2013•2014
FACULTY OF SCIENCES
Master of Statistics

Master's thesis

A new model for multivariate current status data

Promotor :
Prof. dr. Marc AERTS
Prof. dr. Niel HENS

Adelino Martins

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics

Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

universiteit hasselt
KNOWLEDGE IN ACTION

Universiteit Hasselt | Campus Hasselt | Martelarenlaan 42 | BE-3500 Hasselt
Universiteit Hasselt | Campus Diepenbeek | Agoralaan Gebouw D | BE-3590 Diepenbeek

universiteit hasselt | Maastricht University

2013•2014
FACULTY OF SCIENCES
Master of Statistics

Master's thesis

A new model for multivariate current status data

Promotor :
Prof. dr. Marc AERTS
Prof. dr. Niel HENS

Adelino Martins

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics

universiteit hasselt | Maastricht University

Certification

This is to certify that this report was written by **Adelino Martins** under our supervision.

Signature:_____ Date:_____
Prof. dr. Marc Aerts **Internal Supervisor**

Signature:_____ Date:_____
Prof. dr. Niel Hens **Internal Supervisor**

Thesis submitted in partial fulfillment of the requirements for the degree of Master of Statistics: Biostatistics.

Signature:_____ Date:_____
Adelino Martins **Student**

Acknowledgments

Thanks to prof. dr. Marc Aerts and prof. dr. Niel Hens (Hasselt University, Center for Statistics) who provided the topic and all necessary material needed throughout the entire master thesis project and their valuable guidance, my classmates and all my lecturers (and other staff members) in the Master of Statistics 2012-2014 (Hasselt University) for providing an intellectually challenging environment during the program. A word of gratitude to all those involved in the Desafio Program (Eduardo Mondlane University), especially to the leader of the project Biostatistics and Modeling, Rafica Abdulrazac, and Flemish Inter-university Council (VLIR UOS) for the scholarship which has enabled me to enroll in this valuable Master program.

Adelino Martins

Hasselt University

September, 2014

Contents

List of Figures

List of Tables

Abstract

Individual heterogeneity in the acquisition of infectious diseases is recognized as a key concept, which allows improved estimation of important epidemiological parameters. Frailty models allow to represent such heterogeneity. Coull *et al.* (2006) introduced a computational tractable multivariate random effects model for clustered binary data. The objective of this report was to apply and modify the proposed model, and compare to the shared and correlated gamma frailty models in the context of the analysis of multivariate current status data. The models were applied to the bivariate current status data on Varicella-Zoster Virus and Parvovirus B19 using different baseline hazard functions for the force of infection. The findings revealed that the proposed model which is called in this report as new correlated gamma frailty model is closely related to existing frailty models. The main difference is the way the multivariate gamma is introduced in the model, and the indirect way to specify the baseline hazard function. In terms of construction, a frailty model is typically formulated based on specification of the proportional hazard function, whereas the new correlated gamma frailty model is built using a classical generalized linear mixed model for clustered binary data. Furthermore, in the new model the variances of the frailties are assumed to be identical, whereas in case of the frailty model, the variances can be different or identical and the correlation is constraint by the ratio of the variances.

Keywords: Multivariate current status data; multivariate serological data; new correlated gamma frailty model; shared and correlated gamma frailty models; individual heterogeneity.

1 Introduction

Current status data arise in studies where the target measurement is the time of occurrence of some event, but observations are limited to indicators of whether or not the event has occurred at the time the sample is collected - only the current status of each individual with respect to event occurrence is observed (Shiboski 1998, Hens *et al.* 2009, and Chen *et al.* 2009).
The term current status data arose in the field of demography (Diamond *et al.* 1986, cited by Shiboski 1998), where interest often focuses on the age that some landmark event occurs, such as weaning. Because such events are rarely observed prospectively, studies must either rely on retrospective information, which may be unreliable, or use the current status information collected when subjects enroll (Shiboski 1998).
Current status data also arise naturally in epidemiological studies of infectious diseases (Shiboski 1998), where serological samples are used to investigate the epidemiology of infectious diseases and to estimate important parameters such as the force of infection (Farrington and Kanaan 2001 and Hens *et al.* 2012).
In this perspective, serological samples taken at a certain time point provide information about whether or not the individual has been infected before that time point (Hens *et al.* 2012). Moreover, for feasibility and economical reasons, serum samples are often tested for more than one antigen (Farrington and Kanaan 2001 and Hens *et al.* 2012). In this way, the (past) disease status of individuals on multiple diseases is known, and allows studying the association in acquisition between several infections. Under the assumption of lifelong immunity and that the epidemic is in a steady state (i.e., at equlibrium), important epidemiological parameters can be estimated from such data (Hens *et al.* 2012).
In the estimation of epidemiological parameters, Hens *et al.* (2012) pointed out that individuals are dissimilar in the way they acquire infectious. Some individuals are more susceptible than others and will experience infections earlier. Thus, individual heterogeneity in the acquisition of infectious diseases is recognized as a key concept on the estimation of such epidemiological parameters (Hens *et al.* 2012). In the context of current status data on infectious diseases, such individual heterogeneity is expressed in terms of the age-dependent force of infection (FOI) through the multiplicative frailty models (Hens *et al.* 2012 and Enki *et al.* 2014).
Hens *et al.* (2009, 2012) considered correlated and shared gamma frailty models in the context of bivariate current status data to represent such heterogeneity. Coull *et al.* (2006) introduced a computationally tractable multivariate random effects model for clustered binary data that is useful when interest focuses on the association structure among clustered observations.
In this report, the objectives were to apply and modify the proposed model and compare it to the gamma frailty models for the estimation of individual heterogeneity in the acquisition of infections.
The remainder of the report is organized as follows: Section 2.1 describes the data. In Section 2.2 some relevant concepts in modeling infectious diseases are introduced. Section 2.3.1 sets up the shared gamma frailty model. In Section 2.3.2, the correlated gamma frailty model is introduced. The parametric distributions for the baseline hazard function are given in Section 2.3.3. The new correlated gamma frailty model and its extension using fractional polynomials are given in Section 2.3.4 and Section 2.3.5, respectively.

Model fitting and comparison between both correlated gamma frailty models are given in Section 2.3.6 and Section 2.3.7, respectively. In Section 3, the methods developed in this report are applied to paired serological data on Varicella-Zoster Virus and Parvovirus (B19). Discussion and conclusion are given in Section 4 and Section 5, respectively. Computation in this report was carried out using software package R version 3.1 (R Development Core Team 2014). All computer code used is available upon request. The forms of the data structure to fit the models are presented in the Appendix.

2 Methodology

2.1 Data

The data used in this report consist of serum samples tested for two infections, Varicella-Zoster Virus (VZV) and Parvovirus (B19). Data from these infections were collected in a survey in Belgium in a period from November 2001 until March 2003. The Varicella-Zoster Virus, also known as human herpes virus 3 (HHV-3), is one of eight herpes viruses known to affect humans and other vertebrates (Thiry *et al.* 2002; Hens *et al.* 2008). Primary Varicella Zoster-Virus infection results in chickenpox (varicella), has a two-week incubation period and is highly contagious by air droplets starting two days before symptoms appear. Infectiousness is known to last up to ten days. Therefore, chickenpox spreads quickly through close social contacts (Hens *et al.* 2008).
Parvovirus B19 was the first human Parvovirus to be discovered, in 1975. In clinical terms Parvovirus B19 is best known for causing a childhood exanthem called fifth disease or erythema infectiosum. The virus is primarily spread by infected respiratory droplets. Parvovirus B19 symptoms begin some six days after exposure and last for about a week. After being infected, patients are infectious for five to seven days and usually develop the illness after an incubation period of four to fourteen days (Broliden *et al.* 2006 and Hens *et al.* 2008).
In total, 3080 sera were tested for Varicella Zoster-Virus and 2657 sera were tested for Parvovirus B19 from which 2382 sera were tested for both Varicella Zoster-Virus and Parvovirus B19. As mentioned by Hens *et al.* (2010) and Hens *et al.* (2012), when modeling the force of infection a first issue is the definition of antibody activity levels as they truly reflect natural infection rather than maternal antibody or vaccine induced activity. An ad hoc approach commonly applied is to delete the observations which are believed to refer to maternal antibodies (in this case corresponding to the first 6 months). Therefore, without loss of generality samples from children under 6 months which corresponds to less than 0.5 years old were omitted. Since in this report attention was restricted to the association between Varicella Zoster-Virus and Parvovirus B19 infections, individuals having serological results for both infections were considered. Hence, from 3374 blood samples 2377 complete cases together with age were used.
Figure 1 shows the observed proportion of sera that tested positive for both Varicella Zoster-Virus and Parvovirus B19 (top left panel), that tested positive on Parvovirus B19 only (top right panel), that tested positive on Varicella Zoster-Virus only (lower left panel), and that tested negative on both viruses (lower right panel). From the figure, it seems that the prevalence of being co-infected with both viruses increases quickly with age and decreases around the age 21 to 30 years old and start to increase again, while the proportion of still being susceptible decreases (almost zero at age 40 years old). It can also be observed that the prevalence of Varicella Zoster-Virus infection is higher than the prevalence of Parvovirus B19 infection.

2.2 Relevant Concepts in Modeling Infectious Diseases

Before turning to the statistical models which are used to estimate the degree of the individual heterogeneity using current status data, some relevant concepts in modeling

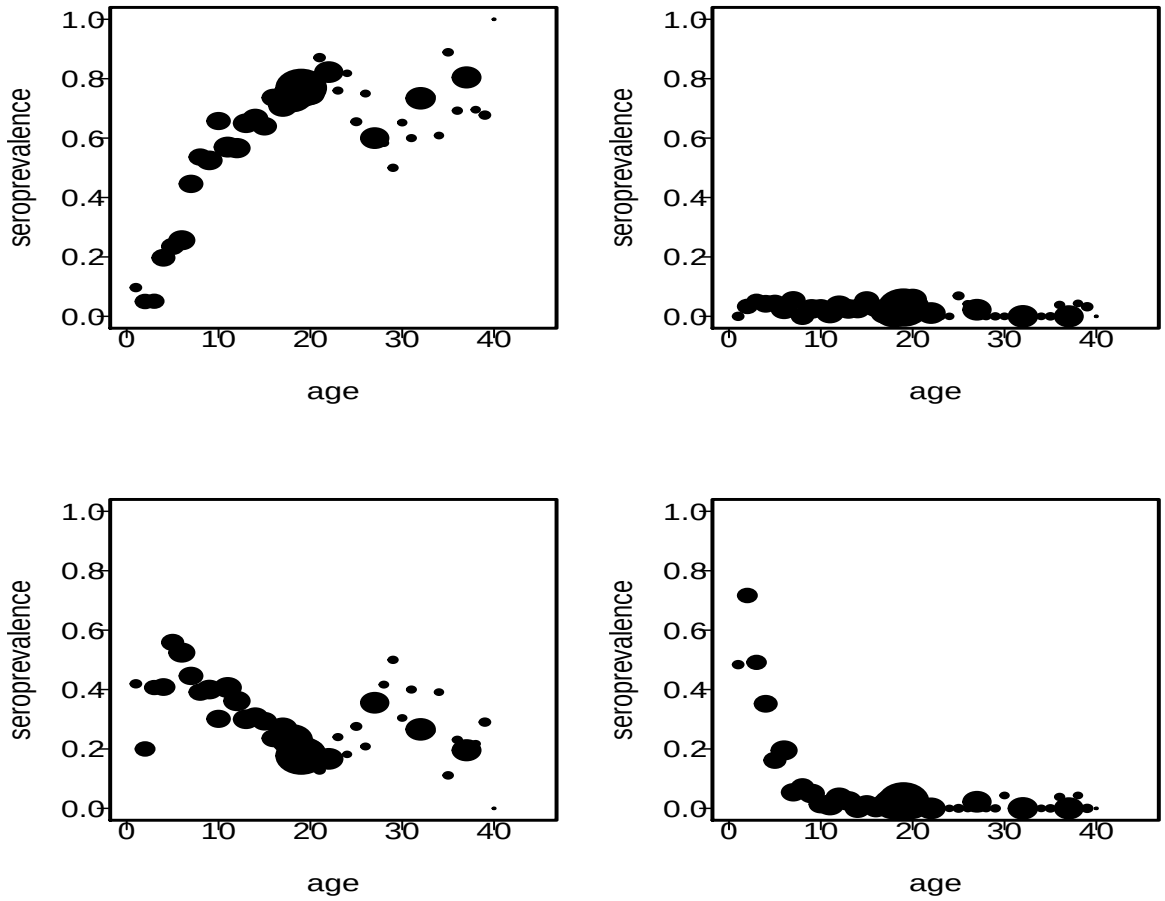

Figure 1: Proportion of samples that tested positive on both Varicella-Zoster Virus and Parvovirus B19 (top left panel), that tested positive on Parvovirus B19 only (top right panel), that tested positive on Varicella-Zoster Virus only (lower left panel), and that tested negative on both viruses with dots proportional to sample size.

infectious diseases are introduced.

There are many drivers behind the spread of diseases (Hannon and Ruth, 2009). The concepts of infectious diseases in this report are referred (Hens *et al.* 2012) to those that are caused by pathogens, which are transmissible between hosts, either directly or indirectly. Suppose that a population P is divide into two groups. One part of the population Z is affected by something and the other part A is not affected. In the context of infectious diseases the population is affected by an infection and therefore Z is the part of the population infected individuals (the infected class) while A is the susceptible part of the population (the susceptible class). In the above definition, each individual in the population belongs only to one part of the population and at each time the individuals move from one class to the others, that means the susceptible class becomes infected and the infected individuals recovers and becomes susceptible again (Keeling and Rohani 2008 and Hens *et al.* 2012). In this process of moving from one class to the other, Ross (1916) assumed that there are three different processes that act simultaneously on the population and proposed a model which represents the population dynamics using a set

of three differential equations.
The main problem of the model proposed by this author is that some of the parameters of the model are unknown and the solution proposed by Ross was to iterate between two modeling frameworks namely, *a posteriori* and *a priori* methods (Hens *et al.* 2012). The former refers to statistical models and the latter to mathematical models. An example of a mathematical model is the Susceptible-Infected-Recovered (SIR) model. This model, is one of the basic compartmental models in infectious diseases epidemiology and is widely used and well suited to model many viral infectious in childhood (Hens *et al.* 2012).
One of the assumption of this SIR model which is the main motivating of the statistical models for this report, is the time homogeneous representation of the model (Hens *et al.* 2012). Under this assumption, the proportion of susceptible individuals at age a is given by

$$s(a) = \frac{S(a)}{N(a)} = \exp\left(-\lambda a\right), \tag{1}$$

where $S(a)$ is the number of individuals in the cohort in the susceptible class at age a, $N(a)$ denotes the number of individuals in the cohort who are still alive at age a, and λ represents the rate at which individuals are infected. As mentioned by Hens *et al.* (2012), when λ depends on age, expression (1) becomes

$$s(a) = \exp\left(-\int_o^a \lambda(u)du\right). \tag{2}$$

From equation (2) it can be clearly observed that there is a relation between the above formulation and survival analysis. In this particular formulation, the time to event can be thought of as the age at infection, $s(a)$ being the survival function (which represents acquiring infection), $\lambda(a)$ denoting the age-specific hazard of infection and $\int_0^a \lambda(u)du$ indicates the cumulative hazard of infection. More detail on the SIR model can be found in Keeling and Rohani (2008) and Hens *et al.* (2012).
Although the concepts of serological and current status data are discussed in more detail in the next section, the connection between cross-sectional serological samples and mathematical model are presented. As defined by Hens *et al* (2012), let Y_i, $i = 1, 2, ..., N$, be an indicator variable representing the disease status for the ith individual in the sample.

$$Y_i = \begin{cases} 1 & \text{when seropositive (previously infected),} \\ 0 & \text{when seronegative (susceptible to infection).} \end{cases}$$

Let $P(Y_i = 1|a_i) = \pi(a_i) = E(Y_i|a_i)$ be the probability to be infected before age a_i. It follows that

$$Y_i \sim \text{Bernoulli}(\pi(a_i)).$$

Observe that $\pi(a_i) = 1 - s(a_i)$ and using expression (1), which relies on the validity of the SIR assumption for the specific infection under consideration, it follows that

$$\pi(a_i) = 1 - \exp\left(-\lambda a_i\right).$$

Hence, in order to estimate the unknown parameter λ one can define a generalized linear model (GLM) for the binary response with complementary log-log link function

$$g(\pi(a_i)) = \log(-\log(1 - \pi(a_i))) = \alpha + \log(a_i),$$

where $\alpha = \log(\lambda)$.

2.3 Statistical Methods

The connection between the SIR model relying on the time homogeneous assumption, serological data, and the estimation of the force of infection has been discussed in the previous section. The concepts of current status, serological data and their relation are described in this section. Current status data arise in studies where the target measurement is the time of occurrence of some event but observations are limited to indicator of whether or not the event has occurred at the time the sample is collected. Only the current status of each individual with respect to event occurrence is observed (Shiboski 1998; Wang and Ding 2000; Balakrishnan and Rao 2004; Sun 2006).

Current status data is also referred to as case I interval-censored data. A more convenient representation of case I interval-censored data is $\{C, \delta = I(T \leq C)\}$, where C denotes the monitoring time, I is the indicator function and δ is the indicator whether the event already occurred before the monitoring time or not. With this definition, clearly it can be noted that this type of data one can only know that the event has occurred between two time points, but not the exact time point (Hens *et al.* 2009).

The bivariate version is defined as follows (Wang and Ding 2000; Balakrishnan and Rao 2004; Unkel and Farrington 2012): consider a study in which interest focuses on the bivariate distribution F of two random survival variables (T_1, T_2) neither of which can be directly measured. Rather, for each individual, we observe at a random monitoring time, C, whether T_j exceeds C or not for each $j = 1, 2$. That is, on each subject, we observe

$$(Y_1 \equiv I(T_1 \leq C), Y_2 \equiv I(T_2 \leq C)),$$

where C is assumed independent of (T_1, T_2).

In this report, current status data that can be used to estimate the degree of the individual heterogeneity in acquisition of infections are considered. Such bivariate data occur naturally in infectious diseases epidemiology, for instance, when T_1 and T_2 represent the ages at the onset of infection by two distinct infectious agents whose onset can only be determined to lie below or above C. In this context, the time scale is age and the defining time point from which times are measured is birth (Unkel and Farrington 2012).

As mentioned by Farrington and Whitaker 2005 cited by Unkel and Farrington 2012, with the above definition the association between the ages T_1 and T_2 may carry information about relevant infection processes and can be examined using paired serological survey data on two infections. In this perspective, serological data are obtained by testing blood serum residues for the presence of antibodies to one or more infections. A positive (negative) result indicates prior infection (susceptibility to infection), giving rise to current status data.

In this report, the data used are bivariate corresponding to the Varicella-Zoster Virus and Parvovirus B19-virus. More precisely, the data consist of observations (y_1, y_2, a), where y_1 is the current status of the Varicella-Zoster Virus at the examination time, y_2 is the current status of Parvovirus B19, and a represents the age at the examination time. Given bivariate binary data on two infectious diseases $(y_1 = VZV, y_2 = B19)$ from a sample of individuals together with their age a, denote the joint probability $\pi_{j1,j2} = P(y_1 = j_1, y_2 = j_2)$, where the index $k = 1, 2$ corresponds to disease 1 and 2, respectively, and $j_k = 1(0)$ indicating past or current infection (susceptibility) for disease $k = 1, 2$. Modeling such multivariate categorical data can be done using conditional or marginal models (Liang *et al.* 1992). In this report attention was restricted in

modeling such data using different conditional models in order to estimate the individual heterogeneity in acquisition of both infections. In particular, the frailty models such as shared and correlated gamma models and a new model based on the clustered binary data model of Coull *et al.* (2006) were considered. This new model was applied, modified and compared to the existent frailty models.

2.3.1 Shared Gamma Frailty Model

A first conditional model that can be considered to estimate the degree of the individual heterogeneity in acquisition of infections is the shared gamma frailty model. Vaupel *et al.* (1979) and Aalen (1988), pointed out that individuals are dissimilar in the way they acquire infections. Some individuals are more susceptible than others and will experience infection earlier. This can be expressed in terms of the age-dependent force of infection (FOI) by $\lambda(a, Z)$ where Z can be an individual-specific covariate or, alternatively, a random variable. Z is often referred to as frailty and expresses to what extent an individual has a lower or higher risk of infection. Thus, the model assumes that every individual is infected differently from the others, that is, the force of infection, $\lambda(a_i, z_i)$, depends on a individual-specific random variable z_i or "frailties" (Del Fava *et al.,* 2011).
In this particular context, the FOI is defined as being the rate at which susceptible individuals acquire infection, and in terminology of survival analysis, the FOI is nothing else than the hazard function (Farrington and Kanaan 2001; Sutton *et al.* 2006; Hens *et al.* 2010; Hens *et al.* 2012). Given the assumption of proportional hazards, the FOI for disease $j = 1, 2$, can be written as

$$\lambda_j(a_i, z_i) = z_i \lambda_{0j}(a_i),$$

where λ_{0j} is the baseline hazard for an individual of age a at the time of test. Therefore, the susceptible proportion for the infection j is given by

$$S_j(a_i|Z_i) = \exp\left(-z_i \int_0^{a_i} \lambda_{0j}(t)dt\right), \quad j = 1, 2, \tag{3}$$

where the unconditional susceptible proportion can be obtained by integrating out the random frailty z_i using the Laplace transform:

$$S_j(a_i) = E(S_j(a_i|Z_i)) = L_j\left(\int_0^{a_i} \lambda_{0j}(t)dt\right), j = 1, 2. \tag{4}$$

The susceptible proportion refers to the proportion of the population that is susceptible (no pathogen is present) to the disease (Keeling and Rohani 2008; Hannon and Ruth 2009). Hence, by (3), assuming a common gamma frailty distribution, and using the Laplace transformation, it follows that the unconditional bivariate susceptible proportion is given by (Hens *et al.* 2012)

$$\pi_{00}(a_i) = [S_1(a_i)^{-\frac{1}{\theta}} + S_2(a_i)^{-\frac{1}{\theta}} - 1]^{-\theta}, \tag{5}$$

where θ denotes the shape parameter of the gamma frailty distribution $Z \sim \Gamma(\theta, \frac{1}{\theta})$. Since $E(Z) = 1$ and variance $Var(Z) = \frac{1}{\theta}$, θ is the parameter describing the heterogeneity in

acquisition of infection. The larger θ, the smaller the heterogeneity and thus the more people are alike in the way they acquire the infection.
Reparameterizing the joint probability in terms of the cumulative FOI,

$$\Lambda_j(a_i) = \int_0^{a_i} \lambda_{0j}(t)dt,$$

gives the following set of equations for the four joint probabilities (Hens *et al.* 2012):

$$\begin{cases} \pi_{00}(a_i) = \left[\exp(\frac{\Lambda_1(a_i)}{\theta}) + \exp(\frac{\Lambda_2(a_i)}{\theta}) - 1\right]^{-\theta}, \\ \pi_{10}(a_i) = \exp(-\Lambda_2(a_i)) - \pi_{00}(a_i), \\ \pi_{01}(a_i) = \exp(-\Lambda_1(a_i)) - \pi_{00}(a_i), \\ \pi_{11}(a_i) = 1 - \pi_{10}(a_i) - \pi_{01}(a_i) - \pi_{00}(a_i), \end{cases} \tag{6}$$

where $\pi_{00}(a)$ is the probability that an individual of age a has been infected by neither viruses and $\pi_{10}(a)$ is the probability that an individual of age a has been infected by infection 1 (VZV) but not infection 2 (B19), $\pi_{01}(a)$ represents the probability that an individual of age a has been infected by infection 2 (B19) but not infection 1 (VZV), and $\pi_{11}(a)$ gives the probability that an individual of age a has been infected by both infections.

2.3.2 Correlated Gamma Frailty Model

The shared gamma frailty model presented in Section 2.3.1 is a special case of the correlated gamma frailty model with correlation between the frailties equal to one. A correlated frailty model can be considered as a mixed (random effects) model in survival analysis, with group and individual variation both included in the distribution of the frailty vector. Therefore, correlated frailty models contain association characteristics of frailty (correlation coefficients) among other parameters. In this model, the frailties of individuals in a cluster are correlated but not necessarily shared (Yashin *et al.* 1995; Duchateau and Janssen 2008; Wienke 2011; Hanagal 2011).
As defined by Hens *et al.* (2009) and Wienke (2011), let k_0, k_1, k_2 be nonnegative real-value numbers, set $\lambda_1 = k_0 + k_1$ and $\lambda_2 = k_0 + k_2$. Let Y_0, Y_1, Y_2 be independent, gamma-distributed random variables with parameters $Y_0 \sim \Gamma(k_0, \lambda_0)$, $Y_1 \sim \Gamma(k_1, \lambda_1)$, and $Y_2 \sim \Gamma(k_2, \lambda_2)$.
Consequently,

$$Z_1 = \frac{\lambda_0}{\lambda_1} Y_0 + Y_1 \sim \Gamma(k_0 + k_1, \lambda_1), \tag{7}$$

$$Z_2 = \frac{\lambda_0}{\lambda_2} Y_0 + Y_2 \sim \Gamma(k_0 + k_2, \lambda_2), \tag{8}$$

and $E(Z_1) = E(Z_2) = 1$, $Var(Z_1) = \frac{1}{\lambda_1} := \sigma_1^2$, $Var(Z_2) = \frac{1}{\lambda_2} := \sigma_2^2$. Then the following relation holds

$$Cov(Z_1, Z_2) = \frac{k_0}{(k_0 + k_1)(k_0 + k_2)}. \tag{9}$$

This leads to the correlation

$$\rho = \frac{k_0}{\sqrt{(k_0+k_1)(k_0+k_2)}}. \tag{10}$$

The explicit expression for the survival function in terms of σ_1, σ_2 and ρ applying the Laplace transform of gamma-distribution random variables is given by (Hens *et al.* 2009 and Wienke 2011):

$$S(t_1,t_2) = \frac{[S_1(t_1)]^{1-\frac{\sigma_1}{\sigma_2}\rho}[S_2(t_2)]^{1-\frac{\sigma_2}{\sigma_1}\rho}}{([S_1(t_1)]^{-\sigma_1^2}+[S_2(t_2)]^{-\sigma_2^2}-1)^{\frac{\rho}{\sigma_1\sigma_2}}}, \tag{11}$$

where the univariate marginal survival functions are given by

$$S_j(t_j) = \left(1+\sigma_j^2\int_0^{t_j}\lambda_{0j}(s)ds\right)^{-\frac{1}{\sigma_j^2}}. \tag{12}$$

It is important to notice that with the above notation in this particular setting, the correlated gamma frailty model is defined without observed covariates. Similarly, the unconditional bivariate susceptible proportion is given by setting $\pi_{00}(a_i,a_i) = S(t_1,t_2)$ and the four joint probabilities are expressed as follows

$$\begin{cases} \pi_{00}(a_i) = \dfrac{[S_1(a_i)]^{1-\frac{\sigma_1}{\sigma_2}\rho}[S_2(a_i)]^{1-\frac{\sigma_2}{\sigma_1}\rho}}{([S_1(a_i)]^{-\sigma_1^2}+[S_2(a_i)]^{-\sigma_2^2}-1)^{\frac{\rho}{\sigma_1\sigma_2}}}, \\ \pi_{10}(a_i) = \left(1+\sigma_2^2\int_0^{a_i}\lambda_{02}(s)ds\right)^{-\frac{1}{\sigma_2^2}} - \pi_{00}(a_i), \\ \pi_{01}(a_i) = \left(1+\sigma_1^2\int_0^{a_i}\lambda_{01}(s)ds\right)^{-\frac{1}{\sigma_1^2}} - \pi_{00}(a_i), \\ \pi_{11}(a_i) = 1-\pi_{10}(a_i)-\pi_{01}(a_i)-\pi_{00}(a_i), \end{cases} \tag{13}$$

where $\lambda_{02}(s)$ is the baseline hazard function for infection 2 (B19), $\lambda_{01}(s)$ represents the baseline hazard function for infection 1 (VZV), and where $\pi_{00}(a_i)$, $\pi_{10}(a_i)$, $\pi_{01}(a_i)$ and $\pi_{11}(a_i)$ are defined as before.

2.3.3 Baseline Hazard Function for the Force of Infection

To fit the correlated gamma frailty models, it is necessary to assume a particular function for the baseline hazard (which represents the force of infection) (Del Fava *et al.* 2011). In this section, under the parametric approach, most commonly adopted distributional assumptions when dealing with parametric proportional hazard models in survival analysis are presented. The possible choice can be for instance: The Exponential, Weibull, Gompertz, Lognormal and Loglogistic distribution (Duchateau and Janssen 2008; Munda *et al.* 2012). Table 1 displays the hazard and cumulative hazard functions for each of these distributions.

It is important to notice that a particular choice of the distributional assumption leads to a specific link function and interpretation under a generalized linear model framework. For instance the Weibull baseline hazard leads to a proportional hazard interpretation with complementary log-log link fucntion. Another example is the Log-logistic which leads to a proportional odds interpretation and logit link function (Fahrmeir and Tutz

Table 1: Hazard function $\lambda_0(a)$, cumulative hazard function $\Lambda_0(a)$ of some distributions used for modeling survival time. With the lognormal, $\phi(.)$ and $\Phi(.)$ respectively denote the probability density and the cumulative distribution functions of a standard normal random variable and a represents age of the individual at time of test.

Distribution	$\lambda_0(a)$	$\Lambda_0(a) = \int_0^a \lambda_0(s)ds$	Parameter space
Exponential	α_j	$\alpha_j a$	$\alpha_j > 0$
Weibull	$\alpha_j \beta_j a^{\beta_j - 1}$	$\alpha_j a^{\beta_j}$	$\alpha_j > 0, \beta_j > 0$
Gompertz	$\alpha_j \exp(\gamma_j a)$	$\frac{\alpha_j}{\gamma_j}(\exp(\gamma_j a) - 1)$	$\alpha_j > 0, \gamma_j > 0$
Lognormal	$\frac{\phi(\frac{\log(a)-\mu}{\gamma})}{\gamma t(1 - \Phi(\frac{\log(a)-\mu}{\gamma}))}$	$-\log(1 - \Phi(\frac{\log(a)-\mu}{\gamma}))$	$\mu_j \epsilon \Re, \gamma_j > 0$
Loglogistic	$\frac{\exp(\phi_j) k_j a^{k_j - 1}}{1 + \exp(\phi_j) a^{k_j}}$	$\log(1 + \exp(\phi_j) a^{k_j})$	$k_j > 0, \phi_j \epsilon \Re$

Source: Duchateau and Janseen 2008; Munda, Rotolo and Legrand 2012.

2001; Allison 2010 and Hens *et al* 2012).
Another model for the force of infection can be assumed. Hens *et al.* (2012), applied a shared gamma frailty model to estimate individual heterogeneity in acquisition of rubella and mumps using a gamma function for the force of infection given by

$$\lambda(a) = \alpha a^{\beta} \exp(-\frac{a}{\gamma}), \tag{14}$$

where a is the age at the time of test, α, β and γ are positive parameters.
In this report, the shared and correlated gamma frailty model were fitted using the baseline hazard functions considered in Table 1. Based on the Akaike's information criteria (AIC), the best baseline hazard for the force of infection was selected. The choice of the parametric baseline hazard was motivated due to the fact that (Hens *et al.* 2009) the correlated gamma frailty model is not identifiable using a nonparametric baseline hazard function for the force of infection.

2.3.4 New Correlated Gamma Frailty Model

A third and new way to estimate the individual heterogeneity in acquisition of infections, is by considering the model proposed by Coull *et al.* (2006). These authors, introduced a computationally tractable multivariate random effects model for clustered binary data. This model can also be seen as a special case of the correlated gamma frailty model with gamma random variables having the same variances but not necessarily the correlation coefficient to be exactly equal to one. In this report, the model is applied and modified to the setting of modeling infectious diseases and related it to correlated gamma frailty models in the context of current status data. In what follows we describe the model and we show how it can be related to the correlated gamma frailty models.
The proposed model is built around multivariate gamma random effects, as defined by Henderson and Shimakura (2003). Let $W_1, ..., W_q$ be independent p-variate Gaussian random variables with standard marginals and common $(p \times p)$ correlation matrix C. Write $W_j = (W_{j1}, ..., W_{jp})^t$, $j = 1, ..., q$ and let $Z_k = \sum_{j=1}^{q} \frac{W_{jk}^2}{q}$, for $k = 1, ..., p$. Then

the vector $Z = (Z_1, ..., Z_p)^t$ is said to be multivariate gamma with marginal $Ga(\frac{q}{2}, \frac{q}{2})$ distributions and Laplace transform

$$E(\exp(-u^t Z)) = |I + \frac{2C\text{diag}(u)}{q}|^{\frac{-q}{2}}, \tag{15}$$

for $u \epsilon \Re^p$ and $C = (c_{jk})$. Bapat (1998) showed that the above Laplace transform defines a proper distribution more generally for non-integer values $q > 0$. With $\zeta = \frac{2}{q}$, we denote this multivariate distribution by

$$Z \sim MG(\zeta, C)$$

with the correlation matrix R describing the association among the gamma components, having elements

$$r_{jk} = c_{jk}^2.$$

Observe that a potential disadvantage of the model is the fact that the multivariate gamma distribution does not accommodate negative correlations (Coull *et al.* 2006). This is not perceived however as a severe limitation for many applications.
The model proposed by Coull *et al.* (2006) is given by

$$\log\{-\log[E(Y_{ij}|a_i, Z_{ij}))]\} = \log(Z_{ij}) + f_j(a_i), \tag{16}$$

for $i = 1, 2..., n$ and $j = 1, 2, ..., p$ and with $f_j(a_i)$ a linear, non-linear or nonparametric function to model the effect of age on the probability to be infected for infectious disease j. In this particular application the function $f_j(a_i)$ has to meet the additional constraints that it is a non-decreasing function of age a with $f_j(a_i) \to -\infty$ as $a_i \to 0$.
Conditional on the random effects (frailties) $\log(Z_{ij})$, the simplest form of the model that can be applied to investigate the effect of age on the probability of being infected by Varicella-Zoster Virus and Parvovirus B19 is given in the following way:

$$\begin{cases} \log\{-\log[E(Y_{i1}|age, Z_{i1})]\} = \log(Z_{i1}) + \beta_{01} + \beta_{11} * age, \\ \log\{-\log[E(Y_{i2}|age, Z_{i2})]\} = \log(Z_{i2}) + \beta_{02} + \beta_{12} * age. \end{cases} \tag{17}$$

For this particular application of the model, the parameters of primary interest are the variance of the gamma components (ζ) and their correlation (ρ).
Note that expression (16) leads to a probability curve given by (Agresti 2002):

$$E(Y_{ij}|a_i, Z_{ij}) = \exp(-\exp(\log(Z_{ij}) + f_j(a_i))).$$

To have the probability curve

$$E(Y_{ij}|a_i, Z_{ij}) = 1 - \exp(-\exp(\log(Z_{ij}) + f_j(a_i))),$$

we rewrite the proposed model in the following way:

$$\log\{-\log[1 - E(Y_{ij}|a_i, Z_{ij}))]\} = \log(Z_{ij}) + f_j(a_i). \tag{18}$$

With the expression (18), the relationship between the response probabilities and covariates can be related to the analysis of survival data (Shiboski 1998, Fahrmeir and Tutz 2001 and Balakrishnan and Rao 2004). In this particular situation for the complementary log-log link function in (18), implies that the susceptible proportion for the infection j can be rewritten as expression (3) with baseline hazard given by

$$\lambda_{0j}(a_i) = \exp(f_j(a_i))f_j'(a_i), \tag{19}$$

where, $f_j'(a_i)$ represents the derivative of $f_j(a_i)$ with respect to a_i, and where a_i is the age of the ith individual at the time of test . Note that by (3) and (19) the proposed model is closely related to the correlated gamma frailty model as defined in Section 2.3.2.

2.3.5 New Correlated Gamma Frailty Model using Fractional Polynomials

In Section 2.3.4 the new correlated gamma frailty model was described and showed how this model can be related to a correlated gamma frailty model in the context of current status data. It was noted that in its natural definition the model uses a linear predictor function. These linear predictor functions can have some disadvantages, for instance their limited flexibility of linear covariate function type. To enhance flexibility of this new correlated gamma frailty model, fractional polynomial functions were used.
Fractional polynomial functions, as a natural extension of polynomial functions (Hens *et al.* 2012) allow flexible parametric shapes by considering not only integer powers of the key continuous covariates, but also fractional powers (Royston and Sauerbrei 2008).
In the context of binary responses, a fractional polynomial of degree m for the linear predictors is defined as

$$\eta(a, \beta, \mathbf{P}, m) = \sum_{i=0}^{m} \beta_i H_i(a),$$

with $\beta = (\beta_0, \beta_1, \beta_2, ..., \beta_m)$ being a vector of coefficients and $\mathbf{P} = (p_0, p_1, p_2, ..., p_m)$ a vector of powers with $p_0 = 0$ and $H_0 \equiv 1$ representing the intercept. The powers $p_1 \leq p_2 \leq ... \leq p_m$ can be positive or negative integers or fractional powers. $H_i(a)$ is a transformation on a continuous variable a defined as

$$H_i(a) = \begin{cases} a^{p_i} & if \quad p_i \neq p_{i-1}, \\ H_{i-1}(a) * \log(a) & if \quad p_i = p_{i-1}, \end{cases} \tag{20}$$

and $a^{p_i} = \log(a)$ if $p_i = 0$. In this report the continuous covariate a represents the age at the time of test for the Varicella-Zoster Virus and Parvovirus B19 infections. Royston and Altman (1994) argued that, in practice, fractional polynomials of degree higher than 2 are rarely needed and suggested to choose the values of the powers from the set $\{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$.
Although Royston and Altman (1994) suggested to choose the value of the powers from the above set, one can extend the family of candidate models by refining the grid of possible powers, such as for instance an equidistant grid on the interval $[-2, max(3, m)]$ with step size 0.1 or even 0.01. In this report, the method was applied and all possible models of the same degree were fitted using maximum likelihood estimation and the model with the smallest deviance was chosen. To decide whether a model of first degree is adequate or a second degree model is needed, the Akaike's Information Criteria (AIC) was used.

In the previous sections we have seen that the force of infection can be easily derived from the serological data under assumption including time homogeneity and lifelong immunity. Under this assumption the probability to be infected (prevalence) should be monotonically increasing with age and therefore implying a positive force of infection (Hens *et al.* 2012). Although fractional polynomials provide a wide range of curve shapes, there is no guarantee that the probability of being infected will be a monotone function of age and therefore fractional polynomials can still result in a negative estimate for the force of infection (Shkedy *et al.* 2006).
In case of a violation of the monotonicity assumption, one solution is to fit the new correlated gamma frailty model under the constraint that the probability to be infected should be monotonically increasing with age. Bollaerts *et al.* (2008), proposed a modified version of the fractional polynomial which satisfy the property of monotonicity within a generalized linear mixed model (GLMM). The modification is defined as proposed by Bollaerts *et al.* (2008). Assume a GLMM given by

$$Y_{ij} \sim \text{Binomial}(N_{ij}, \pi_{ij}),$$

$$g(\pi_{ij}) = \beta_0 + \beta_1 x_{ij} + b_i = \eta_{ij} \tag{21}$$

with x_{ij} being a continuous covariate, and $b_i \sim N(0, \sigma)$ the random intercepts. The fractional polynomial of degree $m = 2$ for the linear predictor is defined as

$$\eta(x, \beta, \mathbf{P}, 2) = \beta_0 + \beta_1 x^{p1} + \beta_2 x^{p2}, \tag{22}$$

with $p_1 < 0$, $p_2 \geq 0$, $\beta_1 < 0$, and $\beta_2 > 0$. According to Bollaerts *et al.* (2008), the restrictions imposed on the powers and the parameter estimates yield a monotonically increasing linear function $\beta_0 + \beta_1 x^{p1} + \beta_2 x^{p2}$ bounded between $-\infty$ and $+\infty$.

2.3.6 Model Fitting

Following the development in the previous sections, all models were fitted by maximum likelihood based on equation:

$$L = \sum_a \left(n_{00a} \log(\pi_{00}(a_i)) + n_{10a} \log(\pi_{10}(a_i)) + n_{01a} \log(\pi_{01}(a_i)) + n_{11a} \log(\pi_{11}(a_i))\right), \tag{23}$$

where n_{00a} denotes the number of individuals of age a with neither infections, n_{10a} is the number of individuals of age a infected by Variella-Zoster Virus but not infected by Parvovirus B19, n_{01a} represents the number of individuals of age a infected by Parvovirus B19 but not infected by Varicella-Zoster Virus and n_{11a} is the number of individuals of age a infected by Varicella-Zoster Virus and Parvovirus B19.
Regarding models computation the shared and correlated gamma frailty models were fitted by creating the log-likelihood function in the R software package. For the new frailty model, the R program developed by the authors which is available from the web at http://www.hsph.harvard.edu/betensky/papers.html was used. Table C1 (new correlated gamma frailty model) and Table C2 (correlated gamma frailty model) in the appendix show how to prepare the data and fit the models using R software. Model comparison was performed with Akaike's Information Criteria (AIC).

2.3.7 Comparison Between Both Correlated Gamma Frailty Models

Following the development of the statistical models in the previous sections, their similarities and differences are discussed in this section. It was observed that the expressions (3) and (19) clearly picture how closely the new correlated gamma frailty model (Coull *et al.* 2006) is related to the correlated gamma frailty model (Hens *et al.* 2012). The main difference is the way the multivariate gamma is introduced in the model, and the indirect way to specify the baseline hazard. Saying that, it can be seen that from (19) the new correlated gamma frailty model in (17) implies a particular baseline hazard function for the force of infection.
Similarly, for the correlated gamma frailty model using a Log-logistic baseline hazard function, by expression (19) results on $f_j(a_i)$ given by

$$f_j(a_i) = \log(\log(1 + \exp(\phi_j))a_i^{k_j}). \tag{24}$$

From (24) it can be observed that a particular choice of baseline hazard function can lead to a nonlinear function of $f_j(a_i)$ for the new correlated gamma frailty model. In this particular setting the function $f_j(a_i)$ has to meet the additional constraints that it is a non-deceasing function of age a with $f_j(a_i) \to -\infty$ as $a_i \to 0$. However, current R program for implementing the new correlated gamma frailty model is limited to a linear predictor function.
Another baseline hazard function that can be fitted within this new correlated gamma frailty model is a Weibull model. With this Weibull model, by (19) the new correlated gamma frailty model can be reformulated with linear predictor given by

$$f_j(a_i) = \log(\beta_{0j}) + \beta_{1j}\log(a_i), \tag{25}$$

where β_{0j} and β_{1j} are the parameters for the infectious disease $j = 1, 2$ and a_i is the age of the ith individual at the time of test. Hence, in order to show how closely the new correlated gamma frailty model is related to correlated gamma frailty model both models were fitted using a Weibull baseline hazard function. In particular, the following new correlated gamma frailty model was fitted

$$\log\left(-\log\left(E(Y_{ij}|\log(age), \log(Z_{ij}))\right)\right) = \begin{cases} \log(Z_{i1}) + \log(\beta_{01}) + \beta_{11}\log(age) & \text{if B19,} \\ \log(Z_{i2}) + \log(\beta_{02}) + \beta_{12}\log(age) & \text{if VZV,} \end{cases}$$

where $\log(Z_{i1})$ and $\log(Z_{i2})$ are the log gamma frailties corresponding to Y_{i1} (Parvovirus B19 infection) and Y_{i2} (Varicella-Zoster Virus infection), respectively.

Since the models are based on correlated gamma frailties, both can be considered as correlated gamma frailty models. One (new model) is built under classical generalized linear model with complementary log-log link function and another (correlated gamma frailty model by Hens *et al.* 2009) is formulated in terms of the proportion hazard function. Beyond their similarities and differences, the models have limitations. For instance, for the correlated gamma frailty model, Hens *et al.* (2009) and Weinke (2011) the possible range of the correlation between frailties depends on the values of σ_1 and σ_2:

$$0 \leq \rho \leq min\left(\frac{\sigma_1}{\sigma_2}, \frac{\sigma_2}{\sigma_1}\right). \tag{26}$$

With this restriction, clearly it can be seen that, if $\sigma_1 \neq \sigma_2$, the correlation between the frailties is constrained to be always less than one. This property can be a serious limitation, especially when the values of σ_1 and σ_2 differ strongly (Weinke 2011).
Note that expression (26) implies that, for the correlated gamma frailty model introduced by Hens *et al.* (2009), the correlation between the frailties is constrained by the ratio of the variances and bounded between zero and one. In case of the new correlated gamma frailty model, the correlation is also bounded between zero and one, but the constraint is done by the construction of the model. In addition, this new model assumes that both variances are identical. Despite their similarities and differences, one remark is that with the notation introduced in Section 2.3.4, the model proposed by Coull *et al.* (2006) is closely related to the correlated gamma frailty model (Hens *et al.* 2009) when the variances of the gamma random variables is assumed to be the same. Further investigation could be done for the new correlated gamma frailty model in order to allow the variances to be different.
Turning to the difference of the specification of the baseline hazard function in the models, we show how both models are related to each other. The marginal probability of a response in the new correlated gamma frailty model is given by (Coull *et al.* 2006):

$$P(Y_{ij} = 1) = |I + \zeta C_i \text{diag}(u_{ij})|^{-\frac{1}{\zeta}}, \tag{27}$$

where I is the identity matrix, $\zeta = \sigma^2$ is the variance of the gamma components, u_{ij} is a vector having $\exp(f_j(a_i))$ in position j and 0 elsewhere, and C_i is defined as before. In this particular setting, in case of $p = 2$, using (27), the marginal probabilities are given by

$$P(Y_{i1} = 1) = (1 + \sigma^2 \exp(f_1(a_i)))^{\frac{-1}{\sigma^2}}, \tag{28}$$

$$P(Y_{i2} = 1) = (1 + \sigma^2 \exp(f_2(a_i)))^{\frac{-1}{\sigma^2}}. \tag{29}$$

Note that expressions (28) and (29) are related to the definition of the univariate survival functions in (12). The only difference, is the way how the baseline hazard is introduced in both models.

3 Application to Multisera Data on Varicella and Parvovirus B19

In this section, the methods developed in the report are applied to the multivariate current status data. In particular, the data consist of paired observations of current status of Varicella-Zoster Virus and Parvovirus B19 together with the age of the individual at the time of test. In this report, it was assumed that both diseases are irreversible, meaning that the immunity is assumed to be lifelong. Under this assumption the probability to be infected (prevalence) should be monotonically increasing with age and therefore implying a positive force of infection. Furthermore, by fitting shared and correlated gamma frailty models the report also was restricted to the parametric baseline hazard models for the force of infection for which the only covariate in the model is the host age. The choice of parametric baseline hazards was motivated due to the fact that in case of current status data without any covariates, the model introduced by Hens *et al.* (2009) is not identifiable using a nonparametric baseline hazard function for the force of infection.
Table 2 shows the results obtained by fitting the various version of the correlated gamma frailty model using different baseline hazard functions for the force of infection. Based on AIC-values, the results revealed that the Log-logistic baseline hazard provides the best fit (AIC=3884.238 compared to others).

Table 2: Parameter estimates of the correlation and variances of the frailties together with their standard errors obtained by fitting the correlated gamma frailty model with different and equal variances and shared gamma frailty model using different baseline hazard functions for the force of infection.

Baseline Hazard		Unrestricted		Equal variances		Shared	
Model for the FOI	Parameter	Estim.	Std.err	Estim.	Std.err	Estim.	Std.err
Weibull	ρ	0.2881	0.0765	0.3803	0.0832	1.0000	(—)
	σ_1	1.0040	0.1503	1.4853	0.1737	0.5239	0.0593
	σ_2	2.0971	0.2620	1.4853	0.1737	0.5239	0.0593
	AIC-value		3891.816		3902.090		3945.665
Gompertz	ρ	0.9998	0.0109	0.9979	0.0363	1.0000	(—)
	σ_1	0.3344	0.0777	0.3342	0.0783	0.3342	0.0777
	σ_2	0.3343	0.0776	0.3342	0.0783	0.3342	0.0777
	AIC-value		3910.977		3908.983		3906.976
Log-Logistic	ρ	0.3517	0.0689	0.7407	0.2580	1.0000	(—)
	σ_1	0.3897	0.0699	0.4602	0.0800	0.4056	0.0571
	σ_2	1.1006	0.0905	0.4602	0.0800	0.4056	0.0571
	AIC-value		**3884.238**		3909.149		3907.765
Log-Normal	ρ	0.4152	0.1208	0.5145	0.1435	1.0000	(—)
	σ_1	0.4994	0.1027	0.6622	0.0798	0.4645	0.0578
	σ_2	1.1657	0.1130	0.6622	0.0798	0.4645	0.0578
	AIC-value		3893.939		3907.835		3912.42

3.1 Results for the Shared and Correlated Gamma Frailty Models

In this section, we present the results obtained by fitting the shared and correlated gamma frailty models using a Log-logistic baseline hazard function for the force of infection. Table 3 shows the parameter estimates and their associated standard errors. From the table, it can be observed that the estimated correlation coefficients differ in all fitted models. While the shared gamma frailty model assumes perfect correlation, in the correlated gamma frailty model with different variances, this correlation was estimated to be 0.3517 with standard error 0.0689. The goodness of fit was measured by means of Akaike's information criterion with the smaller its value the better the fit. As can be observed, the correlated gamma frailty model with different variances provides best fit (AIC-value=3884.238 compared to 3909.149 and 3907.765).
Graphically representation of the marginal prevalence and force of infection curves obtained by fitting correlated gamma with different variances, equal variances and shared gamma frailty model using Log-logistic baseline hazard are displayed in Figure 2. Although the observed prevalence for Parvovirus B19 decreases around the age 21 to 30 years old and afterward starts to increase it seems that the correlated gamma frailty model with different variances fits to the data well. The figure also shows that the predicted force of infection is positive and higher at lower age group. This is an indication that there is high risk of being infected at yearly age. In addition, the plot of their joint probabilities that tested positive for both Varicella-Zoster Virus and Parvovirus B19 (top left panel), that tested positive on B19 only (top right panel), that tested positive on Varicella-Zoster Virus only (lower left panel), and that tested negative on both viruses (lower right panel) is shown in Figure 3. From the figure it can also visually be observed a good fit.

Table 3: Parameter estimates and standard errors for the correlated gamma frailty model with different and equal variances and shared gamma frailty model using the best baseline hazard function (Log-logistic baseline hazard), VZV(j=1) and B19(j=2).

	Correlated Gamma Model				Shared Gamma Model	
	Unrestricted		Equal variances		$\sigma_1 = \sigma_2$ and $\rho = 1$	
Parameter	Estimate	Std.err	Estimate	Std.err	Estimate	Std.err
			VZV			
α_1	-2.0076	0.2793	-2.1459	0.3229	-2.0031	0.2668
β_1	2.1840	0.2150	2.3664	0.2850	2.1986	0.1898
σ_1	0.3897	0.0699	0.4602	0.0800	0.4056	0.0571
			B19			
α_2	-4.3897	0.5009	-2.6794	0.2272	-2.6112	0.2060
β_2	2.5014	0.3173	1.3028	0.1041	1.2574	0.0852
σ_2	1.1006	0.0905	0.4602	0.0800	0.4056	0.0571
ρ	0.3517	0.0689	0.7406	0.2580	1.0000	(—)
-2l	3870.238		3897.149		3897.765	
AIC	3884.238		3909.149		3907.765	

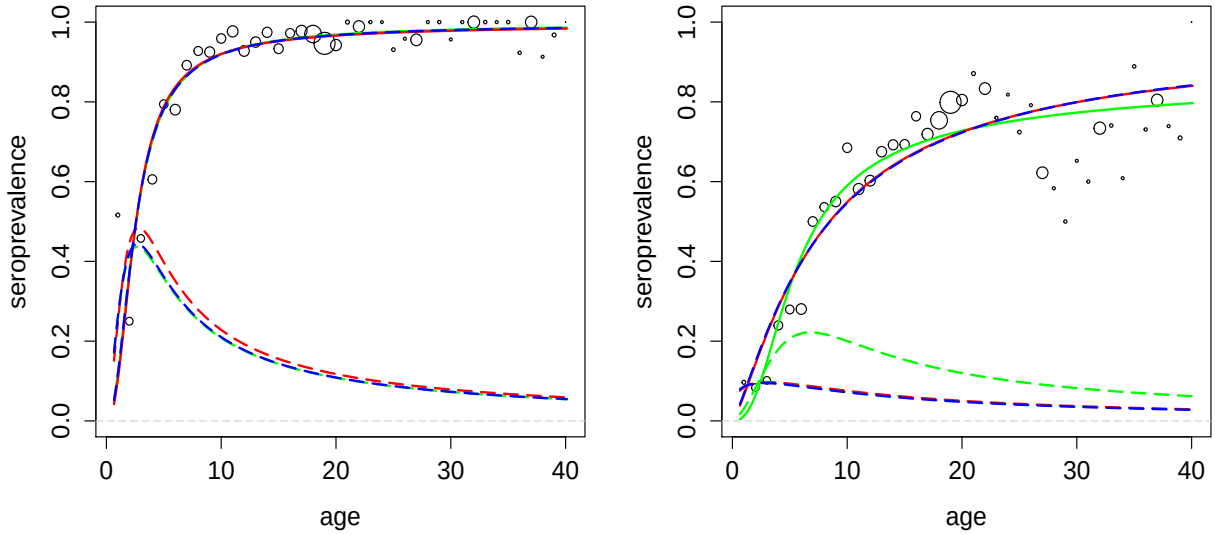

Figure 2: The marginal prevalence and baseline FOI curves for the VZV (*first panel - solid line and dashed line, respectively*) and Parvovirus B19 (*second panel - solid line and dashed line, respectively*) obtained by fitting correlated gamma with different variances (*green curve*), equal variances (*red curve*) and shared gamma frailty model (*blue curve*) using Log-logistic baseline hazard with dots proportional to sample size.

3.2 Results for the New Correlated Gamma Frailty Model

Turning to the new correlated gamma frailty model, the structure of input data was transformed to allow bivariate modeling framework. Mainly, it consists in the integration of the binary responses for the Varicella-Zoster Virus and Parvovirus B19 in the same response vector and an indicator variable indicating the response variable concerned (Indicator=1 for Parvovirus B19 and Indicator = 0 for Varicella-Zoster Virus). The simplest form of the model is given by

$$\log\{-\log[E(Y_{ij}|I_i, age_i, Z_{ij})]\} = \log(Z_{ij}) + \beta_o + \beta_1 I_i + \beta_2 age_i + \beta_3 age_i * I_i$$

or, alternatively,

$$\log\left(-\log\left(E(Y_{ij}|I_i, age_i, Z_{ij})\right)\right) = \begin{cases} \log(Z_{i1}) + \beta_0^* + \beta_1^* age_i & \text{if I=1} \\ \log(Z_{i2}) + \beta_0 + \beta_2 age_i & \text{if I=0} \end{cases}$$

where, $\beta_0^* = \beta_0 + \beta_1$ and $\beta_1^* = \beta_2 + \beta_3$ are the parameters to be estimated, $\log(Z_{i1})$ and $\log(Z_{i2})$ are the log gamma frailties corresponding to Y_{i1} (Parvovirus B19 infection) and Y_{i2} (Varicella-Zoster Virus infection), respectively, I is the indicator variable (I=1 for B19 and I=0 for VZV), and $i = 1, 2, ..., n$.
The predicted marginal prevalence curves for Varicella-Zoster Virus and Parvovirus B19 obtained by fitting the above model resulted on poor fit. Furthermore, a comparison of the AIC-values of the models revealed that the correlated gamma frailty model using a

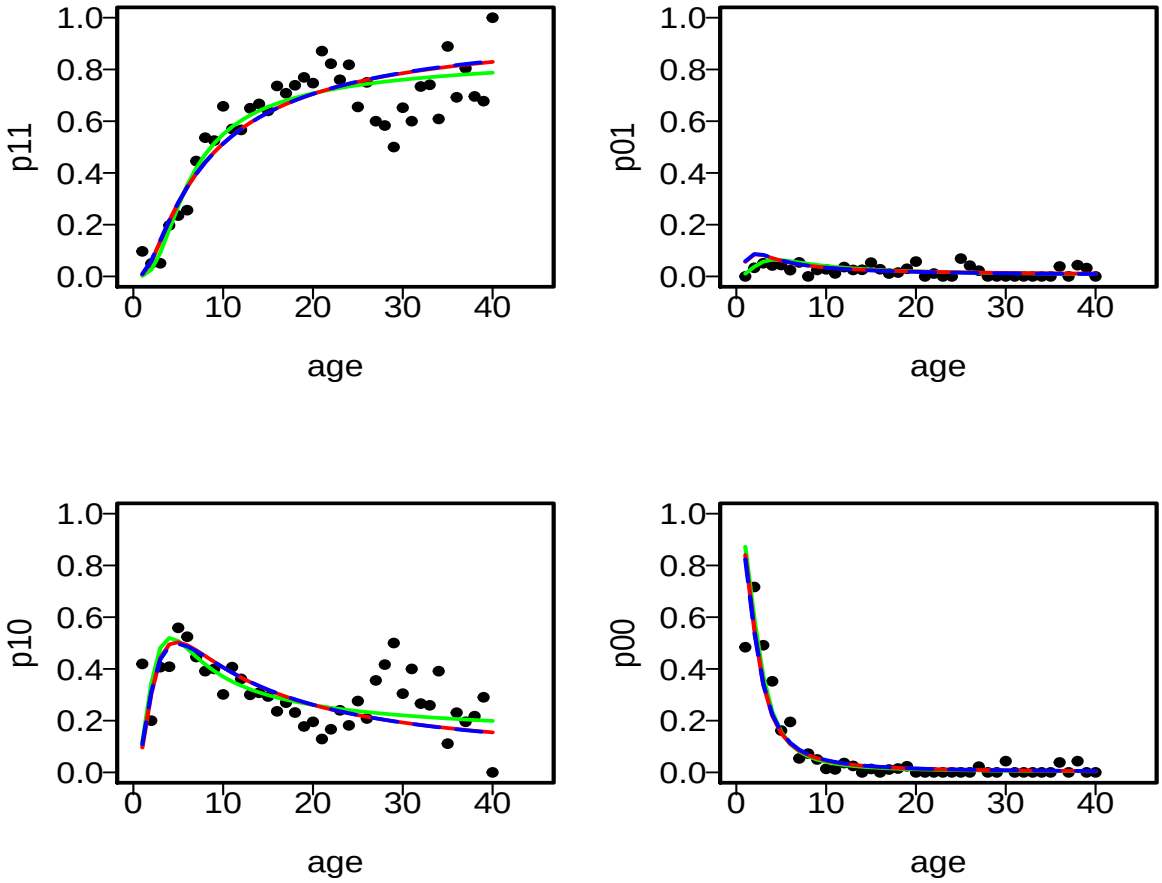

Figure 3: Plot of the joint probabilities of VZV and Parvovirus B19 and the correlated gamma frailty fit with different variances (*green curves*), equal variances (*red curves*), and shared gamma frailty model (*blue curves*) using Log-Logistic baseline hazard. p11 refers to past infection for both viruses (top left panel); p01 represents to no past and past infection for VZV and Parvovirus B19 (to right panel), respectively; p10 refers to past and no past infection for VZV and Parvovirus B19 (lower left panel, respectively); and p00 refers to the joint probability of no past infection for either virus (lower right panel).

Log-logistic is to be preferred (Table A1 and Figure A1 in the Appendix A). However, for the above new correlated gamma frailty model, the implied baseline hazard function for the force of infection differs from Log-logistic model. As mentioned before, a particular choice of Log-logistic baseline hazard implies a non-linear predictor function for the new model. Moreover, current R-program for implementation of the model is limited to a linear predictor function.

Another baseline hazard that can be fitted within this new correlated gamma frailty model framework is a Weibull. Hence, in order to get insight on their similarities and differences both models were fitted using a Weibull baseline hazard function for the force of infection. Parameter estimates and their associated standard errors obtained by fitting both models are presented in Table 4.

From the table, it can be observed that using a Weibull baseline hazard function, the

Table 4: Parameter estimates and their standard errors obtained by fitting the *new** and *original*$^+$ correlated gamma frailty models using a Weibull baseline hazard function.

				Correlated Gamma Frailty					
New Correlated Gamma				$\sigma_1 \neq \sigma_2$		$\sigma_1 = \sigma_2$		Shared Gamma	
Parameter	Est.	s.e	Par.	Est.	s.e	Est.	s.e	Est.	s.e
Intercept	1.179	0.180	α_1	0.210	0.060	0.119	0.042	0.381	0.007
Indicator	0.708	0.212	β_1	1.700	0.332	2.947	0.555	0.918	0.006
log(age)	-1.548	0.087	α_2	0.015	0.009	0.047	0.017	0.138	0.003
log(age)*Ind.	0.572	0.095	β_2	2.790	0.560	1.685	0.267	0.789	0.004
σ_1	0.687	0.095	σ_1	1.004	0.150	1.485	0.174	0.524	0.059
σ_2	0.687	0.095	σ_2	2.097	0.262	1.485	0.174	0.524	0.059
ρ	0.999	0.005	ρ	0.288	0.077	0.380	0.083	1.000	(—)
Deviance	3887.5			3877.8		3890.1		3935.7	
AIC	3899.5			3891.8		3902.1		3945.7	

- *new** refers to the model proposed by Coull *et al.* (2006).
- *original*$^+$ refers to the correlated gamma frailty model introduced by Hens *et al.* (2012).
- Par. = Parameter; Est. = Estimate; s.e = Standard Error

Akaike's information criterion supported the correlated gamma frailty model with different variances (AIC-value= 3891.8 compared to 3899.5 and 3945.7). Although both models were fitted using the same baseline hazard, the resulted correlation between the gamma components is different. This is not surprising however, within correlated gamma frailty model the correlation is constrained by the ratio of the variances, whereas in the new correlated gamma frailty model the correlation is restricted by the construction of the model.
Figure 4 displays the marginal prevalence curves obtained by fitting the new model, shared and correlated gamma frailty models. Although, the models yielded different correlations, the plots are overlapped indicating that both models can produce similar results on the marginal prevalence while the correlation being different. Similarly fit is observed on their resulted baseline force of infection (Figure 5 and Figure 6). In addition, similar results were also observed on their predicted joint probabilities that tested positive on both Varicella-Zoster Virus and Parvovirus B19, that tested positive on Parvovirus B19 only, that tested positive on Varicella-Zoster Virus only, and that tested negative on both viruses (Figure B1 in the Appendix B).

3.3 Results for the Unrestricted New Correlated Gamma Frailty Model Using Fractional Polynomials

In the previous section it was observed that the simplest form of the new correlated gamma frailty model did not fit well to the multisera data on Varicella-Zoster Virus and Parvovirus B19. As mentioned in the development of the methods used in this report, this new model in its natural definition it allows a linear predictor function. This linear predictor function can have some disadvantages, for instance their limited flexibility of linear covariate function type. To enhance flexibility, the model was modified using

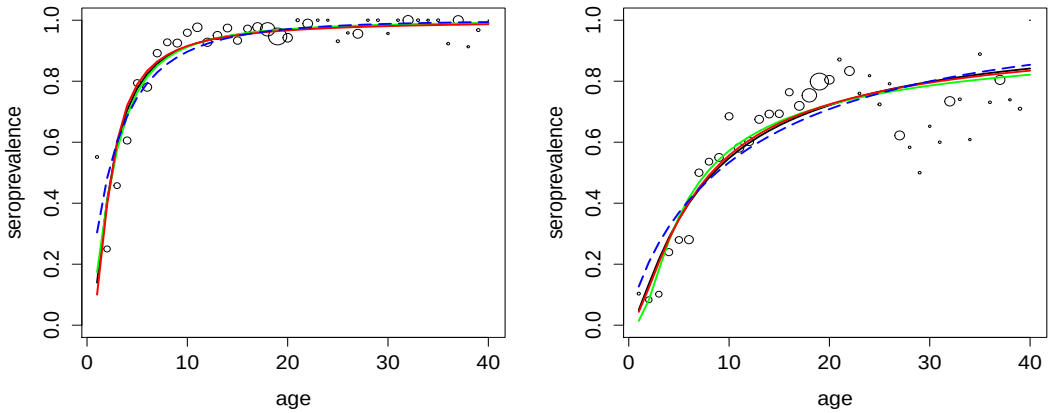

Figure 4: The marginal prevalence curves for VZV (*first panel*) and B19 (*second panel*) obtained by fitting new model (black curve), original correlated gamma with $\sigma_1 \neq \sigma_2$ (*green curve*), with $\sigma_1 = \sigma_2$ (*red curve*) and shared gamma (*dashed blue curve*) using a Weibull baseline hazard.

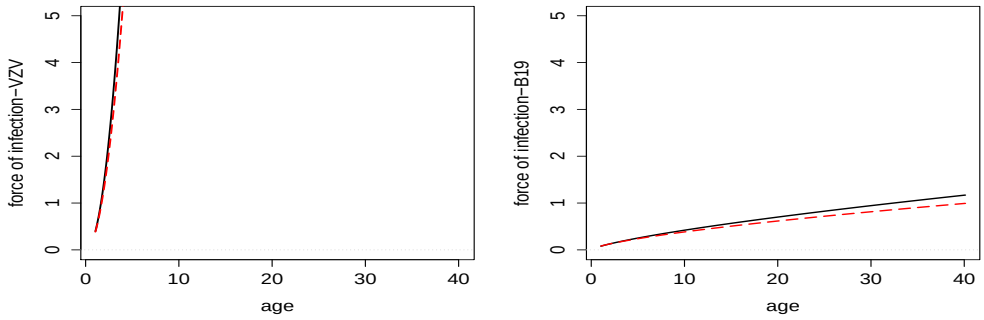

Figure 5: Plot of the baseline FOI curves of VZV (*left panel*) and Parvovirus B19 (*right panel*) obtained by fitting new model (*black curve*) and correlated gamma frailty model with $\sigma_1 = \sigma_2$ (*red curve*) using a Weibull baseline hazard.

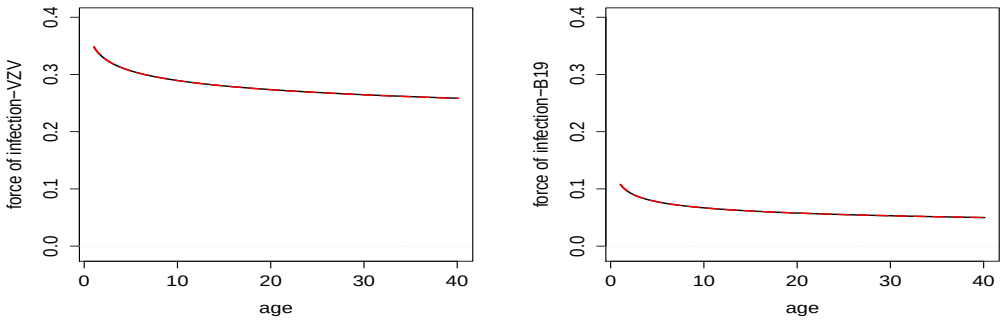

Figure 6: Plot of the baseline FOI curves of VZV (*left panel*) and Parvovirus B19 (*right panel*) obtained by fitting new model (*black curve*) and shared gamma frailty model (*red curve*) using a Weibull baseline hazard (Figures on the top of each other).

fractional polynomial functions without taken into consideration that the probability of being infected should be monotonically increasing with age. In particular, the following new correlated gamma frailty model using fractional polynomial of age with degree two was fitted

$$\log\left(-\log\left(E(Y_{ij}|ter1_j, ter2_j, \log(Z_{ij}))\right)\right) = \begin{cases} \log(Z_{i1}) + \beta_0^* + \beta_1^* ter1_1 + \beta_2^* ter2_1 & \text{if I=1,} \\ \log(Z_{i2}) + \beta_0 + \beta_2 ter1_2 + \beta_3 ter2_2 & \text{if I=0,} \end{cases}$$

where $ter1_j$ and $ter2_j$ (ter1 and ter2 stand by term1 and term2, respectively) are defined using (20) and where the powers p_1 and $p_1^{'}$ were selected from equidistant grid of powers on the interval $[-2, 3]$ with stepsize 0.1, $\log(Z_{i1})$ and $\log(Z_{i2})$ are the log gamma frailties corresponding to Y_{i1} (Parvovirus B19 infection) and Y_{i2} (Varicella-Zoster Virus infection), respectively.

This fractional polynomial with degree two was fitted with powers p_1 and p_2 for the Varicella-Zoster Virus and powers $p_1^{'}$ and $p_2^{'}$ for Parvovirus B19. Since the computation time for fractional polynomial models increases exponentially with the number of powers (Bollaerts *et al.* 2008), the powers p_1 and $p_1^{'}$ were chosen to be the same for both infectious diseases while the powers p_2 and $p_2^{'}$ were chosen to be common or different.

The parameter estimates and their associated standard errors for the best unconstrained model are shown in Table 5. From a set of all possible fractional polynomial models with degree two, the model with powers $p_1 = p_1^{'} = -1.0$ and $p_2 = p_2^{'} = -1.0$, had a smallest AIC-value (3848.577) and thus fits the data best. All fixed effects are statistically significant at 5% level of significance. Although the new correlated gamma frailty model was not restricted to the monotonicity assumption, the estimated correlation is almost 1.000 (0.9999) with standard error 0.0014.

Furthermore, a comparison of the AIC-values of this model and correlated gamma frailty models using a Log-logistic and Weibull baseline hazard functions, revealed that without forcing the new correlated gamma frailty model to satisfy the monotonicity assumption, the model provides best fit for Varicella-Zoster Virus and Parvovirus B19 (AIC-value=3848.577 compared to 3884.238 and 3891.8). However, graphically representation of the marginal prevalence (dashed blue curve) curves for the Varicella-Zoster Virus and Parvovirus B19 displayed in Figure 7 (blue dashed curves), indicate that the predicted baseline force of infection obtained from the model is negative. These results lead to non-meaningful epidemiological interpretation. Therefore, in the next section we fit the model under the constraint that the probability of being infected should be monotonically increasing with age.

Table 5: Parameter estimates and their standard errors for the best unconstrained new correlated gamma frailty model using fractional polynomial with degree two.

Model	Parameter	Estimate	Standard error
Fractional polynomial	Intercept	-5.3285	0.2485
degree two	Indicator	3.2364	0.2688
with powers	term1	4.3970	0.3920
VZV: $p_1 = -1.0, p_2 = -1.0$	term2	10.1126	0.9189
B19: $p_1^{'} = -1.0, p_2^{'} = -1.0$	Indicator*term1	-1.1970	0.5422
	Indicator*term2	-4.5461	1.0743
	ζ	0.3691	0.1372
	ρ	0.9999	0.0014
Deviance			3832.577
AIC			3848.577

3.4 Results for the Restricted New Correlated Gamma Frailty Model using Fractional Polynomials

In this section we present the results of the new correlated gamma frailty model using fractional polynomial of age under the constraint that the probability of being infected should be monotonically increasing with age. In particular, the fitted model is similar to that fractional polynomial with degree two presented in Section 3.3. Similarly, the powers p_1 and p_2 for Varicella-Zoster Virus and powers $p_1^{'}$ and $p_2^{'}$ for Parvovirus B19 were defined in the same way.
For fractional polynomial with degree one, the following model was fitted

$$\log\left(-\log\left(E(Y_{ij}|term_j, \log(Z_{ij}))\right)\right) = \begin{cases} \log(Z_{i1}) + \beta_0^* + \beta_1^* term_1 & \text{if I=1,} \\ \log(Z_{i2}) + \beta_0 + \beta_1 term_2 & \text{if I=0,} \end{cases}$$

where $term_1$, $term_2$, $\log(Z_{i1})$, and $\log(Z_{i2})$ are defined as before.
In Table 6, parameter estimates and standard errors are given for the best new correlated gamma frailty model using fractional polynomials with degree one and two. As can be observed, from a set of all possible fractional polynomial models with degree one, the model with powers $p_1 = -0.6$ and $p_1^{'} = -0.6$ fits the data best, whereas for degree two the best model was the one with powers $p_1 = p_1^{'} = -1.2$ and $p_2 = p_2^{'} = 0$. A comparison of the AIC-values of the models showed that the new correlated gamma frailty model using fractional polynomial with degree one provides best fit for Varilla-Zoster Virus and Parvovirus B19 data (AIC=3890.736 compared to AIC=3893.472). Furthermore, the model fits slightly better than the correlated gamma frailty models with equal variances using Log-logistic and Weibull baseline hazard functions.
In Figure 7, the marginal prevalence curves for the Varicella-Zoster Virus and Parvovirus B19 obtained by fitting the new correlated gamma frailty model using fractional polynomials with degree one and two are shown. From the figure, it can be seen that allowing for more flexibility in the new correlated gamma frailty model, resulted an improved behavior of the marginal prevalence curves.

Table 6: Parameter estimates and their standard errors for the best constrained new correlated gamma frailty models using fractional polynomials of age with degree one and two.

Model	Parameter	Estimate	Standard err.
Fractional polynomial	Intercept	-5.1062	0.2821
degree 1	Indicator	2.5876	0.2713
with powers	I(term1*(1-Indicator))	10.4864	1.0880
$p_1 = -0.6(VZV), p_1^{'} = -0.6(B19)$	I(Indicator*term2)	11.1415	1.6326
	ζ	2.6637	0.6095
	ρ	0.6632	0.0978
Deviance			3878.736
AIC			3890.736
Fractional polynomial	Intercept	1.9844	0.4418
degree two	Indicator	-0.7094	0.5235
with powers	I(term1*(1-indicator))	-1.2589	0.7570
VZV: $p_1 = -1.2, p_2 = 0$	I(term1*indicator)	1.8095	1.0621
B19: $p_1^{'} = -1.2, p_2^{'} = 0$	I(term2*(1-indicator))	-1.8522	0.1746
	I(term3*indicator)	-0.7816	0.1089
	ζ	0.5292	0.1691
	ρ	0.9981	0.0314
Deviance			3877.472
AIC			3893.472

$term1 = ter1_1 = ter1_2$, $term2 = ter2_1$, and $term3 = ter2_2$.

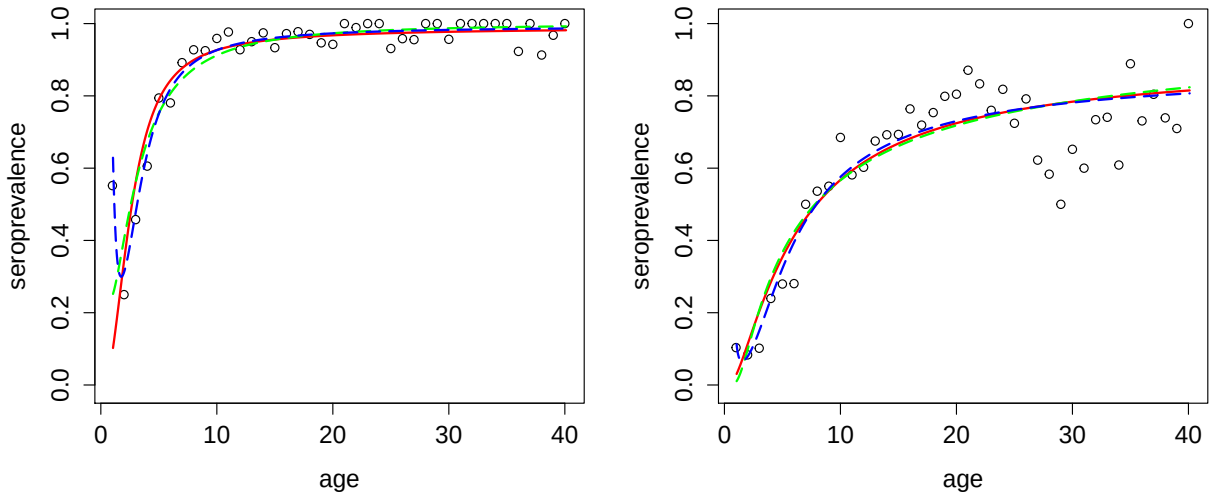

Figure 7: The marginal prevalence curves for Varicella-Zoster Virus (*first panel*) and Parvovirus B19 (*second panel*) obtained by fitting the best constrained new correlated gamma frailty model using fractional polynomial with degree one (*red curve*), degree two (*dashed green curve*) and the best unconstrained fractional polynomial with degree two (*blue dashed curve*).

3.5 Results for the Comparison Between Both Correlated Gamma Frailty Models

In order to get more insights on similarities and differences in both correlate gamma frailty models, we fit using the baseline hazard function obtained from the best fractional polynomial with degree one. In Table 7, parameter estimates and standard errors for the new model and correlated gamma frailty model are presented. Since the new correlated gamma frailty model assumes common variances for the gamma components, the comparison was made with correlated gamma with equal variances. As can be seen, the AIC-values showed that with this baseline hazard function, the new correlated gamma frailty model provides the best fit for Varicella-Zoster Virus and Parvovirus B19 (AIC-value=3890.7 compared to 3893.8, 3896 and 3894.1). In contrast, using a Weibull baseline hazard the correlated gamma frailty model was preferred. It was also observed that the baseline hazard obtained from fractional polynomial with degree one, produced close results on the estimated correlation (new model: $\rho = 0.663$ with s.e equal to 0.098; correlated gamma with equal variances: $\rho = 0.840$ with s.e of 0.569), whereas using a Weibull baseline hazard the estimated correlations differ in both models.
Observe that visually inspection of the marginal prevalence, predicted joint probabilities and baseline force of infection curves for Varicella-Zoster Virus and Parvovirus B19 are similar for both correlated gamma frailty models with equal variances (Figure 8 and Figure 9). These results were also obtained by fitting both models using a Weibull baseline hazard function for the force of infection. These results in combination with the findings in Section 3.2, we conclude that both correlated gamma frailty models can produce similar marginal prevalence curves, but the correlation between the gamma components and their variances being different.

Table 7: Parameter estimates and their standard errors obtained by fitting the *new** and *original*$^+$ correlated gamma frailty models using baseline hazard function obtained from the best fractional polynomial with degree one.

				Correlated gamma frailty model					
New correlated gamma				$\sigma_1 \neq \sigma_2$		$\sigma_1 = \sigma_2$		Shared gamma	
Parameter	Est.	s.e	Par.	Est.	s.e	Est.	s.e	Est.	s.e
Intercept	-5.106	0.282	α_1	2.320	0.180	2.317	0.390	2.235	0.142
Indicator	2.588	0.271	β_1	-4.607	0.372	-4.600	0.660	-4.471	0.327
term1.VZV	10.486	1.088	α_2	2.115	0.509	1.241	0.179	1.207	0.090
term1.B19	11.142	1.633	β_2	-7.151	1.114	-5.318	0.488	-5.245	0.359
σ_1	1.632	0.187	σ_1	0.414	0.082	0.415	0.188	0.372	0.069
σ_2	1.632	0.187	σ_2	0.978	0.220	0.415	0.188	0.372	0.069
ρ	0.663	0.098	ρ	0.424	0.100	0.840	0.569	1.000	(—)
Deviance	3878.7			3879.8		3884.0		3884.1	
AIC	3890.7			3893.8		3896.0		3894.1	

- *new** refers to the model proposed by Coull *et al.* (2006).
- *original*$^+$ refers to the correlated gamma frailty model introduced by Hens *et al.* (2012).
- Par. = Parameter; Est. = Estimate; s.e = Standard Error

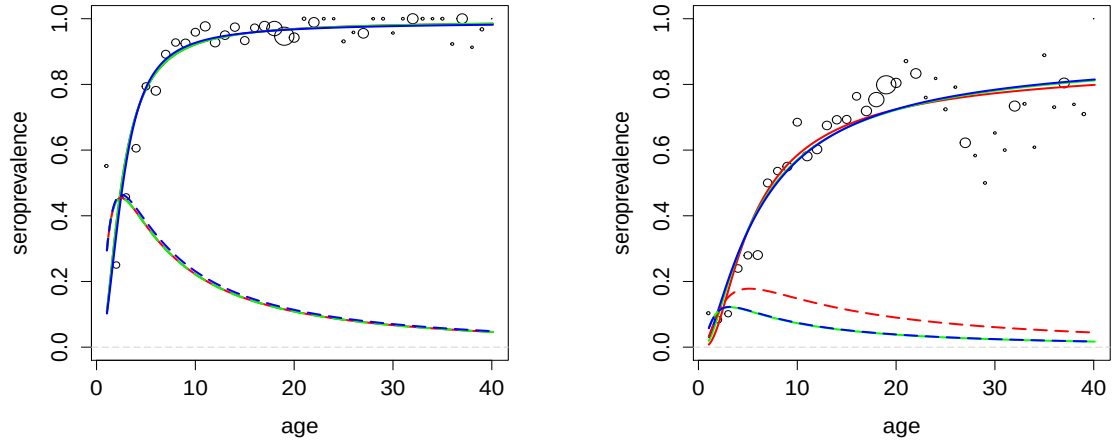

Figure 8: The marginal prevalence and baseline FOI curves for VZV (*first panel - solid line and dashed line, respectively*) and B19 (*second panel - solid line and dashed line, respectively*) obtained by fitting new model (*blue curves*), original correlated gamma with $\sigma_1 \neq \sigma_2$ and with $\sigma_1 = \sigma_2$ (red and green curves, respectively) using baseline hazard function obtained from the best fractional polynomial with degree one (Figures on the top of each other).

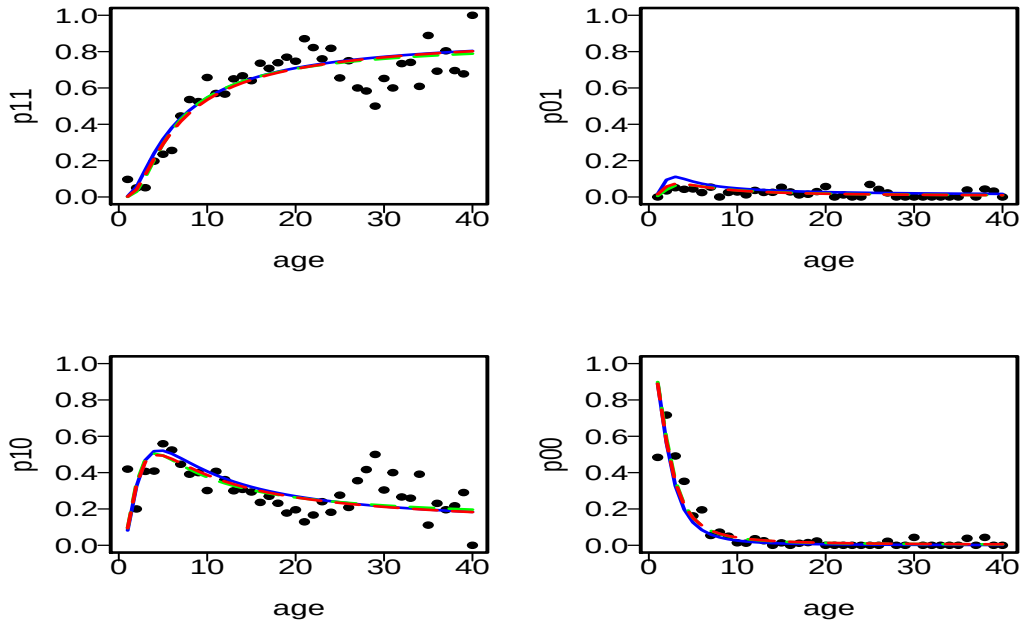

Figure 9: Plot of the joint probabilities of VZV and B19 and the correlated gamma frailty model fit with $\sigma_1 \neq \sigma_2$ (*green curves*), with $\sigma_1 = \sigma_2$ (*red curves*), and new correlated gamma frailty model (*blue curves*) using baseline hazard function obtained from the best fractional polynomial with degree one. p11 refers to past infection for both viruses (top left panel); p01 represents to no past and past infection for VZV and B19 (to right panel), respectively; p10 refers to past and no past infection for VZV Virus and B19 (lower left panel, respectively); and p00 refers to the joint probability of no past infection for either virus (lower right panel).

4 Discussion

In modeling infectious diseases, individual heterogeneity in the acquisition of infectious diseases is recognized as a key concept, which allows improved estimation of important epidemiological parameters. In this perspective, serological samples taken at certain time point provide information about whether or not the individual has been infected before that time point. From these serological samples, a positive (negative) result indicates prior infection (susceptible to infection), giving rise to current status data.
Under the assumption of lifelong immunity and that the epidemic is in steady state (i.e., at equilibrium), these epidemiological parameters can be estimated from serological samples. The traditional statistical models that allow representing such heterogeneity are the correlated gamma frailty models. In 2006, Coull, Houseman and Betensky introduced a computational tractable multivariate random effects model for clustered binary data.
In this report, the proposed model was introduced as a new correlated gamma frailty model, and it was applied and modified to the setting of modeling infectious diseases, and relate it to existing frailty models applied to current status data. The application of these models, was done using a bivariate current status data on Varicella-Zoster Virus and Parvovirus B19. Firstly, the theoretical similarities, limitations and differences of the models are discussed. Secondly, the results obtained from the shared and correlated gamma frailty models and the findings from the new correlated gamma frailty model are also discussed.
Following the development of the models, it was showed that the new model is closely related to the correlated gamma frailty model and both models can be seen as a correlated gamma frailty models since they are based on correlated gamma frailties. The main difference is the way the multivariate gamma is introduced in the new model, and the indirect way to specify the baseline hazard function. In terms of constructions, a correlated gamma frailty model is typically formulated based on specification of the proportional hazard function, whereas the new correlated gamma frailty model is built using a classical generalized linear mixed model with complementary log-log link function for clustered binary data. Since the frailty models are defined under the assumption of time homogeneous representation of Susceptible-Infected-Recovered (SIR) model and that implying a monotonically increasing relationship between the probability to be infected and age, in this particular application, the new correlated gamma frailty model is required to satisfy such property as well.
Going to the parameter which describes the association between multiple infections (i.e., the correlation coefficient), both models share similar limitation on positive correlation between the gamma components. While the correlated gamma frailty model the correlation is bounded between zero and the ratio of the variances of the frailties, in the new correlated gamma frailty model the correlation is limited by construction of the model. In addition, the shared gamma is a special case of the correlated gamma frailty model with correlation between the frailties equal to one (perfect correlation), while the new correlated gamma frailty model can be seen as a special case of the correlated gamma frailty model with common variances but the correlation is not necessarily to be equal to one.
Turning to the results, the shared and correlated gamma frailty models were fitted using different parametric baseline hazard functions for the force of infection. The choice of

parametric baseline hazards was motivated due to the fact that in case of current status data without any covariates, the model introduced by Hens *et al.* (2009) is not identifiable using a nonparametric baseline hazard function for the force of infection. The Log-logistic baseline hazard function provided the best fit for bivariate current status data on Varicella-Zoster Virus and Parvovirus B19. With this baseline hazard function, the correlated gamma frailty model with different variances was preferred.
Fitting the new correlated gamma frailty model, it was observed that the simplest form of the model did not adequately fit to the data as compared with the correlate gamma frailty model with equal variances using Log-logistic baseline hazard. However, the fitted new correlated gamma indirectly implied a particular baseline hazard which was different from the Log-logistic hazard function. Since the new correlated gamma in its natural definition it uses a linear predictor function and that this linear predictor can have some disadvantages on their limited flexibility on linear covariate function type, the model was extended using fractional polynomials. The findings supported that the new correlated gamma frailty model using fractional polynomial with degree one provided the best fit for the Varicella-Zoster Virus and Parvovirus B19 and the estimated correlation was found to be 0.6632 with standard error of 0.0978.
In order to get insight on their similarities, both models were fitted using Weibull baseline hazard and the baseline hazard obtained from the fractional polynomial with degree one. The choice of the Weibull baseline hazard function was motivated due to the fact that, this baseline hazard can be used within the new correlated gamma frailty model framework. Although, using the Weibull baseline hazard the models produced different correlation, the marginal prevalence and baseline force of infection curves were similar. In contrast, when baseline hazard function obtained from fractional polynomial with degree one was used, the correlation was not much different as observed using Weibull baseline hazard and both models provided similar fit as well. As mentioned before, this is not surprising, however in the correlated gamma frailty model, the correlation coefficient is constraint by the ratio of the variances, whereas in the new model the restriction is made by construction of the model.

5 Conclusion and Further Research

This report has presented the model proposed by Coull *et al.* (2006) as a new correlated gamma frailty model and the original correlated gamma frailty model in the context of bivariate current status data to estimate the degree of individual heterogeneity in acquisition of infections. This new correlated gamma frailty model, in its natural definition it allows a linear predictor function and also assumes that the frailties (random effects) have common variances. In this particular setting, the new model has to satisfy the assumption of monotonicity that is a non-decreasing probability of being infected as function of age. Furthermore, the proposed model is closely related to the original correlated gamma frailty model. The main difference is the way the multivariate gamma is introduced in the model, and the indirect way to specify the baseline hazard function. Future investigation could be done on this new correlated gamma frailty model in order to allow the random effects to have different variances and also to extent the model for non-linear or nonparametric smoothed functions.

6 References

Aalen O. (1988). Heterogeneity in survival analysis. Stat Med 7:1121-1137.

Agresti A. (2002). *Categorical Data Analysis.* 2nd ed. John Wiley & Sons, Inc., New York.

Allison P. D. (2010). *Survival Analysis Using SAS: A Practical Guide.* Second Edition. SAS Institute Inc., Cary, NC, USA.

Bapat R. B. (1989). Infinite divisibility of multivariate gamma distributions and M-matrices. Sankyha A51:73-8.

Balakrishnan N., Rao C. R. (2004). *Handbook of Statistics: Advances in Survival Analysis.* north Holland.

Bollaerts K., Aerts M., Faes C., Grijspeerdt K., Dewulf J., Mintiens K. (2008). Human Salmonellosis: Estimations of Dose-Illness from outbreak Data. Risk Analysis, Vol, 28, N0. 2. DOI: 10.1111/j.1539-6924.2008.01038.x.

Broliden K., Tolfvens T., Norbeck O. (2006). Clinical aspects of parvovirus B19 infection. Jornal of Internal Medicine, 260:285-304. DOI:10.1111/j.1365-2796.2006.01697.x.

Chen M. H., Tong X., Sun J. (2009). A frailty model approach for regression analysis of multivariate current status data. Statist. Med., 28:3424-3436.

Coull B. A., Houseman E. A., Betensky R. A. (2006). A computationally tractable multivariate random effects model for clustered binary data. Biometrika, 93, 587-599.

Del Fava E., Shkedy Z., Hens N., Aerts M., Suligoi B., Camoni L., Vallejo F., Wiessing L., Kretzschmar M. (2011). Joint modeling of HCV and HIV co-infection among injecting drug users in Italy and Spain using individual cross-sectional data. Statistical communication in Infectious Diseases; Vol. 3: Iss. 1, Article 3. DOI: 10.2202/1948-4690.1010.

Diamond I. D., McDonald J. W., Shah I. H. (1986). Proportional hazards models for current status data: application to the study of differentials in age at weaning in Pakistan. Demography, 23:607-620.

Duchateau L., Janssen P. (2008). *The Frailty Model.* Springer, New York.

Enki D. G., Noufaily A., Farrington C. P. (2014). A time-varying shared frailty model with application to infectious diseases. Appl Stat 8(1):430-447. DOI: 10.1214/13-AOAS693.

Fahrmeir L., Tutz G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models.* Second Edition. Springer, New York.

Farrington C., Kanaan M., Gay N., (2001). Estimation of the basic reproduction number for infectious diseases from age-stratified serological survey data. Appl Stat 50:251-292.

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

Hanagal D. D. (2011). *Modeling Survival Data Using Frailty Models.* Chapman and Hall/CRC: Boca Raton, FL 33487-2742.

Hannon B., Ruth M. (2009). *Dynamic Modeling of Diseases and Pests.* Springer, New York.

Henderson R., Shimakura S. (2003). A serially correlated gamma frailty model for longitudinal count data. Biometrika, 90:355-66.

Hens N., Aerts M., Shkedy Z., Theeten H., Damme P. V., Beutels Ph. (2008). Modelling multisera data: The estimation of new joint and conditional epidemiological parameters. Stat. Med., 27(14): 2651-64.

Hens N., Wienke A., Aerts M., Molenberghs G. (2009). The correlated and shared gamma frailty model for bivariate current status data: An illustration for cross-sectional serological data. Statist. Med., 28:2785-28900.

Hens N., Aerts M., Shkedy Z., Damme P. V., Beutel P. (2010). Seventy-five years of estimation the force of infection from current status data. Epidemiol Infect 138(6):802-812. DOI: 10.1017/S0950268809990781.

Hens N., Shkedy Z., Aerts M., Faes C., Damme P. V., Beutels P. (2012). *Modeling Infectious Disease Parameters Based on Serological and Social Contact Data. A Modern Statistical Perspective.* Springer, New York.

Keeling M., J., Rohani P. (2008). *Modeling Infectious Diseases in Humans and Animals.* Princeton University Press, United Kingdom.

Liang K. Y., Zeger S. L., Qaqish B. (1992). Multivariate regression analysis for categorical data. J.R. Statist. Soc. 54(1):3-40.

Munda M., Rotolo F., Legrand C. (2012). parfm: Parametric frailty models in R. Jornal of Statistical Software, Volume 51, Issue 11.

Roston, P., Sauerbre W. (2008). *Multivariable Model-Building: A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables.* John Wiley & Sons, Inc., England.

Ross R. (1916). An application of the theory of probabilities to the study of a priori

pathometry. In: proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and physical Character., 92:204-230.

Royston P., Altman D. (1994). Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling. Appl Stat 43(3):429-467.

Shiboski S.C. (1998). Generalized additive models for current status data. Lifetime Data Anal. 4, 29–50.

Shkedy Z., Aerts M., Molenberghs G., Beutels Ph., Damme P. V. (2006). Modelling age-dependent force of infection from prevalence data using fractional polynomials. Statist. Med. 25:1577-1591. DOI: 10.1002/sim.2291.

Sun J. (2006). *The Statistical Analysis of Interval-Censored Failure Time Data.* Springer, New York.

Sutton A., Gay N., Edmunds W., Hope V., Gill O., Hickman M. (2006). Modeling the force of infection for hepatitis B and C in injecting drug users in England and Wales. BMC Infect Dis 6:93. DOI: 101186/1471-2334-6-93.

Thiry N., Beutels P., Shkedy Z. (2002). The seroepidemiology of primary varicella-zoster virus infection in Flanders (Belgium). Eur J Pediatr, 161:588-593. DOI: 10.1007/800431-002-1053-2.

Unkel S., Farrington CP (2012). A new measure of time-varying association for shared frailty models with bivariate current status data. Biostatistics, 13(4):665-79. DOI: 10.1093/biostatistics/kxs010.

Vaupel J., Manton K., Stallard E., (1979). The impact of heterogeneity in individual frailty in the dynamics of mortality . Demography, 16(3):439-454.

Wang W., Ding A. A. (2000). On assessing the association for bivariate current status data. Biometrika, 87(4):879-893.

Wienke, A. (2011). *Frailty Models in Survival Analysis.* Chapman and Hall/CRC Biostatistics Series 37. CRC Press, Boca Raton, FL.

Yashin A. I., Vaupel J. W., Iachine I. A. (1995). Correlated Individual Frailty: An advantageous approach to survival analysis of bivariate data. Mathematical Population Studies, 5(2):145-159.

7 Appendices

7.1 Appendix A: Additional Output for the New Correlated Gamma Frailty Model

Table A1. Parameter estimates and standard errors for the new correlated gamma frailty model with $\rho > 0$ and $\rho = 0$.

New correlated gamma ($\rho > 0$)			New correlated gamma ($\rho = 0$)	
Parameter	Estimate	Standard Error	Estimate	Standard Error
Intercept	-0.2529	0.1218	-0.0454	0.1347
Age	-0.1678	0.0124	-0.1819	0.0132
Indicator	0.6532	0.1556	0.6567	0.1610
Indicator*Age	0.1049	0.0132	0.1150	0.0140
ζ	0.4623	0.1549	1.000	$(----)$
ρ	0.9999	0.0016	0	$(----)$
Deviance		4065.037		4111.921
AIC		4077.037		4119.921

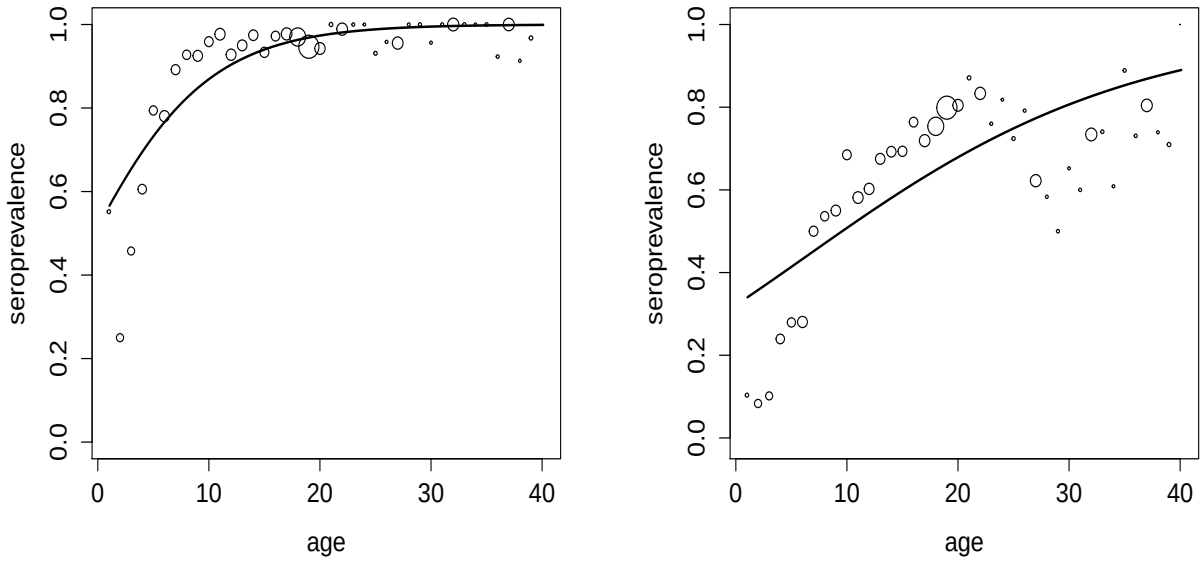

Figure A1. The marginal prevalence curves for the Varicella-Zoster Virus (*first panel*) and parvovirus B19 (*second panel*) obtained by fitting the new correlated gamma frailty model.

7.2 Appendix B: The Plot of the Joint Probabilities - New and Original Correlated Gamma Frailty Models Using a Weibull Baseline Hazard Function

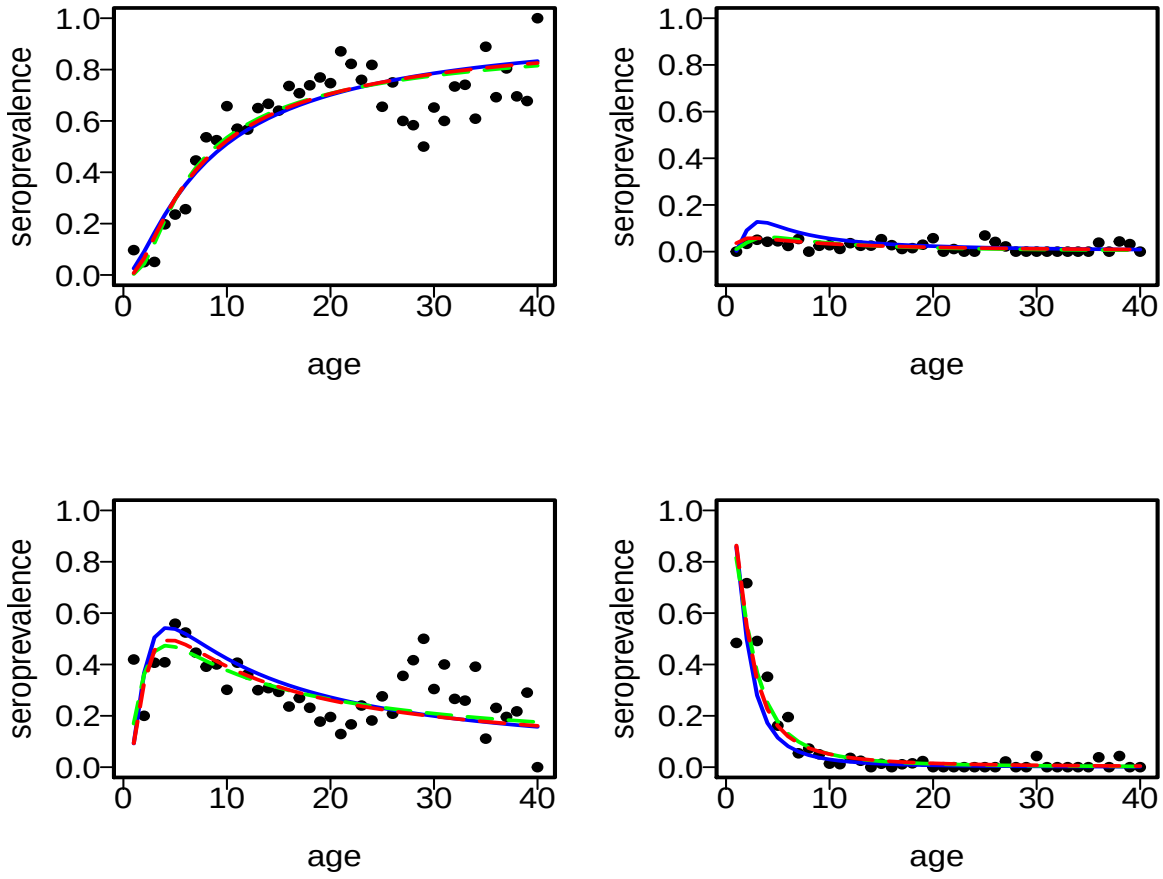

Figure B1. Plot of the joint probabilities of Varicella-Zoster Virus and Parvovirus B19 and the correlated frailty fit with different variances (*green curve*), equal variances (*red curve*) and new gamma frailty (*blue curve*) using a Weibull baseline hazard. p11 refers to past infection for both viruses (top left panel); p01 represents to no past and past infection for Varicella-Zoster Virus and Parvovirus B19 (to right panel), respectively; p10 refers to past and no past infection for Varicella-Zoster Virus and Parvovirus B19 (lower left panel, respectively); and p00 refers to the joint probability of no past infection for either virus (lower right panel).

7.3 Appendix C: Data Structure to Fit Both Correlated Gamma Frailty Models

Table C1. Form of the data structure for the new correlated gamma frailty model.

ID	Gender	Age	Response	Time	Indicator
1	1	18	1	1	1
1	1	18	1	2	0
2	0	22	0	1	1
2	0	22	1	2	0
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
4763	0	9.75	1	1	1
4763	0	9.75	1	2	0

Table C2. Form of the data structure for the correlated gamma frailty model.

NN	NP	PN	PP	a
0	0	0	1	1.05
1	0	1	0	1.07
.	.	.	.	.
.	.	.	.	.
0	1	7	35	19.00
0	0	1	0	19.01
.	.	.	.	.
.	.	.	.	.
0	0	0	1	39.98
0	0	0	1	40.12

7.4 Appendix D: R-code to Fit the Correlated Gamma Frailty Model to Current Status Data

```
-------------------------------------------------------------------------
--Correlated gamma frailty model with different variances using a ------------
--------------------- Log-Logistic baseline hazard ---------------------------
library(stats4)
C.G.F.M1 <- function(alphaeta1=-1.68,betaeta1=-0.46,alphaeta2=-0.065,betaeta2=
                     -.96,k0eta=-2,k1eta=-2,k2eta=-2){
  alpha1 <- alphaeta1
  beta1  <- exp(betaeta1)
  alpha2 <- alphaeta2
  beta2  <- exp(betaeta2)

  k0=exp(k0eta)
  k1=exp(k1eta)
  k2=exp(k2eta)
  sigma1=1/sqrt(k0+k1)
  sigma2=1/sqrt(k0+k2)
  rho=k0/sqrt((k0+k1)*(k0+k2))

  Lambda.VZV <- log(1+exp(alpha1)*data$a^beta1)
  Lambda.B19 <- log(1+exp(alpha2)*data$a^beta2)

  Sa1.VZV <- (1 + (sigma1^2)*Lambda.VZV)^(-1/sigma1^2)
  Sa2.B19 <- (1 + (sigma2^2)*Lambda.B19)^(-1/sigma2^2)
  S12a <- (Sa1.VZV^(-sigma1^2)+Sa2.B19^(-sigma2^2)-1)^(-rho/(sigma1*sigma2))

  p00 <- (Sa1.VZV^(1-(sigma1/sigma2)*rho))*(Sa2.B19^(1-(sigma2/sigma1)*rho))
         *S12a
  p10 <- Sa2.B19 - p00
  p01 <- Sa1.VZV - p00
  p11 <- 1 - p00 - p01 - p10
  return(-sum(PP*log(p11)+PN*log(p10)+NP*log(p01)+NN*log(p00)))
}
Fit1 <- mle(C.G.F.M1,start=list(alphaeta1=-1.68,betaeta1=-0.46,alphaeta2=
                        -0.065,betaeta2=-.96,k0eta=-2,k1eta=-2,k2eta=-2))

AIC1 <- -2*logLik(Fit1) + 2*length(coef(Fit1)); AIC1
beta1 <- exp(coef(Fit1)[2]); round(beta1,4)
s.e.beta1 <- beta1*sqrt(vcov(Fit1)[2,2]);round(s.e.beta1,4)
beta2 <- exp(coef(Fit1)[4]);round(beta2,4)
s.e.beta2 <- beta2*sqrt(vcov(Fit1)[4,4]);round(s.e.beta2,4)
k0 <- exp(coef(Fit1)[5]); k1 <- exp(coef(Fit1)[6]); k2 <- exp(coef(Fit1)[7])
sigma1 <- 1/sqrt(k0+k1);sigma1 ; sigma2 <- 1/sqrt(k0+k2);sigma2
-------------------------------------------------------------------------
```

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
A new model for multivariate current status data

Richting: **Master of Statistics-Biostatistics**
Jaar: **2014**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Martins, Adelino

Datum: **10/09/2014**