenoten en met hun (klein)kinderen of (groot)ouders. Er is echter nog ruimte voor verbetering, aangezien het model geen rekening houdt met nul-inflatie, *digit preference* of clustering van het aantal contacten, hoewel de laatste twee aspecten in rekening worden gebracht in de niet-parametrische *bootstrap* procedure voor de overdrachtsintensiteiten. Naast de variabiliteit die voortkomt uit de serologische data, erkent deze procedure immers ook die variabiliteit die voortkomt uit de contactgegevens. Dit heeft een duidelijk effect op de precisie van de parameterschattingen.

De sociale-contact-hypothese van Wallinga *et al.* (2006) dat de WAIFW matrix recht evenredig is aan het contactoppervlak, kan in vraag gesteld worden. De contacten die gerapporteerd worden in de enquête gelden immers als benadering voor die gebeurtenissen waarbij een infectie via de lucht overgedragen zou kunnen worden en zijn zeker niet alomvattend. Verder kan het zijn dat er leeftijdsspecifieke karakteristieken bestaan met betrekking tot vatbaarheid en besmettelijkheid die niet vervat zitten in de contactfrequenties, zoals het aantal dagen dat men besmettelijk is, afscheiding van slijmen en persoonlijke hygiëne. Een verbeterde fit voor de seroprevalentie van het waterpokkenvirus wordt verkregen via een nieuwe methode waarbij de WAIFW matrix ontrafeld wordt in twee leeftijdsspecifieke variabelen: het contactoppervlak en een leeftijdsspecifieke evenredigheidsfactor $q(a, a')$.

Ondanks het feit dat de methode gebaseerd op sociale contactgegevens de belangrijkste nadelen van de traditionele Anderson and May (1991) methode aanpakt, blijven er twee aspecten van onzekerheid bestaan: de keuze van het soort contact dat de eigenlijke overdracht van infectieziekten drijft, en de keuze van een (parametrisch) model dat het contactoppervlak relateert tot de overdrachtsintensiteiten. Van vijf pre-gedefinieerde soorten van contacten bleken contacten die langer duren dan 15

minuten en gepaard gaan met een fysieke aanraking, het best in staat te zijn om het leeftijdsspecifiek serologisch profiel voor het waterpokkenvirus te beschrijven. Conditioneel op dit soort contact, resulteerden verschillende modellen voor $q(a, a')$ in een gelijkaardige fit, doch met verschillende schattingen voor het basisreproductiegetal R_0. Concepten van multi-model inferentie werden toegepast om dit probleem van modelselectie-onzekerheid te overbruggen, waarbij een model-gemiddelde schatting werd berekend voor R_0.

Verder werd in deze thesis de 'sociale contact'-methodologie uit Hoofdstuk 5 uitgebreid om fundamentele immunologische processen voor parvovirus B19 (PVB19) te kunnen bestuderen. Algemeen wordt aangenomen dat de IgG antilichamen, die door de mens aangemaakt worden na een infectie met PVB19, levenslange bescherming bieden (MSIR model). In dat geval zou het geobserveerd percentage seropositieven monotoon moeten toenemen over de leeftijd. Verschillende databronnen vertonen echter een seroprevalentie waarbij een steile, monotone stijging over de leeftijd gevolgd wordt door een dal of plateau voor volwassenen tussen 20 en 40 jaar. In Hoofdstuk 6 hebben we aangetoond dat andere compartimentele modellen meer plausibel zijn voor de geobserveerde leeftijdsspecifieke serologische profielen in vier van de vijf bestudeerde Europese landen. Enerzijds betreft dit het MSIRW model waarbij antilichamen langzaam afnemen over de tijd mogelijks gevolgd door een natuurlijke boosting van het immuunsysteem door contact met iemand die besmet is met PVB19. Anderzijds betreft dit het MSIRS model waarbij men na een periode van bescherming terug vatbaar wordt voor een PVB19 infectie. Deze modellen zijn meer plausibel in vergelijking met de hypothese van levenslange immuniteit.

Op basis van één seroprevalentiestudie is het echter moeilijk te zeggen of een scenario van boosting van lage immuniteit (MSIRW) al dan niet een scenario van reïnfecties (MSIRS) het meest waarschijnlijk is voor PVB19. Nochtans is dit belangrijk gezien de impact op de geschatte leeftijdsspecifieke infectiedruk en het daaraan gerelateerd risico van een infectie tijdens de zwangerschap. De geschatte frequentie van een PVB19 infectie tijdens de zwangerschap en het jaarlijks aantal vruchtdoden dat daaraan te wijten is, verschillen niet sterk voor een MSIRW en een MSIR model. Zo variëren de schattingen voor het jaarlijks aantal vruchtdoden tussen 23 en 31 voor België in 2003. Maar gebaseerd op een MSIRS scenario, waarbij iemand gedurende zijn/haar leven meerdere infecties kan ondergaan, wordt het risico van een PVB19 infectie tijdens de zwangerschap wel veel hoger geschat, tot 77 jaarlijkse vruchtdoden in België (2003). Dit is mogelijk omdat de meeste secundaire infecties waarschijnlijk zonder specifieke of met atypische symptomen verlopen en dus niet opgemerkt worden door traditionele rapporteringssystemen.

Tenslotte hebben we in Hoofdstuk 7 de bestaande methodes van Gay (2000) en Altmann and Altmann (2000) om vaccinatiecouvertures te schatten vanuit trivariate seroprevalentiegegevens - voor infectieziekten waarvoor een trivalente vaccinatiestrategie bestaat - besproken en geïllustreerd. Hiertoe werden twee serologische datasets voor België en Ierland met testresultaten voor mazelen, bof en rubella (MBR), gebruikt. Het basisidee komt van Gay (2000): "hoe groter de trivalente vaccinatiecouverture in een bepaald leeftijdscohort, des te groter de mate waarmee de seropositiviteit voor de drie infectieziekten overeenkomt binnen een individu." Hij stelde modelvergelijkingen op in termen van vaccinatiecouverture, kansen van seroconversie en blootstellingskansen. Hoewel de algebraïsche methode van Altmann and Altmann (2000) elegante, exacte oplossingen voor deze vergelijkingen oplevert, is het vanuit statistisch oogpunt minder interessant omdat het de variabiliteit afkomstig uit de data negeert en kan leiden tot waarden die biologisch gezien onrealistisch zijn.

In deze thesis hebben we Gay's schattingsmethode veralgemeend door de afhankelijkheid tussen het oplopen van mazelen, bof en rubella expliciet in rekening te brengen. Deze afhankelijkheid tussen de blootstellingskansen vloeit voort uit het feit dat sociale contacten aan de basis liggen van de effectieve overdracht van deze infectieziekten, wat we als hypothese voorop stelden. Om de associaties te modelleren werd het Bahadur model voor trivariate binaire data gebruikt, hetgeen een daling in de geschatte MBR vaccinatie couverture en een stijging in de overeenkomstige geschatte variabiliteit teweegbracht. Omwille van de beperkte parameterruimte van het Bahadur model, verkennen we momenteel ook alternatieven zoals het trivariate Dale model. Verder werd Gay's verzadigd model voor de blootstellingskansen in functie van de leeftijd vervangen door een eenvoudiger doch flexibel model met beperkte kubische *splines*, wat een verbetering van het model opleverde.

Er zijn een aantal belangrijke, gerelateerde aspecten die we niet bestudeerd hebben in deze thesis, zoals bijvoorbeeld: testmisclassificatie voor serologische gegevens, de invloed van cyclische epidemieën op seroprevalentieschaal, het gebruik van andere soorten databronnen voor infectieziekten zoals incidentiegegevens, de veronderstelling van een constant contactpatroon over de tijd, een realistisch parametrisch model voor het contactoppervlak en een vergelijking van netwerkmodellen met modellen steunend op de wet van massa-actie. Dit zouden interessante onderwerpen kunnen zijn voor toekomstig onderzoek.