The statistical analysis of infectious disease data from the PIENTER-2 project

20-wöchige Abschlussarbeit im Rahmen der Prüfung im Studiengang Angewandte Statistik an der Universität Göttingen

Funding Acknowledgement: This work was supported by the Deutsche Forschungsgemeinschaft (DFG; German Research Foundation, grant number UN 400/2-1).

vorgelegt am:	16. August 2019
von:	Anna-Maria Kloidt
Matrikelnummer:	21741851
aus:	Eschwege
Referent:	PD Dr. Steffen Unkel
Korreferent:	Dr. Andreas Leha

Contents

Lists of tables and figures **II**

1 Introduction **1**

2 The PIENTER-2 project **3**

3 Univariate analysis **6**
3.1 Measles . . . 6
3.2 Mumps . . . 7
3.3 Rubella . . . 8
3.4 Varicella / Herpes Zoster . . . 9
3.5 Cytomegalovirus disease . . . 10
3.6 Toxoplasmosis . . . 11
3.7 Hepatitis A . . . 12
3.8 Hepatitis B . . . 13
3.9 Human Papilloma Virus infection . . . 14

4 Bivariate analysis **17**
4.1 Heterogeneity . . . 17
4.2 Measures of associations . . . 17
4.2.1 Odds ratio . . . 18
4.2.2 An alternative measure $\phi(x)$ . . . 18
4.3 Modelling of associations . . . 22
4.3.1 Pairs of infections with similar mode of transmission . . . 26
4.3.2 Pairs of infections with different mode of transmission . . . 42

5 Estimation of key epidemiological parameters **50**
5.1 Basic reproduction number R_0 . . . 50
5.2 Social contact data . . . 50
5.3 Estimation of R_0 . . . 51

6 Conclusion **56**

References **58**

Appendices **61**

List of Tables

1 Overview number of invited individuals and participants for the PIENTER-2 study. . . . 4
2 Main route of transmission and information about available vaccine for analysed infections. . . . 16
3 Associations ($\bar{\phi}$) between paired infections over all age groups (0-79 years). . . . 21
4 Associations ($\bar{\phi}$) between paired infections from age 21. . . . 21
5 Fitting results for Measles and Mumps infection data for the nationwide sample. 27
6 Fitting results for Measles and Rubella infection data. . . . 30
7 Fitting results for Mumps and Rubella infection data for the nationwide sample. 31
8 Fitting results for Human Papilloma Virus 16 and 18 infection data. . . . 33
9 Fitting results for Human Papilloma Virus 16 and 45 infection data. . . . 35
10 Fitting results for Toxoplasma and Hepatitis A Virus (HAV) infection data for the nationwide sample. . . . 38
11 Fitting results for HAV and Hepatitis B Virus (HBV) infection data for the nationwide sample from age class 9. . . . 43
12 Fitting results for Toxoplasma and Cytomegalovirus infection data for the nationwide sample. . . . 44
13 Fitting results for Toxoplasma and Varicella Zoster Virus (VZV) infection data for the nationwide sample. . . . 47
14 Estimates for the proportionality factor q and the basic reproduction number R_0. 55

List of Figures

1 Participating municipalities in the PIENTER-2 study. . . . 3
2 Number of participants per age stratum in the nationwide sample, stratified by gender, created with 'ggplot' from the ggplot2 package (Wickham, 2016). . . . 5
3 Seroprevalence plots of (a) the nationwide sample and (b) the low vaccination coverage sample for Measles. . . . 7
4 Seroprevalence plots of (a) the nationwide sample and (b) the low vaccination coverage sample for Mumps. . . . 8
5 Seroprevalence plots of (a) the nationwide sample and (b) the low vaccination coverage sample for Rubella. . . . 9
6 Seroprevalence plot of the nationwide sample for Varicella. . . . 10
7 Seroprevalence plot for CMV disease separated into individuals with Dutch nationality and non-Western nationality. . . . 11
8 Seroprevalence plot of the nationwide sample for Toxoplasmosis. . . . 12

9 Seroprevalence plot of the nationwide sample for Hepatitis A. 13
10 Seroprevalence plot of the nationwide sample for Hepatitis B. 14
11 Seroprevalence plots of the nationwide sample for (a) HPV 16 infection and (b) HPV 45 infection. 15
12 Seroprevalence plot of the nationwide sample for HPV 18 infection. 15
13 Observed and fitted associations between the infections Measles and Mumps. . 26
14 Observed seroprevalences of Measles and Mumps and fit of the 2-component frailty model to the data. 28
15 Observed and fitted associations between the infections Measles and Rubella. . 29
16 Observed seroprevalences of Measles and Rubella and fit of the 2-component Dirichlet-multinomial frailty model to the data. 31
17 Observed and fitted associations between the infections Mumps and Rubella. . 32
18 Observed seroprevalences of Mumps and Rubella and fit of the 2-component frailty model to the data. 33
19 Observed and fitted associations between the infections HPV 16 and 18. 34
20 Observed seroprevalences of HPV 16 and 18 and fit of the 1-component frailty model to the data. 35
21 Observed and fitted associations between the infections HPV 16 and 45. 36
22 Observed seroprevalences of HPV 16 and 45 and fit of the (a) inverse Gaussian and (b) 1-component frailty model to the data. 37
23 Standard deviation of the frailty (——) with 95% CIs obtained from fitting the 1-component gamma model to (a) HPV 16 and 18 and (b) HPV 16 and 45. . . . 37
24 Observed and fitted associations between the infections Toxoplasma and HAV. 38
25 Observed seroprevalences of Toxoplasmosis and Hepatitis A and fit of the 1-component frailty model to the data. 39
26 Standard deviation of the frailty, similar route of transmission. 40
27 Pointwise absolute deviances for the multinomial model and its compound Dirichlet-multinomial counterpart I. 41
28 Observed and fitted associations between the infections HAV and HBV. 42
29 Observed seroprevalences of Hepatitis A and B and fit of the 1-component frailty model to the data. 44
30 Observed and fitted associations between the infections Toxoplasma and CMV. 45
31 Observed seroprevalences of Toxoplasmosis and CMV disease and fit of the 1-component Dirichlet-multinomial frailty model to the data. 46
32 Observed and fitted associations between the infections Toxoplasma and VZV. 46
33 Observed seroprevalences of Toxoplasmosis and Varicella and fit of the 1-component Dirichlet-multinomial frailty model to the data. 47

34 Pointwise absolute deviances for the multinomial model and its compound Dirichlet-multinomial counterpart II. 48
35 Standard deviation of the frailty, different route of transmission. 49
36 Contact matrix for the Dutch population. 51
37 Observed and fitted proportions of the 1-component frailty model for HPV 16 and 18. 63
38 Observed and fitted proportions of the 1-component frailty model for HAV and HBV. 63
39 Observed and fitted (one-component Dirichlet-multinomial) cumulative force of infection for Toxoplasmosis and CMV. 64
40 Observed and fitted (one-component Dirichlet-multinomial) cumulative force of infection for Toxoplasmosis and VZV. 64

Abbreviations

NIP Nationwide Immunisation Program

PIENTER Peiling Immunisatie Effect Nederland Ter Evaluatie van het Rijksvaccinatieprogramma

LVCS low vaccination coverage sample

CMV Cytomegalovirus

MMR Measles-Mumps-Rubella

VZV Varicella Zoster Virus

HPV Human Papilloma Virus

HAV Hepatitis A Virus

HBV Hepatitis B Virus

IU/ml International Units per millilitre

IU/l International Units per litre

RU/ml Relative Units per millilitre

CRF cross-ratio function

OR odds ratio

VIF variance inflating factor

AIC Akaike information criterion

CI confidence interval

WAIFW Who Acquires Infection From Whom

1 Introduction

Infectious diseases, caused by bacteria, viruses, fungi or parasites, are one of the leading causes of death worldwide (Bundesgesundheitsministerium, 2019). They can be transmitted through different routes, such as droplets or sexual intercourse. Infection does not always lead to an outbreak of a disease but the infected individuals can still function as a carrier, infecting other individuals (Bundesgesundheitsministerium, 2019). Highly communicable infectious diseases can lead to devastating epidemics. For instance, the Ebola outbreak in West Africa from 2014 to 2016 which started in Guinea and spread across land borders to Sierra Leone and Liberia (WorldHealthOrganisation, 2019). To prevent such outbreaks knowledge of the patterns of infectious disease spread is needed. To gain insight into the proportion of individuals in a population who are infected and vaccinated, respectively (called seroprevalence) serological surveys are a key resource. These surveys collect blood serums to measure the levels of antibodies against infectious diseases to get a better understanding which groups are at risk and how well a population is protected. These data are used to evaluate the effectiveness or the need of a vaccination.

A key factor for the spread of an infectious disease is the individuals "activity level", also called heterogeneity. For instance, an individual with changing sexual partners is more likely to acquire a sexual transmitted infectious disease than an individual with only one sexual partner. These heterogeneities are important to take into account when working with statistical and mathematical models of infectious diseases. When using such models, a contact rate between individuals needs to be specified. This can be difficult because relevant heterogeneities may be unknown or hard to measure. For example, for sexual transmitted diseases it is easy to define a contact, "having sexual intercourse", and to quantify the degree of heterogeneity. Nevertheless, for most infections, including foodborne or waterborne infections, no one event can be clearly defined as a contact. For the analysis it is not enough to get data on infected individuals but researchers also need data from the individuals who infected them and other individuals with whom they may have come into contact with. One possible method is the use of diaries to estimate the number of social contacts between individuals and to get insight into the spread of airborne infections. This method is very costly and unknown heterogeneities cannot be considered in research design. Another possibility is to use correlations between infections in individuals for quantifying relevant heterogeneities which is the main topic of this thesis. The advantage is that no exact definition of the heterogeneity is needed. Instead an association measure is used, for a better understanding of the transmission routes. Shared frailty models can be used to estimate the degree of heterogeneity.

In this thesis serological data from a survey taken in the Netherlands, called the PIENTER-2

study, a Dutch acronym for "Peiling Immunisatie Effect Nederland Ter Evaluatie van het Rijksvaccinatieprogramma" are used. Serological data from a representative sample for many infectious diseases are available. Furthermore, additional information such as age, gender and nationality is given. In addition, social contact data are available which are used to estimate a main threshold parameter, the basic reproduction number R_0. The main used data are the dichotomized version of the levels of antibodies which describes whether an individual has experienced the infection or not. These data are called multivariate current status data which means that the exact time of an event is unknown, it is only known that the event did or did not take place at a given time in point. These data are interpreted in this analysis and the main focus lies on the application of different frailty models to the given data. Afterwards, the results are compared to previous studies.

Section 2 of this thesis introduces the PIENTER-2 project and the data collection is illustrated. An univariate analysis of selected infectious diseases and their seroprevalences in the Dutch population follows in Section 3. The alternative measure of associations and the use of frailty models are further discussed and applied in Section 4 on pairs of infections. In Section 5 the basic reproduction number R_0 is estimated for some infectious diseases using serological and social contact data. A conclusion of all results is given in Section 6.

2 The PIENTER-2 project

The PIENTER-2 project is a serosurveillance study in the Netherlands. It was performed from February 2006 to June 2007 and it is the second serosurveillance study in the Netherlands. The first study was carried out in 1995/1996. The aim of these studies is to analyse the seroprevalence and immunity of infectious diseases in the Dutch population and to evaluate the effectiveness of the Nationwide Immunisation Program (NIP). For a representative sample of the general population 40 municipalities were randomly selected and 380 individuals aged between 0 and 79 years were invited from each cluster. In twelve of these municipalities an additional sample of non-Western migrants was taken to analyse this group separately and to learn more about their immunity against vaccine preventable diseases. Moreover a sample from eight municipalities with low immunisation coverage was taken. Individuals from those municipalities are orthodox reformed and refuse vaccination for religious reasons. In these communities the seroprevalence is lower compared to other communities, so the risk of outbreaks is higher, such as a Measles outbreak in 1999/2000, a Rubella outbreak in 2004 and a Mumps outbreak in 2007/2008 (Van der Klis et al., 2009). The randomly selected municipalities are shown in Figure 1. All in all 24147 individuals were invited, 17341 individuals as part of the nationwide sample, 2574 individuals as part of the non-Western migrants sample and 4376 as part of the low vaccination coverage sample (LVCS) (Table 1) (Mollema et al., 2010).

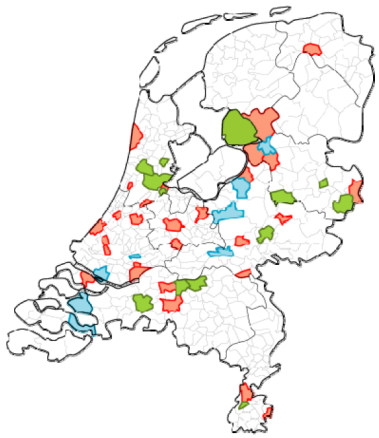


Figure 1: Participating municipalities: Red and green municipalities participated in the nationwide sample, in the green ones an oversampling of migrants was taken and the low vaccination coverage sample was taken from the blue municipalities (Mollema et al., 2010, p. 23).

Each individual was asked to fill in a questionnaire at home and give a blood sample at the clinic. Two versions of the questionnaire were available, one for children between 0 and 14 years with questions of opinion on vaccination related topics and one for 15-79 year-olds with questions of sexual history and sexual related diseases. Both questionnaires contained questions on personal details, vaccination history, health status and diseases in the past and activities

that could be possibly related to infectious diseases. At the clinic three tubes of 10 ml blood were taken from adults and adolescents and less blood was taken from children. In total 7904 blood samples were used in the analysis, 6386 of which were part of the nationwide sample (oversampling of migrants included) and 1518 serums of the LVCS. A participant was defined as an individual who donated blood and filled in the questionnaire. In the nationwide sample 6386 individuals (including the extra sampling of 646 individuals from migrant background) participated. In the LVCS 1518 individuals participated. In total 7904 (33%) individuals were part of the PIENTER-2 project (Table 1) (Mollema et al., 2010).

Table 1: Overview number of invited individuals and participants for the PIENTER-2 study.

Sample	Number of invited individuals	Number of participants	Participants diary
Nationwide	17341	5740 (33.1%)	824
Non-western migrants	2574	646 (25.1%)	0
LVCS	4376	1518 (34.7%)	0

In both samples the number of female respondents was slightly higher. Only in the age strata from 0-9 years and 70-74 years more men participated in the nationwide sample, see Figure 2, the LVCS was more balanced. Each age stratum in the nationwide sample included more than 260 individuals. The age strata 35-39 years and 50-54 years in the LVCS had the lowest response rate with only 69 individuals and the age stratum 1-4 years had the highest with 196 individuals.

The blood samples were tested for antibodies against infectious diseases. The concentration of antibodies, also called titer, gives information on whether an individual had an infection or not. Different cut-off values exist for antibody levels of different infectious diseases to indicate an infection. If the cut-off value is above a certain value an infection or a vaccination is present (Van der Klis et al., 2009). Therefore, for interpreting the test results it is important to have information on the vaccination history of each individual. So, participants were asked to bring their vaccination booklet. If no vaccination booklet was available the information was retrieved from the local authorities for registration of vaccinations (PEAs) (Mollema et al., 2010).

An additional study of the PIENTER-2 project, which is part of the European modelling project POLYMOD was incorporated. The aim of this study is to collect social contact data by means of a diary because information in social contacts is important for determining the spread of airborne infections. The participants (Table 1) were asked to record every individual they had contact with. It was distinguished between physical (kiss, handshake) or non-physical (conversation) contacts. In addition, information about the place where the contact took place was recorded (Mossong et al., 2008).

Unfortunately, the data from those diaries are not available for this analysis. Nevertheless, there are other social contact data available which were part of the questionnaire. Participants were asked to fill in the questionnaire how many conversation partners they had within a day and to

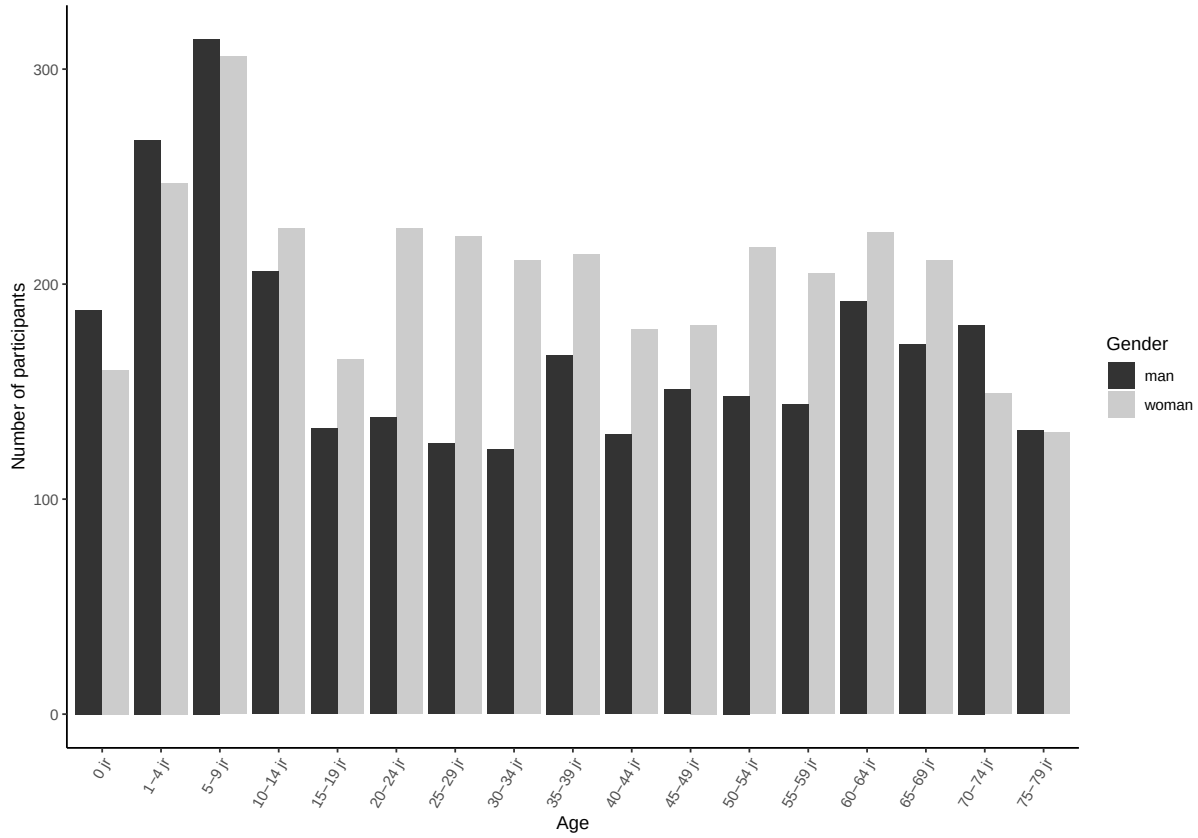


Figure 2: Number of participants per age stratum in the nationwide sample, stratified by gender, created with 'ggplot' from the ggplot2 package (Wickham, 2016).

specify the age classes of their contacts. The information of the place and the exact age of the contact is missing. Moreover, these statements are only a rough number of social contacts and not an exact number as the contacts getting from a diary (Mollema et al., 2010).

In this thesis special attention focusses on the analysis of the blood serums and the consequential results of whether an individual had an infection or not. All calculations are done with the statistical program R (R Core Team, 2018).

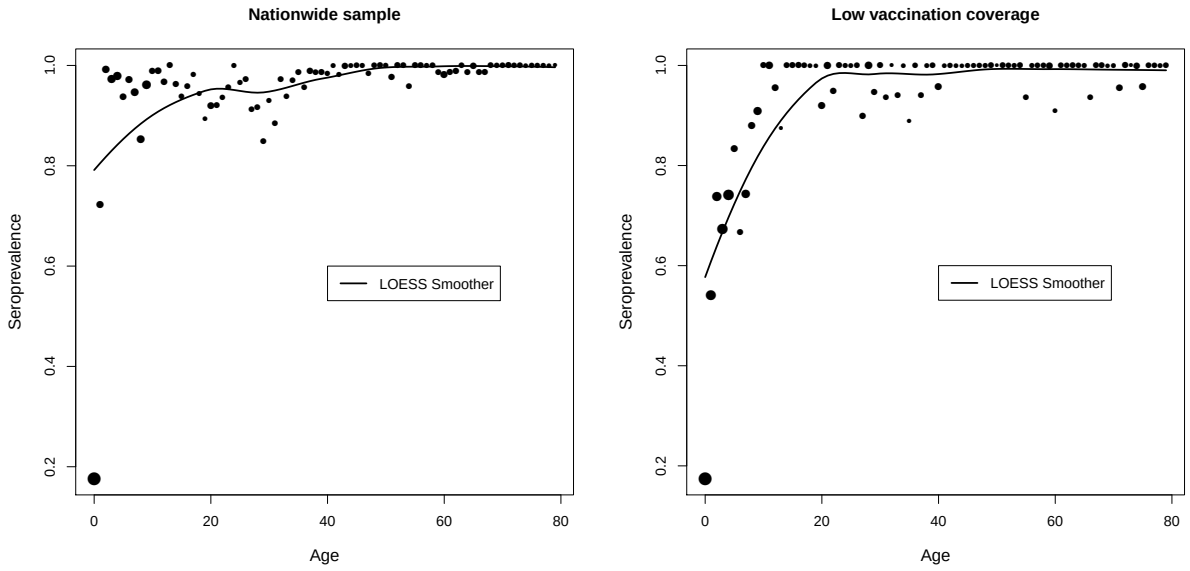
3 Univariate analysis

This section introduces the infectious diseases which are analysed in this thesis and discusses their seroprevalences in the Dutch population. In addition, the nationwide and the low vaccination coverage sample are compared and the vaccination status is illustrated.

The data for these analyses are so called current status data, also known as case I interval-censored data. These data are characterized by the fact that while the exact time of an event is unknown, it is only known that the event did or did not take place at a given time point (Unkel and Farrington, 2012, p. 665). In this study only the information whether an individual had an infection or not for a given point in time is known and the given time point is age. There is no information about when exactly the individual got infected.

3.1 Measles

Measles are an acute, highly communicable viral disease caused by the Measles virus, a member of the genus *Morbillivirus* of the family *Paramyxoviridae*. It is one of the most highly communicable infectious diseases. The transmission route is airborne by droplet, through direct contact with nasal or throat secretions of the infected individual. An indirect spread through contaminated objects is less common. The period of communicability is from four days before rash onset to four days after rash appearance and is minimal after the second day of rash. The infection starts with low-grade fever, cold (coryza) and sore throat. In the early stage the so called Koplik spots, small spots with white or bluish-white centers on an erythematous base on the buccal mucosa, are typical of Measles. At the main stage the fever rises and the rash spreads over the whole body, beginning behind the ears. Only the palms and the soles of the feet are not involved. All individuals who have not had the disease or been successfully immunised are susceptible. The acquired immunity after illness is permanent. Since 1987 a combination vaccine with Mumps and Rubella (Measles-Mumps-Rubella (MMR)) is available in the Netherlands (Waaijenborg et al., 2013) which successfully immunises individuals after two dose of vaccine (Heymann et al., 2008, pp. 389–397).
In this study the concentration of the antibodies of Measles is measured in International Units per millilitre (IU/ml) and the cut-off value is 0.2 IU/ml. Looking at the whole sample 91.8% are seropositive and 3031 individuals (38.4%) received a vaccination. Comparing both groups, vaccinated and unvaccinated, it is noticeable that the antibody titer of the vaccinated individuals is lower compared to those who had a natural infection with the Measles virus. The median in the vaccinated group is 1.45 IU/ml and 2.91 IU/ml in the unvaccinated group. In Figure 3 the age-specific seroprevalences of the nationwide sample (including the migrants sample) and the LVCS with smoothed trends (——) and the points porportional to the sample size are shown. The seroprevalences of age zero are in both samples similar. For the next age groups the



(a) *(b)*

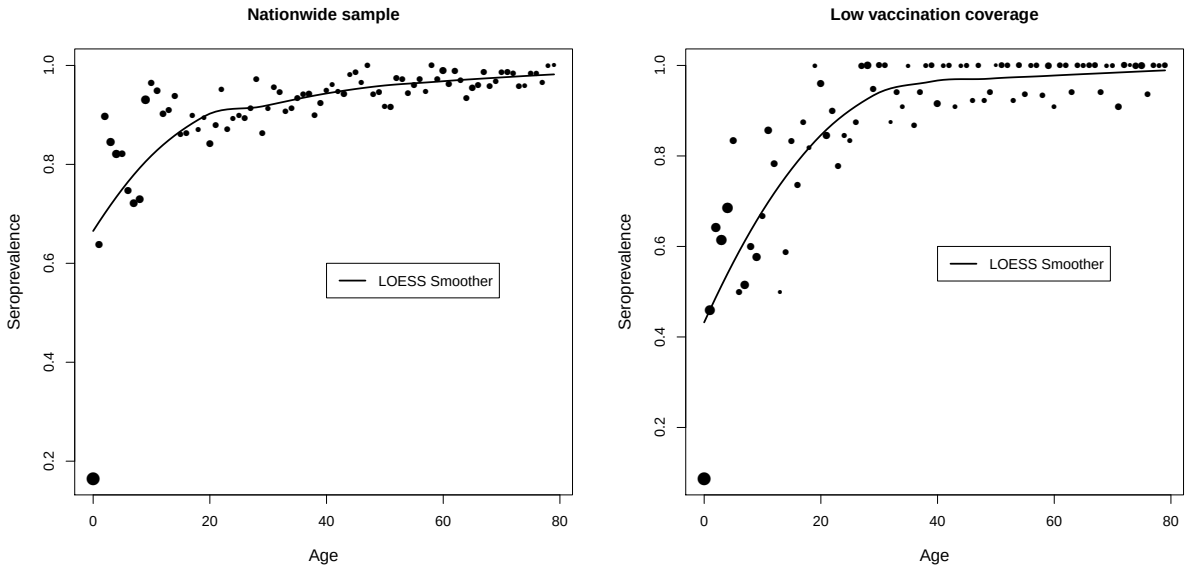
Figure 3: Seroprevalence plots of (a) the nationwide sample and (b) the low vaccination coverage sample for Measles.

seroprevalence in the nationwide sample is almost higher than 80% and for some age groups it is even 100%. In the LVCS the seroprevalences in the first age groups are lower compared to those in the nationwide sample. One reason could be that parents did not let their children get vaccinated. From the age of 20 the age-specific seroprevalences in both samples behave similar, they are all more than 80%. For more information about this infection in the Dutch population see Waaijenborg et al. (2013).

3.2 Mumps

Mumps is an acute viral disease caused by the Mumps virus, a member of the family *Paramyxoviridae*, genus *Rubulavirus*. The transmission is via droplet infection or direct contact with the saliva (e.g. kissing) of an infected individual. The infectiousness occurs from seven days before the onset of parotitis to nine days afterwards, the maximum infectiousness occurs between two days before and five days afterwards. The infection is characterised by fever, swelling and tenderness of one or more salivary glands, usually the parotid. Parotitis may be unilateral or bilateral and typically lasts 7-10 days. In countries where Mumps vaccine has not been introduced, the incidence of Mumps remains high, mostly affecting children between 5-9 years of age. Immunity after infection is generally long-lasting and a vaccine is available (see subsection Measles) (Heymann et al., 2008, pp. 419–423).

The concentration of antibodies of Mumps is measured in Relative Units per millilitre (RU/ml) and the cut-off value is 45 RU/ml. Compared to Measles the overall seroprevalence is lower with 86.3% seropositive and less individuals are vaccinated against Mumps, 2698 (34.2%). The



(a) *(b)*

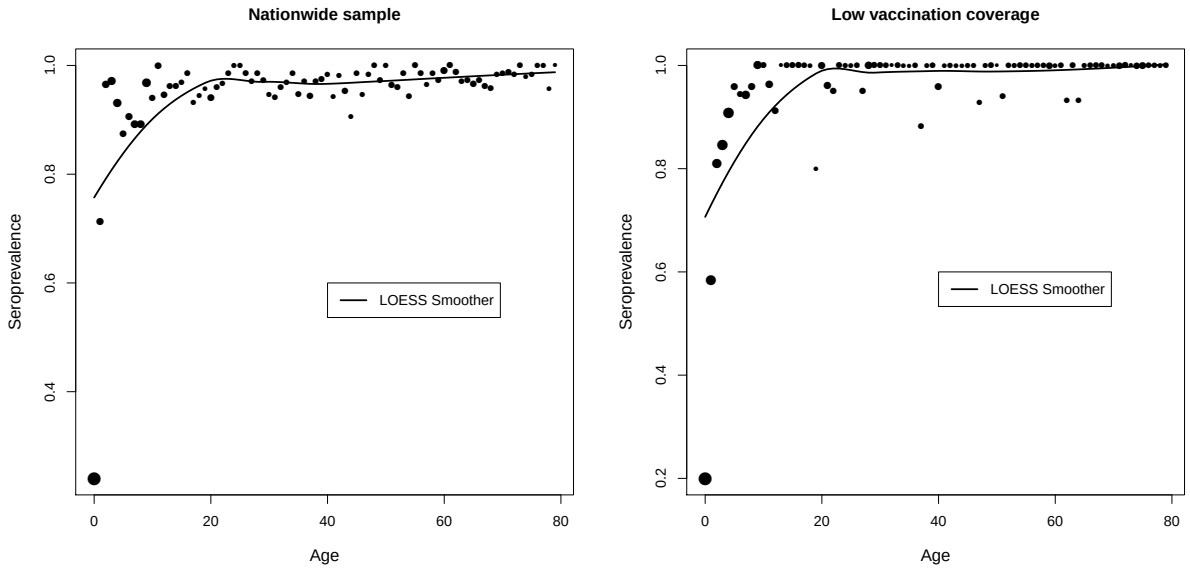
Figure 4: Seroprevalence plots of (a) the nationwide sample and (b) the low vaccination coverage sample for Mumps.

vaccination rate is higher in the nationwide sample (35.2%) than in the LVCS (29.7%). Again the higher titer is noticeable in the group with a natural infection of Mumps. The median in this group is 299.89 RU/ml and in the vaccinated group it is only 179.79 RU/ml. The age-specific seroprevalences for both samples are represented in Figure 4. In the first 20 age groups the seroprevalences are higher in the nationwide sample which may also have something to do with the refusal of vaccination. After this the seroprevalences in both samples are similar in all age groups and almost more than 80%. In the work of Smits et al. (2013) more information about Mumps in the Netherlands is available.

3.3 Rubella

Rubella is most often a mild febrileviral disease caused by the Rubella virus (family *Togaviridae*, genus *Rubivirus*). The transmission is via droplet infection through contact with nasopharyngeal secretions of infected people and it is highly communicable. The period of communicability is about one week before and at least four days after onset of rash. Prodromal symptoms are low-grade fever, headache, mild coryza and conjunctivitis but it is clinically indistinguishable from febrile rash illness. Typical of Rubella is a diffuse punctate and maculopapular rash. All individuals who have not had the disease or been successfully immunised are susceptible. After natural infection the immunity is permanent and after immunisation it is long term, probably lifelong (Heymann et al., 2008, pp. 527–532).

The titer is measured in IU/ml and the cut-off value is 10 IU/ml. The overall seroprevalence is very high with 92.1% and is also very high in both samples (nationwide 92.1% and LVCS



(a) *(b)*

Figure 5: Seroprevalence plots of (a) the nationwide sample and (b) the low vaccination coverage sample for Rubella.

91.8%). In total, 3102 individuals (39.3%) got a vaccination against Rubella, but the proportion in the LVCS is as well smaller compared to the nationwide sample. In the LVCS 33.2% of the subjects got a vaccination and 40.7% in the nationwide sample. The concentration of the titer is lower in the group of vaccinated individuals. The median in this group is 65.33 IU/ml and in the group with a natural infection it is 106.24 IU/ml. The age-specific seroprevalences (Figure 5) are nearly similar in both samples, only in the first five age groups the seroprevalences in the LVCS are lower. For all other age groups the seroprevalences are very high and the Dutch population is well protected. For more detailed information see Waaijenborg et al. (2013).

3.4 Varicella / Herpes Zoster

Varicella is a highly contagious viral disease caused by the human (alpha) herpesvirus 3, a member of the *Herpesvirus* group, the VZV. A primary infection mainly occurs in childhood (chickenpox) and is transmitted via droplet infection, through contact with infected secretions of the respiratory tract in the air or smear infection, through direct contact with vesicle fluid of skin lesions. It is characterized by fever and a rash typically consisting of 250-300 lesions in varying stages of development. After infection the virus may reactivate in a later stage, resulting in Herpes Zoster, commonly known as shingles. All individuals who have not had the disease or been vaccinated are at high risk. After infection the immunity is permanent. A vaccine is available and a vaccination in the first years of life is recommended (Heymann et al., 2008, pp. 669–675).

The titer is measured in IU/ml as well and the cut-off value for an infection or vaccination is

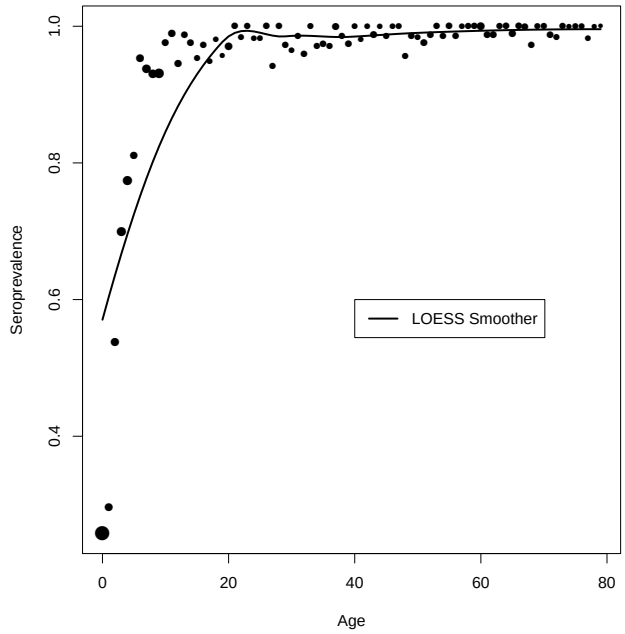


Figure 6: Seroprevalence plot of the nationwide sample for Varicella.

0.26 IU/ml. The seroprevalence of Varicella in the Dutch population is very high with 90.8% seropositives in the nationwide sample. As already mentioned a vaccine is available but the vaccination is not included in the NIP. That is because Varicella is usually a mild disease and so the acceptance of this vaccine might be low. So there are no differences between the nationwide sample and the LVCS and only the nationwide sample is analysed. The age-specific seroprevalences (Figure 6) are similar to those of Measles and Rubella. In the first age groups the seroprevalence increases sharply and from age group five it is more than 80%. In some age groups the seroprevalence is even 100%. More details about Varicella in the Netherlands can be found in the work of Waaijenborg et al. (2013).

3.5 Cytomegalovirus disease

Infection with the Cytomegalovirus (CMV) is common and often passes undiagnosed as a febrile illness without specific characteristics. The cytomegalovirus is a human (beta) herpersvirus 5, a member of the subfamily Betaherpesvirus of the family *Herpesviridae*. The virus is transmitted via secretions and excretions, including saliva, cervical secretions, semen and breast milk. A transmission via blood transfusions and organ donation is possible as well. The virus is excreted in urine and saliva for many months and may persist or be episodic for several years following primary infection. Infants infected in utero show signs and symptoms of severe generalized infection, especially involving the central nervous system and the liver. Survivors can show for example motor disabilities and hearing loss. A primary infection later in life is usually asymptomatic but may cause a syndrome similar to Epstein-Barr virus mononucleosis and Hepatitis. For this infection no vaccine is available. (Heymann et al., 2008, pp. 141–144).

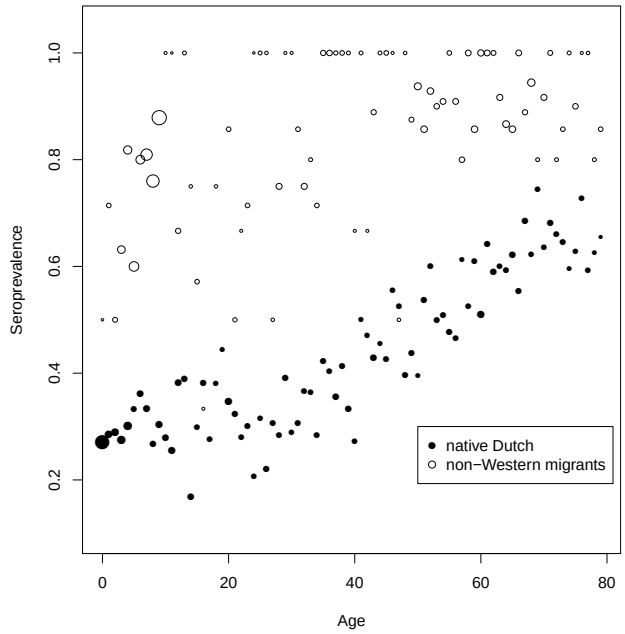


Figure 7: Seroprevalence plot for CMV disease separated into individuals with Dutch nationality and non-Western nationality.

The concentration of antibodies is measured in IU/ml and the cut-off value is 0.4 IU/ml. The seroprevalence of CMV disease depends on geographic and socio-economic factors and is generally higher in developing countries (Korndewal et al., 2015). The total seroprevalence in the nationwide sample, excluding the oversampling of non-Western migrants is 43.6%. The seroprevalence in the native Dutch population (41.7%) is much lower compared to the seroprevalence of non-Western migrants (84.3%). In Figure 7 the age-specific seroprevalences for the native Dutch population and the non-Western migrants are shown. For both groups the seroprevalences increase with age but even for the first age groups in the non-Western migrants sample the seroprevalence is more than 50%. Whereas, in the native Dutch sample the seroprevalence is lower than 50% in age group younger than 41. The highest seroprevalence is in the age group 69 with 74.5% compared to the non-Western migrants sample where the seroprevalence is higher than 90% in many age groups. For more information about CMV disease in the Dutch population and its risk factors see Korndewal et al. (2015).

3.6 Toxoplasmosis

Toxoplasmosis is a zoonosis (animal-to-human and human-to-animal transmission) caused by *Toxoplasma gondii*, an intracellular protozoan. Cats and other *felidae* are the definitive hosts, in their intestinal epithelium the parasite completes its sexual life cycle phase. There are two main transmission routes, eating raw or undercooked infected meat containing tissue cysts and inhalation of sporulated oocysts (contact with cats). The pathogen can survive several months in the soil (feline feces) and infection can occur while gardening or playing in sandboxes and

playgrounds. Two more rarely transmission routes are through blood transfusion or organ transplantation and congenital transmission to the fetus. The infection is in most cases asymptomatic. Many individuals do not detect the infection, otherwise there are flu-like symptoms with fever and lymphadenopathy. Among immunodeficient individuals (including HIV-infected) the infection can cause severe or life-threatening illness. Duration and degree of immunity are unknown but they are assumed to be long-lasting or permanent. A vaccine against this infectious disease is not available (Heymann et al., 2008, pp. 614–617).

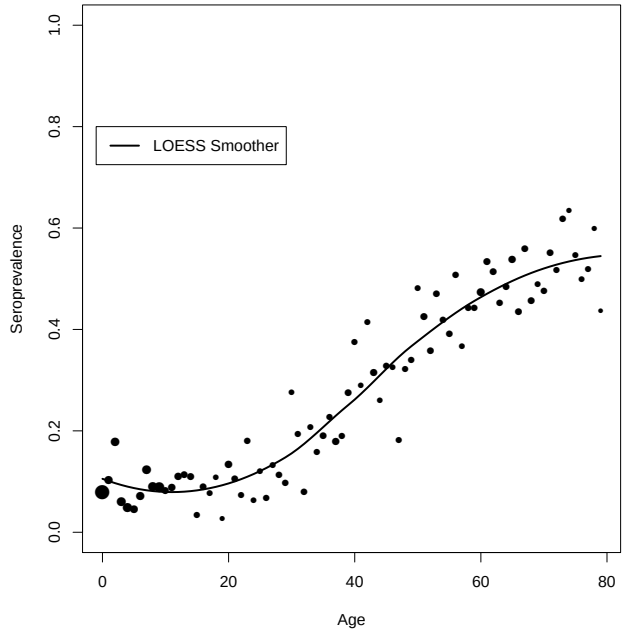


Figure 8: Seroprevalence plot of the nationwide sample for Toxoplasmosis.

Testing 5539 serums the overall seroprevalence in the nationwide sample is 25.7%. The seroprevalence increases with age (Figure 8). For the first age groups the seroprevalence is between 2.1% and 20% and at the higher age groups it is more than 50%. A detailed analysis of Toxoplasmosis in the Netherlands can be found in the work of Hofhuis et al. (2011).

3.7 Hepatitis A

Hepatitis A is an acute liver inflammation caused by the HAV, a member of the family *Picornaviridae*. The infection is transmitted via person-to-person contact by the faecal-oral route and levels of endemicity are related to hygienic and sanitary conditions. Most children have asymptomatic or unrecognised infections and play an important role in HAV transmission and serve as a source of infections for others. In most developing countries the infection is an asymptomatic or mild illness in childhood. Onset of illness in adults is usually abrupt, with fever, anorexia and nausea, followed within a few days by jaundice. Individuals living in high or intermediate endemicity areas and susceptible individuals travelling to or working in HAV-endemic coun-

tries are the main risk groups. Homologous immunity after infection probably lasts for life and long-term protection after two vaccines is possible (Heymann et al., 2008, pp. 253–257).

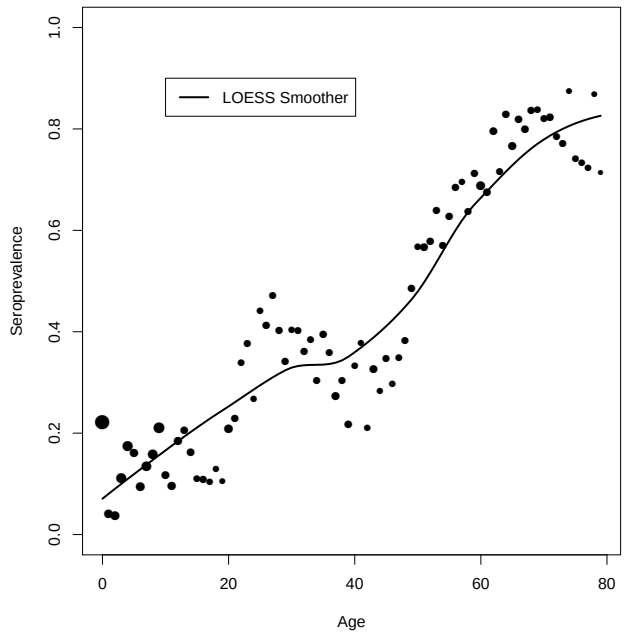


Figure 9: Seroprevalence plot of the nationwide sample for Hepatitis A.

The titer is measured in International Units per litre (IU/l) and the cut-off value is set to 10 IU/l. The overall seroprevalence in the nationwide sample is 40.5% which is associated with an increasing of vaccinations of travellers and an increased amount of immigrants. From 6229 individuals 786 of them are vaccinated against HAV. Without the vaccinated subjects the total seroprevalence is 31.1%. In Figure 9 the age-specific seroprevalences are shown. The seroprevalence increases rapidly with age and it is in the higher age groups more than 70%. For more details of Hepatitis A and its risk factors see Verhoef et al. (2011).

3.8 Hepatitis B

Hepatitis B is a liver inflammation caused by the HBV. This virus is one of the most common viral Hepatitis worldwide. Body substances that include blood and blood products are capable of transmitting HBV, e.g. semen and vaginal secretions. Major modes of HBV transmission include sexual or close household contact with an infected individual. The virus is stable on environmental surfaces for at least seven days, so a transmission via contaminated and inadequately sterilized syringes and needles is possible as well. There are two types of Hepatitis B, the acute and the chronic Hepatitis B infection. The acute one starts usually insidious, with anorexia, vague abdominal discomfort, nausea and vomiting and sometimes jaundice. A chronic Hepatitis B infection has unspecific symptoms like tiredness, myalgia and anorexia. An estimated 15% - 25% of individuals with chronic HBV infection die prematurely of either cirrhosis or hepatocellular carcinoma (HCC). About 50% of theses cases globally are attributable

to chronic Hepatitis B infection. Houshold contact and/or intercourse with infected individuals is very infectious. Also workers in health care or public safety who perform tasks involving contact with blood or blood-contaminated body fluids are at higher risk. Effective Hepatitis B vaccines are available and lasts for at least 20 years and may be lifelong (Heymann et al., 2008, pp. 257–264).

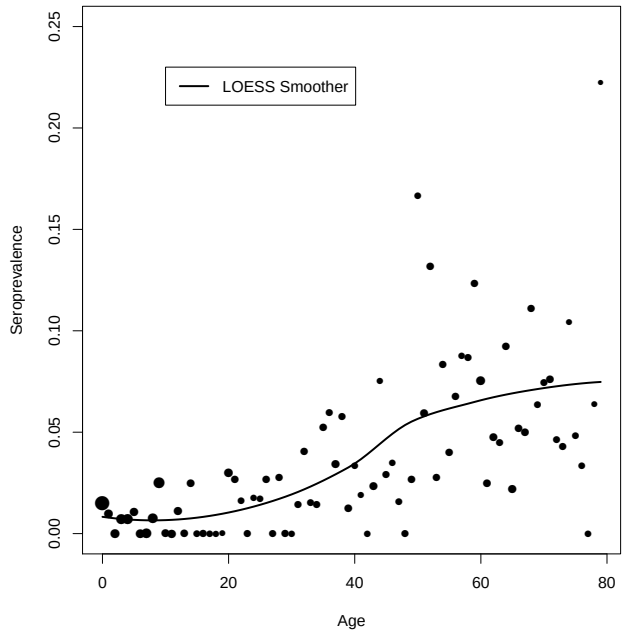


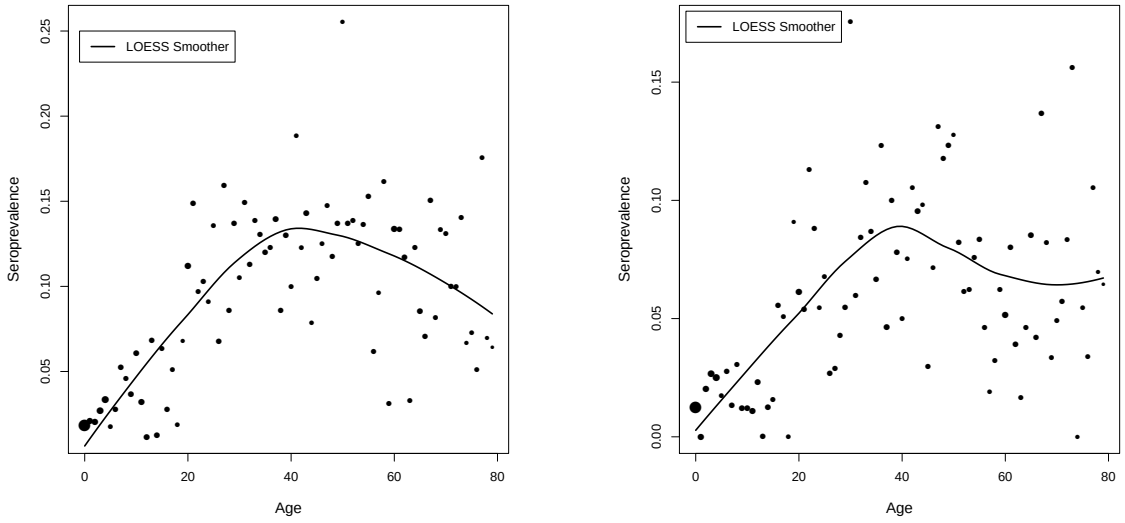
Figure 10: Seroprevalence plot of the nationwide sample for Hepatitis B.

The seroprevalence is determined by two types of antibodies. The first are the antibodies to the HBV surface antigen and the second are the antibodies to the HBV core antigen. In this thesis the seroprevalence of the antibodies to the HBV surface antigen is analysed. The overall seroprevalence is generally low in the Dutch population with 3.4%. In all, 347 individuals got a vaccination against HBV. A slight increase of the seroprevalence with age can be seen in Figure 10. For more detailed information see Hahné et al. (2012).

3.9 Human Papilloma Virus infection

The Human Papilloma Virus (HPV) infection is one of the most sexual transmitted infection and causes warts. In this study seven types of HPV were analysed, all of them are certain cancer-causing and type 16 and 18 account for more than 70% of cervical cancers. The infections with the high risk types are in most cases transient and after 1-2 years unverifiable. It is possible to be infected with more than one type simultaneously. The warts are transmitted by direct skin-to-skin contact, usually through sexual intercourse. An infection is possible at birth from mother to child as well, but rather rare. Genital warts are most frequently seen in sexually active young adults. After an infection a reinfection with the same or another type of HPV is possible, so

there is no immunity (Heymann et al., 2008, pp. 298–302). A vaccine for girls was introduced in the Netherlands in 2009 (Van der Klis et al., 2009), so it is not relevant for this analysis.



(a) *(b)*

Figure 11: Seroprevalence plots of the nationwide sample for (a) HPV 16 infection and (b) HPV 45 infection.

In this thesis the three most prevalent types, 16, 45 and 18 are analysed. HPV 16 infection has the highest seroprevalence in the Dutch population with 9.0%. The seroprevalences for HPV 45 and 18 infection are quite similar with 5.6% and 5.4%. The age-specific seroprevalences of these three types have an almost identical structure (Figure 11, 12). A sharp increase of

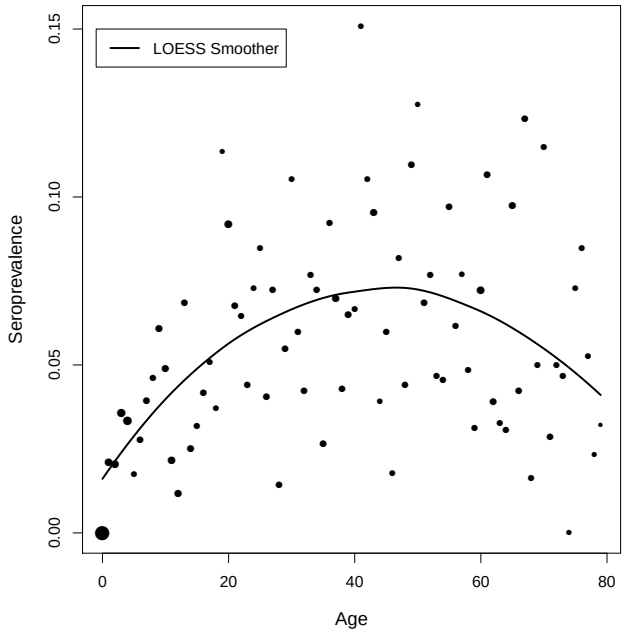


Figure 12: Seroprevalence plot of the nationwide sample for HPV 18 infection.

the seroprevalence during adolescence and an increasing trend up to the age group 40 is ob-

served. From 50 years of age onwards a declining trend can be observed. In the higher age groups the variability in the age-specific seroprevalences is much higher than in the smaller age groups. A detailed analysis of the HPV infection can be found in Scherpenisse et al. (2012).

In the next section the distinction between infectious diseases with similar or different transmission routes becomes important. Moreover, vaccinated individuals must be excluded from further analysis. Table 2 provides a short overview of the infectious diseases, their main routes of transmission and if a vaccine is available.

Table 2: Main route of transmission and information about available vaccine for analysed infections.

Infectious disease	Main transmission route	Vaccine available
Measles	Airborne, respiratory droplets or direct contact (nasal and throat secretions)	Yes
Mumps	Respiratory droplets or direct contact (saliva)	Yes
Rubella	Respiratory droplets or direct contact, aerosol	Yes
Varicella	Airborne, droplets, contact with respiratory secretions	Yes (not included in NIP)
Hepatitis A	Fecal-oral, foodborne	Yes
Hepatitis B	blood/sexual	Yes
HPV infection	direct contact (sexual intercourse)	Yes (but only since 2010)
CMV disease	Mucosal contact with any body fluid	No
Toxoplasmosis	Foodborne, oral ingestion of feline feces	No

4 Bivariate analysis

In this Section two infectious diseases are analysed together and the advantages of this method are elucidated. In Section 4.1, the term heterogeneity and its relationship to shared frailty models is explained. Different measures of associations and several frailty models are presented in Section 4.2 and 4.3 and applied to the seroprevalence data from the PIENTER-2 study.

4.1 Heterogeneity

Individuals show variation in factors that are relevant to the transmission of infectious diseases. This variation is also called heterogeneity. There are factors such as age, gender or sexual activity that are measurable without great effort. Information, such as the number of sexual partners could be relevant for research concerning sexually transmitted infections. Other factors such as personal hygiene, which may be relevant for infections transmitted by faecal-oral route or the number of social contacts, which can be connected to infections transmitted by airborne droplets are more difficult to measure. It is even hard to decide what constitutes a contact or what are relevant heterogeneities. Nevertheless, it is important to take these heterogeneities into account because they have an impact on the estimation of key epidemiological parameters such as the basic reproduction number R_0 (see Section 5). Ignoring them leads to biased estimates and could have a negative effect on mass vaccination programs (Unkel et al., 2014).
The idea is to analyse paired infectious disease data and use the correlation between infections in individuals. Having two infections with the same route of transmission and an individual with a high "activity level" related to the transmission route, this individual might be more likely to get both infections than an individual with a lower "activity level". So the degree of heterogeneity is shown by the correlation of the two infections. The advantage of this method is that no exact definitions of the relevant heterogeneities are needed. To estimate the extent of heterogeneities shared frailty models (Duchateau and Janssen, 2007), (Hougaard, 2012) and appropriate measures of association which are relevant for finding suitable frailty models can be used (Unkel et al., 2014), (Farrington et al., 2013).

4.2 Measures of associations

There are several measures of associations. The most common measure which evaluates time-varying dependence is Clayton's local cross-ratio function (CRF), developed by Clayton (1978). Let X be the monitoring time and T_j $(j = 1, 2)$ are the two event times of interest which can lie below, above or at the given time point and (T_1, T_2) are independent from X. This information

can be represented in short as $\{X, \delta_1, \delta_2\}$ with

$$\delta_j = \begin{cases} 1 & \text{if } T_j \leq X, \\ 0 & \text{if } T_j > X, \end{cases}$$

for $j = 1, 2$ (Unkel and Farrington, 2012). Let $S_j(t_j) = P(T_j > t_j)$ $(j = 1, 2)$ be the marginal survivor functions and $S(t_1, t_2) = P(T_1 > t_1, T_2 > t_2)$ the joint survival function of T_j. The CRF is defined as

$$\theta^*(t) = \frac{S(t) D_1 D_2 S(t)}{[D_1 S(t)][D_2 S(t)]} \tag{4.1}$$

where $t = (t_1, t_2)$ and D_j represents the derivative operator $\partial / \partial t_j$. This measure can be interpreted as the ratio of the hazard of T_1 given $T_2 = t_2$ over the hazard of T_1 given $T_2 > t_2$ (Oakes, 1989).

4.2.1 Odds ratio

The CRF has a local odds ratio (OR) interpretation (J. E. Anderson et al., 1992). But the disadvantage of the CRF (Equation 4.1) is that it cannot be calculated directly from current status data, so the same applies to the corresponding local OR. Instead a nonlocal OR at (t_1, t_2) can be used. It is defined as

$$\text{OR}(t_1, t_2) = \frac{\pi_{00}\pi_{11}}{\pi_{10}\pi_{01}} = \frac{\text{odds}(T_1 \leq t_1 | T_2 \leq t_2)}{\text{odds}(T_1 \leq t_1 | T_2 > t_2)},$$

with $\pi_{00} = P(T_1 > t_1, T_2 > t_2) = S(t_1, t_2)$, $\pi_{10} = P(T_1 \leq t_1, T_2 > t_2)$, $\pi_{01} = P(T_1 > t_1, T_2 \leq t_2)$ and $\pi_{11} = P(T_1 \leq t_1, T_2 \leq t_2)$. For infectious disease data $\pi_{00}(x)$ is the probability that an individual of age x has not been infected by both infections, $\pi_{10}(x)$ is the probability that the individual has only been infected by infection 1 and is still susceptible to infection 2 and $\pi_{01}(x)$ and $\pi_{11}(x)$ defined in the same way. With this information one can calculate the association between the two infections by $\text{OR}(x) = \frac{\pi_{00}(x)\pi_{11}(x)}{\pi_{10}(x)\pi_{01}(x)}$ for each age. Although this measure is easy to calculate it has two main disadvantages. Even if there are no time-specific heterogeneities it can vary with time and it is not applicable for the choice of suitable frailty models. For more detailed information see the work of Unkel and Farrington (2012).

4.2.2 An alternative measure $\phi(x)$

The new association measure presented by Unkel and Farrington (2012) can be estimated from current status data and provides a simple interpretation. It has properties similar to the CRF (Equation 4.1) and can be used for finding appropriate frailty models. The CRF for shared

frailty models can be represented as

$$\theta^*(t_1, t_2) = 1 + a^*(t_1, t_2),$$

with

$$a^*(t_1, t_2) = \frac{\text{var}(U|T_1 > t_1, T_2 > t_2)}{E(U|T_1 > t_1, T_2 > t_2)^2}$$

being the coefficient of variation of the relative frailty distribution, $U/E(U|T_1 > t_1, T_2 > t_2)$ (J. E. Anderson et al., 1992). It is called the relative frailty variance and it describes the heterogeneity of the hazard functions of survivors over time. As already mentioned the CRF cannot be estimated directly for current status data. As an alternative Unkel and Farrington (2012) use the variance of a gamma-distributed frailty. For a shared gamma frailty having mean one and shape parameter $\theta > 0$ for the frailty with reparametrization $\phi = \ln\Big(1 + \frac{1}{\theta}\Big)$, so $\theta = 1/(e^\phi - 1)$ the following holds

$$S(x, x) = \max\Big\{0,\ (S_1(x)^{1-e^\phi} + S_2(x)^{1-e^\phi} - 1)^{1/(1-e^\phi)}\Big\},$$

Let

$$f(\phi, S(x, x), S_1(x), S_2(x)) = (S_1(x)^{1-e^\phi} + S_2(x)^{1-e^\phi} - 1)^{1/(1-e^\phi)} - S(x, x).$$

The equation

$$f(\phi, S(x, x), S_1(x), S_2(x)) = 0$$

has a unique root $\phi_0(S(x, x), S_1(x), S_2(x))$ if $S(x, x)$ lies between $\max\{0, S_1(x) + S_2(x) - 1\}$ and $\min\{S_1(x), S_2(x)\}$. The new measure of association is defined as this root:

$$\phi(x) \equiv \phi_0(S(x, x), S_1(x), S_2(x)).$$

For further information see the work of Unkel and Farrington (2012).
This new measure can be estimated by using paired serological data, where n_x is the total number of individuals of age x in the sample, n_{00x} is the number of individuals of age x who have not been infected by both infections, n_{10x} is the number of individuals who have only been infected by infection 1 and are still susceptible to infection 2 and n_{01x} and n_{11x} defined in the same way. The root of the implicit function

$$f(\phi(x), \hat{S}(x, x), \hat{S}_1(x), \hat{S}_2(x)) = (\hat{S}_1(x)^{1-e^{\phi(x)}} + \hat{S}_2(x)^{1-e^{\phi(x)}} - 1)^{1/(1-e^{\phi(x)})} - \hat{S}(x, x), \quad (4.2)$$

with $\hat{S}(x, x) = \frac{n_{00x}}{n_x}$, $\hat{S}_1(x) = \frac{n_{0+x}}{n_x}$ and $\hat{S}_2(x) = \frac{n_{+0x}}{n_x}$ ($n_{0+x} = n_{00x} + n_{01x}$ and $n_{+0x} = n_{00x} + n_{10x}$) is the estimate of $\phi(x)$. For finding the value $\hat{\phi}(x)$ the bisection algorithm, implemented in the function "uniroot" in R (R Core Team, 2018), is used. With the Delta method the estimated

asymptotic standard errors for $\hat{\phi}(x)$ can be calculated (Unkel and Farrington, 2012). The value $\phi(x)$ can be interpreted as follows, $\phi(x) = 0$ is equivalent to no association, $\phi(x) > 0$ is equivalent to a positive association, according to heterogeneity and $\phi(x) < 0$ is equivalent to a negative association, possible because of cross-immunisation (Farrington et al., 2013).
The subsequent summary measure of association over all age groups $x = 0, 1, .., \text{J}$ can be used:

$$\bar{\phi} = \frac{\sum_{x=1}^{J} p_x \hat{\phi}(x)}{\sum_{x=1}^{J} p_x}, \quad \text{Var}(\bar{\phi}) = \frac{1}{\sum_{x=1}^{J} p_x}, \tag{4.3}$$

with $\hat{\phi}(x)$ being the estimated value of $\phi(x)$ and p_x its precision, the reciprocal of the variance (Farrington et al., 2013).
For paired serological data it is possible that within the 4-tuples (n_{00x}, n_{01x}, n_{10x}, n_{11x}) of each age counts are zero. It is suggested to deal with zeroes as follows. If only the cells are zero but all four margins (n_{0+x}, n_{+0x}, n_{1+x} and n_{+1x}) are greater than zero, add 0.5 to all cells (Agresti, 2002). If one margin is zero the point should be combined with the data for age $x - 1$ and assigned to the average of the ages for the combined data. If there is more than one margin zero, this data point should be deleted (Unkel and Farrington, 2012).

For several paired infectious diseases $\hat{\phi}(x)$ and its 95% confidence interval (CI) are computed. For the computation only the individuals who are not vaccinated against the infectious diseases (if there is a vaccine available) are included. The pairs are divided into two groups, pairs of infections with shared main transmission routes and with different main transmission routes. In previous works, see Farrington et al. (2013) the following results were observed. The associations for pairs of infections with the same main route of transmission are generally higher compared to pairs of infections with different main routes of transmission.
The associations for pairs of infections not sharing a main transmission route are often not positive except the association in childhood. In many analysed pairs of infections, regardless of whether the transmission routes were the same or different, a positive association in childhood can be observed which declines to a positive constant value in adulthood for same routes of transmission and declines towards zero for different transmission routes, respectively. The associations for pairs of infections transmitted via the respiratory route have a propensity to be lower than between infections transmitted via other routes.
In Table 3 the summary measure $\bar{\phi}$ and its 95% CI over all age groups for different pairs of infections using data from the PIENTER-2 study are presented. The measure $\bar{\phi}$ is for all pairs of infections with same transmission routes higher than for infections with different routes of transmission, except for HAV and HBV. The analysed lower association between infections transmitted via the respiratory route by Farrington et al. (2013) can be observed in these data as well (Measles, Mumps and Rubella). Whereas, for pairs of infections where transmission occurs through sexual intercourse $\bar{\phi}$ is very high (HPV 16, 18 and 45). For some pairs of

infections with different routes of transmission even a negative $\bar{\phi}$ is observed, e.g. Toxoplasma and VZV.

Table 3: Associations ($\bar{\phi}$) between paired infections over all age groups (0-79 years).

Infection pair	$\bar{\phi}$	95% CI
Shared main route of transmission likely		
Measles and Mumps	0.255	0.141, 0.370
Measles and Rubella	0.254	0.156, 0.352
Rubella and Mumps	0.275	0.185, 0.366
HPV 16 and 18	2.809	2.661, 2.957
HPV 16 and 45	2.049	1.900, 2.199
Toxoplasma and HAV	0.126	0.039, 0.214
Shared main route of transmission unlikely		
HAV and HBV	0.836	0.612, 1.060
Toxoxplasma and CMV	0.080	0.002, 0.159
Toxoxplasma and VZV	-0.060	-0.164, 0.045
Measles and CMV	0.072	-0.029, 0.172
Mumps and CMV	0.049	-0.021, 0.119
Rubella and CMV	-0.071	-0.142, 0.001
HAV and VZV	-0.129	-0.215, 0.043

Due to the observed positive association in childhood which decreases to a positive constant value in adulthood and towards zero, respectively the summary measure $\bar{\phi}$ and its 95% CI only over the age groups from 21 to 79 for different pairs are calculated (Table 4).

Table 4: Associations ($\bar{\phi}$) between paired infections from age 21.

Infection pair	$\bar{\phi}$	95% CI
Shared main route of transmission likely		
Measles and Mumps	0.226	0.109, 0.342
Measles and Rubella	0.205	0.100, 0.310
Rubella and Mumps	0.242	0.150, 0.334
HPV 16 and 18	2.604	2.440, 2.768
HPV 16 and 45	1.996	1.836, 2.156
Toxoplasma and HAV	0.085	-0.006, 0.175
Shared main route of transmission unlikely		
HAV and HBV	0.810	0.578; 1.041
Toxoxplasma and CMV	0.057	-0.024, 0.138
Toxoxplasma and VZV	-0.032	-0.153, 0.088
Measles and CMV	0.035	-0.070 0.140
Mumps and CMV	0.027	-0.044, 0.098
Rubella and CMV	-0.094	-0.167, -0.021
HAV and VZV	-0.101	-0.201, -0.002

For pairs of infections with the same route of transmission the 95% CI of $\bar{\phi}$ does not include zero, except for the pair Toxoplasma and HAV. For pairs of infections with different routes of transmission the 95% CI of $\bar{\phi}$ includes the zero or for some pairs the CI is negative. Only

for the infection pair HAV and HBV the CI is positive which was already shown with the high value of $\bar{\phi}$ in Table 3.

These results support the observed features of the alternative association measure $\phi(x)$. But only to analyse the overall measure is a bit crude. For better understanding of the age-specific associations it is possible to plot the association parameter $\phi(x)$ at each age x. This representation is used in the following section to find appropriate models for the data.

4.3 Modelling of associations

This Section is mainly based on the work of Unkel et al. (2014).
As already mentioned in Section 4.1 shared frailty models can be used to estimate the extent of heterogeneity. Let age x be the only measured factor and the unmeasured factors can be described by a random variable $U > 0$ with density $f(U)$ and mean one. Shared in this context means that the frailty U is relevant to transmission of both infections. These models can be used for the force of infection j ($j = 1, 2$, referring to infections 1 and 2) for an individual of age x with shared activity level U:

$$\lambda_j(x, U) = U\,\lambda_{0j}(x), \tag{4.4}$$

with $\lambda_{0j}(x)$ being the baseline forces of infection and independent of U. The baseline force of infection describes the age effect and the frailty U causes association between the failure times (Farrington, Kanaan, and Gay, 2001). Nevertheless, this model ignores the variation in the extent of heterogeneity with age. Including age dependence in the heterogeneity leads to the model with $U(x) = w(x, Z_1, ..., Z_q)$, which can vary with age:

$$\lambda_j\{x, U(x)\} = U(x)\,\lambda_{0j}(x), \tag{4.5}$$

with w being some known function and $Z_1, ..., Z_q$ being independent time-invariant frailties. The mean $E(U)$ or $E\{U(x)\}$ of the frailties is always one and the focus lies on the frailty variance. It represents the degree of the unmeasured individual heterogeneity.
In this thesis two time-invariant frailty models are used. The first one is the shared gamma frailty model where the frailty term U is gamma distributed with $\Gamma(\theta, 1/\theta)$. The second one is the inverse Gaussian frailty model with variance θ. With these models the variance of the frailty is required to be constant. Using time-varying models the variance var$\{U(x)\}$ is supposed to be time dependent. For infections with the same transmission route var$\{U(x)\}$ is expected to be non-zero at all ages. For infections with different transmission routes var$\{U(x)\}$ is expected to be non-zero in childhood for non-sexual transmitted diseases because of the intensive and close contact between children. In adulthood var$\{U(x)\}$ is expected to be zero, for example, caused by homogeneous increasing social distance.

There exist several time-varying frailty models with different correlation structures. Farrington, Unkel, and Anaya-Izquierdo (2012) use piecewise frailty models where

$$U(x) = \sum_{j=1}^{q} Z_j I_j(x),$$

with disjoint age intervals $I_j = (x_{j-1}, x_j]$ for $(j = 1, ..., q)$ with $x_0 = 0$ and $I_j(x) = 1$ if $x \in I_j$. The Z_j are independent gamma distributed with unit mean and variance σ_j^2. The variance var$\{U(x)\}$ is constant if $\sigma_j^2 = \sigma^2$ or by supposing $\sigma_j^2 = \sigma^2 \exp[-\{(m_j - m_1)/\rho\}^k]$ a declining variance is required, with m_j being the midpoint of I_j, $\rho > 0$ describing the rate of decline and k being some positive integer, here $k = 2$. The piecewise frailty models have the strong assumption that the frailties in each age group j are independent. By contrast, Farrington, Unkel, and Anaya-Izquierdo (2012) introduced a frailty model where the frailties are perfectly correlated:

$$U(x) = 1 + (Z - 1)h(x),$$

with Z being a time invariant frailty of mean one and $0 < h(x) < 1$. The function $h(x)$ is defined as

$$h(x) = \exp\{-(x/\rho)^k\}, \tag{4.6}$$

which represents the decrease in heterogeneity in childhood. The variance of this frailty model is var$\{U(x)\} = h(x)^2$ var(Z). So for this model with $h(x)$ chosen as in Equation 4.6 a decrease of heterogeneity to zero with increasing age is assumed. For associations that are greater than zero at larger x a two-component multiplicative model can be used:

$$U(x) = \prod_{j=1}^{q} \{1 + (Z_j - 1)h_j(x)\}, \qquad 0 \leq h_j(x) \leq 1, \tag{4.7}$$

with $q = 2$ and $h_1(x)$ defined as in Equation 4.6 and $h(x) = 1$, which describes the constant positive heterogeneity in adulthood. The variance of this model is

$$\text{var}\{U(x)\} = h_1(x)^2 \text{var}(Z_1) + \text{var}(Z_2) + h_1(x)^2 \text{var}(Z_1)\text{var}(Z_2)$$

Let the probabilities $\pi_{00}(x), \pi_{01}(x), \pi_{10}(x)$ and $\pi_{11}(x)$ be defined as in Section 4.2.1. Using a shared time-varying frailty model (Equation 4.5) for the force of infection j $(j = 1, 2)$ the probabilities $\pi_{00}(x), \pi_{01}(x), \pi_{10}(x)$ and $\pi_{11}(x)$ at age x can be calculated as

$$\pi_{00}(x) = E\Big(\exp\Big[-\int_0^x U(x)\{\lambda_{01}(y) + \lambda_{02}(y)\}\, dy\Big]\Big) \tag{4.8}$$

$$\pi_{01}(x) = E\Big[\exp\Big\{-\int_0^x U(x)\lambda_{01}(y)\, dy\Big\}\Big] - \pi_{00}(x) \tag{4.9}$$

$$\pi_{10}(x) = E\Big[\exp\Big\{-\int_0^x U(x)\lambda_{02}(y)\,dy\Big\}\Big] - \pi_{00}(x) \tag{4.10}$$

and

$$\pi_{11}(x) = 1 - \pi_{01}(x) - \pi_{10}(x) - \pi_{00}(x) \tag{4.11}$$

with respect to the chosen frailty model $U(x)$ and U respectively for a time-invariant model (4.4). For example, the three probabilities in Equation 4.8 - 4.10 for the 1-component gamma age-dependent frailty model with $h(x)$ as defined in Equation 4.6 and $k = 2$ are

$$\pi_{00} = \exp\{H^1(x) + H^2(x) - \Lambda_{01}(x) - \Lambda_{02}(x)\}\left[1 + \frac{H^1(x) + H^2(x)}{\theta}\right]^{-\theta},$$

$$\pi_{01} = \exp\{H^1(x) - \Lambda_{01}(x)\}\left[1 + \frac{H^1(x)}{\theta}\right]^{-\theta} - \pi_{00}(x),$$

$$\pi_{10} = \exp\{H^2(x) - \Lambda_{02}(x)\}\left[1 + \frac{H^2(x)}{\theta}\right]^{-\theta} - \pi_{00}(x),$$

with $H^j(x) = \int_0^x h(t)\lambda_{0j}(t)\,dt$ $(j = 1, 2)$. The probabilities for the shared gamma, the inverse Gaussian and the two-component multiplicative double gamma frailty model are given in the Appendix.

Using paired serological data leads to a multinomial observation $(n_{00x}, n_{01x}, n_{10x}, n_{11x})$ for each age class x, which is already defined in Section 4.2.1. Specifying $U(x)$ and U respectively and λ_{0j} $(j = 1, 2)$ the shared frailty model in Equation 4.5 and 4.4 respectively is fitted by maximizing a product multinomial likelihood. The multinomial log-likelihood kernel is

$$l = \sum_x \sum_{i,j=0,1} n_{ijx} \ln\{\pi_{ij}(x)\},$$

with ln representing the natural logarithm. The data are modelled with a compound Dirichlet-multinomial distribution with dispersion parameter ν and $0 < \nu < 1$, which allows for overdispersion (if there is more variability in the data than it is expected from the model). The log-likelihood kernel for the compound Dirichlet-multinomial distribution is

$$\begin{aligned} l_{DM} = \sum_x \Big(&\ln\Big\{\frac{\Gamma(\psi)}{\Gamma(n_x + \psi)}\Big\} + \ln\Big[\frac{\Gamma\{n_{00x} + \psi\,\pi_{00}(x)\}}{\Gamma\{\psi\,\pi_{00}(x)\}}\Big] + \ln\Big[\frac{\Gamma\{n_{01x} + \psi\,\pi_{01}(x)\}}{\Gamma\{\psi\,\pi_{01}(x)\}}\Big] \\ &+ \ln\Big[\frac{\Gamma\{n_{10x} + \psi\,\pi_{10}(x)\}}{\Gamma\{\psi\,\pi_{10}(x)\}}\Big] + \ln\Big[\frac{\Gamma\{n_{11x} + \psi\,\pi_{11}(x)\}}{\Gamma\{\psi\,\pi_{11}(x)\}}\Big]\Big), \end{aligned} \tag{4.12}$$

with $\psi = (1 - \nu)/\nu$, so $\nu = 1/(1 + \psi)$. The multinomial component variances are inflated with the dispersion parameter ν by the factor $1 + \nu(n_x - 1)$.

All models are fitted by the same procedure. One gets the baseline hazards $\lambda_{0j}(x)$ $(j = 1, 2)$

then one computes the probabilities in Equation 4.8 - 4.11, calculates the log-likelihood and the saturated log-likelihood to compute the deviance $-2(l - l_{Sat})$ and then one iterates until convergence. For the Dirichlet-multinomial model the log-likelihood (Equation 4.12) is maximized first and the estimated value for ψ is used to compute the deviance $-2(l_{DM} - l_{DMSat})$, with l_{DMSat} being the saturated log-likelihood defined as

$$l_{DMSat} = \sum_x \Big(\ln\Big\{ \frac{\Gamma(\hat{\psi})}{\Gamma(n_x + \hat{\psi})} \Big\} + \ln\Big[\frac{\Gamma\{n_{00x} + \hat{\psi}\, s_{00}(x)\}}{\Gamma\{\hat{\psi}\, s_{00}(x)\}} \Big] + \ln\Big[\frac{\Gamma\{n_{01x} + \hat{\psi}\, s_{01}(x)\}}{\Gamma\{\hat{\psi}\, s_{01}(x)\}} \Big] + \ln\Big[\frac{\Gamma\{n_{10x} + \hat{\psi}\, s_{10}(x)\}}{\Gamma\{\hat{\psi}\, s_{10}(x)\}} \Big] + \ln\Big[\frac{\Gamma\{n_{11x} + \hat{\psi}\, s_{11}(x)\}}{\Gamma\{\hat{\psi}\, s_{11}(x)\}} \Big] \Big), \tag{4.13}$$

with $s_{ij}(x) = n_{ijx}/n_x$ $(i, j = 0, 1)$. For all models piecewise constant baseline hazards chosen on epidemiological grounds are used. A division into eight age groups (0-4, 5-9, 10-19 20-20, 30-39, 40-49, 50-64, 65-79) is meaningful. Other choices of the baselines, for example the Gompertz hazard showed worse fits. The function `nlm` in R version 3.5.2 (R Core Team, 2018) is used to maximize the log-likelihood and the deviance respectively. The probabilities in Equation 4.8 - 4.11 for the two-component multiplicative model (Equation 4.7) cannot be calculated in closed form, therefore the function `integrate` for a numerical integration is used. The different models are compared by using the Akaike information criterion (AIC), defined as AIC$= 2k - 2\ln(\hat{l})$, with k the number of estimated parameters and $\hat{l}$ the value of the log-likelihood. A goodness-of-fit test is performed as well. For the best fitting model approximate 95% confidence intervals (CIs) for the parameters of interest are computed by simulating from a multivariate normal distribution with the function `rmvnorm` from the mvtnorm package (Genz et al., 2019) with covariance matrix set equal to a numerical estimate of the observed Fisher information matrix.

All models are fitted to different paired infections, six pairs have similar transmission routes and three pairs have different ones. Only the bivariate data for the analysis are used because including the univariate data leads to worse fits. For the best fitted model the Dirichlet-multinomial distribution is used and its necessity is assessed. If all models show bad fits the Dirichlet-multinomial distribution is used for all of them. In addition, the model without frailty terms is fitted to the data. For all pairs of infections this model shows worse fits compared to the applied shared frailty models.

4.3.1 Pairs of infections with similar mode of transmission

Measles and Mumps

For 3376 unvaccinated individuals of 0-79 years of age from the nationwide sample bivariate serological data on both infections are available. Because the vaccination rate for children is very high the number of individuals in the first age groups might be a bit too small. This fact could lead to some biased estimations for the first age groups. Using the whole sample leads to even worse fits. Measles and Mumps are both transmitted via respiratory droplets and direct contact, so heterogeneity in the number of contacts is likely to result in association between the two infections. Figure 13 shows the observed associations between the two infections. The areas of the points are proportional to the precision p_x and the curve (——) is a precision-weighted scatter plot smoothing curve to capture trends. There is a strong heterogeneity in early childhood which is declining with age towards a positive constant in adulthood.

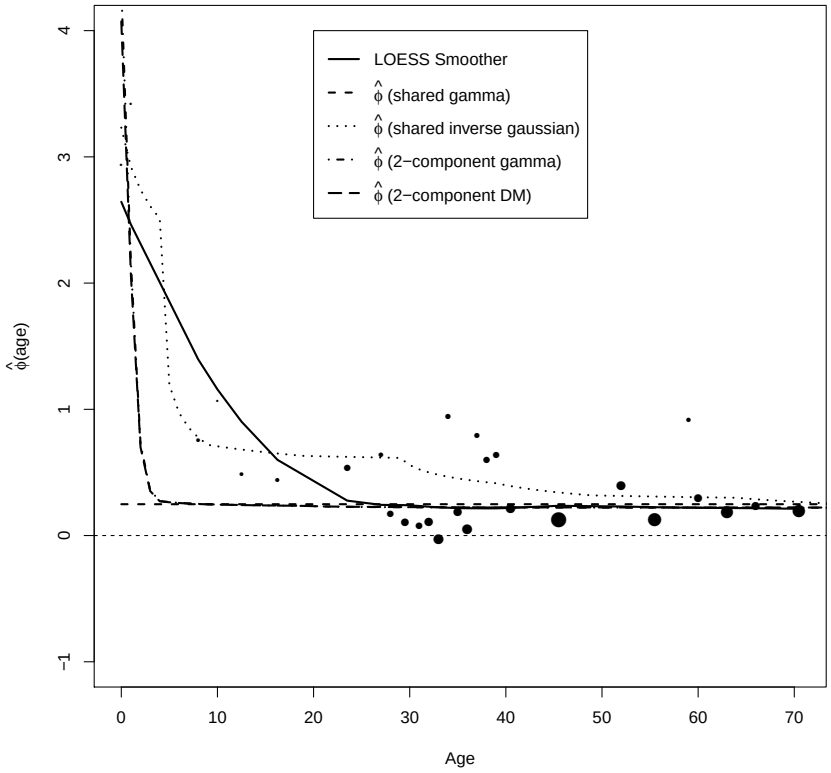


Figure 13: Observed and fitted associations between the infections Measles and Mumps.

All introduced shared frailty models are fitted to the data and the results are presented in Table 5. The piecewise constant frailty model with a declining variance fits the data worst. The inverse Gaussian frailty model shows quite good fits but the best model (lowest AIC and deviance) is the two-component multiplicative model with gamma distributed frailties. Including an overdispersion parameter in the two-component model does not improve the model fit which can be seen in Figure 27 (a). It shows the pointwise deviances which can be used to analyse

what each observation contributes to the overall goodness of fit. Only for the age zero the deviance is higher for the multinomial model. All other pointwise deviances have nearly the same values. The two-component model supports the declining measured association in child-

Table 5: Fitting results for Measles and Mumps infection data for the nationwide sample.

Frailty model	Parameterization of $h(x), h_1(x)$ and $h_2(x)$	Parameter estimates	Deviance	Degrees of freedom	p-value	AIC
$U \sim \Gamma(\theta, 1/\theta)$		$\hat{\theta} = 3.547$	228.84	223	0.3799	1717.34
$U \sim InvG(1, \theta)$		$\hat{\theta} = 0.006$	206.32	223	0.7819	1694.82
Piecewise independent gamma	$\sigma_j^2 = \sigma^2\ (j = 1, ..., 8)$	$\hat{\theta} = 0.066$	207.74	223	0.7606	1696.23
Piecewise independet gamma	$\sigma_j^2 = \sigma^2 \cdot \exp[-\{(m_j - m_1)/\rho\}^2]$ $(j = 1, ..., 8)$	$\hat{\theta} = 0.117$ $\hat{\rho} = 5.72$	283.76	222	0.0032	1774.26
1-component gamma	$h(x) = \exp\{-(x/\rho)^2\}$	$\hat{\theta} = 0.014$ $\hat{\rho} = 3.26$	215.80	222	0.6046	1706.30
2-component multiplicative double gamma	$h_1(x) = \exp\{-(x/\rho)^2\}$, $h_2(x) = 1$	$\hat{\theta}_1 = 0.013$ $\hat{\rho} = 2.98$ $\hat{\theta}_2 = 4.07$	200.00	221	0.8415	1692.49
2-component multiplicative double gamma (Dirichlet multinomial)	$h_1(x) = \exp\{-(x/\rho)^2\}$ $h_2(x) = 1$	$\hat{\theta}_1 = 0.018$ $\hat{\rho} = 2.97$ $\hat{\theta}_2 = 2.02$ $\hat{\nu} =< 0.0001$	198.29	221	0.8616	1690.79

hood with the first frailty Z_1 and a constant association in adulthood with the second frailty Z_2. Figure 13 shows the fitted values of the association measure ϕ for selected models as well. The time invariant gamma frailty model predicts a constant association. The inverse Gaussian model describes the associations satisfactorily but values between 25 and 40 years of age are a bit too high. The two-component model and its counterpart are virtually identical and closely resemble the observed pattern for the higher age groups. For the smaller age groups the fitted associations might be a bit too low which could be attributed to the fact that not enough data for the smaller age groups are available. Nevertheless, the two-component model estimates the seroprevalences for both infections largely well except for the first age groups (Figure 14).
Approximate 95% CIs for the parameters of interest of this model are as follows: for θ_1, (0.0066, 0.0274), for ρ, (0.5327, 5.2212) and for θ_2, (2.8114, 5.8960). Because the extent of unmeasured heterogeneity is represented by the variance of the frailty $U(x)$ the estimate of the frailty standard deviation with 95% CIs is given in Figure 26 (a). For the first age groups the fitted frailty standard deviation is very high and declines with higher age groups. At the age of eleven it reaches a constant value of 0.4958.

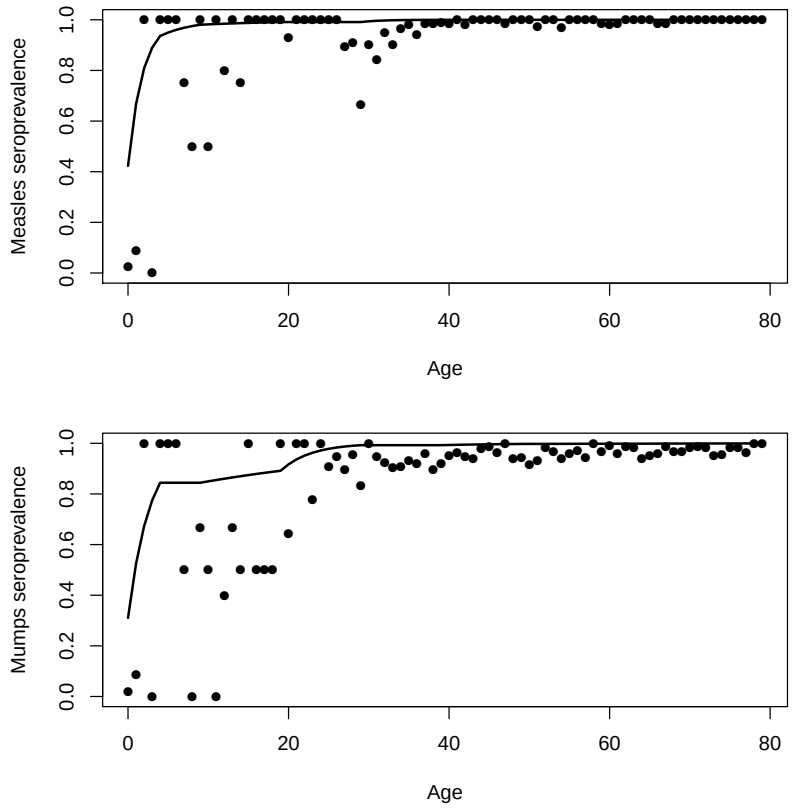


Figure 14: Observed seroprevalences of Measles and Mumps and fit of the 2-component frailty model to the data.

Measles and Rubella

For 4448 unvaccinated individuals of 0-79 years of age bivariate serological data on both infections are available. For this pair of infections the whole sample need to be used because otherwise there are not enough individuals in the younger age groups. Measles and Rubella are also two infections transmitted by direct contact and respiratory droplets, so heterogeneity in the number of contacts is likely to result in association between the two infections. Figure 15 shows the observed associations between the two infections. Again a stronger heterogeneity in early childhood which is declining with age towards a positive constant in adulthood can be observed. However, the associations are lower compared to the previous example.
Table 6 presents the results of the fitted frailty models. The time invariant models and also the piecewise constant frailty models can not fit the data well. The best fit is obtained by fitting the two-component Dirichlet-multinomial which represents the positive declining association in childhood and a positive constant value in adulthood. The Dirichlet-multinomial distribution leads to a reduction in the pointwise deviance for many of the first observed 4-tupels (Figure 27(b)). The maximum variance inflating factor is 1.48 with an average of 1.12, so the multinomial component variances are increased by more than 12%, on average. The fitted associations in Figure 15 for the one- and two-component multinomial models are nearly identical and es-

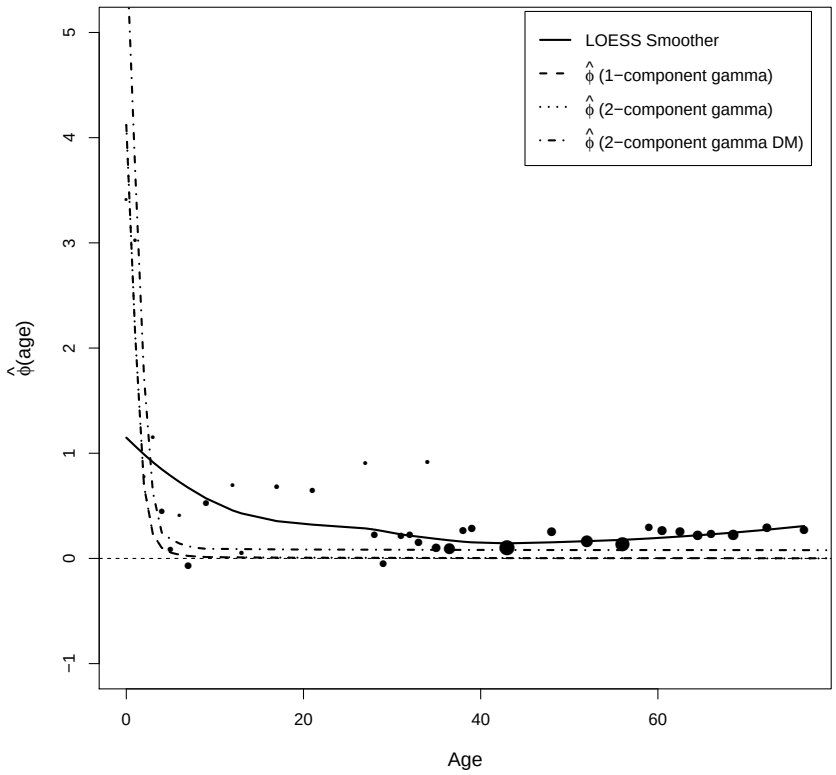


Figure 15: Observed and fitted associations between the infections Measles and Rubella.

timate associations equal zero in adulthood. The two-component Dirichlet-multinomial model estimates the observed pattern better because in adulthood there are still positive values of association. The seroprevalences for Measles and Rubella (Figure 16) are well fitted with this model.

The approximate 95% CIs for the parameters of interest of this model are as follows: for θ_1, (0.0097, 0.0430), for ρ, (0.8052, 4.8137) and for θ_2, (2.0140, 15.1444). The estimate of the frailty standard deviation with 95% CIs is given in Figure 26 (b). For the first age groups the fitted frailty standard deviation is very high and decreases with higher age groups. At the age of nine it reaches a constant value of 0.4177.

Mumps and Rubella

For the analysis of Mumps and Rubella data of 3210 unvaccinated individuals from the nationwide sample are available. Here again, the small number of individuals in the first age groups probably because of the high vaccination rate could lead to worse fits in the first age groups. Mumps and Rubella are both transmitted via respiratory droplets and close contact as well. Figure 17 shows a transient heterogeneity in childhood and a smaller positive constant association in adulthood.

The results of the fitted frailty models are presented in Table 7. The shared gamma frailty model fits the data worst, all other models show quite good fits. Again, the best fit is obtained by fitting

Table 6: Fitting results for Measles and Rubella infection data.

Frailty model	Parameterization of $h(x), h_1(x)$ and $h_2(x)$	Parameter estimates	Deviance	Degrees of freedom	p-value	AIC
$U \sim \Gamma(\theta, 1/\theta)$		$\hat{\theta} = 8.970$	287.89	223	0.0022	2027.08
$U \sim InvG(1, \theta)$		$\hat{\theta} = 1.597$	284.97	223	0.0032	2024.16
Piecewise independent gamma	$\sigma_j^2 = \sigma^2$ $(j = 1, ..., 8)$	$\hat{\theta} = 0.851$	276.29	223	0.0087	2015.48
Piecewise independet gamma	$\sigma_j^2 = \sigma^2 \cdot \exp[-\{(m_j - m_1)/\rho\}^2]$ $(j = 1, ..., 8)$	$\hat{\theta} = 2.840$ $\hat{\rho} = 4.10$	346.48	222	<0.0001	2087.67
1-component gamma	$h(x) = \exp\{-(x/\rho)^2\}$	$\hat{\theta} = 0.017$ $\hat{\rho} = 2.27$	230.83	222	0.3282	1972.02
2-component multiplicative double gamma	$h_1(x) = \exp\{-(x/\rho)^2\}$, $h_2(x) = 1$	$\hat{\theta}_1 = 0.017$ $\hat{\rho} = 2.27$ $\hat{\theta}_2 = 1744.65$	230.79	221	0.3118	1973.98
1-component gamma (Dirichlet multinomial)	$h(x) = \exp\{-(x/\rho)^2\}$	$\hat{\theta}_1 = 0.017$ $\hat{\rho} = 2.44$ $\hat{\nu} = 0.0023$	215.97	222	0.6014	1957.17
2-component multiplicative double gamma (Dirichlet multinomial)	$h_1(x) = \exp\{-(x/\rho)^2\}$ $h_2(x) = 1$	$\hat{\theta}_1 = 0.021$ $\hat{\rho} = 2.68$ $\hat{\theta}_2 = 5.73$ $\hat{\nu} = 0.0022$	213.51	221	0.6287	1956.70

the two-component model which represents the positive declining heterogeneity in childhood and a positive constant value in adulthood. The two-component Dirichlet-multinomial model leads to a reduction in the pointwise deviance only for the first observed 4-tupel (Figure 27(c)), so no better fit with this model can be achieved. This can also be seen in the estimated value ν which is smaller than 0.0001. The fitted associations in Figure 17 for the two-component multinomial and Dirichlet-multinomial model are virtually nearly identical. They closely resemble the observed pattern for the higher age groups. The association in childhood decreases a bit too fast which could be attributed to the fact that not enough data for the smaller age groups are available. The piecewise constant frailty model with declining frailty variance can closely resemble the observed pattern as well. The reason for the stair-step configuration is the piecewise frailty on the age intervals. The two-component multinomial model fits the seroprevalences for Mumps and Rubella for the higher age groups quite well but the estimated seroprevalences in childhood are too high (Figure 18).

Approximate 95% CIs for the parameters of interest of the two-component multinomial model are as follows: for θ_1, (0.0040, 0.0171), for ρ, (0.3460, 10.4771) and for θ_2, (3.2913, 7.6932). The estimate of the frailty standard deviation with 95% CIs is given in Figure 26 (c). Similar to the other two pairs the fitted frailty standard deviation is quite high in childhood and decreases with higher age groups. At the age of 15 it reaches a constant value of 0.4505.

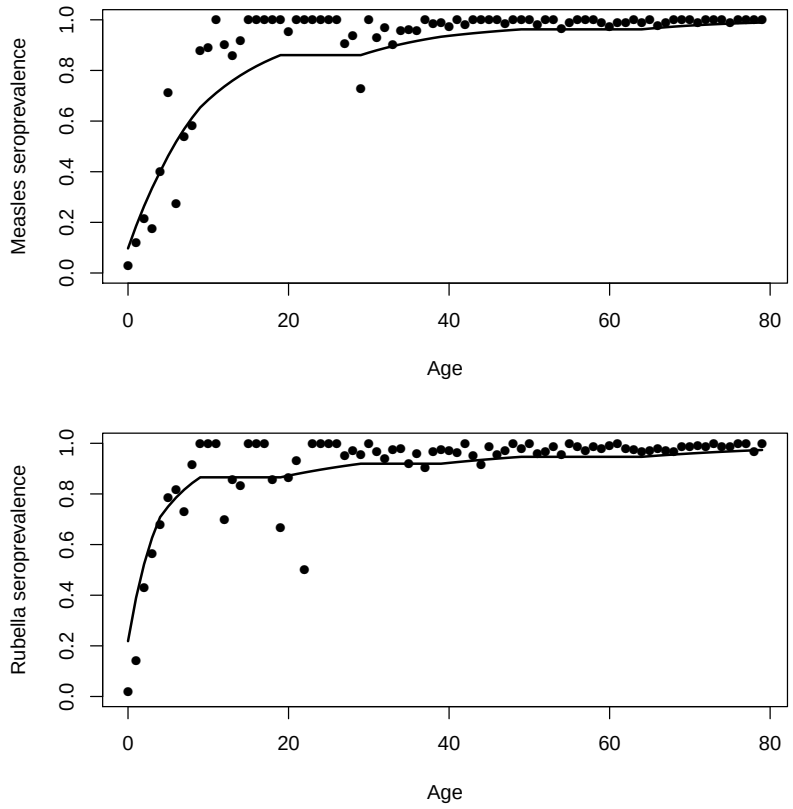


Figure 16: Observed seroprevalences of Measles and Rubella and fit of the 2-component Dirichlet-multinomial frailty model to the data.

Table 7: Fitting results for Mumps and Rubella infection data for the nationwide sample.

Frailty model	Parameterization of $h(x), h_1(x)$ and $h_2(x)$	Parameter estimates	Deviance	Degrees of freedom	p-value	AIC
$U \sim \Gamma(\theta, 1/\theta)$		$\hat{\theta} = 4.273$	266.51	223	0.0244	2041.38
$U \sim InvG(1, \theta)$		$\hat{\theta} < 0.0001$	239.14	223	0.2183	2014.01
Piecewise independent gamma	$\sigma_j^2 = \sigma^2$ $(j = 1, ..., 8)$	$\hat{\theta} = 0.026$	231.52	223	0.3337	2006.39
Piecewise independet gamma	$\sigma_j^2 = \sigma^2 \cdot$ $\exp[-\{(m_j - m_1)/\rho\}^2]$ $(j = 1, ..., 8)$	$\hat{\theta} = 0.023$ $\hat{\rho} = 42.04$	226.70	222	0.4000	2003.58
1-component gamma	$h(x) = \exp\{-(x/\rho)^2\}$	$\hat{\theta} = 0.008$ $\hat{\rho} = 5.72$	240.64	222	0.1860	2017.52
2-component multiplicative double gamma	$h_1(x) = \exp\{-(x/\rho)^2\}$, $h_2(x) = 1$	$\hat{\theta}_1 = 0.008$ $\hat{\rho} = 4.64$ $\hat{\theta}_2 = 4.92$	223.54	221	0.4395	2002.42
2-component multiplicative double gamma (Dirichlet multinomial)	$h_1(x) = \exp\{-(x/\rho)^2\}$ $h_2(x) = 1$	$\hat{\theta}_1 = 0.010$ $\hat{\rho} = 4.70$ $\hat{\theta}_2 = 2.45$ $\hat{\nu} =< 0.0001$	222.95	221	0.4505	2001.83

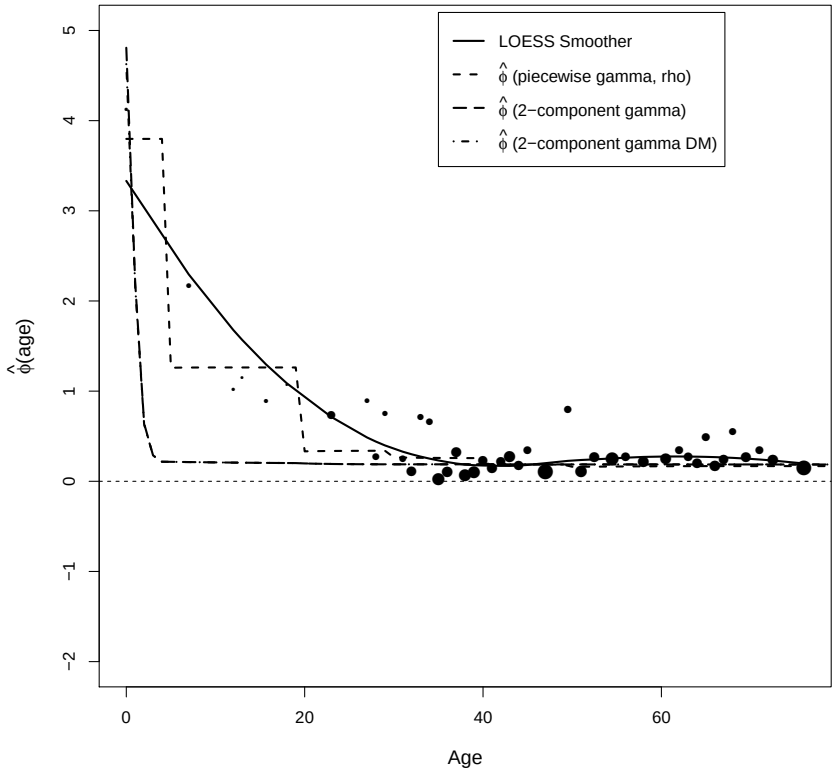


Figure 17: Observed and fitted associations between the infections Mumps and Rubella.

HPV 16 and 18

To have enough information in all age groups 7902 individuals from the whole sample are included. HPV 16 and 18 are both sexual transmitted infectious diseases, so heterogeneity in the number of sexual partners is likely to result in association between the two infections. Figure 19 gives the association between times of infection for the two infections. Again, the typical decreasing association in childhood and a positive constant association in adulthood can be observed but the associations in adulthood have higher values compared to those of the MMR pairs.

Table 8 presents the results of the fitted frailty models. The time invariant frailty models can not fit the data well. The piecewise constant frailty model with constant variance fits the data better but the best fitting model is the one-component multinomial model. A second frailty term is not needed (two-component model) because the parameter $\rho = 68.70$ is high enough to model the positive association in adulthood. Including an overdispersion parameter in the one-component model does not improve the model fit ($\hat{\nu} < 0.0001$). The fitted associations in Figure 19 for the one- and two-component multinomial models are virtually identical. Both closely resemble the observed pattern. Nevertheless, the models fit the seroprevalences not satisfyingly, all fitted seroprevalences are too high (Figure 20) although the fitted proportions ($f_{00}, f_{01}, f_{10}, f_{11}$) fit the observed proportions ($s_{00}, s_{01}, s_{10}, s_{11}$) quite well (see Appendix Figure 37). Almost all frailty models show satisfying fits of the observed proportions ($s_{00}, s_{01}, s_{10}, s_{11}$) even if they

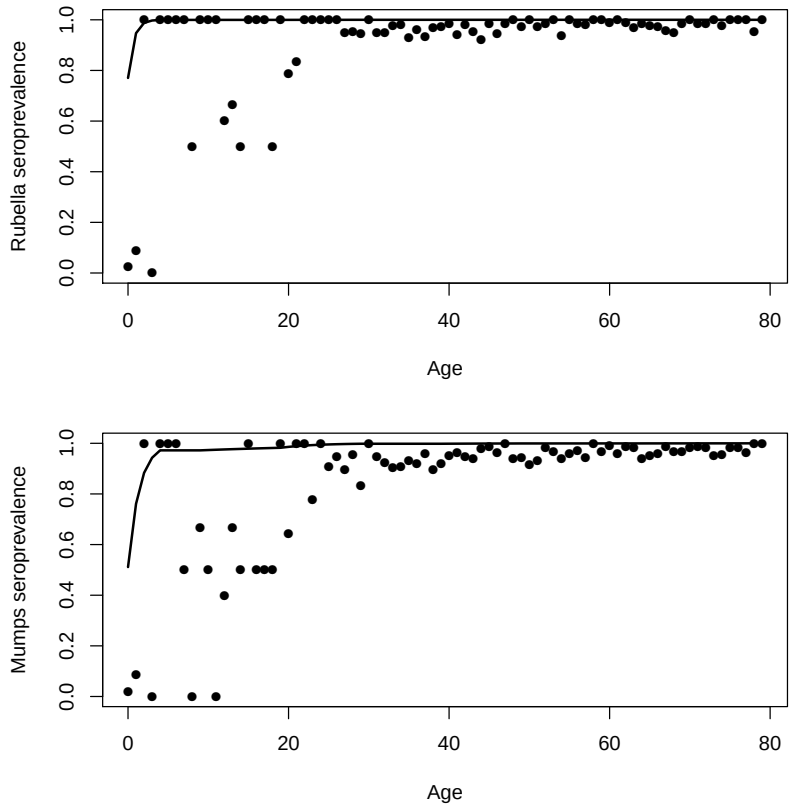


Figure 18: Observed seroprevalences of Mumps and Rubella and fit of the 2-component frailty model to the data.

have high deviances. Therefore, these plots provide no information on how well the models fit the observed association patterns (Farrington, Unkel, and Anaya-Izquierdo, 2012).

Table 8: Fitting results for Human Papilloma Virus 16 and 18 infection data.

Frailty model	Parameterization of $h(x), h_1(x)$ and $h_2(x)$	Parameter estimates	Deviance	Degrees of freedom	p-value	AIC
$U \sim \Gamma(\theta, 1/\theta)$		$\hat{\theta} = 0.067$	295.87	223	0.0008	6387.13
$U \sim InvG(1, \theta)$		$\hat{\theta} = 0.004$	257.46	223	0.0564	6348.71
Piecewise independent gamma	$\sigma_j^2 = \sigma^2$ $(j = 1, ..., 8)$	$\hat{\theta} = 0.014$	245.76	223	0.1413	6337.01
Piecewise independet gamma	$\sigma_j^2 = \sigma^2 \cdot$ exp$[-\{(m_j - m_1)/\rho\}^2]$ $(j = 1, ..., 8)$	$\hat{\theta} = 0.009$ $\hat{\rho} = 24.55$	247.59	222	0.1147	6340.84
1-component gamma	$h(x) = \exp\{-(x/\rho)^2\}$	$\hat{\theta} = 0.018$ $\hat{\rho} = 68.70$	243.23	222	0.1566	6336.48
2-component multiplicative double gamma	$h_1(x) = \exp\{-(x/\rho)^2\}$, $h_2(x) = 1$	$\hat{\theta}_1 = 0.018$ $\hat{\rho} = 69.03$ $\hat{\theta}_2 = 107.58$	243.23	221	0.1456	6338.49
1-component gamma (Dirichlet multinomial)	$h(x) = \exp\{-(x/\rho)^2\}$	$\hat{\theta}_1 = 0.018$ $\hat{\rho} = 68.67$ $\hat{\nu} =< 0.0001$	243.21	222	0.1568	6336.46

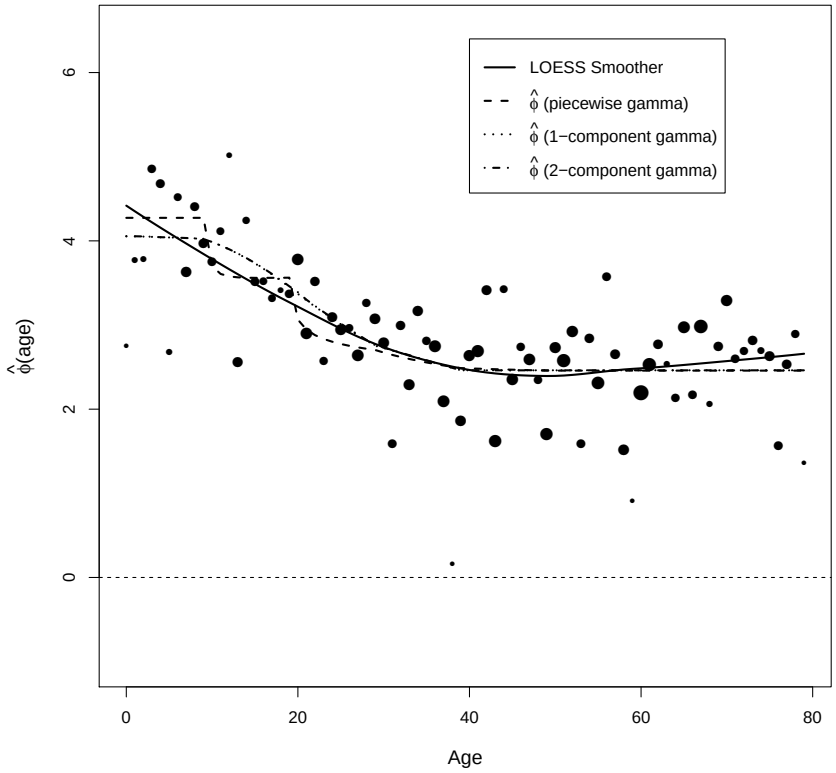


Figure 19: Observed and fitted associations between the infections HPV 16 and 18.

For the one-component multinomial model, approximate 95% CIs for the parameters of interest are as follows: for θ_1, (0.0133, 0.0241) and for ρ, (50.2035, 86.4496). The estimate of the frailty standard deviation with 95% CIs is given in Figure 23 (a). The frailty standard deviation is really high in childhood and decreases with age but has still a high value in adulthood (more than 2).

HPV 16 and 45

For the paired infections HPV 16 and 45 data from 7902 individuals are available. Both infections are transmitted via sexual intercourse, so heterogeneity in the number of sexual partners is likely to result in association between the two infections. Figure 21 shows the observed associations between the two infections. The association is similar to the pair of HPV 16 and 18 but more constant for all age groups.
The results of the fitted frailty models are presented in Table 9. All models show good results. The inverse Gaussian and the one-component multinomial model show the best fits. In this case, a second frailty term is not needed (two-component model) because the parameter $\rho = 51.21$ of the one-component model is high enough to model the positive association in adulthood. The fitted associations in Figure 21 for the one-component multinomial and the Dirichlet-multinomial model are virtually identical. Both closely resemble the observed pattern and the inverse Gaussian model shows good fits as well. The inverse Gaussian and the one-

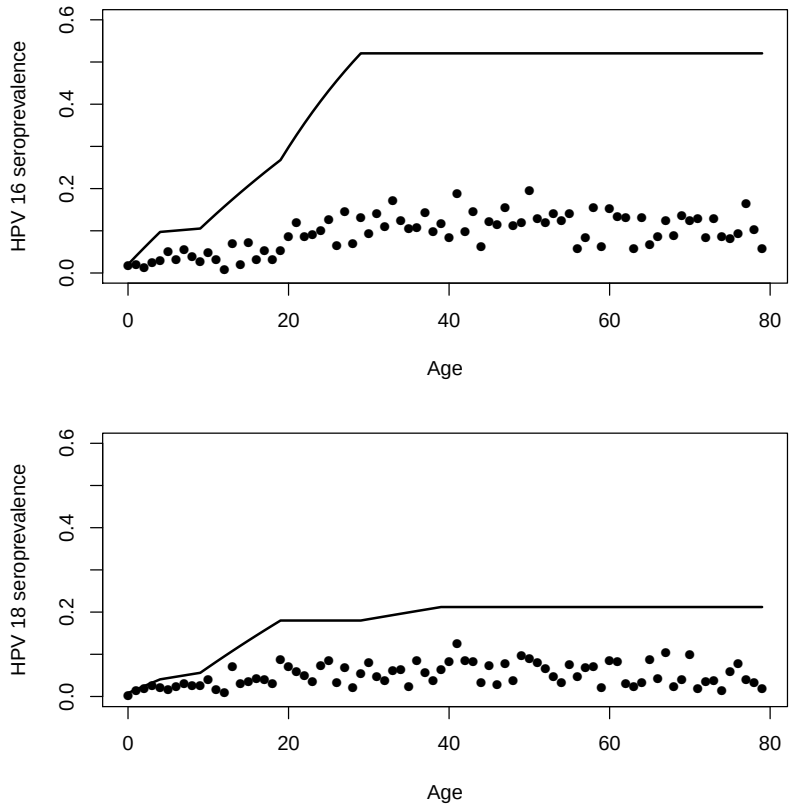


Figure 20: Observed seroprevalences of HPV 16 and 18 and fit of the 1-component frailty model to the data.

component model fit the seroprevalences of HPV 16 and 45 satisfactorily only for the higher age groups the seroprevalences are estimated a bit too high (Figure 22).

Table 9: Fitting results for Human Papilloma Virus 16 and 45 infection data.

Frailty model	Parameterization of $h(x), h_1(x)$ and $h_2(x)$	Parameter estimates	Deviance	Degrees of freedom	p-value	AIC
$U \sim \Gamma(\theta, 1/\theta)$		$\hat{\theta} = 0.176$	224.91	223	0.4516	7373.47
$U \sim InvG(1, \theta)$		$\hat{\theta} = 0.095$	222.17	223	0.5032	7370.73
Piecewise independent gamma	$\sigma_j^2 = \sigma^2$ $(j = 1, ..., 8)$	$\hat{\theta} = 0.071$	226.43	223	0.4235	7374.99
Piecewise independet gamma	$\sigma_j^2 = \sigma^2 \cdot$ $\exp[-\{(m_j - m_1)/\rho\}^2]$ $(j = 1, ..., 8)$	$\hat{\theta} = 0.081$ $\hat{\rho} = 39.91$	229.89	222	0.3440	7380.45
1-component gamma	$h(x) = \exp\{-(x/\rho)^2\}$	$\hat{\theta} = 0.111$ $\hat{\rho} = 51.21$	222.15	222	0.4845	7372.71
2-component multiplicative double gamma	$h_1(x) = \exp\{-(x/\rho)^2\}$, $h_2(x) = 1$	$\hat{\theta}_1 = 0.111$ $\hat{\rho} = 51.22$ $\hat{\theta}_2 = 1389.31$	222.15	221	0.4656	7374.71
1-component gamma (Dirichlet multinomial)	$h(x) = \exp\{-(x/\rho)^2\}$	$\hat{\theta}_1 = 0.111$ $\hat{\rho} = 51.21$ $\hat{\nu} =< 0.0001$	222.15	222	0.4846	7372.71

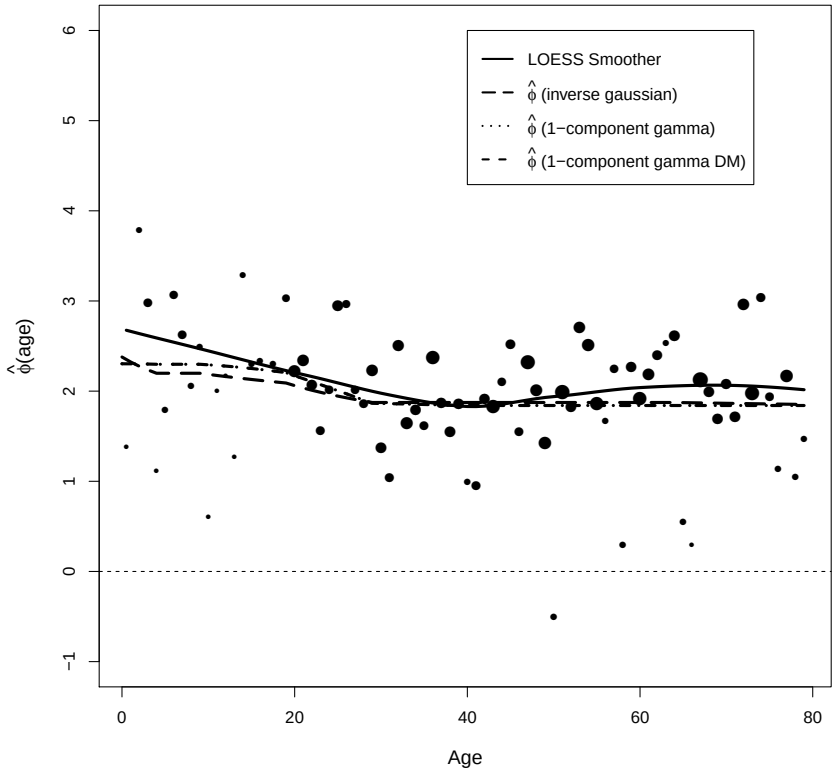
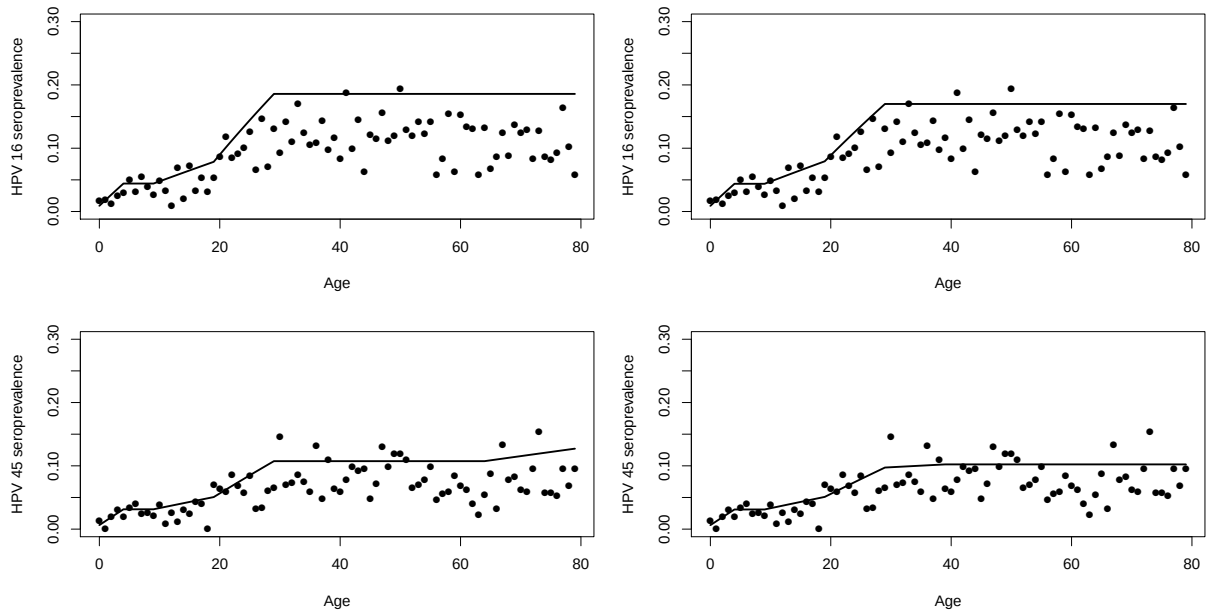


Figure 21: Observed and fitted associations between the infections HPV 16 and 45.

Approximate 95% CIs for the parameters of interest of the one-component model are as follows: for θ_1, (0.0786, 0.1574) and for ρ, (32.7840, 69.7808). The estimate of the frailty standard deviation with 95% CIs is given in Figure 23 (b). The frailty standard deviation is high in childhood and has a steeper decrease with age compared to the pair HPV 16 and 18 but has still a positive value in adulthood. The frailty standard deviation for the inverse Gaussian model is constant for all age groups with value 0.3082.

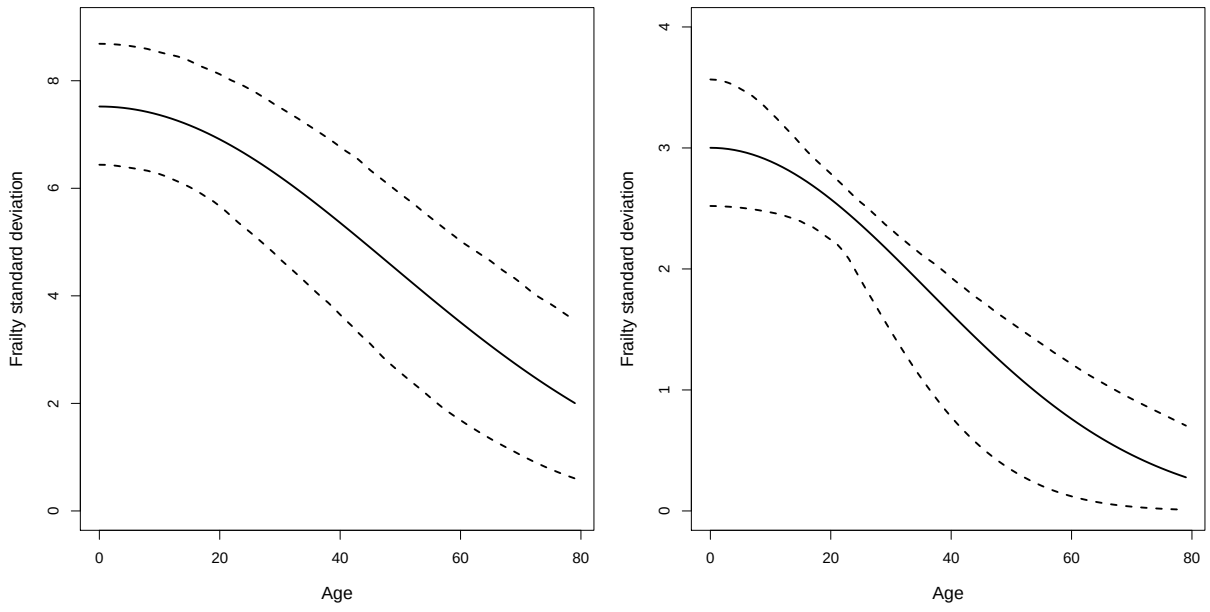
Toxoplasma and HAV

For Toxoplasmosis and Hepatitis A data from 4160 unvaccinated individuals from the nationwide sample are available. Both infections can be transmitted via oral ingestion of contaminated objects. Heterogeneity in hygiene is likely to result in association between these two infections. Figure 24 shows the observed associations between the two infections. Again a stronger heterogeneity in childhood which is declining with age can be observed. However, the heterogeneity in adulthood tails off to zero with a moderate increase at the higher age groups.
Table 10 presents the results of the fitted frailty models. Because all multinomial models show bad fits the Dirichlet-multinomial models are used. Since the heterogeneity in adulthood tails off to zero and hence the second component is not needed the one-component Dirichlet-multinomial is the most satisfying model. Nevertheless, in general all models show good fits. The use of the Dirichlet-multinomial distribution leads to a reduction in the pointwise deviances



(a) *(b)*

Figure 22: Observed seroprevalences of HPV 16 and 45 and fit of the (a) inverse Gaussian and (b) 1-component frailty model to the data.



(a) *(b)*

Figure 23: Standard deviation of the frailty (——) with 95% CIs obtained from fitting the 1-component gamma model to (a) HPV 16 and 18 and (b) HPV 16 and 45.

for almost all observed 4-tupels (Figure 27(d)). The biggest difference between the pointwise deviances is at the age of zero, the multinomial model has a value of 195.33 and the Dirichlet-multinomial model only a value of 49.65. The maximum variance inflating factor is 3.24 with an average of 1.56, so the multinomial component variances are increased by more than 56%, on average. The fitted associations in Figure 24 for the time invariant models are virtually nearly

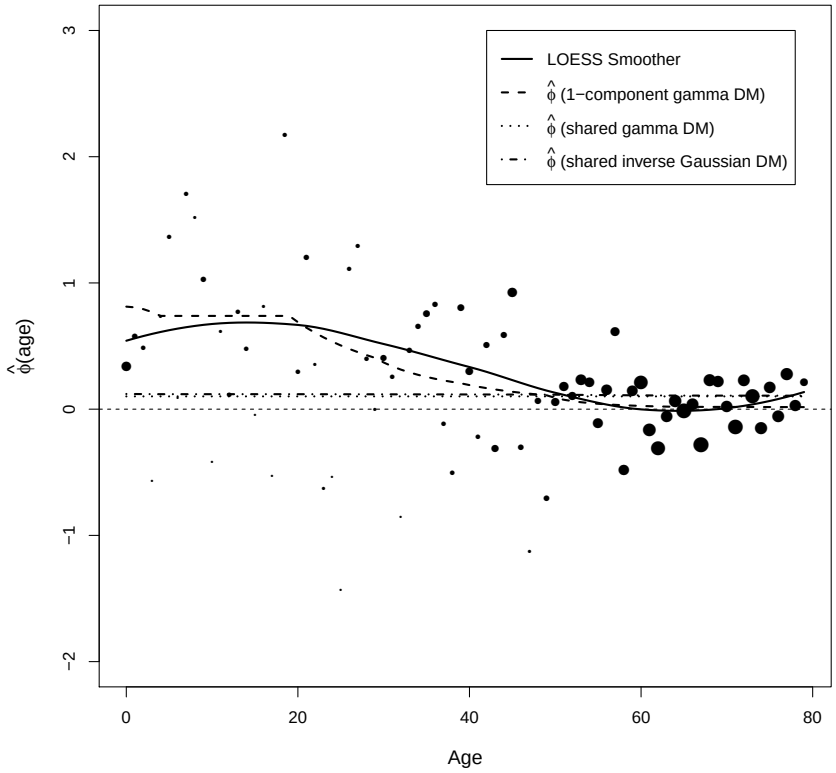


Figure 24: Observed and fitted associations between the infections Toxoplasma and HAV.

identical and predict a constant association. The one-component Dirichlet-multinomial model resembles the observed pattern most likely and also the seroprevalences are fitted satisfyingly with this model (Figure 25).

Table 10: Fitting results for Toxoplasma and HAV infection data for the nationwide sample.

Frailty model	Parameterization of $h(x), h_1(x)$ and $h_2(x)$	Parameter estimates	Deviance	Degrees of freedom	p-value	AIC
$U \sim \Gamma(\theta, 1/\theta)$ (Dirichlet multinomial)		$\hat{\theta} = 9.39$ $\hat{\nu} = 0.0112$	214.57	223	0.6453	8572.73
$U \sim InvG(1, \theta)$ (Dirichlet multinomial)		$\hat{\theta} = 7.79$ $\hat{\nu} = 0.0112$	214.30	223	0.6502	8572.46
1-component gamma (Dirichlet multinomial)	$h(x) = \exp\{-(x/\rho)^2\}$	$\hat{\theta} = 0.798$ $\hat{\rho} = 10.00$ $\hat{\nu} = 0.0109$	207.18	222	0.7542	8567.34
2-component multiplicative double gamma (Dirichlet multinomial)	$h_1(x) = \exp\{-(x/\rho)^2\}$, $h_2(x) = 1$	$\hat{\theta}_1 = 1.006$ $\hat{\rho} = 25.07$ $\hat{\theta}_2 = 103.44$ $\hat{\nu} = 0.0110$	208.15	221	0.7230	8570.31

For the one-component Dirichlet-multinomial model approximate 95% CIs for the parameters of interest are as follows: for θ_1, (0.3548, 1.7154) and for ρ, (0.5232, 33.3142). The estimate of the frailty standard deviation with 95% CIs is given in Figure 26 (d). The frailty standard

deviation is higher in childhood and decreases with age. By age 32 it is smaller than 0.0001.

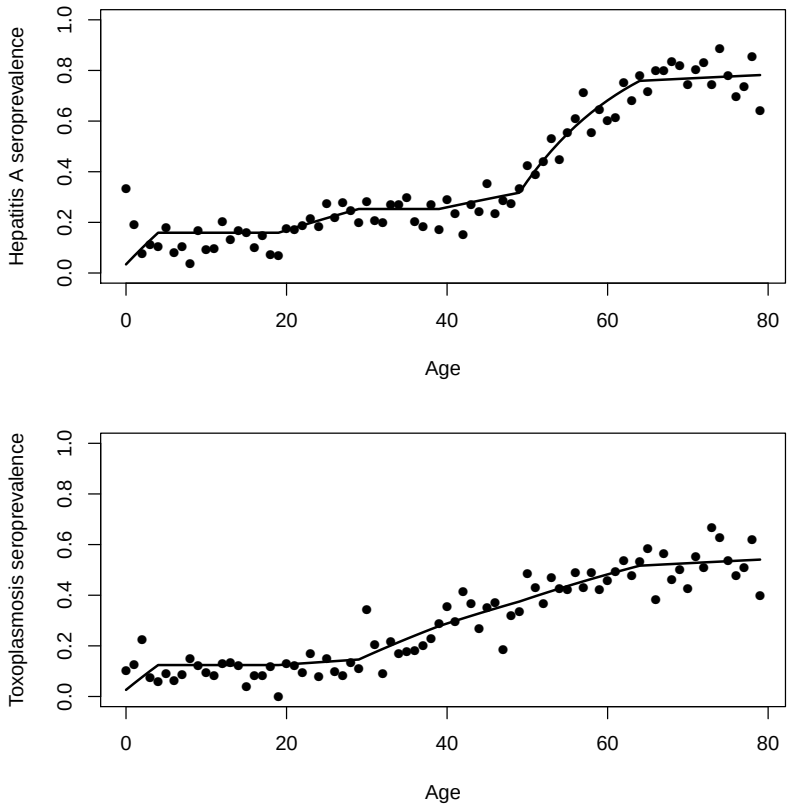


Figure 25: Observed seroprevalences of Toxoplasmosis and Hepatitis A and fit of the 1-component frailty model to the data.

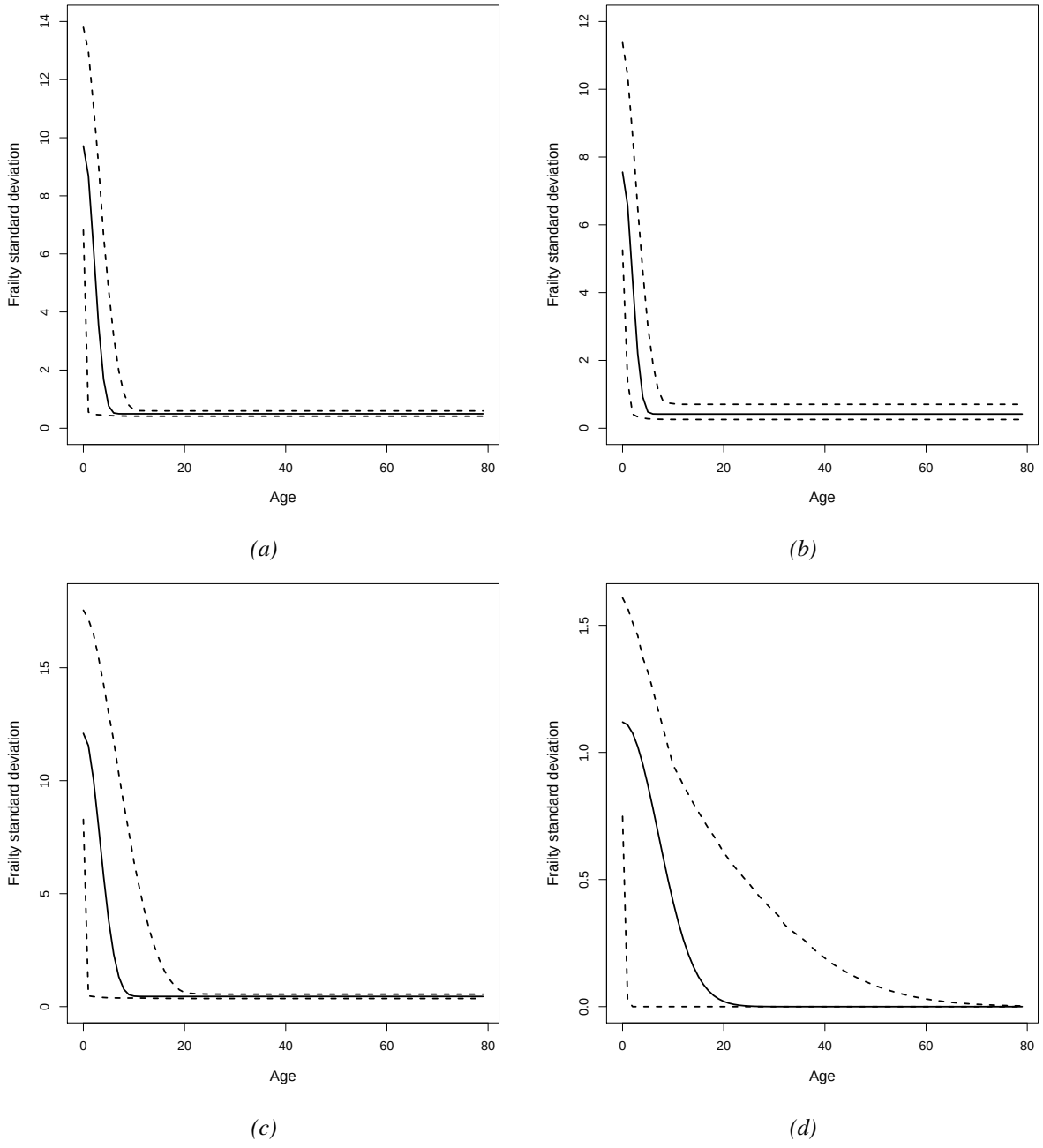
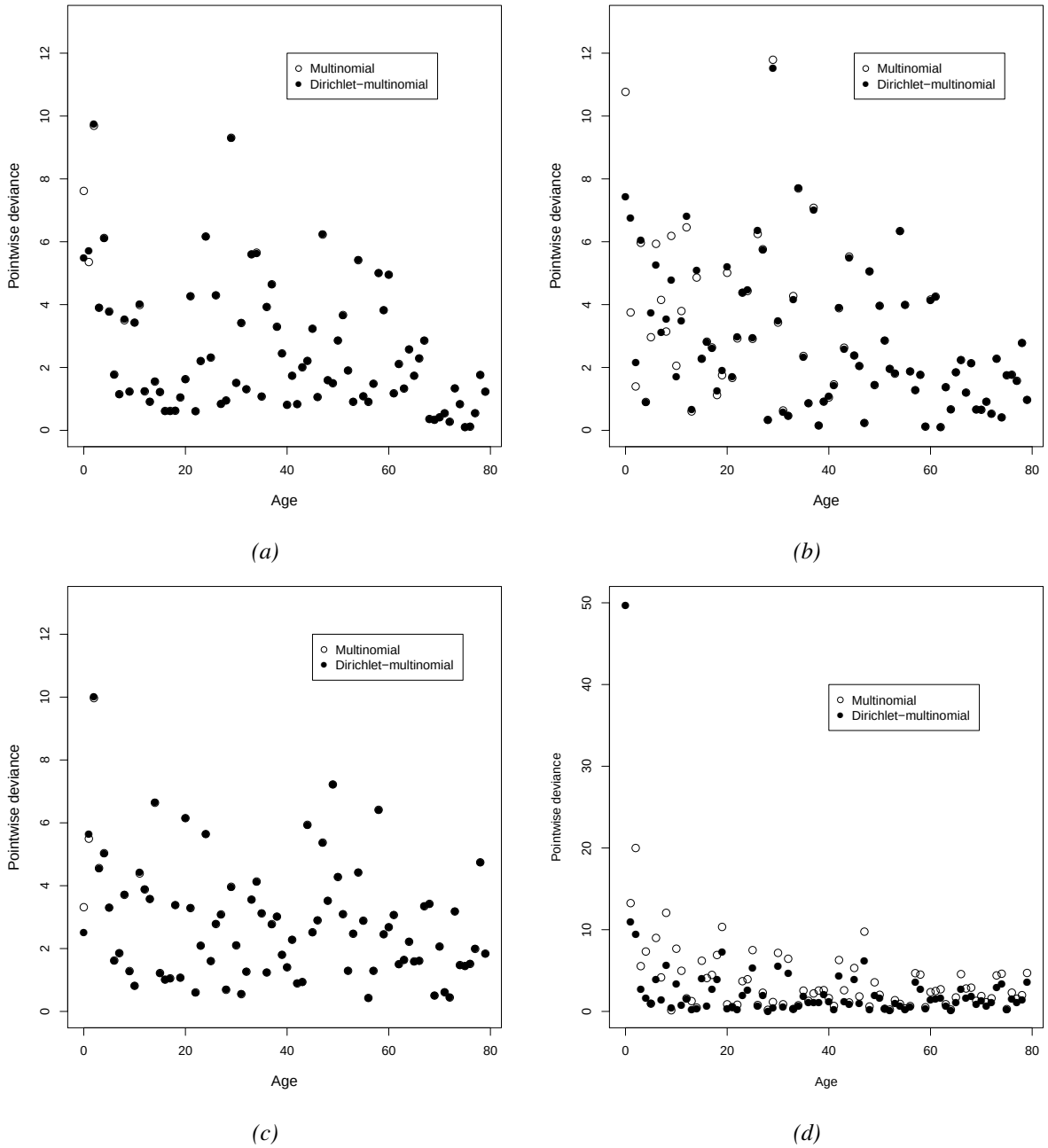


Figure 26: Standard deviation of the frailty (——) with 95% CIs obtained from fitting (a) the 2-component gamma model to Measles and Mumps, (b) the 2-component gamma Dirichlet-multinomial model to Measles and Rubella, (c) the 2-component gamma model to Mumps and Rubella and (d) the 1-component gamma Dirichlet-multinomial model to Toxoplasma and HAV.



(a) *(b)*

(c) *(d)*

Figure 27: Pointwise absolute deviances for the multinomial model and its compound Dirichlet-multinomial counterpart applied to (a) Measles and Mumps, (b) Measles and Rubella, (c) Mumps and Rubella and (d) Toxoplasma and HAV.

4.3.2 Pairs of infections with different mode of transmission

HAV and HBV

For the infection pair HAV and HBV not enough data for the smaller age groups are available, leading to bad fits. For this reason only data from the age classes 9 to 79 for the analysis are used. In total, data from 3861 unvaccinated individuals are available. HAV is transmitted via the fecal-oral route and HBV via sexual or close contacts. Because the routes of transmission are distinct a positive association in childhood which tails off to zero in adulthood could be expected. As already shown in Section 4.2 the associations of these two infections behave a bit different. Figure 28 shows the observed associations. A lower heterogeneity in childhood can be observed which increases up to the age of 30 and then decreases towards zero.

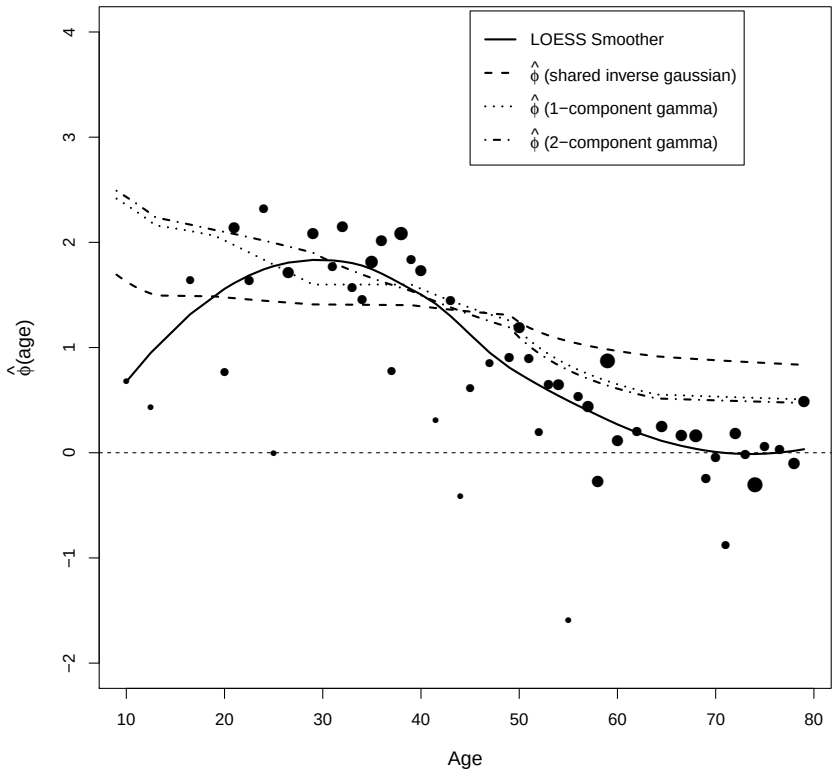


Figure 28: Observed and fitted associations between the infections HAV and HBV.

The results of the fitted frailty models are presented in Table 11. The time invariant frailty models cannot fit the data well. The best fitting model is the one-component multinomial model. Including an overdispersion parameter in the one-component model does not improve the model fit. The fitted associations in Figure 28 for the inverse Gaussian model are too constant. The one- and two-component model can capture the declining structure in adulthood but not the structure in the smaller age groups because they are not made to model such structures. Nevertheless, the seroprevalences can be fitted well with the one-component model except the fitted prevalences for Hepatitis A which are a bit too high (Figure 29). Even the fitted proportions

($f_{00}, f_{01}, f_{10}, f_{11}$) describe the observed proportions ($s_{00}, s_{01}, s_{10}, s_{11}$) quite satisfyingly (see Appendix Figure 38).

Table 11: Fitting results for HAV and HBV infection data for the nationwide sample from age class 9.

Frailty model	Parameterization of $h(x), h_1(x)$ and $h_2(x)$	Parameter estimates	Deviance	Degrees of freedom	p-value	AIC
$U \sim \Gamma(\theta, 1/\theta)$		$\hat{\theta} = 0.697$	240.28	198	0.0216	5057.21
$U \sim InvG(1, \theta)$		$\hat{\theta} = 0.201$	225.33	198	0.0888	5042.26
1-component gamma	$h(x) = \exp\{-(x/\rho)^2\}$	$\hat{\theta} = 0.083$ $\hat{\rho} = 37.77$	209.04	197	0.2650	5027.97
2-component multiplicative double gamma	$h_1(x) = \exp\{-(x/\rho)^2\}$, $h_2(x) = 1$	$\hat{\theta}_1 = 0.078$ $\hat{\rho} = 40.22$ $\hat{\theta}_2 = 806.76$	207.49	196	0.2732	5028.43
2-component multiplicative double gamma (Dirichlet multinomial)	$h_1(x) = \exp\{-(x/\rho)^2\}$ $h_2(x) = 1$	$\hat{\theta}_1 = 0.073$ $\hat{\rho} = 49.03$ $\hat{\theta}_2 = 823.15$ $\hat{\nu} =< 0.0001$	207.97	196	0.2655	5028.90
1-component gamma (Dirichlet multinomial)	$h(x) = \exp\{-(x/\rho)^2\}$	$\hat{\theta}_1 = 0.078$ $\hat{\rho} = 40.31$ $\hat{\nu} =< 0.0001$	207.48	197	0.2903	5026.41

Approximate 95% CIs for the parameters of interest of the one-component model are as follows: for θ_1, (0.0471, 0.1634) and for ρ, (29.2783, 55.6205). The estimate of the frailty standard deviation with 95% CIs is given in Figure 35 (a). The frailty standard deviation is higher in childhood and decreases with age. At the age of 79 it is still more than zero.

Toxoplasma and CMV

For Toxoplasma and CMV data from 4907 individuals from the nationwide sample are available. Toxoplasma is transmitted via oral ingestion of contaminated objects, whereas CMV is transmitted by mucosal contact with any body fluid. Again, a heterogeneity in childhood which tails off to zero in adulthood is expected. Figure 30 shows the observed associations.
Table 12 presents the results of the fitted frailty models. Because all multinomial models show bad fits the Dirichlet-multinomial models are used. The two-component Dirichlet-multinomial is not fitted because its counterpart already showed no better fits than the one-component model. In general, all fitted models show good fits. The use of the Dirichlet-multinomial distribution leads to a reduction in the pointwise deviance for almost all observed 4-tupels (Figure 34(a)). The maximum variance inflating factor is 2.99 with an average of 1.45, so the multinomial component variances are increased by more than 45%, on average. The fitted associations in Figure 30 for the time invariant models are virtually nearly identical and predict a constant association. The one-component Dirichlet-multinomial model resembles the observed pattern most likely and the seroprevalences are fitted satisfactorily with this model as well (Figure 31). The use of a piecewise constant force of infection is confirmed by comparing the observed and

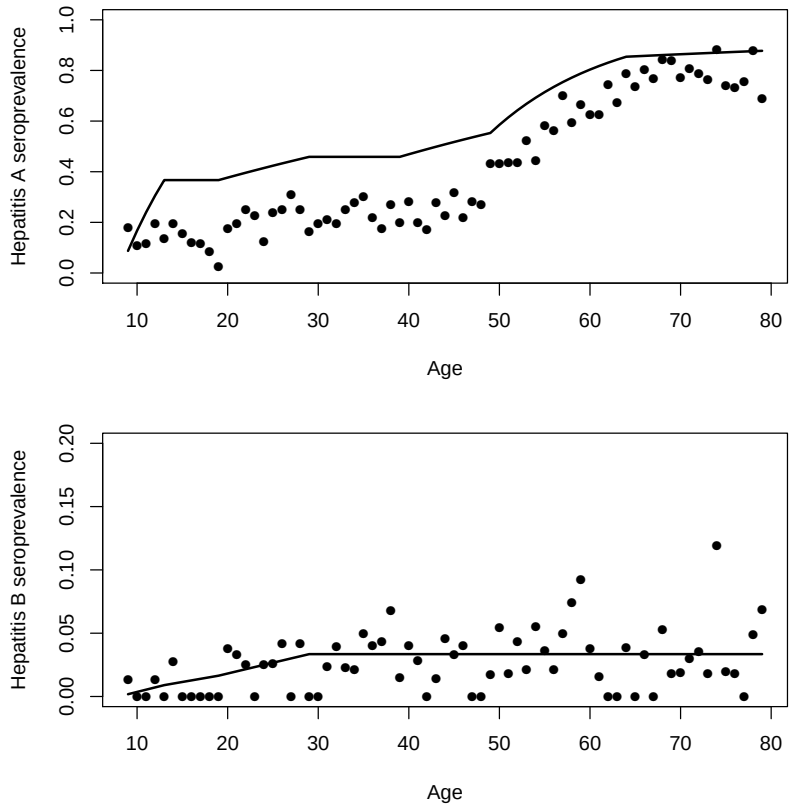


Figure 29: Observed seroprevalences of Hepatitis A and B and fit of the 1-component frailty model to the data.

fitted cumulative forces of infection for both infections (see Appendix Figure 39). The fitted values describes the observed cumulative forces of infection very closely.

Table 12: Fitting results for Toxoplasma and Cytomegalovirus infection data for the nationwide sample.

Frailty model	Parameterization of $h(x), h_1(x)$ and $h_2(x)$	Parameter estimates	Deviance	Degrees of freedom	p-value	AIC
$U \sim \Gamma(\theta, 1/\theta)$ (Dirichlet multinomial)		$\hat{\theta} = 16.07$ $\hat{\nu} = 0.0077$	230.90	223	0.3441	11244.01
$U \sim InvG(1, \theta)$ (Dirichlet multinomial)		$\hat{\theta} = 14.43$ $\hat{\nu} = 0.0077$	230.81	223	0.3456	11243.93
1-component gamma (Dirichlet multinomial)	$h(x) = \exp\{-(x/\rho)^2\}$	$\hat{\theta} = 2.288$ $\hat{\rho} = 20.96$ $\hat{\nu} = 0.0074$	228.54	222	0.3672	11243.66

For the one-component Dirichlet-multinomial model approximate 95% CIs for the parameters of interest are as follows: for θ_1, (1.1909, 4.2169) and for ρ, (1.4408, 48.7863). The estimate of the frailty standard deviation with 95% CIs is given in Figure 35(b). The frailty standard deviation is high in childhood and decreases with age towards zero.

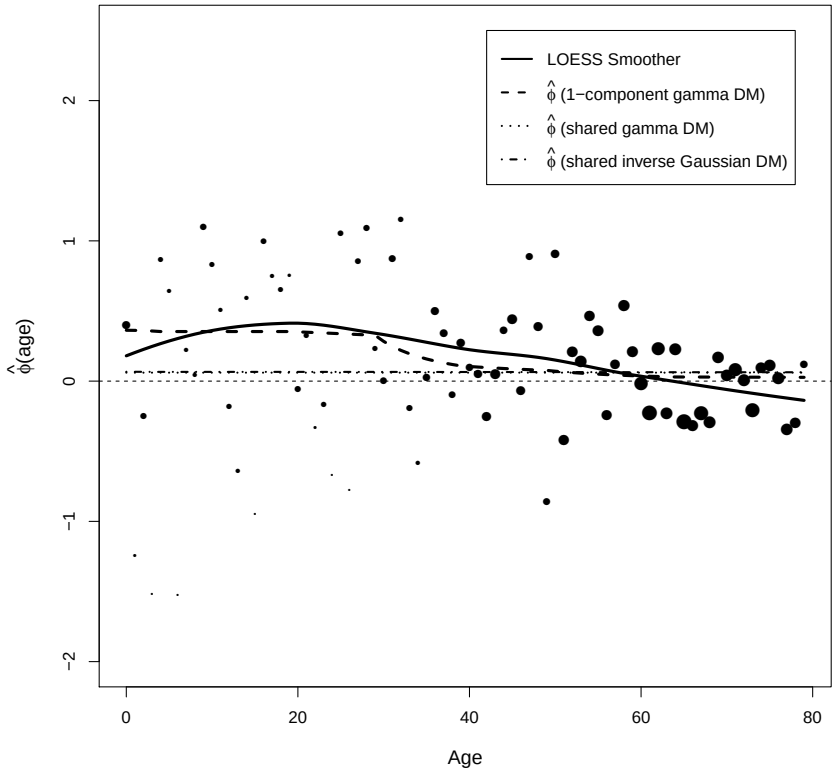


Figure 30: Observed and fitted associations between the infections Toxoplasma and CMV.

Toxoplasma and VZV

For Toxoplasma and VZV data from 4906 individuals from the nationwide sample are available. Toxoplasma is transmitted via oral ingestion of contaminated objects, whereas VZV is transmitted by close contact with respiratory secretions. Figure 32 shows the observed associations. A higher association in the first years of age can be observed which decreases to a negative value and then increases again to zero.

The results of the fitted frailty models are presented in Table 13. All multinomial frailty models cannot fit the data well. Including an overdispersion parameter in the models improves the model fits. The use of the Dirichlet-multinomial distribution leads to a reduction in the pointwise deviance for almost all observed 4-tupels (Figure 34(b)). The maximum variance inflating factor is 2.20 with an average of 1.27, so the multinomial component variances are increased by more than 27%, on average. The fitted associations in Figure 32 for the inverse Gaussian model are constant. The one-component Dirichlet-multinomial model instead can capture the transient heterogeneity in childhood and no association in adulthood but it is not made to fit the negative association from 15 to 35 years of age. Nevertheless, the seroprevalences are fitted quite well with this model (Figure 33) and also the fitted cumulative force of infection with this model can describe the observed one well (see Appendix Figure 40).

Approximate 95% CIs for the parameters of interest of the one-component Dirichlet-multinomial model are as follows: for θ_1, (0.2282, 2.8268) and for ρ, (0.4420, 2.4769). The estimate of the

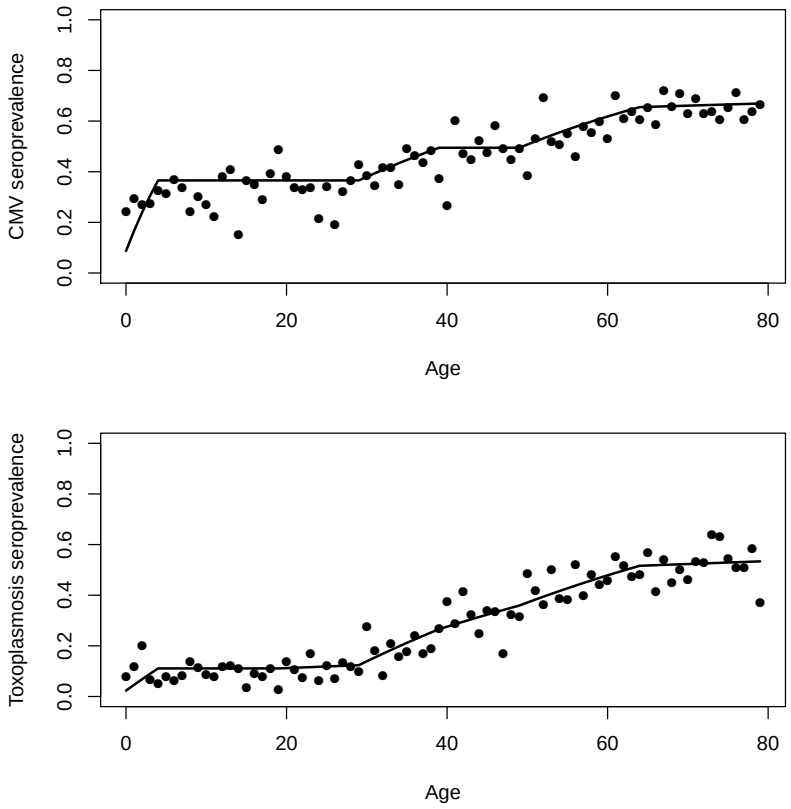


Figure 31: Observed seroprevalences of Toxoplasmosis and CMV disease and fit of the 1-component Dirichlet-multinomial frailty model to the data.

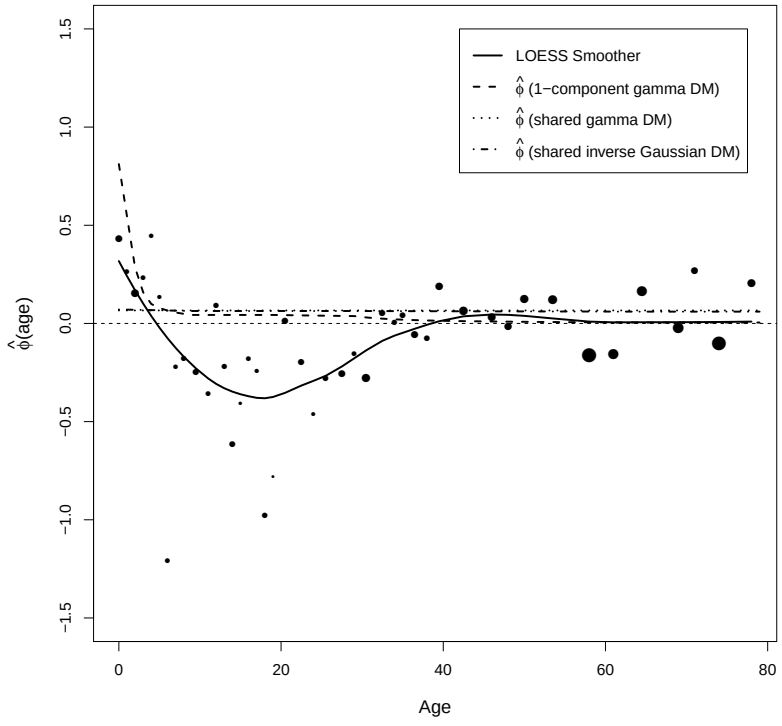


Figure 32: Observed and fitted associations between the infections Toxoplasma and VZV.

Table 13: Fitting results for Toxoplasma and VZV infection data for the nationwide sample.

Frailty model	Parameterization of $h(x), h_1(x)$ and $h_2(x)$	Parameter estimates	Deviance	Degrees of freedom	p-value	AIC
$U \sim \Gamma(\theta, 1/\theta)$		$\hat{\theta} = 11.90$	254.30	223	0.0737	6382.17
$U \sim InvG(1, \theta)$		$\hat{\theta} = 10.59$	254.37	223	0.0732	6382.24
Piecewise independent gamma	$\sigma_j^2 = \sigma^2$ $(j = 1, ..., 8)$	$\hat{\theta} = 5.85$	254.71	223	0.0712	6382.58
Piecewise independet gamma	$\sigma_j^2 = \sigma^2 \cdot \exp[-\{(m_j - m_1)/\rho\}^2]$ $(j = 1, ..., 8)$	$\hat{\theta} = 6.867$ $\hat{\rho} = 16.94$	255.29	222	0.0620	6385.16
1-component gamma	$h(x) = \exp\{-(x/\rho)^2\}$	$\hat{\theta} = 1.332$ $\hat{\rho} = 1.46$	253.01	222	0.0750	6382.88
2-component multiplicative double gamma	$h_1(x) = \exp\{-(x/\rho)^2\}$, $h_2(x) = 1$	$\hat{\theta}_1 = 1.552$ $\hat{\rho} = 1.35$ $\hat{\theta}_2 = 17.62$	252.35	221	0.0725	6384.22
$U \sim \Gamma(\theta, 1/\theta)$ (Dirichlet multinomial)		$\hat{\theta} = 14.61$ $\hat{\nu} = 0.0044$	209.00	223	0.7407	6336.87
$U \sim InvG(1, \theta)$ (Dirichlet multinomial)		$\hat{\theta} = 13.41$ $\hat{\nu} = 0.0045$	208.21	223	0.7532	6336.08
1-component gamma (Dirichlet multinomial)	$h(x) = \exp\{-(x/\rho)^2\}$	$\hat{\theta}_1 = 0.799$ $\hat{\rho} = 1.45$ $\hat{\nu} = 0.0044$	206.81	222	0.7599	6336.68

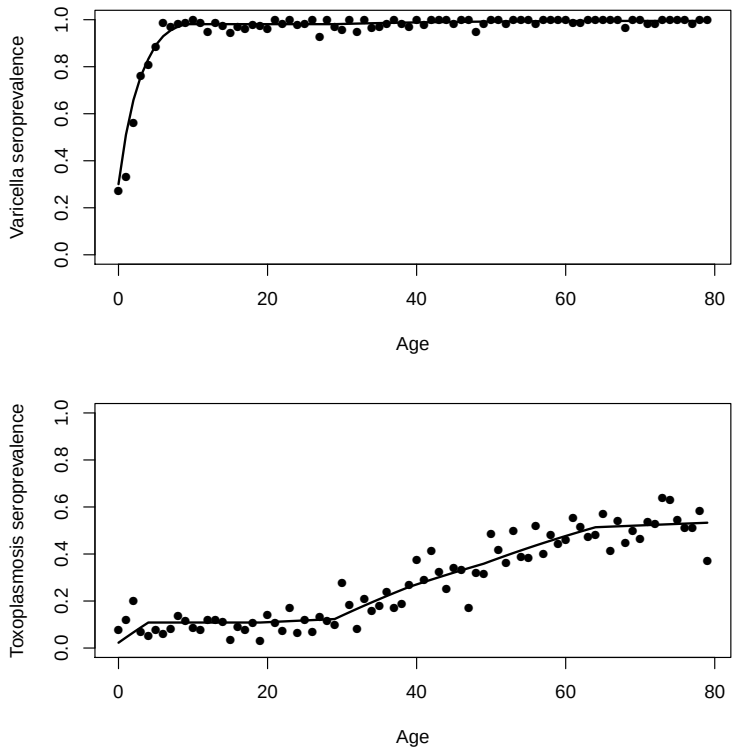
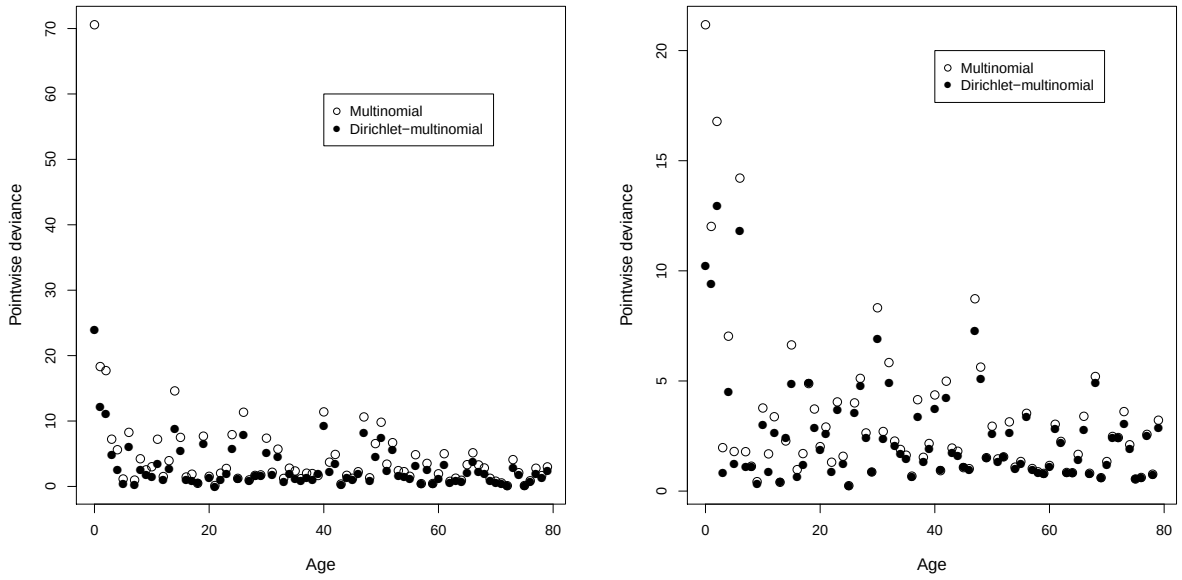


Figure 33: Observed seroprevalences of Toxoplasmosis and Varicella and fit of the 1-component Dirichlet-multinomial frailty model to the data.

frailty standard deviation with 95% CIs is given in Figure 35 (b). The frailty standard deviation is a bit higher in childhood and strongly decreases with age towards zero.



(a) *(b)*

Figure 34: Pointwise absolute deviances for the multinomial model and its compound Dirichlet-multinomial counterpart applied to (a) Toxoplasma and CMV and (b) Toxoplasma and VZV.

These analyses show that shared frailty models are very useful for fitting paired current status data and to estimate the extent of heterogeneity. The association measure $\phi(x)$ can be used to find suitable frailty models. For pairs of infections with the same route of transmission a positive association in childhood which decreases to a constant positive association in adulthood is mostly observed. The two-component frailty model is advisable for these data. If the association is large for all age groups (for example HPV) the estimated value ρ is large, too, so a one-component model is adequate for these data. For pairs of infections with different transmission routes the one-component frailty model seems to be satisfactory. This model represents the transient heterogeneity in childhood which tails off to zero. One possible explanation is the heterogeneous behaviour concerning household size and visiting nursery-school at young ages and a more homogeneous behaviour at older ages, one reason the compulsory school attendance, which leads to a reduced heterogeneity (Farrington et al., 2013).
Some exceptions could be observed. For instances, HAV and HBV show an atypical structure of associations with smaller associations in childhood which increase up to the age of 30 and then decrease towards zero. Another structure of the function $h(x)$ could be used to model those trends. Nevertheless, these analyses achieve similar results as Unkel et al. (2014). For the pair Toxoplasma and CMV with different transmission routes the typical structure of positive associations in childhood decreasing with age could not be observed. But this behaviour was already observed by Farrington et al. (2013).

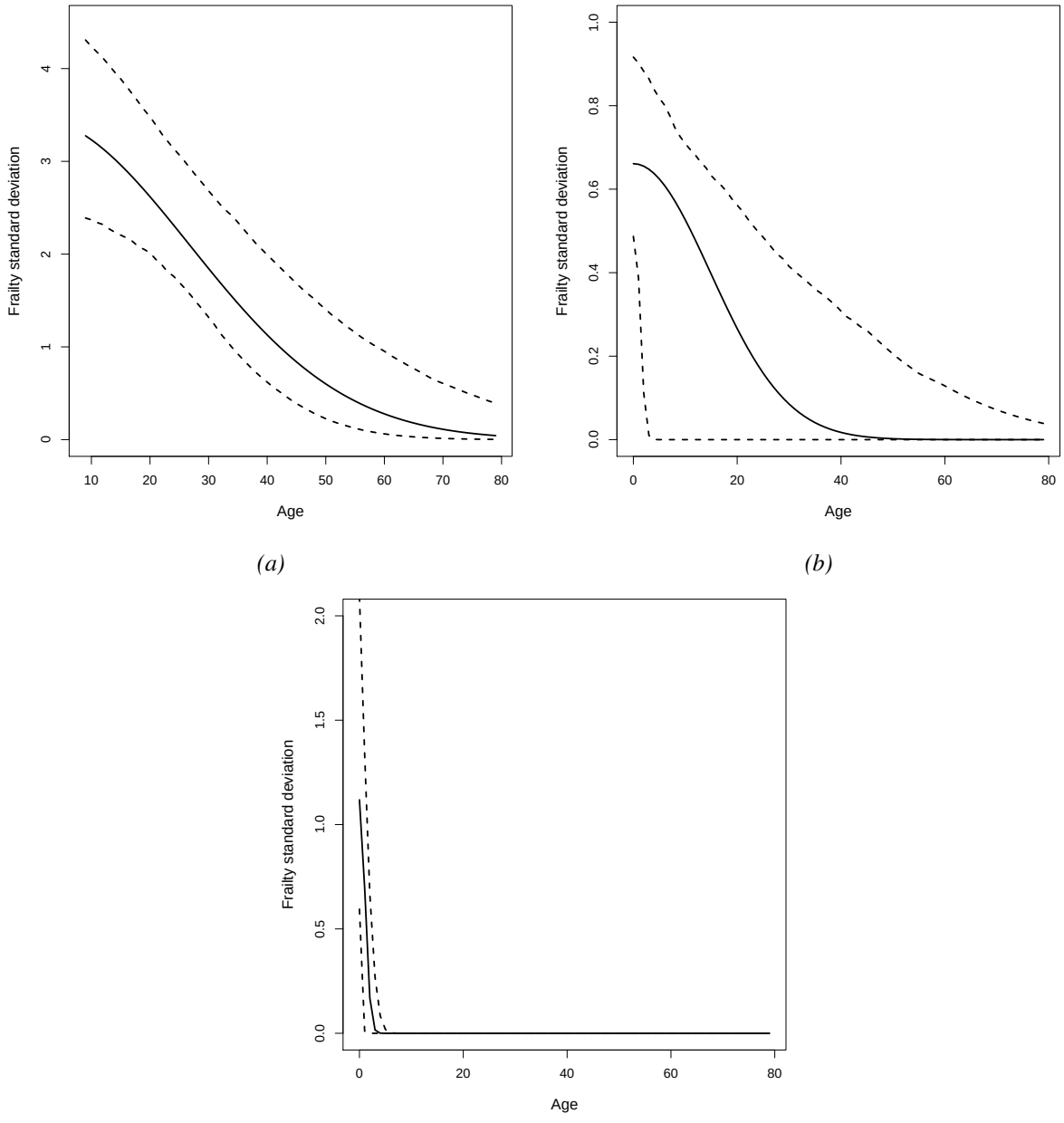


Figure 35: Standard deviation of the frailty (——) with 95% CIs obtained from fitting (a) the 1-component gamma model to HAV and HBV, the 1-component gamma Dirichlet-multinomial model to (b) Toxoplasma and CMV and (c) Toxoplasma and VZV.

However, the use of frailty models suffers from two disadvantages. First, only the sources of heterogeneity which are shared between infections are incorporated and not the unshared. Second, the source of heterogeneity can not be determined (Farrington et al., 2013).

5 Estimation of key epidemiological parameters

In this Section one of the most useful threshold parameters for infectious diseases, the basic reproduction number R_0, is introduced (Section 5.1). The use of social contact data and a description of these data for the Dutch population follow in Section 5.2. In Section 5.3, an estimation with social contact data for the infectious diseases Measles, Mumps, Rubella and Varicella is presented.

5.1 Basic reproduction number R_0

The basic reproduction number R_0 is one of the most useful threshold parameters for infectious diseases and describes the average number of new infections in a completely susceptible population if one infected individual enters this population. If R_0 is greater than 1, the infection spreads out and if R_0 is smaller than 1 the infection will die out. This number is useful for immunisation programs to evaluate the effort required to eliminate an infection (Farrington, Kanaan, and Gay, 2001).
Several methods of estimating R_0 exist according to the given type of data. A rough classification in three categories would be: Using parameters describing the contacts in the population, for example, numbers of sexual contacts for sexual transmitted diseases. The second category uses data of susceptible individuals in a population and the amount of cases in a given time interval. The third method needs quantities that are estimated from the endemic equilibrium (Farrington, Kanaan, and Gay, 2001).
In this thesis the latter method is used. Therefore, serological data and a social contact pattern are needed. One method is to assume social contact pattern with different mixing patterns. However, this approach does not consider the underlying contact mixing pattern in the analysed population. Another approach uses social contact data to estimate R_0 for airborne infections, which is used in this thesis (Hens et al., 2012, p. 233).

5.2 Social contact data

Social contact data can be obtained by means of a diary which is part of the POLYMOD study (see Section 2). As already mentioned this social contact data are not available. Instead the contact numbers given in the questionnaires are used. Individuals with no information are excluded from the analysis and for individuals with a total number of contacts the missing values in the other age classes are replaced with zeros. For this analysis data from 6790 individuals are available. Figure 36 shows the calculated average number of contacts reported per participant with subjects of different age classes. One special characteristic is the strong diagonal element. Individuals of similar age have contacts favoured with others of similar age. The number of contacts decreases with age. Most contacts per day has the group of those aged 10-19 years

with an average number of twelve contacts per day. The group of those aged 70-79 years has the smallest number of contacts with their age class. Two further features are the mixing pattern between children and their adults and the further the age classes of respondents and contacts are apart the less contact they have. These features lead to a slight "rocket" shape. This shape is more dominant in the work of Mossong et al. (2008) where they used smoothed contact matrices.

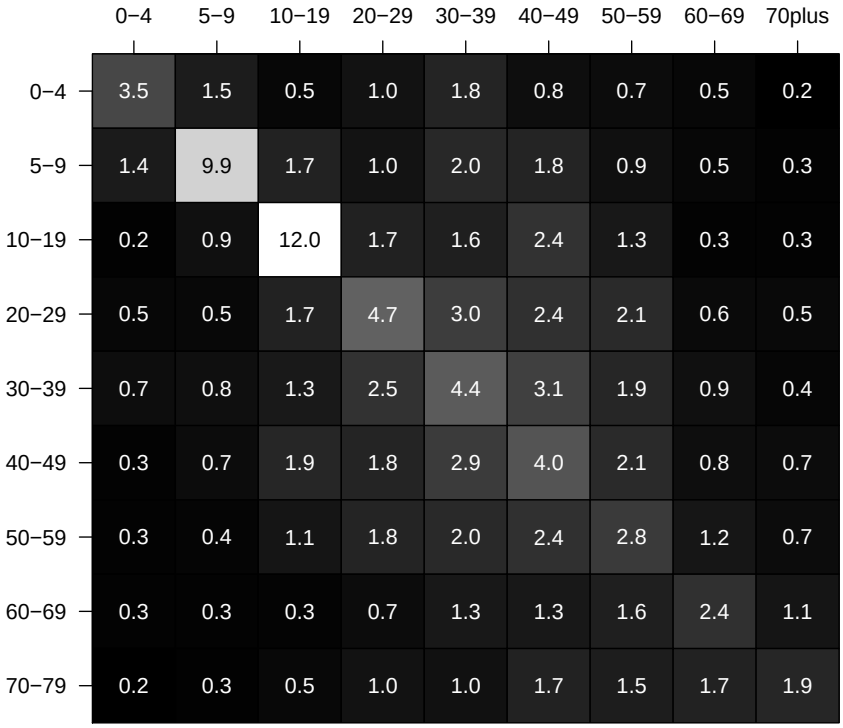


Figure 36: Contact matrix for the Dutch population estimated from the questionnaire data, y-axis: respondents and x-axis: contacts, created with 'color2D.matplot' from the plotrix package (Lemon, 2006).

5.3 Estimation of R_0

This Section is mainly based on the work of Goeyvaerts et al. (2010).
Consider a large population of size N with a constant mortality rate $\mu(a) = L^{-1}$, with a being the age (0-79) and L the life expectancy. As well, assume a lifelong immunity of recovered individuals and a short mean duration of infectiousness D (Farrington et al., 2001). With the

age-specific rate $\gamma(a)$ of losing maternal antibodies, defined as

$$\exp\Big(- \int_0^a \gamma(a)\, ds\Big) = \begin{cases} 1 & \text{if } a \leq A, \\ 0 & \text{if } a > A, \end{cases}$$

the fact that in the first months of life newborns are protected by maternal antibodies is taken into account. After the certain age A they are instantaneously susceptible. The proportion of susceptible individuals is then given by

$$s(a) = \exp\Big(- \int_A^a \lambda(s)\, ds\Big), \quad \text{if } a > A,$$

with $\lambda(a)$ being the age-specific force of infection, and $s(a) = 0$ if $a \leq A$. The approximated force of infection if D is short is:

$$\lambda(a) = \frac{ND}{L} \int_A^\infty \beta(a, a')\lambda(a')x(a')\, da',$$

where $\beta(a, a')$ describes the transmission rate, the per capita rate at which an individual of age a' makes effective contacts with individuals of age a (R. M. Anderson, B. Anderson, and May, 1991).

Given the first age interval $(a_{[1]}, a_{[2]})$ and the j-th age interval $[a_{[j]}, a_{[j+1]})$, $j = 2, ..., J$, with $a_{[1]} = A$ and $a_{[J+1]} = L$, a constant force of infection is assumed in each age group. The approximated prevalence of immune individuals of age a in the j-th age interval is (R. M. Anderson, B. Anderson, and May, 1991):

$$\pi(a) = 1 - \exp\Big(- \sum_{k=1}^{j-1} \lambda_k(a_{[k+1]} - a_k) - \lambda_j(a - a_{[j]})\Big). \tag{5.1}$$

Moreover, the force of infection for age class i ($i = 1, ..., J$) is given by (Hens et al., 2012):

$$\begin{aligned} \lambda_i = \frac{ND}{L} \sum_{j=1}^{J} \beta_{ij} \frac{\lambda_j}{\lambda_j + \mu_j} \Bigg[&\exp\Bigg(- \sum_{k=1}^{j-1} (\lambda_k + \mu_k)(a_{[k+1]} - a_{[k]})\Bigg) \\ &- \exp\Bigg(- \sum_{k=1}^{j} (\lambda_k + \mu_k)(a_{[k+1]} - a_{[k]})\Bigg)\Bigg], \end{aligned} \tag{5.2}$$

with β_{ij} being the per capita rate as defined previously. The effective contact rates β_{ij} are thus a $J \times J$ matrix , called Who Acquires Infection From Whom (WAIFW) matrix. The basic reproduction number R_0 is the dominant eigenvalue of the $J \times J$ next generation matrix with elements ($i, j = 1, ..., J$):

$$\frac{ND}{L} \beta_{ij} M,$$

where M is the diagonal matrix with entries

$$\int_{a_{[i]}}^{a_{[i+1]}} \exp\Big\{ -\int_0^a \mu(s)\, ds \Big\} da. \tag{5.3}$$

The transmission rate $\beta(i,j)$ can be estimated by using social contact data. The transmission rate is assumed to be proportional to the contact rate $c(i,j)$, the per capita at which an individual of age j makes contact with an individual of age i:

$$\beta(i,j) = q \cdot c(i,j), \tag{5.4}$$

where q is a constant proportionality factor (Wallinga, Teunis, and Kretzschmar, 2006). To estimate the contact rates $c(i,j)$ assume a random variable Y_{ij} to be the number of contacts in age group j during one day as reported by a participant in age group i $(i,j = 1,...,J)$ with observed values $y_{ij,t}$, $t = 1,...,T_i$ with T_i being the number of respondents in age class i in the contact survey. The "social contact matrix", a $J \times J$ matrix, has entries $m_{ij} = E(Y_{ij})$ being the mean number of contacts per one day in age group j as reported by a participant in age group i. So the contact rates can be presented as:

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

with w_i being the size of the population in age group i (Goeyvaerts et al., 2010). The reciprocal nature of contacts must to be taken into account for estimating the social contact matrix. This means that the total number of contacts from age group i to age group j must equal the total number of contacts from age group j to age group i (Wallinga, Teunis, and Kretzschmar, 2006):

$$m_{ij} w_i = m_{ji} w_j.$$

The observed data usually do not fulfil this relationship exactly. Nevertheless, to take the reciprocal nature of contacts into account the following equation is used (Funk, 2018):

$$m'_{ij} = \frac{1}{2w_i}(m_{ij} w_i) + m_{ji} w_j.$$

To estimate the proportionality factor q an iterative procedure is used. A plausible starting value for q is assumed and Equation 5.2 is solved. This is repeated until the Bernoulli log-likelihood

$$\sum_{i=1}^{n} \{ y_i \ln[\pi(a_i)] + (1 - y_i) \ln[1 - \pi(a_i)] \},$$

has reached its maximum. Where n is the number of all participants and $y_i = 1$ if individual i was infected before age a_i and $y_i = 0$ if individual i is still susceptible at age a. Equation 5.1 is

used for the prevalence $\pi(a_i)$ (Goeyvaerts et al., 2010).

To estimate the transmission rates information of the population size w_i in each age group i ($i = 1, .., 9$) is used from UNdata (2019). For the calculation the code of Hens et al. (2012) from Chapter 15 is used and modified so that it can be used for the given data.
The population size in the Netherlands in 2007 was 16,357,209 people, this number is used for N. The life expectancy for 2007 was 80 years of age, so L is set to 80 and the mean duration of infectiousness D is set to $5/365$. The loss of maternal antibodies is assumed to be with age $A = 0.5$. For the analysis, the population is divided into nine age classes: (0.5, 5), [5, 10), [10, 20), [20, 30), [30, 40), [40, 50), [50, 60), [60, 70) and [70, 80). Again, a piecewise constant force of infection is assumed.

Measles

Removing the unvaccinated participants data from 4446 individuals can be used for the analysis of Measles. Applying the method described above to the data the estimated proportionality factor $\hat{q}$ equals 0.078 and the resulting estimated basic reproduction number $\hat{R}_0$ has the value 4.64 (see Table 14), meaning that if one infected individual enters a completely susceptible population 4.64 individuals would be infected, on average. Compared to other analyses, for example see (Guerra et al., 2017), this estimated $\hat{R}_0$ is quite small. One reason could be the underestimation of effective contacts caused by the possible biased contact matrix.

Mumps

Including only the unvaccinated individuals data from 4779 individuals can be used to estimate R_0 for Mumps. The proportionality factor $\hat{q} = 0.051$ is a bit smaller compared to Measles and also $\hat{R}_0 = 3.01$ is smaller (Table 14). The estimated number is compared to other studies a bit smaller. For example, Wallinga, Teunis, and Kretzschmar (2006) used data from Utrecht in 1986. In this analyses $\hat{q} = 0.16$ and $\hat{R}_0 = 7.68$ were estimated. Edmunds et al. (2000) estimated R_0 with a value of 4.3 for the Netherlands.

Rubella

Removing the unvaccinated participants data from 4375 individuals can be used for the analysis of Rubella. The estimated proportionality factor $\hat{q}$ equals 0.070 and the resulting estimated basic reproduction number $\hat{R}_0$ has the value 4.17 (Table 14). This estimated number is compared to other analyses a bit smaller. For example, Edmunds et al. (2000) estimated R_0 with a value of 6.4 for the Netherlands.

Varicella

For the calculation of R_0 for Varicella data from 7475 individuals are available. The estimated proportionality factor $\hat{q}$ equals 0.130 and the resulting estimated basic reproduction number $\hat{R}_0$ has the value 7.69 (Table 14). Comparing these estimated numbers with the numbers of the neighbouring country Belgium which where analysed by Goeyvaerts et al. (2010), they most likely match with the results from contact matrix in Belgium with all close contacts longer than 15 minutes ($\hat{q} = 0.173$ and $\hat{R}_0 = 8.86$).

Table 14: Estimates for the proportionality factor q and the basic reproduction number R_0.

Infectious disease	Proportionality factor $\hat{q}$	Basic reproduction number $\hat{R}_0$
Measles	0.078	4.64
Mumps	0.051	3.01
Rubella	0.070	4.17
VZV	0.130	7.69

For all analysed infectious diseases R_0 is greater than one meaning that the required efforts to eliminate the infectious diseases from the population need to be maintained. Another idea of estimating key epidemiological parameters is to use shared frailty models and so to take the heterogeneity into account. Assuming contact patterns or using social contact data obtaining from diaries are restrictive and costly to collect, respectively. Nevertheless, it has long been understood that it is important to take heterogeneities into account. Frailty models are an alternative to improve estimation of epidemiological parameters, such as R_0 (Farrington, Kanaan, and Gay, 2001). So further analyses should estimate R_0 with frailty models and compare the results with the results of this thesis.

6 Conclusion

This thesis is based on the available data from the PIENTER-2 project. An univariate analysis of several infectious diseases and their seroprevalences in the Dutch population was conducted. For Measles, Mumps, Rubella and Varicella a steep increase of the seroprevalence with age towards one is observed. The seroprevalence for Cytomegalovirus depends on socio-economic factors and is higher in developing countries. Dividing the data into native Dutch individuals and non-Western migrants the overall seroprevalence in the first group (41.7%) is much lower compared to the second group (84.3%). For Toxoplasmosis and Hepatitis A and B an increase of the seroprevalence is observed. The overall seroprevalence of Toxoplasmosis in the Dutch population is 25.7%, for Hepatitis A 40.5% and for Hepatitis B 3.4%. The overall seroprevalence for Human Papilloma Virus 16 which is the most common type is 9%, for type 45 5.6% and type 18 5.4%.

The main focus was on the analysis of paired current status data of different infectious diseases. For the analysis infections with similar and different routes of transmission were differentiated. The association measure $\phi(x)$ and its summary measure $\bar{\phi}$ were used to measure the association between pairs of infections. For most of the pairs of infections with a similar route of transmission a strong heterogeneity in childhood could be observed, for Measles and Mumps and Measles and Rubella only in early childhood, which is decreasing in adulthood to a positive constant value. For most of the pairs of infections with different transmission routes a positive heterogeneity in childhood was detected, however, this heterogeneity decreases with age towards zero.

Shared frailty models were used to estimate the extent of heterogeneity. Time invariant (gamma and inverse Gaussian) as well time varying frailty models, with frailty terms being gamma distributed were applied to the data. In many cases the time invariant models showed worse fits. The time varying one- and two-component models fitted the data for all pairs of infections quite well and are advocated to use. The one-component model is able to model the transient decline of heterogeneity in childhood that tails off to zero in adulthood which is observed for pairs of infections with different routes of transmission. Whereas, the two-component model is appropriate for pairs of infections with the same transmission route in which the first frailty term represents the decreasing heterogeneity in childhood and the second term represents the persistent heterogeneity in adulthood.

Another part of the analysis was the estimation of the basic reproduction number R_0 with the use of serological and social contact data. Unfortunately, only the data of an estimated number of contact for one day were available. The basic reproduction number was calculated for four

infectious diseases which are transmitted via airborne, Measles, Mumps, Rubella and Varicella. For all infectious diseases the estimated R_0 were smaller compared to other analyses. One reason could be the possible biased social contact matrix. For Measles the estimated basic reproduction number is 4.64, meaning that if one infected individual enters a completely susceptible population 4.64 individuals would be infected, on average. For Mumps it is 3.01, for Rubella 4.17 and for Varicella 7.69. Because all numbers are greater than one, the required efforts to eliminate infectious diseases from the population need to be upheld.

Shared frailty models seem to be adequate for modelling current status data. Nevertheless, it is not easy to find the "best" model. There are many assumptions that must be made beginning with choosing the form of the baseline hazards or the distribution of the frailty. Instead of using gamma or inverse Gaussian distributions other distributions might be more adequate for some pairs of infections. For example, for sexual transmitted diseases discrete frailty distributions (Poisson distribution) could be used (Unkel et al., 2014). Other potential research designs might use correlated frailties instead of shared frailties (Unkel et al., 2014) or to analyse more than two infections together.

Data of many other infectious diseases are available in the PIENTER-2 dataset. Additional analyses of further pairs of infections should be performed to study the heterogeneity and to find suitable frailty models.

References

Agresti, A. (2002). *Categorical data analysis*. 2nd edition. John Wiley & Sons.

Anderson, J. E. et al. (1992). "Time-dependent association measures for bivariate survival distributions". In: *Journal of the American Statistical Association* 87.419, pp. 641–650.

Anderson, R. M., Anderson, B., and May, R. M. (1991). *Infectious diseases of humans: dynamics and control*. Oxford university press.

Bundesgesundheitsministerium (2019). *Gesundheitsgefahren - Infektionskrankheiten*. URL: https://www.bundesgesundheitsministerium.de/themen/praevention/gesundheitsgefahren/infektionskrankheiten.html (visited on 08/15/2019).

Clayton, D. G. (1978). "A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence". In: *Biometrika* 65.1, pp. 141–151.

Duchateau, L. and Janssen, P. (2007). *The frailty model*. Springer Science & Business Media.

Edmunds, W. J. et al. (2000). "The pre-vaccination epidemiology of measles, mumps and rubella in Europe: implications for modelling studies". In: *Epidemiology & Infection* 125.3, pp. 635–650.

Farrington, C. P. et al. (2013). "Correlated infections: quantifying individual heterogeneity in the spread of infectious diseases". In: *American journal of epidemiology* 177.5, pp. 474–486.

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". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 50.3, pp. 251–292.

Farrington, C. P., Unkel, S., and Anaya-Izquierdo, K. (2012). "The relative frailty variance and shared frailty models". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74.4, pp. 673–696.

Funk, S. (2018). *socialmixr: Social Mixing Matrices for Infectious Disease Modelling*. R package version 0.1.3. URL: https://CRAN.R-project.org/package=socialmixr.

Genz, A. et al. (2019). *mvtnorm: Multivariate Normal and t Distributions*. R package version 1.0-11. URL: https://CRAN.R-project.org/package=mvtnorm.

Goeyvaerts, N. et al. (2010). "Estimating infectious disease parameters from data on social contacts and serological status". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 59.2, pp. 255–277.

Guerra, F. M. et al. (2017). "The basic reproduction number (R0) of measles: a systematic review". In: *The Lancet Infectious Diseases* 17.12, e420–e428.

Hahné, S. J. M. et al. (2012). "Prevalence of hepatitis B virus infection in The Netherlands in 1996 and 2007". In: *Epidemiology & Infection* 140.8, pp. 1469–1480.

Hens, N. et al. (2012). *Modeling infectious disease parameters based on serological and social contact data: a modern statistical perspective*. Vol. 63. Springer Science & Business Media.

Heymann, D. L. et al. (2008). *Control of communicable diseases manual.* Ed. 19. American Public Health Association.

Hofhuis, A. et al. (2011). "Decreased prevalence and age-specific risk factors for Toxoplasma gondii IgG antibodies in The Netherlands between 1995/1996 and 2006/2007". In: *Epidemiology & Infection* 139.4, pp. 530–538.

Hougaard, P. (2012). *Analysis of multivariate survival data*. Springer Science & Business Media.

Korndewal, M. J. et al. (2015). "Cytomegalovirus infection in the Netherlands: seroprevalence, risk factors, and implications". In: *Journal of Clinical Virology* 63, pp. 53–58.

Lemon, J. (2006). "Plotrix: a package in the red light district of R". In: *R-News* 6.4, pp. 8–12.

Mollema, L. et al. (2010). *PIENTER 2-project: second research project on the protection against infectious diseases offered by the national immunization programme in the Netherlands*.

Mossong, J. et al. (2008). "Social contacts and mixing patterns relevant to the spread of infectious diseases". In: *PLoS medicine* 5.3, e74.

Oakes, D. (1989). "Bivariate survival models induced by frailties". In: *Journal of the American Statistical Association* 84.406, pp. 487–493.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org/`.

Scherpenisse, M. et al. (2012). "Seroprevalence of seven high-risk HPV types in The Netherlands". In: *Vaccine* 30.47, pp. 6686–6693.

Smits, G. et al. (2013). "Seroprevalence of mumps in The Netherlands: dynamics over a decade with high vaccination coverage and recent outbreaks". In: *PLoS One* 8.3, e58234.

UNdata (2019). *UNdata - A world of information*. URL: `http://data.un.org/Data.aspx?d=POP&f=tableCode%3A22` (visited on 08/15/2019).

Unkel, S. et al. (2014). "Time varying frailty models and the estimation of heterogeneities in transmission of infectious diseases". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 63.1, pp. 141–158.

Unkel, S. and Farrington, C. P. (2012). "A new measure of time-varying association for shared frailty models with bivariate current status data". In: *Biostatistics* 13.4, pp. 665–679.

Van der Klis, F. R. et al. (2009). "Second national serum bank for population-based seroprevalence studies in the Netherlands". In: *Neth J Med* 67.7, pp. 301–308.

Verhoef, L. et al. (2011). "Changing risk profile of hepatitis A in The Netherlands: a comparison of seroprevalence in 1995–1996 and 2006–2007". In: *Epidemiology & Infection* 139.8, pp. 1172–1180.

Waaijenborg, S. et al. (2013). "Waning of maternal antibodies against measles, mumps, rubella, and varicella in communities with contrasting vaccination coverage". In: *The Journal of infectious diseases* 208.1, pp. 10–16.

Wallinga, J., Teunis, P., and Kretzschmar, M. (2006). "Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents". In: *American journal of epidemiology* 164.10, pp. 936–944.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN: 978-3-319-24277-4. URL: http://ggplot2.org.

WorldHealthOrganisation (2019). *Ebola virus disease*. URL: https://www.who.int/health-topics/ebola/#tab=overview (visited on 08/15/2019).

Appendices

Shared gamma frailty model

The three probabilities in Equation 4.8 - 4.10 are

$$\pi_{00}(x) = \Big(1 + \frac{(\Lambda_{01}(x) + \Lambda_{02}(x))}{\theta}\Big)^{-\theta},$$

$$\pi_{01}(x) = \Big(\Big(1 + \frac{\Lambda_{01}(x)}{\theta}\Big)^{-\theta}\Big) - \pi_{00}(x),$$

$$\pi_{10}(x) = \Big(\Big(1 + \frac{\Lambda_{02}(x)}{\theta}\Big)^{-\theta}\Big) - \pi_{00}(x).$$

Shared inverse Gaussian frailty model

The three probabilities in Equation 4.8 - 4.10 are

$$\pi_{00}(x) = \exp\Big(-\theta\Big(\sqrt{1 + 2\frac{(\Lambda_{01}(x) + \Lambda_{02}(x))}{\theta}} - 1\Big)\Big),$$

$$\pi_{01}(x) = \exp\Big(-\theta\Big(\sqrt{1 + 2\frac{\Lambda_{01}(x)}{\theta}} - 1\Big)\Big) - \pi_{00}(x),$$

$$\pi_{10}(x) = \exp\Big(-\theta\Big(\sqrt{1 + 2\frac{\Lambda_{02}(x)}{\theta}} - 1\Big)\Big) - \pi_{00}(x).$$

Two-component multiplicative double gamma frailty model

The three probabilities in Equation 4.8 - 4.10 are

$$\pi_{00}(x) = \exp\{H_1^1(x)+H_2^1(x)+H_1^2(x)+H_2^2(x)-H_{12}^1(x)-H_{12}^2(x)-\Lambda_{01}(x)-\Lambda_{02}(x)\}\times K_{12}(x),$$

$$\pi_{01}(x) = \exp\{H_1^1(x) + H_2^1(x) - H_{12}^1(x) - \Lambda_{01}(x)\} \times K_1(x) - \pi_{00}(x),$$

$$\pi_{10}(x) = \exp\{H_1^2(x) + H_2^2(x) - H_{12}^2 - \Lambda_{02}(x)\} \times K_2(x) - \pi_{00}(x),$$

where $H_{ij}^k(x) = \int_0^x h_i(t)h_j(t)\lambda_{0k}(t)dt$ $(i, j, k = 1, 2)$ and

$$\begin{aligned} K_{12}(x) = \mathrm{E}(\exp\{&-U_1[H_1^1(x) + H_1^2(x) - H_{12}^1(x) - H_{12}^2(x)] \\ &-U_2[H_2^1(x) + H_2^2(x) - H_{12}^1(x) - H_{12}^2(x)] \\ &-U_1U_2[H_{12}^1(x) + H_{12}^2(x)]\}), \end{aligned}$$

$$K_1(x) = \mathrm{E}(\exp\{-U_1[H_1^1(x) - H_{12}^1(x)] - U_2[H_2^1(x) - H_{12}^1(x)] - U_1U_2H_{12}^1(x)\}),$$

$$K_2(x) = \mathrm{E}(\exp\{-U_1[H_1^2(x) - H_{12}^2(x)] - U_2[H_2^2(x) - H_{12}^2(x)] - U_1U_2H_{12}^2(x)\}).$$

Bivariate analysis

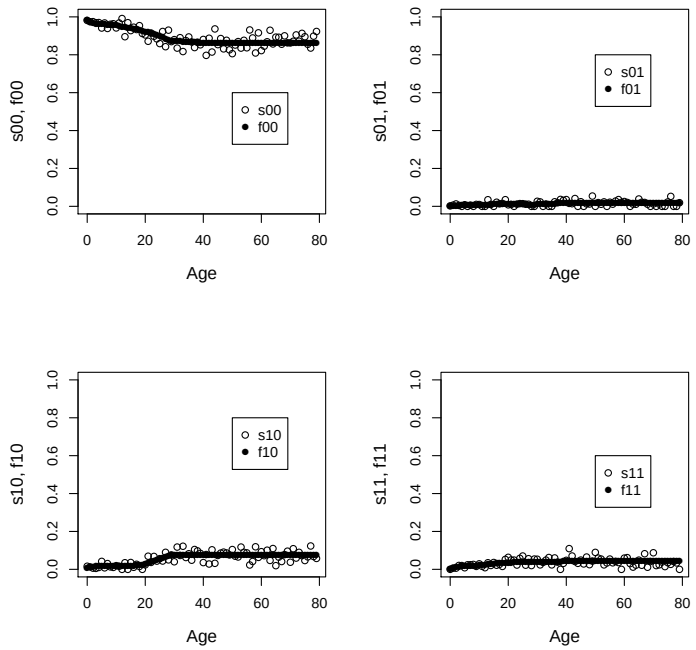


Figure 37: Observed $(s_{00}, s_{01}, s_{10}, s_{11})$ and fitted $(f_{00}, f_{01}, f_{10}, f_{11})$ proportions of the 1-component frailty model for HPV 16 and 18, where s_{00} denotes the proportion of individuals who are susceptible to both infections.

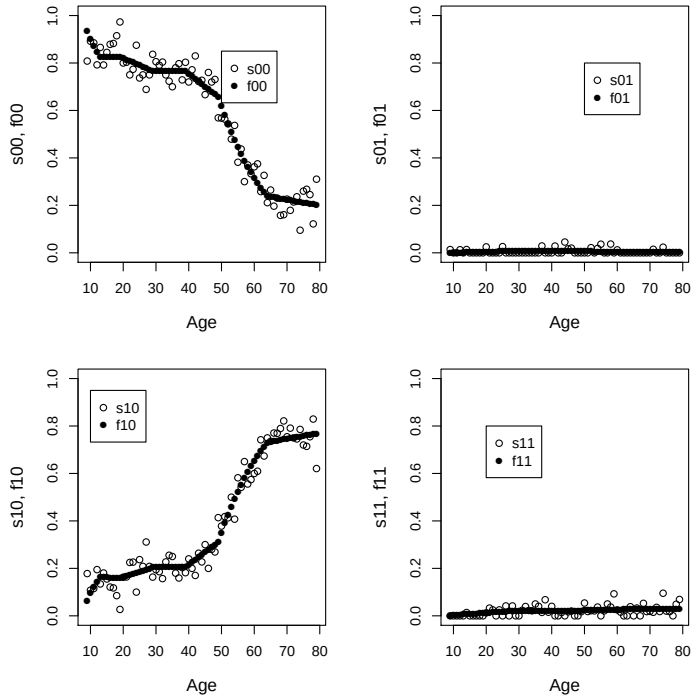


Figure 38: Observed $(s_{00}, s_{01}, s_{10}, s_{11})$ and fitted $(f_{00}, f_{01}, f_{10}, f_{11})$ proportions of the 1-component frailty model for HAV and HBV.

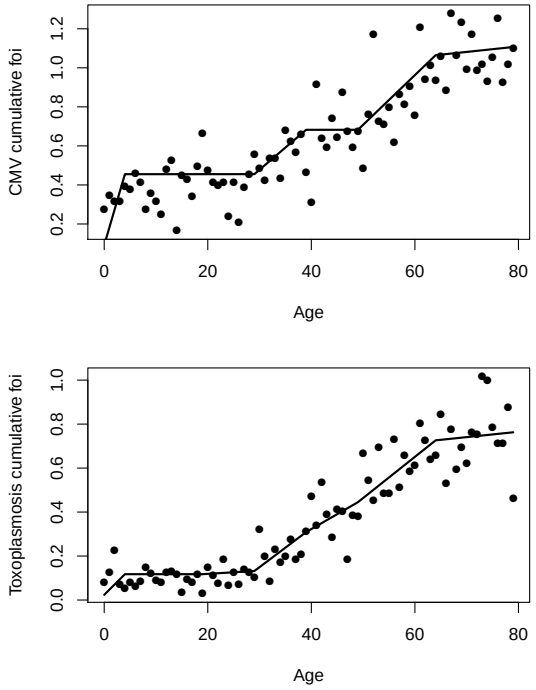


Figure 39: Observed and fitted (one-component Dirichlet-multinomial) cumulative force of infection for Toxoplasmosis and CMV.

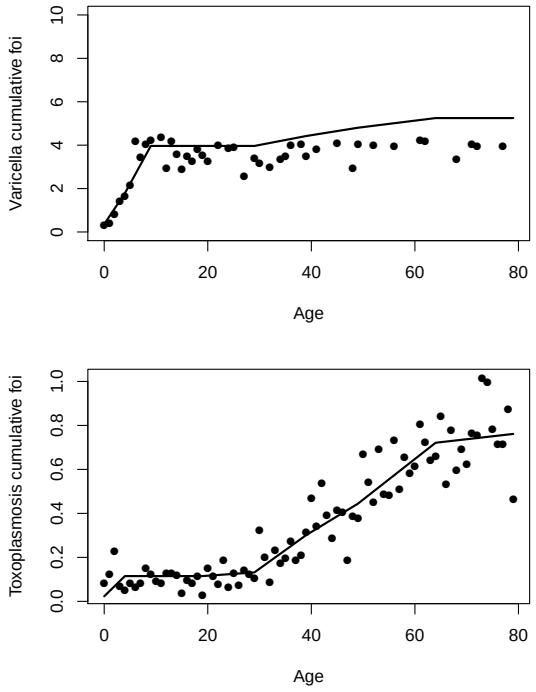


Figure 40: Observed and fitted (one-component Dirichlet-multinomial) cumulative force of infection for Toxoplasmosis and VZV.

Danksagung

Ganz besonderer Dank geht an meinen Betreuer Herrn PD Dr. Steffen Unkel. Einmal für die Vergabe des Themas, die sehr gute Betreuung und die ständige Erreichbarkeit und auch dafür, dass er mir indirekt zu meiner zukünftigen Stelle verholfen hat, indem er mich für eine studentische Hilfskraftstelle empfohlen hat.
Außerdem möchte ich meiner Familie und vor allen Dingen meiner Mama danken, die mich immer meine Entscheidungen frei treffen lassen hat und mich jederzeit darin bestärkt hat. Ein weiterer großer Dank geht an meinen Freund, der schon mein ganzes Studium an meiner Seite steht, mich motiviert und unterstützt, gerade in den letzten Monaten war dies eine Herausforderung.

Selbstständigkeitserklärung

Ich versichere, dass ich die Arbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen oder anderen Quellen entnommen sind, sind als solche kenntlich gemacht. Die schriftliche und elektronische Form der Arbeit stimmen überein. Ich stimme der Überprüfung der Arbeit durch eine Plagiatssoftware zu.

Ort, Datum: _____

Unterschrift: _____

(Anna-Maria Kloidt)