2010
2011

FACULTY OF SCIENCES

Master of Statistics: Epidemiology & Public Health Methodology

Masterproef

The trivariate correlated gamma frailty model for current status data

Promotor :
Prof. dr. Niel HENS

Steven Abrams

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Epidemiology & Public Health Methodology

Maastricht University

universiteit hasselt
UNIVERSITEIT VAN DE TOEKOMST

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

2010
2011

FACULTY OF SCIENCES

Master of Statistics: Epidemiology & Public Health Methodology

Masterproef

The trivariate correlated gamma frailty model for current status data

Promotor :
Prof. dr. Niel HENS

Steven Abrams

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Epidemiology & Public Health Methodology

Maastricht University

universiteit hasselt
UNIVERSITEIT VAN DE TOEKOMST

Preface

This master thesis investigates the performance of the trivariate correlated frailty model in the context of serological current status data. The existing frailty methodology used in multivariate survival analysis to model individual heterogeneity is reformulated to be applicable in the field of infectious disease modelling. Furthermore, traditional frailty methodology is generalized in order to be applicable in case of nonimmunizing infections. The thesis was performed under the supervision of my promotor prof. dr. Niel Hens.

I would like to thank everybody who made it possible to complete my master thesis. First of all, I would like to thank prof. dr. Niel Hens to organise this thesis project and for his continuous assistance throughout the entire master thesis project. His input and ideas during our meetings were of crucial importance in the completion of this work. Furthermore, I like to thank prof. dr. Andreas Wienke of the Institut für Medizinische Epidemiologie, Biometrie und Informatik in Halle (Saale) for his valuable contributions and viewpoints with respect to frailty modelling.

To conclude, I really appreciate the guidance and input of my promotor prof. dr. Niel Hens during the entire period in which I worked on this interesting topic.

Contents

Abstract

Frailty models have become increasingly popular in multivariate survival analysis to study individual heterogeneity. Although the shared frailty models are widely applied, the correlated frailty model has recently gained attention since it is quite more flexible as compared to the shared model. In fact, the correlated frailty model elevates the restriction of unobserved factors to act similarly within clusters. The estimation of model specific parameters is often complicated due to the presence of censoring. Within this master thesis, trivariate frailty models are formulated in line with recent work by Hens *et al.* (2009) and relying on serological current status data. Furthermore, refinements of the traditional frailty models in order to cope with recurrent events are suggested and illustrated on a combination of serology and social contact data. Applying the extended models to serological data on nonimmunizing infections such as CMV and PVB19 has shown to improve the model fit considerably. Although further research is certainly required, a first attempt is made to model individual heterogeneity in the absence of lifelong immunity which differs substantially from the current state-of-the-art methods used in the field of infectious disease modelling.

Keywords: Frailty models; Univariate frailty model; Shared frailty model; Correlated frailty model; Trivariate current status data; SIRS infections; Social contact data.

List of Tables

List of Figures

List of Abbreviations and Symbols

Abbreviations

A.D.	Anno Domini
AIC	Akaike Information Criterion
CMV	Cytomegalovirus
HAV	Hepatitis A virus
ODE	Ordinary Differential Equation
PDE	Partial Differential Equation
PVB19	Parvovirus B19
TTE	Time-to-event
RC	Right-censored
CS	Current status
CP	Constant Proportionality
IG	Inverse Gaussian distribution
Re	Real part of complex number
IgG	Immunoglobulin G
GP	Gompertz distribution
PS	Positive Stable distribution
PVF	Power Variance Function distribution
Bin	Binomial distribution
Var	Variance of a random variable
W	Weibull distribution
SARS	Severe Acute Respiratory Syndrome
SIRS	Susceptible-Infected-Recovered-Susceptible transmission model
ELISA	Enzyme-Linked Immunosorbent Assay
i.i.d.	independent and identically distributed
p.d.f.	probability density function
c.d.f.	cumulative density function

Symbols

C	random variable for the censoring time
E	expectation of a random variable
D	mean duration of infectiousness
N	population size
T^*	random variable representing the true event time
T	random variable representing the observed event time
f	symbol for the probability density function of T^*

F	symbol for the cumulative distribution function of T^*
g	symbol for the probability density function of C
G	symbol for the cumulative distribution function of C
S	symbol for the survival function
I	fraction of infected individuals
R	fraction of recovered individuals
S^*	number of susceptible individuals
I^*	number of infected individuals
R^*	number of recovered individuals
Y_l	additive components in correlated frailty models
X	symbol for an arbitrary random variable
Z	frailty variable
Δ	indicator variable for the occurrence of infection in the past
Γ	symbol for the Gamma function
θ	symbol for vector of unknown model parameters
μ	symbol for the hazard function, natural death rate
ρ	symbol for the Pearson product-moment correlation coefficient
α	parameter of the Weibull distribution
β	parameter of the Weibull distribution
a	parameter of the Gompertz distribution, individual age
b	parameter of the Gompertz distribution
ψ	parameter of the gamma distribution
k	parameter of the gamma distribution
ϕ	parameter of the PVF distribution
ζ	parameter of the PVF distribution
η	parameter of the PVF distribution
γ	symbol for the recovery rate
σ	symbol for the replenishment rate
σ^2	symbol for the (frailty) variance
σ_f^2	symbol for the univariate (frailty) variance for SIRS infections
λ	symbol for the force of infection
λ_0	symbol for the baseline force of infection
π	number pi = 3.14159265
$\pi_{...}$	symbol for the seroprevalence
q	constant proportionality factor
n	arbitrary number of random variables
M	symbol for the cumulative hazard function
M_0	symbol for the cumulative baseline force of infection
$\mathbf{L}$	Laplace transform
$\mathbf{L}'$	first derivative of the Laplace transform
$\mathbf{L}''$	second derivative of the Laplace transform
L	likelihood function, likelihood contribution
L_j	likelihood contribution for observation j
L^*	life expectancy of an individual
$\Gamma(.,.)$	generic symbol for the gamma distribution
$\beta(.,.)$	(augmented) effective contact function
$\beta_0(.,.)$	baseline effective contact function
$c(.,.)$	contact function or rate

Chapter 1

Introduction

Infectious diseases remain today a leading cause of morbidity and mortality worldwide. Infectious diseases are illnesses in humans, animals or plants resulting from the presence of microbial pathogens, like viruses, bacteria and parasites (Goeyvaerts, 2011). New pathogens continue to emerge which is demonstrated by the SARS epidemic of 2003 and the swine flu pandemic in 2009. Although Hippocrates (458-377 A.D.) already documented epidemics, mathematical modelling of infections was only introduced in 1760 by the work of Daniel Bernoulli (1760) on the mortality caused by smallpox infections. In his publication, Bernoulli tried to demonstrate the benefits of variolation against smallpox for the population of France using a mathematical transmission model. Deterministic transmission models describe the dynamics of infectious diseases by partitioning the population into different disease states or compartments. In general, mathematical models are applied increasingly to elicidate the transmission of infections and to evaluate the impact of control strategies in reducing morbidity and mortality. In addition to mathematical modelling, statistical models have supplemented the mathematical compartmental models as they allow to estimate important disease parameters from different types of data. One of the most important parameters which drives the disease mechanism is the so-called force of infection. The force of infection is the equivalent of the hazard function in survival analysis and represents the rate at which susceptible individuals acquire the infection. Hugo Muench (1934, 1959) was the first to model the force of infection as a key parameter in mathematical models. Muench's work initiated the development of many parametric and non-parametric methods to estimate the force of infection based on incidence and serological data. An historical overview of the statistical methods that have been applied by many other authors in order to estimate the force of infection, is addressed in a paper by Hens *et al.* (2009).

Frailty models are becoming increasingly popular and important in multivariate survival analysis. Although frailty models have been studied extensively in multivariate survival analysis, the concepts are to be further implemented and extended in the field of infectious disease modelling. Despite the importance of integrating individual heterogeneity in the statistical analysis of infectious diseases as shown by Wienke (2010), frailty models are often applied to data concerning a single infection or at most two diseases. Coutinho *et al.* (1999) were the first to treat individual heterogeneity systematically in the assessment of the force of infection for different infections. A commonly used approach to tackle the problem of modelling dependent multivariate data is to specify independence among observations conditional on a set of latent variables. These latent variables are often referred to as random effects. Frailty models for multivariate survival data rely on the assumption of conditional independence when considering latent frailties acting multiplicatively on the baseline hazard. Therefore, the concept of conditional independence provides a straightforward extension of the univariate frailty model (Vaupel *et al.*, 1979) in order to take dependence of observations into account in the analysis of survival data. Two important frailty models are commonly used in multivariate

survival analysis, the shared frailty model and the correlated frailty model. The shared frailty model is widely applied and assumes the frailty to be common for individuals within a cluster. The shared frailty term creates the dependency between different observations (see e.g. Wienke, 2010). Despite the strong limitations of shared frailty models, they are most often used in practice. To overcome the disadvantages of the shared frailty models, numerous correlated frailty models were established during the last decade as mentioned by Wienke (2005). The correlated frailty model extends the shared frailty model by assuming the frailties of individuals in a cluster to be correlated, but not necessarily shared. In contrast to the shared frailty model in which all the correlation parameters are set equal to one, additional parameters in the correlated model implicitly allow the correlation structure to be different.

Hens *et al.* (2009) already considered the correlated and shared gamma frailty models in the context of bivariate current status data regarding hepatitis A and B infection. In this master thesis, the bivariate models proposed by Hens *et al.* (2009) are extended in order to model trivariate current status data. In the statistical analysis incorporated in this thesis, current status data with respect to cytomegalovirus, parvovirus B19 and hepatitis A infection is analyzed to illustrate the applicability of the derived frailty models. Although in many applications, the Gompertz baseline hazard function seems to be a valuable candidate in describing seroprevalence data, sensitivity with respect to the baseline hazard function is investigated. Since Wienke (2010) summarizes the wide variety of available frailty distributions used in frailty model related publications, the trivariate frailty models are formulated for other frailty distributions as well. In addition to modelling the seroprevalence, a simulation study was performed to investigate the performance of the trivariate correlated gamma frailty model. The latter model shows to be the most flexible one to account for individual heterogeneity as the shared frailty model suffers from severe limitations.

In addition to the specification of different frailty models for infectious disease modelling, we are also faced with problems concerning infections that do not confer lifelong immunity. As the presented trivariate frailty models are derived under the assumption of immunizing infections, some additional refinements of the models are made in order to be applicable as well when reinfections with the pathogen are possible. A mathematical SIRS model is used to derive an expression for the extended survival function and serological data is augmented with social contact data to estimate the parameters of a model reflecting a realistic transmission scenario. The relationship between social contact data and serology is based on the mass action principle as formulated in Farrington *et al.* (2001) and touched upon in Appendix A.1.

The master thesis is organized as follows. First of all, in Chapter 2, the data sources are introduced which are used to illustrate the use of the frailty models presented in this thesis. In Chapter 3, a brief overview of some of the most important concepts of survival analysis are included with special emphasis on the elements relevant in frailty modelling. Chapter 4 focuses on trivariate frailty models in the context of serological current status data. Univariate, shared and correlated frailty models are formulated using the gamma frailty distribution among other candidate frailty distributions. Furthermore, the likelihood functions for time-to-event, right censored and trivariate current status data are derived which are used to obtain maximum likelihood estimates for the unknown frailty model parameters in a fully parametric approach. Chapter 5 highlights potential misspecifications in the frailty models. Especially, difficulties with respect to nonimmunizing infections are considered and an alternative approach is suggested in which serology is augmented with social contact data. In addition, the results of the statistical analysis regarding the seroprevalence of cytomegalovirus, parvovirus B19 and hepatitis A in the Belgian population are incorporated in Chapter 6. Finally, some concluding remarks and potential ideas for further research are written down in Chapter 7.

Chapter 2

Data Sources

In this master thesis, two sources of data are used in order to evaluate the performance of the discussed models. The first data source is serological data on cytomegalovirus (CMV), parvovirus B19 (PVB19) and hepatitis A virus (HAV) which represents the age-specific prevalence of past infection in a population in the absence of an immunization program. Secondly, serological data is augmented with social contact data to be able to estimate the transmission rates for the specific infections under study. In the present chapter, an overview is given of the different diseases discussed in this thesis. In addition, some key notions with respect to the varying data sources enables the reader to have an idea about the data structure before shifting towards the specification of frailty models and the introduction of some useful concepts in univariate survival analysis.

2.1 Serological Data

Serological data consists of cross-sectional sets of residual blood samples which are tested for infection-specific immunoglobulin G (IgG) antibodies using a so-called *enzyme-linked immunosorbent assay* (ELISA) test. Blood samples for 3379 individuals were collected in Belgium between 2001 and 2003 and were tested for CMV, PVB19 and HAV among other infections. In serological data, past infection with a certain pathogen is determined based on the antibody level with respect to a cut-off value pre-specified by the manufacturer of the ELISA-test. Individuals having an antibody level above the cut-off value are classified as being seropositive, and below as seronegative. The serological status of an individual is a direct measure of immunity against the disease, at least if serological protection is agreed upon. In this thesis, our focus is on dichotomized serological data which are in fact type I interval-censored data or current status data. More details on current status data are presented in Sections 3.3 and 4.5. In addition to the serological status of individuals, the age at the time of data collection is obtained as well. In modelling the seroprevalence in the statistical analysis incorporated in this thesis, the focus is on individuals characterized with a complete serological profile concerning the infections under study. The latter implies that data on a total of 2890 blood samples is used in our data applications.

Cytomegalovirus (CMV) is a member of the herpes family which is experienced by many individuals during their lives. The infection is very common during puberty and adolescence, the latter corresponding to the start of sexual activity. In the population at large, primary infection occurs by direct close personal contact via exposure to body fluids such as saliva, tears, urine, stool, semen, and breast milk (Ho, 1990). The primary infection is often inapparent and associated disease is therefore an exceptional event in normal individuals. However, in immunosuppressed patients, the infection provokes several disparate outcomes. As is the case for other herpes viruses (e.g. Varicella Zoster Virus), CMV remains latent within the human host until the host's immune system is compromised.

In addition, CMV is not highly contagious and has an incubation period of about three to twelve weeks (Taylor, 2003).

Parvovirus B19 (PVB19) was the first human parvovirus to be discovered in 1975 (Goeyvaerts, 2011). Most cases of parvovirus B19 infection are asymptomatic. PVB19 causes a range of diseases of which one of the most common clinical presentations is childhood exanthem called fifth disease or erythema infectiosum (Anderson and Cherry, 2004). The childhood exanthem is characterized by a *slapped cheek* rash (Young and Brown, 2004). In children and teenagers, the disease is usually mild whereas in adults, and especially in women, the disease is often complicated by acute arthritis which may persist in some cases (Cohen, 1995). The disease is primarily spread by infected respiratory droplets and infection with PVB19 during pregnancy has been associated with intrauterine fetal death, fetal anemia and hydrops fetalis (Tolfvenstam *et al.*, 2001). The clinical symptoms of the disease start to manifest about six days after exposure to the pathogen and last for approximately a week. The development of the disease occurs after an incubation period of four to fourteen days and after infection, the patients are infectious (and therefore able to infect other individuals) for five to seven days (Hens *et al.*, 2008). Although under development, a vaccine for PVB19 is currently not available.

Hepatitis A (HAV) is one of the oldest diseases known to humankind. The disease is a significant cause of morbidity and socio-economic losses in many countries all over the world. Transmission of HAV is mainly by faeco-oral contact. The infections usually occur early in time in areas where the sanitation is rather poor and living conditions are crowded. In the Western world, the infections are delayed due to improvements in sanitation and hygiene. The virus interferes with the liver's functions which induces the immune system to produce a specific reaction to combat and possibly eradicate the infectious agent. Consequently, the liver becomes inflamed due to pathological damage (Melnick, 1995). The clinical symptoms of the disease are fever, exhaustion, loss of appetite, nausea and abdominal discomfort, dark urine and jaundice. Hepatitis A has an incubation period of 15 to 50 days (Fiore, 2004).

2.2 Social Contact Data

Social contact data is obtained from the European POLYMOD survey conducted between May 2005 and September 2006. Prospective surveys of social contacts were held in eight European countries such as Belgium, Polen, Germany and others. For an extensive description with respect to the survey methodology and the obtained results, we refer to Mossong *et al.* (2008). In the applications presented in this thesis, the contact rates for Belgium with respect to close contacts with total contact time per day exceeding 15 minutes are included. Although different estimation strategies are available for the contact rates, the elements of the social contact matrix are estimated using a bivariate smoothing approach as described by Wood (2006). The average number of contacts is modelled as a two-dimensional continuous function of the age of the respondent and the age of the contacted person. The latter gives rise to a so-called *contact surface*. Details with respect to the bivariate smoothing approach are out of the scope of this thesis. The interested reader is referred to Goeyvaerts (2011). An exploratory investigation regarding the collected social contact data is found as well in Mossong *et al.* (2008). The general idea in this thesis is to supplement the serology with data on contacts among individuals in the population. As the spread of an infectious agent is greatly determined by the number and types of contacts between subjects, it is interesting to use this kind of data to describe infection dynamics.

Chapter 3

Basic Concepts in Survival Analysis

As frailty models gained popularity in multivariate survival analysis over the years, it is interesting to explore the usefulness of these models in the context of infectious disease modelling. Before implementing these frailty models to identify infection characteristics, one needs to consider the basic concepts and notions of frailty modelling. In fact, since these models originate from survival analysis, some of the most relevant and important statistical aspects of univariate survival data are summarized in this chapter. For more detailed information with respect to survival data, we refer to Wienke (2010), Hougaard (1999), and Keiding and Andersen (2006).

3.1 Hazard Function

Survival data requires a special statistical theory due to the unique properties of the response variable. The outcomes of interest in survival data are event times which are certainly not measured in the same way as other variables. Consider a nonnegative random variable T^* and let T^* represent the time from a well-defined starting point until the occurrence of an event (e.g. occurrence of disease). In survival analysis, T^* is often referred to as the survival time which results from the fact that death is the major event studied within this field. However, the term survival time will also be used when referring to occurrence of infection in the remainder of this master thesis. Most often the survival time T^* is assumed to follow a continuous distribution on the interval $[0,\infty)$. The probability density function (p.d.f.) of T^* is denoted by f, and the cumulative distribution function (c.d.f.) by F. From basic probability theory, one obtains:

$$F(t) = P(T^* \leq t) = \int_0^t f(s)ds.$$

In survival analysis, the survival function is defined as $S(t) = 1 - F(t)$. One of the most important concepts in survival analysis is the so-called hazard function $\mu(t)$. It specifies the instantaneous event (e.g. infection) rate at time t, given that the individual has not experienced the event before that point in time. Hence, the hazard function (or hazard rate) is defined as:

$$\mu(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)} \tag{3.1}$$

Sometimes, it is useful to deal with the cumulative or integrated hazard function $M(t)$ which is computed by integrating the hazard function over a finite interval:

$$M(t) = \int_0^t \mu(s)ds. \tag{3.2}$$

From equation (3.1), one can easily derive the relationship between the cumulative hazard function $M(t)$ and the survival function $S(t)$:

$$M(t) = \int_0^t \mu(s)ds = \int_0^t \frac{f(s)}{1-F(s)}ds = -\ln(1-F(t)) = -\ln(S(t)),$$

from which the main exponential formula in survival statistics is obtained, namely:

$$S(t) = 1 - F(t) = e^{-\int_0^t \mu(s)ds} = e^{-M(t)}. \tag{3.3}$$

The exponential formula characterizes the survival function $S(t)$ in terms of the hazard function $\mu(t)$. In infectious disease modelling, the event of interest is infection with a certain pathogen. The hazard function, often called the hazard of infection or force of infection, is a very important epidemiological parameter in practice. The force of infection is assumed to be time-varying and age-dependent. Nevertheless, in order to estimate the force of infection from serological data, one often relies on the untestable assumption of time homogeneity. The latter implies that the infection is assumed to be in steady-state. For the purposes of this thesis, the force of infection is related to the transmission rate or effective contact function by means of the mass action principle. The derivation of these relationships are included in Appendix A.1. In Chapter 5, the mass action principle will be discussed in somewhat more detail.

3.2 Laplace Transform

The concept of the Laplace transform is essential in frailty modelling. It gives the means to derive the unconditional survival functions quite easily if a closed-form expression for the Laplace transform is available and infections are assumed to confer lifelong immunity. The explanation of the latter idea is postponed to Chapter 4 and a general introduction to the Laplace transform is formulated in the present section. In probability theory, the Laplace transform of a random variable X is defined as follows:

$$\mathbf{L}(u) = E(e^{-uX}) = \begin{cases} \int_{-\infty}^{\infty} e^{-ux} f(x)dx & \text{if X is continuous,} \\ \sum_x e^{-ux} P(X = x) & \text{if X is discrete.} \end{cases}$$

For the nonnegative continuous random variable T^* in univariate survival data, the Laplace transform is represented by (Wienke, 2010):

$$\mathbf{L}(u) = E(e^{-uT^*}) = \int_0^{\infty} e^{-ut^*} f(t^*)dt^*. \tag{3.4}$$

Different distributional assumptions with respect to the random variable T^* will produce different expressions for the Laplace transform (see also Section 3.4). If u is replaced by -u, one obtains the so-called moment-generating function of the random variable T^*. The moment-generating function uniquely defines the distribution of a random variable as is the case for the p.d.f. and c.d.f..

3.3 Censored Data

Survival analysis differs from other fields of statistics due to the presence of censoring. A censored observation is in fact an incomplete observation from which only partial information about the event time is available. Such censored observations are the result of, for example, patients that are followed over some period of time in which the event did not occur. For the described situation, one only knows that the true event time exceeds the observed censoring time. There are different types of censoring such as left, right and interval censoring. As it is the most common type of censoring in survival data, we first consider type I right-censored (RC) data. Consider n i.i.d. random variables T_1^*, T_2^*, ..., T_n^*

representing the survival times of n individuals. The c.d.f. and the censoring times are denoted by F and C_1, C_2, ..., C_n, respectively. The censoring times are i.i.d. random variables with c.d.f. G. In addition, let f and g refer to the p.d.f. with respect to F and G, respectively. In practice, one is only able to observe the data (T_1,Δ_1), (T_2,Δ_2), ..., (T_n,Δ_n) where $T_j = \min\{T_j^*,C_j\}$, $j = 1, ..., n$, equals the observation time of subject j. The random variables Δ_j indicate whether the event is observed or censored for individual j, i.e.

$$\Delta_j = \begin{cases} 1 & \text{if } T_j^* \leq C_j, \\ 0 & \text{if } T_j^* > C_j. \end{cases}$$

The p.d.f of the right-censored survival data (T_j,Δ_j) is given by:

$$f(t_j, \delta_j) = (f(t_j)(1 - G(t_j)))^{\delta_j}(g(t_j)(1 - F(t_j)))^{1-\delta_j}, \tag{3.5}$$

under the assumption of independence between censoring and survival times which is assumed throughout derivations in the entire thesis project. The derivation of equation (3.5) is included in Appendix B.1.

The likelihood function for univariate right-censored survival data is now presented below. In the remainder of this thesis, we consider only parametric situations in which the distribution of the survival times T_j^* is assumed to be known up to an unknown parameter vector $\boldsymbol{\theta}$. In Section 3.4, some parametric models often applied in survival analysis are presented, some of which are used here as well in the context of frailty modelling. We assume the censoring to be noninformative which implies that the censoring distribution does not contain information on the survival distribution. Moreover, since we are not interested in estimating the parameters of the censoring distribution, the terms $g(t_j)$ and $G(t_j)$ in the joint density function of (T_j,Δ_j) (equation (3.5)) become additive constants in the log-likelihood function. These components are therefore dropped from the likelihood function as they do not contribute to the derivative of the log-likelihood function. The log-likelihood function is to be maximized in the likelihood framework in order to obtain maximum likelihood estimates for the unknown model parameters $\boldsymbol{\theta}$ (Pawitan, 2001).

The likelihood contribution of right-censored survival data (t_j,δ_j) $(j = 1, ..., n)$ is derived from equation (3.5) after removing the components associated with the censoring distribution as stated previously. The dependence of the density function and survival function on the parameter vector $\boldsymbol{\theta}$ is suppressed from notation for simplicity in the following equation.

$$L_j(\boldsymbol{\theta}) = f(t_j)^{\delta_j}(1 - F(t_j))^{1-\delta_j} = f(t_j)^{\delta_j} S(t_j)^{1-\delta_j}, \tag{3.6}$$

where $S(t_j)$ is the survival function defined in Section 3.1. From equation (3.6), the likelihood function for a sample of independent right-censored survival times (t_1,δ_1), (t_2,δ_2), ..., (t_n,δ_n), equals:

$$L(\boldsymbol{\theta}) = \prod_{j=1}^{n} f(t_j)^{\delta_j} S(t_j)^{1-\delta_j}. \tag{3.7}$$

The likelihood function in equation (3.7) simplifies to the product of the density values $f(t_j)$ in case of uncensored survival data (i.e. T_j^* equals T_j, $j = 1, ..., n$) as in the case of standard situations without censoring. However, a right censored observation induces an evaluation of the survival function to be used as contribution in the likelihood function. We refer to uncensored survival data by means of the term time-to-event (TTE) data in the subsequent chapters.

Sometimes, event times are only known to lie in a specific time interval. Such kind of data arises when study subjects are not under continuous supervision and is called interval censored

data. Furthermore, one can easily understand that right-censored data in general is a special case of interval-censored data. Although some of the methods for RC data are readily applicable for interval-censored data, one often faces some difficulties in analyzing interval-censored data. For the purposes of this master thesis, we consider an important special type of interval-censored data, namely current status (CS) data (type I interval-censored data). The term current status data has its origines in the field of demography as mentioned in Wienke (2010). Individual event times are known to be in an interval containing either zero or infinity and these observations are usually found when subjects are examined only once. In the case of serological data in the field of infectious disease modelling, the only available information is whether the pathogen has infected the individual in the past or not. Based on a cut-off value for the amount of antibodies present in the serum sample of a person, one classifies the patient as either being infected or not before the monitoring time. Therefore, let T_j denote the monitoring time for individual j (j = 1, ..., n), T_j^* the unknown true point in time at which infection occurred and Δ_j the random variable representing the immunological status of individual j. This means that Δ_j indicates whether the infection has occurred before the monitoring time T_j or not. However, note that the specification of Δ_j depends on the specified cut-off value for the amount of antibodies present in blood sera.

The likelihood function function of a sample of univariate (serological) current status data (t_1,δ_1), (t_2,δ_2), ..., (t_n,δ_n) can be written as (Sun, 2006):

$$L(\boldsymbol{\theta}) = \prod_{j=1}^{n} (1 - S(t_j))^{\delta_j} S(t_j)^{1-\delta_j}. \tag{3.8}$$

The likelihood functions in case of trivariate TTE, RC and CS data are included in Chapter 4.

3.4 Parametric Models

In the present section, we will summarize some of the distributional assumptions often made for the event times in survival analysis. The selected distribution needs to fulfill the requirement of having zero mass on the negative part of the real axis. Hence, any distribution for nonnegative random variables can describe the time until an event is going to occur. In survival literature, some distributions arise repeatedly due to their simple nature and good performance in general. These distributions are the exponential, Weibull and Gompertz distributions. In addition, although less convenient as survival time distribution in survival analysis, the gamma distribution shows to be of great value in frailty modelling as illustrated later on. The gamma distribution is one of several candidate distributions for the frailty variables in the presented models. Other frailty distributions will be considered in the subsequent chapters. The event times T_j^* (j = 1, ..., n) are assumed to be i.i.d. random variables.

3.4.1 Exponential Distribution

One of the simplest distributions to assume for the event times T_j^* is the exponential distribution. The exponential distribution is the fundamental distribution in survival analysis even though event time data is rarely following this distribution. An exponential model has only one positive parameter λ and assumes a constant hazard over time (which implies the famous and restrictive *lack of memory* property of the exponential distribution), i.e. $\mu(t_j^*) = \lambda$. Consider $T_j^* \sim \text{Exp}(\lambda)$ then the p.d.f. (for $t_j^* > 0$) is given by:

$$f(t_j^*) = \lambda e^{-\lambda t_j^*}, \quad \lambda > 0.$$

Consequently, the survival function $S(t_j^*)$ equals $e^{-\lambda t_j^*}$. The exponential distribution is a special case of the Weibull, Gompertz and gamma distributions discussed next.

3.4.2 Weibull Distribution

The Weibull model (Weibull, 1939) is a generalization of the one-parameter exponential model with two positive parameters α and β. The second parameter increases the flexibility of the model such that different shapes of the hazard function can be modelled. For $\alpha = \lambda$ and $\beta = 1$, one obtains the exponential distribution as a special case of the Weibull distribution. A Weibull distributed random variable T_j^* ($T_j^* \sim W(\alpha,\beta)$) has probability density function (for $t_j^* > 0$):

$$f(t_j^*) = \alpha\beta t_j^{*\beta-1} e^{-\alpha t_j^{*\beta}}, \quad \alpha > 0, \beta > 0.$$

The survival function $S(t_j^*)$ simplifies to $e^{-\alpha t_j^{*\beta}}$ which leads to $\mu(t_j^*) = \alpha\beta t_j^{*\beta-1}$ for the hazard function. Therefore, the hazard function is either constant or a monotone function of time.

3.4.3 Gompertz Distribution

The Gompertz distribution, in 1825 introduced by the British actuary Benjamin Gompertz, is widely used in biological and demographic applications. A random variable T_j^* follows a Gompertz distribution with parameters a and b, i.e. $T_j^* \sim GP(a,b)$, if the p.d.f. is given by (Gompertz, 1825):

$$f(t_j^*) = ae^{bt_j^*} e^{-\frac{a}{b}(e^{bt_j^*}-1)}, \quad a > 0,$$

and the survival function equals $S(t_j^*) = e^{-\frac{a}{b}(e^{bt_j^*}-1)}$. The hazard function $\mu(t_j^*) = ae^{bt_j^*}$ is increasing starting from a at time zero. If $b < 0$ then the hazard function is decreasing, and the cumulative hazard function $M(t_j^*)$ defined in Section 3.1 converges to the value -a/b for $t_j^* \to \infty$. Hence, not all individuals in the population experience the event under investigation given these conditions.

3.4.4 Gamma Distribution

Finally, the well-known gamma distribution is briefly introduced here. The gamma distribution is also an extension of the exponential distribution and is characterized by a shape parameter k and an inverse scale parameter ψ. A gamma distributed random variable T_j^*, $T_j^* \sim \Gamma(k,\psi)$ has the following density function:

$$f(t_j^*) = \frac{\psi^k}{\Gamma(k)} t_j^{*k-1} e^{-\psi t_j^*}, \quad k > 0, \psi > 0, \tag{3.9}$$

where $\Gamma(.)$ represents the gamma function. The gamma function can be expressed as a definite integral for all complex numbers with positive real part:

$$\Gamma(u) = \int_0^\infty t^{u-1} e^{-t} dt, \quad Re(u) > 0.$$

Furthermore, in general we have $\Gamma(u+1) = u\Gamma(u)$. The mean and variance of T_j^* are equal to k/ψ and k/ψ^2, respectively. Since the gamma distribution does not have closed-form expressions for the survival and hazard function, it is less appealing in modelling event data. However, the gamma distribution is a frequently used frailty distribution in frailty models as discussed in Chapter 4. The Laplace transform plays an important role in the computations associated with frailty modelling as stated earlier. Therefore, the Laplace transform, defined in Section 3.2, for a gamma distributed random variable reduces to:

$$\mathbf{L}(u) = (1 + \frac{u}{\psi})^{-k}. \tag{3.10}$$

In Appendix B.2, details regarding the derivation of the Laplace transform of a gamma distributed random variable are presented. For k equal to one, the gamma distribution simplifies to the exponential distribution in Subsection 3.4.1.

Chapter 4

Frailty Models for Trivariate Current Status Data

In basic survival models, one deals with independent and identically distributed survival data. The underlying assumption in all these models is that the population under study is homogeneous up to some observed covariates. However, it is clear that individuals differ greatly in their reaction to, for example, drugs, treatments, acquisition of infections, etc. Individual heterogeneity is an important source of variability and needs to be taken into account in the statistical analysis of survival data. In that way, frailty models are introduced into survival analysis and are found to be valuable as well in modelling the seroprevalence in case of current status data. The aim of this chapter is to extend existing frailty models to the multivariate setting. Especially, frailty models are formulated to handle trivariate serological data. First of all, the simplest univariate frailty model is described to introduce the concepts of frailty variables and to clarify the importance of the Laplace transform in frailty modelling. In addition, the most commonly used shared frailty model and the far more flexible correlated frailty model are presented in this chapter. Special attention is directed towards the gamma frailty distribution due to the simplicity of the Laplace transform for a gamma distributed frailty variable. However, since other frailty distributions are also frequently used in survival analysis, the frailty models are formulated as well assuming an underlying inverse Gaussian, positive stable and power variance function frailty distribution.

4.1 Univariate Frailty Models

Individuals in a population differ greatly with respect to the acquisition of infections. This unobserved heterogeneity is of great importance and is implicitly taken into account in frailty models. In order to address the problem of unobserved heterogeneity resulting from unobserved covariates, a random effects model was suggested by Beard (1959), and later independently from each other Vaupel *et al.* (1979) and Lancaster (1979). These random effects models are called frailty models whereas the random effects are named individual frailties. The key idea in frailty modelling is that individuals have different frailties, and that the most frail individuals suffer from the event, in our case acquire the infection, earlier in time. Univariate frailty models assume independence between frailties associated with different diseases. Therefore, let $\lambda_i(t_i,Z_i)$ represent the hazard function for infection i at time t_i conditional on the frailty Z_i ($i = 1, 2, 3$). The conditional survival function $S_i(t_i|Z_i)$ ($i = 1, 2, 3$) for immunizing infections equals (Hens *et al.*, 2009):

$$S_i(t_i|Z_i) = e^{-\int_0^{t_i} \lambda_i(s,Z_i)ds}. \tag{4.1}$$

The most frequently applied frailty models rely on a proportional hazard assumption in which the frailty Z_i acts multiplicatively on the baseline hazard function $\lambda_{i0}(t_i)$, i.e. $\lambda_i(t_i,Z_i) = Z_i\lambda_{i0}(t_i)$. Under

the proportional hazard assumption, the conditional survival function can be written as follows:

$$S_i(t_i|Z_i) = e^{-Z_i \int_0^{t_i} \lambda_{i0}(s)ds} = e^{-Z_i M_{i0}(t_i)}, \tag{4.2}$$

where $M_{i0}(t_i) = \int_0^{t_i} \lambda_{i0}(s)ds$ is the cumulative baseline hazard function ($i = 1, 2, 3$) as introduced in Chapter 3. Consequently, the unconditional survival functions can be computed by integrating out the random frailties Z_i. The Laplace transform $\mathbf{L}_i$ of Z_i can be used to write down the unconditional survival functions $S_i(t_i)$ ($i = 1, 2, 3$):

$$S_i(t_i) = E(S_i(t_i, Z_i)) = E(e^{-Z_i M_{i0}(t_i)}) = \mathbf{L}_i(M_{i0}(t_i)). \tag{4.3}$$

From equation (4.3), one can observe that the choice of the univariate frailty distribution determines the expression for the unconditional survival functions. Indeed, the choice of the frailty distribution influences the expression of the Laplace transform. In addition, depending on the selected frailty distribution, the Laplace transform has either an explicit closed-form expression or one requires numerical integration techniques to obtain the unconditional survival functions. Relation (4.3) underlines the key role of the Laplace transform in frailty modelling. Furthermore, the derivatives of the Laplace transform are used to obtain general results about the unconditional survival functions $S_i(t_i)$ ($i = 1, 2, 3$). For example, the density function and hazard of infection can be characterized by the Laplace transform in a convenient way (Wienke, 2010):

$$f_i(t_i) = -\lambda_{i0}(t_i)\mathbf{L}'(M_{i0}(t_i)). \tag{4.4}$$

$$\lambda_i(t_i) = -\lambda_{i0}(t_i)\frac{\mathbf{L}'(M_{i0}(t_i))}{\mathbf{L}(M_{i0}(t_i))}. \tag{4.5}$$

The gamma distribution is widely applied as a frailty distribution from an analytical and computational point of view. Its convenience results from the easy to derive closed-form expressions of unconditional survival, cumulative density and hazard functions, which are entirely due to the simplicity of the Laplace transform. Some background information on the gamma distribution is included in Subsection 3.4.4. As a consequence, the gamma distribution has been extensively used in most of the applications published to date. In general, the Laplace transform of a gamma distributed random variable Z with parameters k and ψ (i.e. $Z \sim \Gamma(k,\psi)$) is given by equation (3.10) and derived in Appendix B.2. For identifiability purposes, one often assumes that the expectation of the frailty variables Z_i equals one. We will refer to this assumption as the standard assumption regarding the frailty expectation. As a result of this assumption, one easily obtains:

$$E(Z_i) = -\mathbf{L}_i^{'}(0) = \frac{k_i}{\psi_i} = 1, \tag{4.6}$$

hence $k_i = \psi_i$. Consequently, $Var(Z_i) = \mathbf{L}_i^{''}(0) - (\mathbf{L}_i^{'}(0))^2 = 1/\psi_i = \sigma_i^2$ and the unconditional survival functions for the univariate gamma frailty model simplifies to:

$$S_i(t_i) = (1 + \sigma_i^2 M_{i0}(t_i))^{-1/\sigma_i^2}. \tag{4.7}$$

In the context of trivariate event data, one can derive bivariate and trivariate unconditional survival functions under the independence assumption of the frailties Z_i. In general, they are equal to the product of the marginal univariate survival functions. The latter result does not longer hold in the case of correlated frailty terms as demonstrated in the following sections.

In the univariate gamma frailty model, the unconditional hazard of infection in equation (4.5) simplifies to:

$$\lambda_i(t_i) = \frac{\lambda_{i0}(t_i)}{(1 + \sigma_i^2 M_{i0}(t_i))} \tag{4.8}$$

In the following section, the restrictive assumption of independence between frailty terms is changed into one of perfect correlation. This gives rise to the popular shared frailty models.

4.2 Shared Frailty Models

The shared frailty model extends the univariate frailty models by relaxing the assumption of independence among the frailty variables Z_i ($i = 1, 2, 3$). Especially, the shared frailty model assumes perfect correlation and equal frailty variances, i.e. $Z_1 = Z_2 = Z_3 = Z$. The marginal unconditional survival functions in a shared frailty model are similar to those obtained in equation (4.3). In general, the unconditional bivariate survival functions are expressed in terms of the Laplace transform $\mathbf{L}$ of the frailty variable Z. Under the assumption of conditional independence, we have for $i = 1, 2$, $j = 2, 3$ and $i \neq j$:

$$S_{ij}(t_i, t_j) = E(e^{-ZM_{i0}(t_i)}e^{-ZM_{j0}(t_j)}) = \mathbf{L}(M_{i0}(t_i) + M_{j0}(t_j)) \tag{4.9}$$

The unconditional trivariate survival function can be written in the same way as the bivariate survival functions. The formula in equation (4.9) is extended in the following way:

$$\begin{aligned} S(t_1, t_2, t_3) &= E(e^{-ZM_{10}(t_1)}e^{-ZM_{20}(t_2)}e^{-ZM_{30}(t_3)}) \\ &= \mathbf{L}(M_{10}(t_1) + M_{20}(t_2) + M_{30}(t_3)). \end{aligned} \tag{4.10}$$

Considering a gamma frailty distribution for the shared frailty Z with parameters $k = 1/\sigma^2$ and $\psi = 1/\sigma^2$ (i.e. the equalities $k = \psi = 1/\sigma^2$ result from the standard assumption with respect to the frailty expectation), the unconditional bivariate survival functions have a closed-form expression derived from equations (4.9) and (3.10):

$$\begin{aligned} S_{ij}(t_i, t_j) &= (1 + \sigma^2(M_{i0}(t_i) + M_{j0}(t_j)))^{-1/\sigma^2} \\ &= (S_i^{-\sigma^2}(t_i) + S_j^{-\sigma^2}(t_j) - 1)^{-1/\sigma^2}. \end{aligned} \tag{4.11}$$

The unconditional trivariate survival function for the shared gamma frailty model corresponds to the expression:

$$\begin{aligned} S(t_1, t_2, t_3) &= (1 + \sigma^2(M_{10}(t_1) + M_{20}(t_2) + M_{30}(t_3)))^{-1/\sigma^2} \\ &= (S_1^{-\sigma^2}(t_1) + S_2^{-\sigma^2}(t_2) + S_3^{-\sigma^2}(t_3) - 2)^{-1/\sigma^2}. \end{aligned} \tag{4.12}$$

Note that σ^2 represents the frailty variance of Z. In fact, the shared frailty model is a constrained version of the general correlated frailty model which is the subject of the following section. For some applications, the shared frailty model shows an improvement in modelling the population heterogeneity in the acquisition of infections as compared to the univariate frailty models presented previously. However, the constraints on the frailty variables associated with the different diseases are quite strong, and therefore need to be relaxed in most of the real-life applications.

4.3 Correlated Frailty Models

Although shared frailty models offer a convenient way of dealing with population heterogeneity, they suffer from lack of flexibibilty as a consequence of the strong assumptions made therein. In contrast, correlated frailty models are a more flexible alternative to account for heterogeneity in the statistical analysis regarding the spread of infections. Yashin *et al.* (1995) introduced the correlated gamma frailty model with an additive decomposition of the frailty variables. These frailties are decomposed into sums of independent gamma distributed random variables. Following this approach, one is able to construct a trivariate frailty distribution for (Z_1,Z_2,Z_3) in which the frailty variables Z_i ($i = 1$, 2, 3) are neither independent nor shared. Although, a decomposition of the frailty variables has the advantage of flexibility over the shared gamma frailty model described in the preceeding section, the model implied correlation between pairs of frailty variables will be constrained. This remark is

illustrated below in the context of the trivariate correlated gamma frailty model.

The additive decomposition of the three frailty variables, as proposed by Yashin *et al.* (1995), for the correlated gamma frailty model assumes the component variables to be independent gamma distributed random variables. Although the gamma frailty distribution is used frequently in practice as stated in Wienke (2010), one can also select other frailty distributions instead which implies a different additive composition in the correlated model. The general decomposition in case of three diseases is a direct generalization of the bivariate correlated gamma frailty model described in Hens *et al.* (2009), and inspired by the seminal work of Yashin *et al.* (1995):

$$\begin{aligned} Z_1 &= \sigma_1^2(Y_0 + Y_1 + Y_4 + Y_5) \\ Z_2 &= \sigma_2^2(Y_0 + Y_2 + Y_4 + Y_6) \\ Z_3 &= \sigma_3^2(Y_0 + Y_3 + Y_5 + Y_6) \end{aligned} \tag{4.13}$$

where Y_l, $l = 0, 1,..., 6$, are independent gamma distributed random variables with parameters $k = k_l$ and $\psi = 1$ (i.e. $Y_l \sim \Gamma(k_l,1)$). Therefore, the mean and variance of the component variables Y_l equal k_l (see Subsection 3.4.4). The constants σ_i^2 are the variance components of the frailties Z_i, $i = 1, 2, 3$. The components Y_l, $l = 1, 2, 3$, represent infection-specific components of the frailties whereas Y_0 reflects a shared part among the three frailty variables. In addition, the remaining variables allow for differing correlation coefficients between pairs of frailties. Under the standard assumption in equation (4.6), one obtains the following relationships with respect to the frailty variances:

$$\begin{aligned} \sigma_1^2 &= (k_0 + k_1 + k_4 + k_5)^{-1} \\ \sigma_2^2 &= (k_0 + k_2 + k_4 + k_6)^{-1} \\ \sigma_3^2 &= (k_0 + k_3 + k_5 + k_6)^{-1} \end{aligned} \tag{4.14}$$

Since the parameters k_l ($l = 0, 1,..., 6$) are real-positive constants, we have $\sigma_i^2 > 0$ ($i = 1, 2, 3$). The additive structure of the frailty variables implies a correlation structure among the frailty variables. Let ρ_{ij} denote the correlation between Z_i and Z_j, $i \neq j$, then we have:

$$\begin{aligned} \rho_{12} &= \frac{k_0 + k_4}{\sqrt{(k_0 + k_1 + k_4 + k_5)(k_0 + k_2 + k_4 + k_6)}} \\ \rho_{13} &= \frac{k_0 + k_5}{\sqrt{(k_0 + k_1 + k_4 + k_5)(k_0 + k_3 + k_5 + k_6)}} \\ \rho_{23} &= \frac{k_0 + k_6}{\sqrt{(k_0 + k_2 + k_4 + k_6)(k_0 + k_3 + k_5 + k_6)}} \end{aligned} \tag{4.15}$$

From expression (4.15), one easily observes that the correlation coefficients are constrained, i.e. $0 \leq \rho_{ij} \leq \min\left\{\frac{\sigma_i}{\sigma_j}, \frac{\sigma_j}{\sigma_i}\right\}$. This disadvantage of the correlated gamma frailty model was already highlighted in the beginning of this section. Although computational unfeasible, a lognormal frailty model with lognormal frailty distribution for Z_i ($i = 1, 2, 3$) and without additive decomposition does not suffer from this limitation. The Laplace transform of a lognormal distributed random variable exhibits no closed-form expression which induces the use of numerical integration techniques to obtain the unconditional survival functions in the likelihood function. Therefore, the choice of the lognormal frailty distribution becomes unpopular, certainly when more than two frailty variables are involved. Other frailty distributions considered in this report do not overcome this restriction on the correlation coefficients.

Obviously, the unconditional univariate survival functions are those presented in equation (4.3). The

unconditional bivariate and trivariate survival functions for the trivariate correlated gamma frailty model are derived under the assumption of conditional independence and can be formulated in terms of the Laplace transforms of the additive components Y_l, $l = 0, 1, ..., 6$. For the correlated gamma frailty model, the expressions are given by:

$$\begin{aligned} S_{12}(t_1, t_2) = &(1 + \sigma_1^2 M_{10}(t_1) + \sigma_2^2 M_{20}(t_2))^{-(k_0+k_4)} \\ &(1 + \sigma_1^2 M_{10}(t_1))^{-(k_1+k_5)} (1 + \sigma_2^2 M_{20}(t_2))^{-(k_2+k_6)} \end{aligned} \tag{4.16}$$

$$\begin{aligned} S_{13}(t_1, t_3) = &(1 + \sigma_1^2 M_{10}(t_1) + \sigma_3^2 M_{30}(t_3))^{-(k_0+k_5)} \\ &(1 + \sigma_1^2 M_{10}(t_1))^{-(k_1+k_4)} (1 + \sigma_3^2 M_{30}(t_3))^{-(k_3+k_6)} \end{aligned} \tag{4.17}$$

$$\begin{aligned} S_{23}(t_2, t_3) = &(1 + \sigma_2^2 M_{20}(t_2) + \sigma_3^2 M_{30}(t_3))^{-(k_0+k_6)} \\ &(1 + \sigma_2^2 M_{20}(t_2))^{-(k_2+k_4)} (1 + \sigma_3^2 M_{30}(t_3))^{-(k_3+k_5)} \end{aligned} \tag{4.18}$$

The unconditional trivariate survival function can be expressed as follows:

$$\begin{aligned} S(t_1, t_2, t_3) = &(1 + \sigma_1^2 M_{10}(t_1) + \sigma_2^2 M_{20}(t_2) + \sigma_3^2 M_{30}(t_3))^{-k_0} \\ &(1 + \sigma_1^2 M_{10}(t_1) + \sigma_2^2 M_{20}(t_2))^{-k_4} \\ &(1 + \sigma_1^2 M_{10}(t_1) + \sigma_3^2 M_{30}(t_3))^{-k_5} \\ &(1 + \sigma_2^2 M_{20}(t_2) + \sigma_3^2 M_{30}(t_3))^{-k_6} \\ &(1 + \sigma_1^2 M_{10}(t_1))^{-k_1} (1 + \sigma_2^2 M_{20}(t_2))^{-k_2} (1 + \sigma_3^2 M_{30}(t_3))^{-k_3} \end{aligned} \tag{4.19}$$

The specification of the frailty models introduced in this chapter is also possible for frailty distributions which differ from the selected gamma frailty distribution. The expressions related to the shared and correlated frailty models using an inverse gaussian, positive stable, or power variance function frailty distribution are incorporated in the following section.

4.4 Other Frailty Distributions

In the present chapter, we include the formulation of the univariate, shared and correlated frailty models when assuming a frailty distribution different from the gamma distribution as proposed in the preceeding sections. As described in Wienke (2010), some other popular choices for the frailty distribution are the inverse Gaussian, the positive stable and the power variance function distributions. Although the lognormal frailty distribution is also often applied in frailty modelling, it is not considered here due to the absence of a closed-form expression for the Laplace transform of a log-normally distributed random variable. Therefore, numerical integration techniques would be required to integrate out the frailty variables which is computer-intensive, especially for the trivariate models included in this master thesis.

4.4.1 Inverse Gaussian Frailty Distribution

A useful alternative for the gamma frailty distribution is the inverse Gaussian frailty distribution. The probability density function for an inverse Gaussian distributed random variable Z with parameters $\phi > 0$ and $\zeta > 0$ is given by:

$$f(z) = \frac{\sqrt{\zeta}}{\sqrt{2\pi z^3}} e^{-\frac{\zeta}{2\phi^2 z}(z-\phi)^2} \qquad (z > 0). \tag{4.20}$$

The density function presented here can be used to obtain the Laplace transform of an inverse Gaussian distributed random variable. The Laplace transform is given by (Wienke, 2010):

$$\mathbf{L}(u) = e^{\frac{\zeta}{\phi}(1-\sqrt{1+\frac{2\phi^2 u}{\zeta}})} \tag{4.21}$$

The derivation of the Laplace transform for an inverse Gaussian distributed random variable is included in Appendix B.3. The expectation and variance of Z can be computed using the derivatives of the Laplace transform (see also equation (4.6)) as follows:

$$E(Z) = -\mathbf{L}^{'}(0) = \phi,$$

$$Var(Z) = \mathbf{L}^{''}(0) - (\mathbf{L}^{'}(0))^2 = \frac{\phi^3}{\zeta}.$$

Assuming $E(Z) = \phi = 1$ and $Var(Z) = \sigma^2 = \frac{1}{\zeta}$, the Laplace transform has the following simplified expression:

$$\mathbf{L}(u) = e^{\frac{1}{\sigma^2}(1-\sqrt{1+2\sigma^2 u})}. \tag{4.22}$$

The univariate inverse Gaussian frailty model for three infections assumes the frailty variables Z_i (i = 1, 2, 3) to be independent inverse Gaussian distributed random variables with parameters $\phi_i = 1$ and $\zeta_i = 1/\sigma_i^2$ (i.e. $Z_i \sim$IG(1,1/σ^2)). The unconditional univariate survival functions $S_i(t_i)$ (i = 1, 2, 3) are equal to (see equation (4.3)):

$$S_i(t_i) = \mathbf{L}_i(M_{i0}(t_i)) = e^{\frac{1}{\sigma_i^2}(1-\sqrt{1+2\sigma_i^2 M_{i0}(t_i)})}, \tag{4.23}$$

where $M_{i0}(t_i)$ is the cumulative baseline force of infection for infection i in t_i and $\mathbf{L}_i$ the Laplace transform of Z_i. In the shared inverse Gaussian frailty model, one restricts the frailty variables to be perfectly correlated. The shared frailty is denoted by Z and one assumes that $Z \sim$IG(1,1/σ^2). The unconditional marginal survival functions are similar to those in equation (4.23) with $\sigma_i^2 = \sigma^2$. In addition, the unconditional bivariate survival functions can be written as:

$$S_{ij}(t_i, t_j) = e^{\frac{1}{\sigma^2}(1-\sqrt{1+2\sigma^2(M_{i0}(t_i)+M_{j0}(t_j))})}, \tag{4.24}$$

for i = 1, 2 and j = 2, 3, $i \neq j$. The trivariate survival function is immediately obtained from the unconditional bivariate survival functions, being the Laplace transform of Z evaluated in $\sum_{i=1}^{3} M_{i0}(t_i)$. Therefore the explicit expression of the unconditional trivariate survival function is omitted here.

Finally, one can formulate the correlated inverse Gaussian frailty model for trivariate data. We propose an additive decomposition of the frailty variables as was already discussed in Section 4.3. The independent component variables are now assumed to follow an inverse Gaussian distribution. In order to satisfy the standard proporty in equation (4.6), one has $Y_l \sim$IG(ϕ_l,ϕ_l^2). Moreover, the frailty variables can be expressed similarly as before:

$$\begin{aligned} Z_1 &= \sigma_1^2(Y_0 + Y_1 + Y_4 + Y_5) \\ Z_2 &= \sigma_2^2(Y_0 + Y_2 + Y_4 + Y_6) \\ Z_3 &= \sigma_3^2(Y_0 + Y_3 + Y_5 + Y_6) \end{aligned} \tag{4.25}$$

Note that the mean and variance of the component variables Y_l are both equal to ϕ_l. As in the case of the gamma frailty distribution, one can write down the expressions for the model implied frailty

variances σ_i^2, $i = 1, 2, 3$:

$$\begin{aligned}\sigma_1^2 &= (\phi_0 + \phi_1 + \phi_4 + \phi_5)^{-1} \\ \sigma_2^2 &= (\phi_0 + \phi_2 + \phi_4 + \phi_6)^{-1} \\ \sigma_3^2 &= (\phi_0 + \phi_3 + \phi_5 + \phi_6)^{-1}\end{aligned} \tag{4.26}$$

The correlation coefficients for the pairs of frailty variables are similar to those included for the correlated gamma frailty model. For completeness, the expressions are included below.

$$\begin{aligned}\rho_{12} &= \frac{\phi_0 + \phi_4}{\sqrt{(\phi_0 + \phi_1 + \phi_4 + \phi_5)(\phi_0 + \phi_2 + \phi_4 + \phi_6)}} \\ \rho_{13} &= \frac{\phi_0 + \phi_5}{\sqrt{(\phi_0 + \phi_1 + \phi_4 + \phi_5)(\phi_0 + \phi_3 + \phi_5 + \phi_6)}} \\ \rho_{23} &= \frac{\phi_0 + \phi_6}{\sqrt{(\phi_0 + \phi_2 + \phi_4 + \phi_6)(\phi_0 + \phi_3 + \phi_5 + \phi_6)}}\end{aligned} \tag{4.27}$$

The constraints imposed by the model specification in the gamma frailty decomposition mentioned in Section 4.3 hold for the inverse Gaussian frailty proposition made in this section. The unconditional bivariate and trivariate survival functions for the trivariate correlated inverse Gaussian frailty model are equal to:

$$\begin{aligned}S_{12}(t_1, t_2) = {} & e^{(\phi_0+\phi_4)\left(1-\sqrt{1+2\left(\sigma_1^2 M_{10}(t_1)+\sigma_2^2 M_{20}(t_2)\right)}\right)} \\ & e^{(\phi_1+\phi_5)\left(1-\sqrt{1+2\sigma_1^2 M_{10}(t_1)}\right)} e^{(\phi_2+\phi_6)\left(1-\sqrt{1+2\sigma_2^2 M_{20}(t_2)}\right)}\end{aligned} \tag{4.28}$$

$$\begin{aligned}S_{13}(t_1, t_3) = {} & e^{(\phi_0+\phi_5)\left(1-\sqrt{1+2(\sigma_1^2 M_{10}(t_1)+\sigma_3^2 M_{30}(t_3))}\right)} \\ & e^{(\phi_1+\phi_4)\left(1-\sqrt{1+2\sigma_1^2 M_{10}(t_1)}\right)} e^{(\phi_3+\phi_6)\left(1-\sqrt{1+2\sigma_3^2 M_{30}(t_3)}\right)}\end{aligned} \tag{4.29}$$

$$\begin{aligned}S_{23}(t_2, t_3) = {} & e^{(\phi_0+\phi_6)\left(1-\sqrt{1+2(\sigma_2^2 M_{20}(t_2)+\sigma_3^2 M_{30}(t_3))}\right)} \\ & e^{(\phi_2+\phi_4)\left(1-\sqrt{1+2\sigma_2^2 M_{20}(t_2)}\right)} e^{(\phi_3+\phi_5)\left(1-\sqrt{1+2\sigma_3^2 M_{30}(t_3)}\right)}\end{aligned} \tag{4.30}$$

The unconditional trivariate survival function can be expressed as follows:

$$\begin{aligned}S(t_1, t_2, t_3) = {} & e^{\phi_0\left(1-\sqrt{1+2\left(\sigma_1^2 M_{10}(t_1)+\sigma_2^2 M_{20}(t_2)+\sigma_3^2 M_{30}(t_3)\right)}\right)} \\ & e^{\phi_4\left(1-\sqrt{1+2\left(\sigma_1^2 M_{10}(t_1)+\sigma_2^2 M_{20}(t_2)\right)}\right)} \\ & e^{\phi_5\left(1-\sqrt{1+2(\sigma_1^2 M_{10}(t_1)+\sigma_3^2 M_{30}(t_3))}\right)} \\ & e^{\phi_6\left(1-\sqrt{1+2(\sigma_2^2 M_{20}(t_2)+\sigma_3^2 M_{30}(t_3))}\right)} \\ & e^{\phi_1\left(1-\sqrt{1+2\sigma_1^2 M_{10}(t_1)}\right)} e^{\phi_2\left(1-\sqrt{1+2\sigma_2^2 M_{20}(t_2)}\right)} e^{\phi_3\left(1-\sqrt{1+2\sigma_3^2 M_{30}(t_3)}\right)}\end{aligned} \tag{4.31}$$

4.4.2 Positive Stable Frailty Distribution

The gamma and inverse Gaussian distribution are very popular choices for the frailty distribution in frailty models since they have nice properties and easy expressions for the Laplace transforms. As illustrated in Chapter 4, the unconditional survival functions are easily obtained from the Laplace transform under the assumption of immunizing infections conferring lifelong immunity. Therefore,

implementation of the frailty models is much more straightforward in case of closed form expressions for the Laplace transform. In addition to the gamma and inverse Gaussian distribution, the positive stable distribution also has a closed-form expression for the Laplace transform. In general, a distribution is stable if the normalized sum of n independent random variables from this distribution has the same distribution as a scale factor multiplied by a single random variable (Wienke, 2010). In order to ensure a distribution on the positive numbers, we restrict ourselves to the positive stable distributions suited for frailty modelling. The p.d.f. for a one-parameter positive stable distribution is given by (Feller, 1971):

$$f(z) = \frac{1}{\pi} \sum_{\kappa=1}^{\infty} (-1)^{\kappa+1} \frac{\Gamma(\kappa\eta + 1)}{\kappa!} z^{-\kappa\eta-1} \sin(\kappa\eta\pi), \quad (z > 0), \tag{4.32}$$

where $0 \leq \eta \leq 1$. In the special case of $\eta = 1$, the frailty distribution becomes degenerated at the point mass $Z = 1$. Although the probability density function of a positive stable distributed random variable with parameter η can only be expressed as an infinite power series, the Laplace transform has an easy closed form. This makes the distribution very attractive in frailty modelling as a frailty distribution as stated by Wienke (2010). The Laplace transform can be deduced easily and is characterized by the expression:

$$\mathbf{L}(u) = e^{-u^{\eta}}. \tag{4.33}$$

All moments of a positive stable distributed random variable are infinite which distinguishes the positive stable distribution from the other frailty distributions proposed until now. Since the expectation equals infinity and the variance does not exist, one can argue that the distribution is useless for frailty variables. However, it was one of the major reasons why the positive stable distribution was introduced as frailty distribution. As the standard assumption in equation (4.6) is invalid, an individual with frailty one can not serve as a reference individual as is the case for the gamma and inverse Gaussian frailty models. The parameter η can be interpreted as a measure of heterogeneity. Values of η close to 1 indicate a small heterogeneity within the population with respect to the infection under consideration whereas values close to zero indicate a large heterogeneity.

Results regarding the unconditional survival functions are presented here for the univariate and shared frailty models. Details with respect to the correlated positive stable frailty model are omitted since it will be clear that the gamma, inverse Gaussian and positive stable frailty distributions are part of the general power variance function frailty distribution discussed in Section 4.4.3. Using the Laplace transform in equation (4.33), the unconditional univariate survival functions $S_i(t_i)$ (i = 1, 2, 3) for immunizing infections are given by:

$$S_i(t_i) = \mathbf{L}_i(M_{i0}(t_i)) = e^{-M_{i0}(t_i)^{\eta_i}}, \tag{4.34}$$

where $Z_i \sim \text{PS}(\eta_i)$. The shared positive stable frailty model can be formulated by assuming the frailty variables to be perfectly correlated and to have equal variances (i.e. $Z_1 = Z_2 = Z_3 = Z$). Letting $Z \sim \text{PS}(\eta)$, the unconditional marginal survival functions presented in formula (4.34) simplify when replacing η_i by η (i = 1, 2, 3). Furthermore, the unconditional bivariate and trivariate survival functions are summarized below. For i = 1, 2, j = 2, 3 and $i \neq j$, one has:

$$S_{ij}(t_i, t_j) = e^{-(M_{i0}(t_i) + M_{j0}(t_j))^{\eta}}, \tag{4.35}$$

and

$$S(t_1, t_2, t_3) = e^{-(M_{10}(t_1) + M_{20}(t_2) + M_{30}(t_3))^{\eta}}. \tag{4.36}$$

4.4.3 Power Variance Function Frailty Distribution

The power variance function (PVF) distribution is a generalized family of frailty distributions that includes the gamma, inverse Gaussian and positive stable distributions. The power variance function distribution is a three-parameter family with parameters ϕ, ζ and $0 \leq \eta \leq 1$. The probability density function for the PVF distribution is given by (see e.g. Hougaard, 2000):

$$f(z) = e^{-\zeta(1-\eta)\left(\frac{z}{\phi}-\frac{1}{\zeta}\right)} \frac{1}{\pi} \sum_{\kappa=1}^{\infty} (-1)^{\kappa+1} \frac{(\zeta(1-\eta))^{\kappa(1-\eta)}\, \phi^{\kappa\eta}}{\eta^{\kappa}} \frac{\Gamma(\kappa\eta+1)}{\kappa!} z^{-\kappa\eta-1} \sin(\kappa\eta\pi), \quad (z>0). \tag{4.37}$$

The Laplace transform of a random variable following the PVF distribution is (Wienke, 2010):

$$\mathbf{L}(u) = e^{\frac{\zeta(1-\eta)}{\eta}\left(1-\left(1+\frac{\phi u}{\zeta(1-\eta)}\right)^{\eta}\right)}. \tag{4.38}$$

The derivation of the Laplace transform is very difficult as shown in Aalen (1992) and is omitted in this thesis. As $E(Z) = \phi$ and $Var(Z) = \phi^2/\zeta = \sigma^2$, one obtains under the standard assumption in frailty modelling, and introducing the frailty variance σ^2 as model parameter, the simplified version of the Laplace transform in equation (4.39):

$$\mathbf{L}(u) = e^{\frac{1-\eta}{\eta\sigma^2}\left(1-\left(1+\frac{\sigma^2 u}{1-\eta}\right)^{\eta}\right)}. \tag{4.39}$$

The latter expression is used in the formulation of the unconditional survival functions for the different frailty models discussed in this section. From the Laplace transform in equation (4.39), one can easily observe that for η equal to 0.5, the power variance function distribution reduces to the inverse Gaussian distribution. For $\eta = 0$, the gamma distribution with parameters $k = 1/\sigma^2$ and $\psi = 1/\sigma^2$ is obtained. The power variance function frailty models are very flexible since they contain many other interesting frailty models as special cases. Therefore, the model is very attractive and often applied in real-life applications.

The univariate power variance function frailty model for trivariate data is specified using the unconditional marginal survival functions $S_i(t_i)$ (i = 1, 2, 3) equal to:

$$S_i(t_i) = e^{\frac{1-\eta_i}{\eta_i\sigma_i^2}\left(1-\left(1+\frac{\sigma_i^2 M_{i0}(t_i)}{1-\eta_i}\right)^{\eta_i}\right)}. \tag{4.40}$$

Details concerning the shared PVF frailty models are omitted since these expressions are straightforward and the derivations are directly obtained using conditional independence of the survival times given the shared frailty Z.

The trivariate correlated PVF frailty model consists of the same additive decomposition of the frailty variables Z_i as described already for the correlated gamma and correlated inverse Gaussian frailty models. The additive decomposition originates from the seminal work of Yashin *et al.* (1995) in the context of the gamma frailty distribution, and is easily extended to other frailty distributions as demonstrated earlier. The independent component variables Y_l (l = 0, 1, ..., 6) are assumed to follow a PVF distribution with parameters ϕ_l, ζ_l and η. Note that the underlying frailty distributions for the three infections are assumed to be equal in order to obtain a well-defined frailty distribution. Moreover, the parameters ζ_l are taken equal to ϕ_l. The additive decomposition is summarized here:

$$\begin{aligned} Z_1 &= \sigma_1^2(Y_0 + Y_1 + Y_4 + Y_5) \\ Z_2 &= \sigma_2^2(Y_0 + Y_2 + Y_4 + Y_6) \\ Z_3 &= \sigma_3^2(Y_0 + Y_3 + Y_5 + Y_6) \end{aligned} \tag{4.41}$$

As a consequence, one derives that $Z_i \sim \text{PVF}(1,1/\sigma_i^2,\eta)$ under the standard assumption presented in equation (4.6). The frailty variances and implicit correlation coefficients for pairs of frailty variables are entirely the same as in the case of the inverse Gaussian frailty model (equations (4.26) and (4.27), respectively). Only the joint unconditional survival function for the trivariate correlated PVF frailty model is presented:

$$\begin{aligned} S(t_1,t_2,t_3) = & e^{\frac{\phi_0(1-\eta)}{\eta}\left(1-\left(1+\frac{\left(\sigma_1^2 M_{10}(t_1)+\sigma_2^2 M_{20}(t_2)+\sigma_3^2 M_{30}(t_3)\right)}{\eta}\right)^{\eta}\right)} e^{\frac{\phi_4(1-\eta)}{\eta}\left(1-\left(1+\frac{\left(\sigma_1^2 M_{10}(t_1)+\sigma_2^2 M_{20}(t_2)\right)}{\eta}\right)^{\eta}\right)} \\ & e^{\frac{\phi_5(1-\eta)}{\eta}\left(1-\left(1+\frac{\left(\sigma_1^2 M_{10}(t_1)+\sigma_3^2 M_{30}(t_3)\right)}{\eta}\right)^{\eta}\right)} e^{\frac{\phi_6(1-\eta)}{\eta}\left(1-\left(1+\frac{\left(\sigma_2^2 M_{20}(t_2)+\sigma_3^2 M_{30}(t_3)\right)}{\eta}\right)^{\eta}\right)} \\ & e^{\frac{\phi_1(1-\eta)}{\eta}\left(1-\left(1+\frac{\left(\sigma_1^2 M_{10}(t_1)\right)}{\eta}\right)^{\eta}\right)} e^{\frac{\phi_2(1-\eta)}{\eta}\left(1-\left(1+\frac{\left(\sigma_2^2 M_{20}(t_2)\right)}{\eta}\right)^{\eta}\right)} e^{\frac{\phi_3(1-\eta)}{\eta}\left(1-\left(1+\frac{\left(\sigma_3^2 M_{30}(t_3)\right)}{\eta}\right)^{\eta}\right)} \end{aligned} \tag{4.42}$$

In the following section, the likelihood function for trivariate current status data is derived which is of importance when implementing the above frailty models within the maximum likelihood framework. In addition, the likelihood functions for trivariate time-to-event (TTE) and right censored (RC) data are formulated as well.

4.5 Maximum Likelihood Estimation

In the thesis, the aim is to fit full parametric frailty models using selected parametric forms for the baseline force of infection. First of all, the likelihood function is formulated for the simplest situation in which the data consists of exactly observed event times for three infections. Time-to-event (TTE) data, as introduced in Section 3.3, refers to the situation in which the point in time at which the event occurs is exactly known. Let T_i^*, $i = 1, 2, 3$, denote the points in time at which an individual acquires the three infections under consideration. The likelihood contribution can be expressed in terms of the unconditional trivariate survival function. The likelihood contribution for an observation (t_1^*,t_2^*,t_3^*) is equal to:

$$L(t_1^*,t_2^*,t_3^*) = f(t_1^*,t_2^*,t_3^*) = -\frac{\partial^3}{\partial t_1^* \partial t_2^* \partial t_3^*} S(t_1^*,t_2^*,t_3^*), \tag{4.43}$$

where $f(.)$ is the joint probability density function for T_1^*, T_2^* and T_3^*, and $S(.)$ refers to the trivariate unconditional survival function (see e.g. equation (4.42)). Equation (4.43) results from the relationship between the survival function S and cumulative density function F in basic probability theory.

As mentioned earlier, the analysis of time-to-event data is often complicated due to the presence of censoring. One can distinguish between different types of censoring such as right censoring (RC) and interval censoring which are of interest here. Right censored (RC) data and interval censored data are introduced in Section 3.3. Trivariate right censored (RC) data is obtained from trivariate time-to-event data by specifying monitoring times C_i and consequently defining the random variables T_i as the minima of T_i^* and C_i ($i = 1, 2, 3$). Furthermore, the random variables Δ_i indicate whether the event occurred at time T_i ($\Delta_i = 1$) or the true time point is censored ($\Delta_i = 0$). The likelihood contribution for trivariate right censored data is a generalization of the contributions in the likelihood function for univariate right censored survival data presented in equation (3.7):

$$
\begin{aligned}
L(t_1,t_2,t_3,&\delta_1,\delta_2,\delta_3) = \\
&\delta_1\delta_2\delta_3(-\frac{\partial^3}{\partial t_1\partial t_2\partial t_3}S(t_1,t_2,t_3)) + \delta_1\delta_2(1-\delta_3)(\frac{\partial^2}{\partial t_1\partial t_2}S(t_1,t_2,t_3)) \\
&+\delta_1(1-\delta_2)\delta_3(\frac{\partial^2}{\partial t_1\partial t_3}S(t_1,t_2,t_3)) + (1-\delta_1)\delta_2\delta_3(\frac{\partial^2}{\partial t_2\partial t_3}S(t_1,t_2,t_3)) \\
&+\delta_1(1-\delta_2)(1-\delta_3)(-\frac{\partial}{\partial t_1}S(t_1,t_2,t_3)) + (1-\delta_1)\delta_2(1-\delta_3)(-\frac{\partial}{\partial t_2}S(t_1,t_2,t_3)) \\
&+(1-\delta_1)(1-\delta_2)\delta_3(-\frac{\partial}{\partial t_3}S(t_1,t_2,t_3)) + (1-\delta_1)(1-\delta_2)(1-\delta_3)S(t_1,t_2,t_3)
\end{aligned}
\tag{4.44}
$$

Interval censoring occurs when event times are only known to be in a specific interval. In the case of serological data, an individual is examined only once. Consequently, one can only determine whether the individual was already infected before the monitoring time. When the individual experienced the disease, based on his/her serological sample, one is not able to pinpoint the exact point in time at which the infection has occurred. Moreover, when the disease did not affect the individual before the monitoring time, it is also not clear whether the individual is going to experience the infection in the near future. Therefore, in either case, the event will take place in an interval containing either zero or infinity. This is an example of type I interval-censored data or current status data. Trivariate CS data is presented in the form (T_i,Δ_i) (i = 1, 2, 3), where T_i equals the monitoring time for infection i and Δ_i represents the immunological status of the individual. Hence, Δ_i indicates whether the infection already occurred before time T_i or not. The observations are denoted by (t_i,δ_i) (i = 1, 2, 3).

Under time homogeneity, lifelong immunity once infected and the lifelong presence of antibodies with respect to the infections considered here, one can estimate the hazard of infection (i.e. the force of infection) from cross-sectionally collected serological data. Therefore, the time unit of interest is the age of the individual under investigation and T_i refers to the age at inspection of a study subject. As a consequence, the force of infection is assumed to be age-dependent. Since serological samples are often tested for more than one antigen, multivariate methods allow for the investigation of the association between the acquisition of different infections.

The likelihood function for trivariate current status data can be expressed in terms of the univariate, bivariate and joint unconditional survival functions (see also Sections 4.1, 4.2 and 4.3). The likelihood contribution for an observation (t_1,t_2,t_3,δ_1,δ_2,δ_3) is given by:

$$
\begin{aligned}
L(t_1,t_2,t_3,&\delta_1,\delta_2,\delta_3) = \\
&\delta_1\delta_2\delta_3(1-S_1(t_1)-S_2(t_2)-S_3(t_3)+S_{12}(t_1,t_2)+S_{13}(t_1,t_3)+S_{23}(t_2,t_3)-S(t_1,t_2,t_3)) \\
&+\delta_1\delta_2(1-\delta_3)(S_3(t_3)-S_{13}(t_1,t_3)-S_{23}(t_2,t_3)+S(t_1,t_2,t_3)) \\
&+\delta_1(1-\delta_2)\delta_3(S_2(t_2)-S_{12}(t_1,t_2)-S_{23}(t_2,t_3)+S(t_1,t_2,t_3)) \\
&+(1-\delta_1)\delta_2\delta_3(S_1(t_1)-S_{12}(t_1,t_2)-S_{13}(t_1,t_3)+S(t_1,t_2,t_3)) \\
&+\delta_1(1-\delta_2)(1-\delta_3)(S_{23}(t_2,t_3)-S(t_1,t_2,t_3)) \\
&+(1-\delta_1)\delta_2(1-\delta_3)(S_{13}(t_1,t_3)-S(t_1,t_2,t_3)) \\
&+(1-\delta_1)(1-\delta_2)\delta_3(S_{12}(t_1,t_2)-S(t_1,t_2,t_3)) \\
&+(1-\delta_1)(1-\delta_2)(1-\delta_3)S(t_1,t_2,t_3)
\end{aligned}
\tag{4.45}
$$

For serological data on the immunological status of individuals, sera are often tested for different diseases at the same monitoring time. Therefore, in case of univariate monitoring times, we have t_1 =

$t_2 = t_3 = t$ in equation (4.45). As stated by Hens *et al.*, frailty models are not identifiable for current status data without any covariates when using a nonparametric baseline hazard function. Therefore, a parametric baseline hazard function is used for the three diseases under study.

One of the possibilities for the baseline force of infection is the Gompertz baseline hazard function, $\lambda_{i0}(t) = a_i e^{b_i t}$, which is often used in practice. The Gompertz baseline hazard function is obtained by assuming that the time until infection with pathogen i follows a Gompertz distribution with parameters a_i and b_i (see Section 3.4.3). Consequently, one can also assume the event times to be exponentially distributed which implies a constant baseline force of infection in contrast to the monotone shape of the Gompertz hazard function. Alternatively, the Weibull baseline force of infection $\lambda_{i0}(t) = \alpha_i \beta_i t^{\beta_i - 1}$ can be applied. As those are the most commonly used baseline hazard functions in survival analysis, the focus will be on these parametric distributions for the remaining of this thesis.

In case of serological current status data, one can estimate the force of infection under the assumption of time homogeneity as mentioned earlier. In fact, one is interested in modelling the seroprevalence of an infection in order to determine the force of infection. In general, the age-specific seroprevalence is defined as the proportion of individuals in the population that are infected in the past with the disease under investigation. For the purposes of this master thesis, the trivariate seroprevalences are denoted by $\pi_{...}$. For example, π_{111} represents the proportion of individuals that suffered from all three diseases before the monitoring time. Since Δ_i ($i = 1, 2, 3$) refers to the immunological status of an individual with respect to infection i, one can define the trivariate seroprevalences as $\pi_{...} = P(\Delta_1 = ., \Delta_2 = ., \Delta_3 = .)$ where . equals 0 or 1 depending on whether the individual is infected in the past or not. Consequently, the additive components of the likelihood contribution in equation (4.45) are by definition equal to the trivariate seroprevalences, e.g. for π_{111} one obtains:

$$\begin{aligned}
\pi_{111} &= P(\Delta_1 = 1, \Delta_2 = 1, \Delta_3 = 1) \\
&= P(T_1 \leq t_1, T_2 \leq t_2, T_3 \leq t_3) \\
&= 1 - S_1(t_1) - S_2(t_2) - S_3(t_3) + S_{12}(t_1, t_2) + S_{13}(t_1, t_3) + S_{23}(t_2, t_3) - S(t_1, t_2, t_3).
\end{aligned}$$

The marginal seroprevalences are expressed as π_{1++}, π_{+1+} and π_{++1} for the proportion of past infections with CMV, PVB19 and HAV, respectively irrespective of the acquisition of the other infections. Therefore, the random variables Δ_i follow a binomial distribution with probability of success equal to their marginal prevalence (e.g. $\Delta_1 \sim \text{Bin}(1,\pi_{1++})$). Consequently, in the trivariate setting discussed in this master thesis the random vector $\Delta = (\Delta_1,\Delta_2,\Delta_3)$ follows a multinomial distribution with parameters $\pi_{...}$.

Chapter 5

Misspecifications in Frailty Models

The frailty models presented in Chapter 4 are applied to describe the disease mechanisms of specific infections while accounting for individual heterogeneity. However, one of the key underlying assumptions in this context is that infection with the pathogen occurs at most once within the human's life for the diseases under study. Nevertheless, some infections are characterized by the possibility of reinfection after recovery from a past infection. The latter implies that the seroprevalence of such diseases exhibits patterns which can not be observed under the assumption of lifelong immunity after experiencing infection. Especially, although the seroprevalence is monotonically increasing with increasing age for immunizing infections, one generally does not observe this pattern for diseases for which reinfections are possible. In this chapter, the age of an individual is denoted by a whereas the calendar time is represented by t which is different from the notation used in the previous chapters.

5.1 Mathematical Models for Infectious Disease Transmission

Deterministic transmission models describe infectious disease dynamics through partitioning of the population into different disease states or compartments. Therefore, these mathematical models are also referred to as compartmental models. The flows between the different disease states within the transmission model are usually described by means of a set of partial differential equations, expressing the time- and age-dependent evolution of the proportion of individuals in each compartment. Although mathematical transmission models are very useful to model the spread of infections in large populations, they are inappropriate to use in case of small populations. For this purpose, stochastic models were developed which are not discussed within this master thesis. Our aim is to present models that estimate the age-related heterogeneity inherent to the spread of infections that allow for reinfection in large populations.

First of all, one of the most basic deterministic models to capture disease transmission is introduced. The so-called SIR transmission model assumes individuals to be in one of the three compartments: S, I and R. All individuals are born into state S in which they are susceptible to infection. As they age from birth onwards, they may become infected and infectious to others (state I). After the infectious individuals are recovered (state R), they are not longer able to transmit the infection and are assumed to be immune for life. The number of individuals in each disease state can be expressed as a function of age and time by $S^*(a,t)$, $I^*(a,t)$ and $R^*(a,t)$. The superscript is added to avoid confusion with the notation used for the survival functions in the preceeding chapters. The total number of individuals of a certain age a at a point in time t is denoted by $N(a,t) = S^*(a,t) + I^*(a,t) + R^*(a,t)$. The set of partial differential equations describing the flows withing the SIR

transmission model is given by:

$$\begin{cases} \frac{\partial S^*(a,t)}{\partial a} + \frac{\partial S^*(a,t)}{\partial t} = -(\lambda(a,t) + \mu(a,t))S^*(a,t), \\ \frac{\partial I^*(a,t)}{\partial a} + \frac{\partial I^*(a,t)}{\partial t} = \lambda(a,t)S^*(a,t) - (\gamma(a,t) + \mu(a,t))I^*(a,t), \\ \frac{\partial R^*(a,t)}{\partial a} + \frac{\partial R^*(a,t)}{\partial t} = \gamma(a,t)I^*(a,t) - \mu(a,t)R^*(a,t). \end{cases} \tag{5.1}$$

In the system of PDEs, the natural mortality rates are represented by $\mu(a,t)$, the force of infection is traditionally denoted by $\lambda(a,t)$ and the age- and time-dependent recovery rates are equal to $\gamma(a,t)$. In general, the disease-specific mortality is neglected which is a valid assumption for most of the childhood diseases in developed countries.

The equations in (5.1) are computationally difficult to work with and therefore some simplifying assumptions are made in order to estimate age-specific transmission dynamics (Goeyvaerts, 2011). First of all, the population is assumed to be in endemic and demographic equilibrium meaning that the disease is in endemic steady state and that the population age distribution is stationary. In addition, the number of births and deaths is taken constant over time and exactly balanced. Under the assumption of endemic equilibrium (i.e. time homogeneity), the time dependency cancels out such that the set of partial differential equations in (5.1) reduces to a system of ordinary differential equations (ODEs):

$$\begin{cases} \frac{dS^*(a)}{da} = -(\lambda(a) + \mu(a))S^*(a), \\ \frac{dI^*(a)}{da} = \lambda(a)S^*(a) - (\gamma(a) + \mu(a))I^*(a), \\ \frac{dR^*(a)}{da} = \gamma(a)I^*(a) - \mu(a)R^*(a). \end{cases} \tag{5.2}$$

Furthermore, the following differential equation describes the stationary age distribution $N(a)$:

$$\frac{dN(a)}{da} = -\mu(a)N(a), \tag{5.3}$$

such that:

$$N(a) = \frac{N}{L^*} e^{-\int_0^a \mu(u)du},$$

where L^* is the life expectancy. More information concerning the derivation of the life expectancy is found in Appendix A.1. The latter expression can be easily derived by solving the differential equation (5.3) under the initial condition that the number of newborns and deaths are equally balanced. Note that $e^{-\int_0^a \mu(u)du}$ corresponds to the survival function introduced in Section 3.1. The system of ODEs can be easily expressed in terms of the proportion of individuals in each disease state. Since the proportion of individuals in the susceptible class corresponds to the survival function in the context of disease events, the proportions are denoted by S, I and R. In that way, we have:

$$\begin{cases} \frac{dS(a)}{da} = -\lambda(a)S(a), \\ \frac{dI(a)}{da} = \lambda(a)S(a) - \gamma(a)I(a), \\ \frac{dR(a)}{da} = \gamma(a)I(a). \end{cases} \tag{5.4}$$

The set of ordinary differential equations can be solved in a straightforward way by means of the method of separation of variables. The expression for the survival function then becomes $S(a)$

$= e^{-\int_0^a \lambda(u)du}$ which is the same formula as the one presented in Chapter 3. The SIR model is just one, though fundamental, example of a mathematical deterministic model used to capture the disease dynamics (Goeyvaerts, 2011). In practice, many extensions of the SIR model are found in which the number and interpretations of the compartments differ. Although the SIR model is the most frequently used model in literature, it may be interesting to relax the assumption of lifelong immunity as mentioned earlier. As it is clear from the previous discussion, the frailty models presented previously support the idea of infections governing an SIR transmission model. An obvious and straightforward extension thereof is the SIRS transmission model, allowing for reinfections with the pathogen. The SIRS model allows individuals that already experienced infection in the past to become susceptible again some time after recovery. Hence, the model is a valuable candidate to describe the behaviour of diseases for which multiple infections can occur during one lifetime.

The SIRS model allows for loss of disease-acquired immunity and potential reinfection as already highlighted. Individuals are assumed to move again from the recovered state to the susceptible state at a rate $\sigma(a)$. Under the same assumptions as formulated above, the set of ordinary differential equations in terms of the fraction of individuals in each compartment simplifies to:

$$\begin{cases} \frac{dS(a)}{da} = -\lambda(a)S(a) + \sigma(a)R(a), \\ \frac{dI(a)}{da} = \lambda(a)S(a) - \gamma(a)I(a), \\ \frac{dR(a)}{da} = \gamma(a)I(a) - \sigma(a)R(a). \end{cases} \tag{5.5}$$

As stated in Goeyvaerts *et al.* (2011), one can solve the corresponding set of equations by assuming that $R(a) \approx 1 - S(a)$ and derive an expression for the proportion of susceptible individuals at a specific age (i.e. the survival function). The set of ODEs is solved using the method of integrating factors under the boundary condition $S(0) = 1$:

$$\begin{aligned} S(a) &= e^{-\int_0^a \{\lambda(u)+\sigma(u)\}du} \left(1 + \int_0^a \sigma(u) e^{-\int_u^0 \{\lambda(v)+\sigma(v)\}dv} du\right) \\ &= e^{-\int_0^a \{\lambda(u)+\sigma(u)\}du} + \int_0^a \sigma(u) e^{-\int_u^a \{\lambda(v)+\sigma(v)\}dv} du. \end{aligned} \tag{5.6}$$

Equation (5.6) gives us the means to refine the frailty models as presented previously and to circumvent the strong assumption of lifelong immunity made therein. However, some remarks are to be formulated with respect to the previous expression. As the simple form of the survival function in the SIR situation is not longer present, one needs to rely on numerical integration methods to evaluate the integral part in equation (5.6). In the following section, we define the univariate frailty model in case of an SIRS-type infection. Multivariate extensions are possible but are not incorporated in this master thesis.

5.2 Univariate Frailty Model for an SIRS Infection

The univariate frailty model discussed in Chapter 4 accounts for individual heterogeneity in the acquisition of a single infection. Let us denote the frailty variable with respect to a specific SIRS infection as Z. As proposed by Farrington *et al.* (2001), the individual frailty Z acts multiplicatively on the baseline force of infection $\lambda_0(a)$ and the survival function derived in equation (5.6) can be extended to include the frailty term Z as follows:

$$S(a|Z) = e^{-\int_0^a \{Z\lambda_0(u)+\sigma(u)\}du} + \int_0^a \sigma(u) e^{-\int_u^a \{Z\lambda_0(v)+\sigma(v)\}dv} du. \tag{5.7}$$

The unconditional survival function has a complicated form due to the presence of the integral term as indicated in the preceeding section. Consequently, one can not write the unconditional survival function as simply the Laplace transform of Z evaluated at the cumulative baseline force of infection (see Section 4.1). Nevertheless, one can integrate out the frailty variable and simplify the resulting expression using the theorem of Fubini:

$$\begin{aligned} S(a) &= e^{-\int_0^a \sigma(u)du} e^{-Z\int_0^a \lambda_0(u)du} + \int_0^\infty \int_0^a \sigma(u) e^{-\int_u^a \{Z\lambda_0(v)+\sigma(v)\}dv} du \\ &= e^{-\int_0^a \sigma(u)du} \mathbf{L}(M_0(a)) + \int_0^a \sigma(u) e^{-\int_u^a \sigma(v)dv} \mathbf{L}(M_0(a) - M_0(u))du. \end{aligned} \tag{5.8}$$

where $\mathbf{L}$ equals the Laplace transform of the frailty variable Z and M_0 represents the cumulative baseline force of infection as before. The implementation of the univariate frailty model in the likelihood framework for an infection with SIRS dynamics requires the specification of a parametric form for the baseline force of infection as well as for the age-dependent replenishment rate $\sigma(a)$. Furthermore, one of the proposed frailty distributions in Chapter 4 for the random variable Z needs to be selected.

5.3 A More Realistic Approach

In the discussion included in this section, we require the definition of the effective contact function or transmission rate $\beta(a,a')$. Farrington *et al.* (2001) identified the effective contact function as one of the most important quantities in infectious disease modelling. The effective contact function drives the infection process and is defined as the per capita rate at which a susceptible individual of age a makes an effective contact with an infectious individual of age a'. In this definition, an effective contact is defined as an event such that an infectious individual of age a' infects a susceptible individual of age a. Farrington *et al.* (2001) suggested a decomposition of the effective contact function into the contact rate $c(a,a')$ and a proportionality factor $q(a,a'|c)$ given a contact which corresponds to degree of infectiousness and susceptibility of the individuals. The mass action principle relates the effective contact function to the force of infection. Under the assumption that the mean duration of infectiousness is short as compared to the time scale on which the transmission and mortality rates change, the mass action principle can be written in terms of the following integral equation:

$$\lambda(a) = \frac{ND}{L^*} \int_0^\infty \beta(a,a')\lambda(a')S(a')e^{-\int_0^{a'} \mu(u)du} da', \tag{5.9}$$

where $\lambda(a)$ equals the age-dependent force of infection, $\beta(a,a')$ is the effective contact function or transmission rate between an individual of age a and a', $S(a')$ is the proportion of susceptible individuals of age a' in the population and $\mu(a)$ is the natural mortality rate as before. Furthermore, the population size is assumed to be constant, denoted by N, D equals the mean duration of infectiousness and L^* is the life expectancy in the population under study. A more elaborate explanation on the important results with respect to the mass action principle and force of infection is found in Appendix A.1. The interested reader is also referred to the excellent paper of Farrington *et al.* (2001) regarding these concepts.

An alternative approach to the problem introduced in Section 5.2 is to use social contact data in combination with seroprevalence data to estimate the force of infection and the replenishment rate (Wallinga *et al.*, 2006). Mathematical models describing the spread of infectious diseases require assumptions concerning the underlying transmission process. Infectious diseases are transmitted via different routes and the acquisition of infections is related to the social contact behaviour of individuals. The frequency and intensity of human social interactions depend on age. Anderson and

May (1991) proposed the use of preconditioned mixing patterns in combination with serological data to estimate the transmission rates in mathematical models, i.e. the age-specific average per capita rate at which two individuals make an effective contact, per time unit. The main disadvantage of this method is the sensitivity of the results with respect to the assumed mixing patterns (Goeyvaerts, 2011).

Therefore, Wallinga *et al.* (2006) argued that serological data can be augmented with social contact data in order to estimate age-dependent transmission parameters. Consequently, it was assumed that the transmission rates are proportional to rates of conservational contact as demonstrated by Farrington *et al.* (2001). In the remaining part of this chapter, the transmission rates $\beta(a,a')$ are decomposed into the contact rate $c(a,a')$ and the proportionality factor $q(a,a'|c)$. For simplicity, one often assumes a constant proportionality factor q. The constant proportionality (CP) assumption is considered throughout this thesis. Using the decomposition of the effective contact function and assuming type I mortality, the integral equation (5.9) simplifies to:

$$\lambda(a) = \frac{ND}{L^*} \int_0^{L^*} qc(a,a')\lambda(a')S(a')da'. \tag{5.10}$$

Type I mortality implies that for $a \leq L^*$, $\mu(a)$ equals 0 and infinity otherwise. In the situation of an SIRS infection, the fraction of susceptible individuals is given by equation (5.6) when individual frailties are ignored. Estimating the transmission rates using seroprevalence data can not be done analytically since the integral equation (5.10) in general has no closed form solution. However, the equation can be solved numerically by turning towards discrete age intervals and assuming a constant force of infection in each interval. Denote the first age interval by $(a_{[1]}, a_{[2]})$ and the jth age interval by $[a_{[j]}, a_{[j+1]})$, $j = 2, ..., J$, where $a_{[1]} = 0$ and $a_{[J+1]} = L^*$. Under the assumption of a constant replenishment rate σ and making use of formula (5.6), the age-dependent proportion of susceptibles can be discretized as follows:

$$\begin{aligned} S(a) = {} & e^{-\sum_{l=1}^{j-1}(\lambda_l+\sigma)\left(a_{[l+1]}-a_{[l]}\right)-(\lambda_j+\sigma)\left(a-a_{[j]}\right)} \\ & + \sum_{l=1}^{j-1} \frac{\sigma}{\sigma+\lambda_l} \left(e^{-\sum_{m=l+1}^{j-1}(\lambda_m+\sigma)\left(a_{[m+1]}-a_{[m]}\right)-(\lambda_j+\sigma)\left(a-a_{[j]}\right)} \left[1 - e^{-(\lambda_l+\sigma)\left(a_{[l+1]}-a_{[l]}\right)}\right]\right) \\ & + \frac{\sigma}{\sigma+\lambda_j}\left[1 - e^{-(\lambda_j+\sigma)\left(a-a_{[j]}\right)}\right], \end{aligned} \tag{5.11}$$

if a belongs to the jth age interval. Furthermore, the integral equation (5.10) can be discretized using equation (5.11) to obtain the force of infection for age class i ($i = 1, ..., J$):

$$\begin{aligned} \lambda_i = {} & \frac{ND}{L^*} \sum_{j=1}^{J} \beta_{ij} \frac{\lambda_j}{\lambda_j+\sigma} \left\{ \left(e^{-\sum_{l=1}^{j-1}(\lambda_l+\sigma)\left(a_{[l+1]}-a_{[l]}\right)} - e^{-\sum_{l=1}^{j}(\lambda_l+\sigma)\left(a_{[l+1]}-a_{[l]}\right)} \right) \right. \\ & + \left[1 - e^{-(\lambda_j+\sigma)\left(a_{[j+1]}-a_{[j]}\right)}\right] \sum_{l=1}^{j-1} \frac{\sigma}{\sigma+\lambda_l} \left\{ e^{-\sum_{m=l+1}^{j-1}(\lambda_m+\sigma)\left(a_{[m+1]}-a_{[m]}\right)} \left[1 - e^{-(\lambda_l+\sigma)\left(a_{[l+1]}-a_{[l]}\right)}\right] \right\} \\ & \left. + \frac{\sigma}{\sigma+\lambda_j}\left[e^{-(\lambda_j+\sigma)\left(a_{[j+1]}-a_{[j]}\right)} - 1\right] + \sigma \right\} \end{aligned} \tag{5.12}$$

In order to estimate the constant proportionality factor q and the constant replenishment rate σ from seroprevalence data, an iterative procedure is applied as described in Farrington *et al.* (2001) and Kanaan and Farrington (2005). First, some plausible starting values for q and σ are selected and one solves the equation (5.12) iteratively for the piecewise constant force of infection λ_i, $i = 1, ..., J$.

Again the estimates for the piecewise constant force of infection are contrasted to the serological data in order to update the proposed values for q and σ. This procedure is repeated under the constraints $q \geq 0$ and $\sigma \geq 0$ until the binomial likelihood function for univariate current status data (equation (3.8)) is maximized (Goeyvaerts, 2011).

The approach presented previously can be extended in order to account for individual heterogeneity. Each individual differs in the rate at which it acquires a specific infection in the population. Therefore, Farrington *et al.* (2001) proposed a straightforward extension of the mass action principle in which the effective contact function is extended with individual frailty terms. The augmented effective contact function is denoted by $\beta(a,u;a',v)$ representing the per capita rate at which an individual of age a and frailty u makes an effective contact with a person of age a' and frailty v. Under the simple assumption of multiplicative frailty terms, one obtains $\beta(a,u;a',v) = uv\beta_0(a,a')$. Furthermore, the frailties act multiplicatively on the baseline force of infection as a consequence of the previous assumption. Given these simplifying conditions, the integral equation in (5.9) can be reformulated as stated by Farrington *et al.* (2001):

$$\begin{aligned} \lambda(a,u) &= \frac{ND}{L^*}\int_0^\infty \int_0^\infty uv\beta_0(a,a')v\lambda_0(a')S(a'|v)e^{-\int_0^{a'}\mu(u)du}da'f(v)dv \\ &= \frac{ND}{L^*}\int_0^\infty v^2 f(v)\int_0^\infty u\beta_0(a,a')\lambda_0(a')S(a'|v)e^{-\int_0^{a'}\mu(u)du}da'dv, \end{aligned} \tag{5.13}$$

where $f(.)$ is the p.d.f. of the frailty variable v and $\lambda(a,u) = u\lambda_0(a)$. The functions $\lambda_0(a)$ and $\beta_0(a,a')$ are called the baseline force of infection and baseline effective contact function, respectively. A thorough explantion on the deduction of the augmented integral equation (5.13) is included in Appendix A.1. concerning the mass action principle. The simplified expression which is used in the thesis is based on type I mortality thereby reducing equation (5.13) to:

$$\lambda_0(a) = \frac{ND}{L^*}\int_0^\infty v^2 f(v)\int_0^{L^*}\beta_0(a,a')\lambda_0(a')S(a'|v)da'dv. \tag{5.14}$$

Analogue to the discretized version of the proportion of susceptible individuals of age a in the absence of individual frailties (see equation (5.11)), the discretized version of the conditional survival function $S(a|v)$ can be obtained by replacing the piecewise constant force of infection λ_j by $v\lambda_j$. Furthermore, the discretized version of the mass action principle including individual heterogeneity describes the piecewise constant baseline force of infection in age class i (i = 1, ..., J):

$$\begin{aligned} \lambda_{0i} = \frac{ND}{L^*}\int_0^\infty vf(v)\Bigg\{\sum_{j=1}^{J}\beta_{0ij}\frac{v\lambda_j}{v\lambda_j+\sigma}\Bigg\{\Big(e^{-\sum_{l=1}^{j-1}(v\lambda_l+\sigma)\left(a_{[l+1]}-a_{[l]}\right)} - e^{-\sum_{l=1}^{j}(v\lambda_l+\sigma)\left(a_{[l+1]}-a_{[l]}\right)}\Big) \\ + \Big[1-e^{-(v\lambda_j+\sigma)\left(a_{[j+1]}-a_{[j]}\right)}\Big]\sum_{l=1}^{j-1}\frac{\sigma}{\sigma+v\lambda_l}\Big\{e^{-\sum_{m=l+1}^{j-1}(v\lambda_m+\sigma)\left(a_{[m+1]}-a_{[m]}\right)}\Big[1-e^{-(v\lambda_l+\sigma)\left(a_{[l+1]}-a_{[l]}\right)}\Big]\Big\} \\ + \frac{\sigma}{\sigma+v\lambda_j}\Big[e^{-(v\lambda_j+\sigma)\left(a_{[j+1]}-a_{[j]}\right)}-1\Big]+\sigma\Bigg\}\Bigg\}dv \end{aligned} \tag{5.15}$$

If one assumes that the baseline effective contact function β_{0ij} equals a constant proportionality factor q_0 times the contact rates c_{ij}, one can estimate the piecewise constant baseline force of infection using the iterative procedure introduced above (Kanaan and Farrington, 2005). Serological data is augmented with social contact data to be able to compute maximum likelihood estimates for the unknown model parameters q_0, σ and the frailty variance σ_f^2. The latter notation is preferred to avoid confusion with the notation of the constant replenishment rate σ.

Chapter 6

Simulation Results and Data Application

In this chapter, the results of our statistical analysis concerning the serological data introduced in Chapter 2 are presented. The main interest is in the comparison between different strategies to account for individual heterogeneity in the acquisition of infections. Univariate, shared and correlated gamma frailty models are considered in order to model the seroprevalence of CMV, PVB19 and HAV, and consequently to estimate the underlying age-dependent force of infection under the assumption of time homogeneity. As the focus of this thesis is on parametric frailty modelling, we use different parametric forms for the baseline force of infection and illustrate the use of different frailty distributions. The Akaike Information Criterion (Akaike, 1973) is used to evaluate the performance of the different frailty models in modelling the seroprevalence. Furthermore, results of the univariate SIRS frailty model are contrasted to those of the univariate SIR frailty model. The statistical analysis presented in this chapter is performed using the free software package R, version 2.11.0 in which the function optim was used for general optimization purposes. In addition, the software package SAS version 9.2. is applied to verify the results obtained with R. Especially, the SAS procedure PROC NLMIXED allowed us to implement the frailty models by specification of the corresponding likelihood functions. Details about optimization algorithms and numerical integration are included in Appendix A.

6.1 Frailty Models for Immunizing Infections

First of all, the univariate gamma frailty model is used to account for individual heterogeneity in the acquisition of CMV, PVB19 and HAV. The univariate gamma frailty model is described in Section 4.1. The baseline force of infection is assumed to be of the Gompertz type for the three infections under study, i.e. $\lambda_{i0}(t_i) = a_i e^{b_i t_i}$, $i = 1, 2, 3$ (see Subsection 3.4.3). The Gompertz baseline hazard function was selected based on its improved fit to the serological data as compared to frailty models with Weibull or exponential baseline hazard functions. The results of the univariate gamma frailty model are incorporated in Table 6.1. The AIC-value corresponding to the univariate gamma frailty model was found to be equal to 9956.7. The correlation coefficients are set equal to zero in the univariate frailty model assuming independence between pairs of frailty variables Z_i, $i = 1, 2, 3$. Furthermore, the parameter estimate for a_1 and its standard error estimate are very large. These extreme parameter values for both a_1 and b_1 induce the estimated seroprevalence to level-off very quickly. This is already a first indication of the limitations with respect to the application of traditional frailty models in the context of infections that do not confer lifelong immunity as stated in Chapter 5.

In Table 6.1, parameter estimates and corresponding standard error estimates are displayed as

Table 6.1: Gamma frailty models with Gompertz baseline hazard functions. Parameter estimates and corresponding standard error estimates of the Gompertz baseline hazard function, estimated frailty variances, estimated correlation coefficients and AIC-values of the frailty models.

	Univariate	Shared	Correlated
Parameter	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)
a_1	11.5988 (57.7456)	0.2837 (0.0941)	0.3286 (0.1391)
b_1	-0.3002 (0.3420)	-0.6816 (0.2311)	-0.7226 (0.3163)
a_2	0.1158 (0.0075)	0.1197 (0.0079)	0.1173 (0.0077)
b_2	-0.0687 (0.0070)	-0.0638 (0.0070)	-0.0665 (0.0072)
a_3	0.0095 (0.0008)	0.0097 (0.0008)	0.0104 (0.0009)
b_3	0.0223 (0.0034)	0.0227 (0.0035)	0.0258 (0.0107)
σ_1^2	15.8959 (16.5598)	0.1215 (0.0327)	0.5505 (0.1578)
σ_2^2	0.0001 (0.0054)	0.1215 (0.0327)	0.0507 (0.0398)
σ_3^2	0.0003 (0.0074)	0.1215 (0.0327)	0.6251 (0.7088)
ρ_{12}	0.0000 (-)	1.0000 (-)	0.2981 (0.1281)
ρ_{13}	0.0000 (-)	1.0000 (-)	0.9363 (0.4221)
ρ_{23}	0.0000 (-)	1.0000 (-)	0.2842 (0.1799)
AIC	9956.7	9937.2	9916.8

well for the shared and correlated gamma frailty models. Based on the Akaike Information Criterion, the unrestricted correlated gamma frailty model seems to outperform the univariate and shared gamma frailty models. Furthermore, it turns out that the estimates for the parameters of the Gompertz baseline hazard functions are similar for the three models, except for those representing the baseline infection hazard of CMV. Moreover, the correlation between the frailty variables for CMV and HAV is approximately equal to one. Therefore, one can conclude that transmission of the pathogens for both CMV and HAV is similar based on the correlated trivariate gamma frailty model. Since the univariate and shared gamma frailty models are in fact different versions of the unrestricted correlated gamma frailty model, restricting the values of the correlation coefficients and frailty variances (see Table 6.1), we enlarge our analysis by investigation of the performance of other restricted correlated gamma frailty models. In Table 6.2, parameter estimates and standard error estimates are summarized for the submodels of the unrestricted correlated gamma frailty model.

The first model assumes that the correlation between the frailty variables of CMV and PVB19 is exactly one and that they have equal variances. In other words, the frailty variable is shared with respect to the acquisition of CMV and PVB19. The unrestricted correlated gamma frailty model as formulated in Section 4.3 is therefore fitted with $Y_1 = Y_2$ and $k_4 = k_5 = k_6 = 0$. The latter restrictions imply that the frailty variances σ_1^2 and σ_2^2 are equal. The AIC-value does not indicate an improvement in model fit as compared to the unrestricted correlated gamma frailty model. Alternatively, the correlation coefficient ρ_{13} is set equal to one in the second model in combination with equal frailty variances. This restricted frailty model gives rise to a smaller AIC-value (AIC-value = 9909.2) compared to the one obtained in case of the unconstrained model. Furthermore, the resulting estimates confirm the conclusions derived from the unconstrained model in which the correlation coefficient between frailties for CMV and HAV was found to be approximately equal to one. Results concerning the model with $\rho_{23} = 1$ and $\sigma_2^2 = \sigma_3^2$ are omitted since this model performed worse (AIC-value = 9944.1) as compared to the other models. The third model assumes that $k_1 = k_3$ and $k_4 = k_5 = k_6 = 0$ which implies the frailty variances of Z_1 and Z_3 are equal. Model 3 outperforms the other submodels presented in Table 6.2 based on the Akaike Information Criterion.

Finally, results are included for the frailty model in which the frailties exhibit only an overal shared component Y_0 and an infection-specific term Y_l ($l = 1, 2, 3$). The parameter estimates for the fourth model are close to the ones from the third model. This conclusion is supported by the fact that the AIC-values of these two models only differ by two, resulting from one extra parameter to be estimated in the fourth model (i.e. $k_1 \neq k_3$).

Table 6.2: Restricted Correlated Gamma frailty models with Gompertz baseline hazard functions. Parameter estimates and corresponding standard error estimates of the Gompertz baseline hazard function, estimated frailty variances, estimated correlation coefficients and AIC-values of the frailty models.

	$\sigma_1^2 = \sigma_2^2$, $\rho_{12} = 1$	$\sigma_1^2 = \sigma_3^2$, $\rho_{13} = 1$	$\sigma_1^2 = \sigma_3^2$	$k_4 = k_5 = k_6 = 0$
Parameter	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)
a_1	0.2737 (0.0783)	0.2820 (0.0743)	0.3000 (0.1129)	0.3160 (0.1169)
b_1	-0.6553 (0.1918)	-0.6156 (0.1666)	-0.6601 (0.2561)	-0.6921 (0.2636)
a_2	0.1197 (0.0079)	0.1173 (0.0077)	0.1170 (0.0077)	0.1170 (0.0077)
b_2	-0.0637 (0.0070)	-0.0672 (0.0073)	-0.0660 (0.0071)	-0.0660 (0.0071)
a_3	0.0097 (0.0008)	0.0105 (0.0009)	0.0105 (0.0009)	0.0104 (0.0009)
b_3	0.0229 (0.0035)	0.0244 (0.0038)	0.0243 (0.0040)	0.0253 (0.0057)
σ_1^2	0.1237 (0.0335)	0.5338 (0.0925)	0.5437 (0.1194)	0.5610 (0.1696)
σ_2^2	0.1237 (0.0335)	0.0371 (0.0481)	0.0540 (0.0375)	0.0530 (0.0417)
σ_3^2	0.1362 (0.0571)	0.5338 (0.0925)	0.5437 (0.1194)	0.5893 (0.3126)
ρ_{12}	1.0000 (-)	0.2636 (0.1351)	0.3113 (0.1174)	0.2844 (0.1542)
ρ_{13}	0.9158 (0.1537)	1.0000 (-)	0.9876 (0.1055)	0.9485 (0.2136)
ρ_{23}	0.9158 (0.1537)	0.2636 (0.1351)	0.3113 (0.1174)	0.2914 (0.1341)
AIC	9942.0	9909.2	9908.9	9910.9

All fitted frailty models indicate a strong correlation between the frailty variables Z_1 and Z_3, and a small frailty variance for Z_2. Moreover, assuming that the frailty variances σ_1^2 and σ_3^2 are equal seems to improve the model fit. In Figure 6.1, the estimated seroprevalences based on the univariate, shared and unconstrained correlated gamma frailty models are plotted. The circles represent the observed seroprevalence for the different serological profiles. The blue, red and green lines are the results obtained from the univariate, shared and correlated gamma frailty model, respectively. Although the estimated seroprevalences roughly follow the general trends in the data, they are not able to describe the bumps in the seroprevalence π_{111}, π_{110} and π_{000} as shown in Figure 6.1. In addition, the marginal seroprevalences for the three infections are depicted in Figure 6.2. From this graph, the choice of the Gompertz baseline force of infection seems inadequate to describe the observed curvature in the marginal seroprevalences of CMV, PVB19 and HAV. However, as pointed out in Chapter 5, the underlying assumption of lifelong immunity after recovery in the presented frailty models is violated for CMV and PVB19. This statement is based on the decline of the observed seroprevalence after an initial increase which clearly does not support the idea of immunizing infections. We will focus on this item in Section 6.3.

Instead of the gamma frailty distribution used previously, one can also apply the power variance function distribution as described in Chapter 4. The PVF distribution is a generalized family of distributions including the gamma, inverse Gaussian and positive stable distribution. Therefore, it is interesting to consider the performance of the correlated PVF frailty model in the context of the

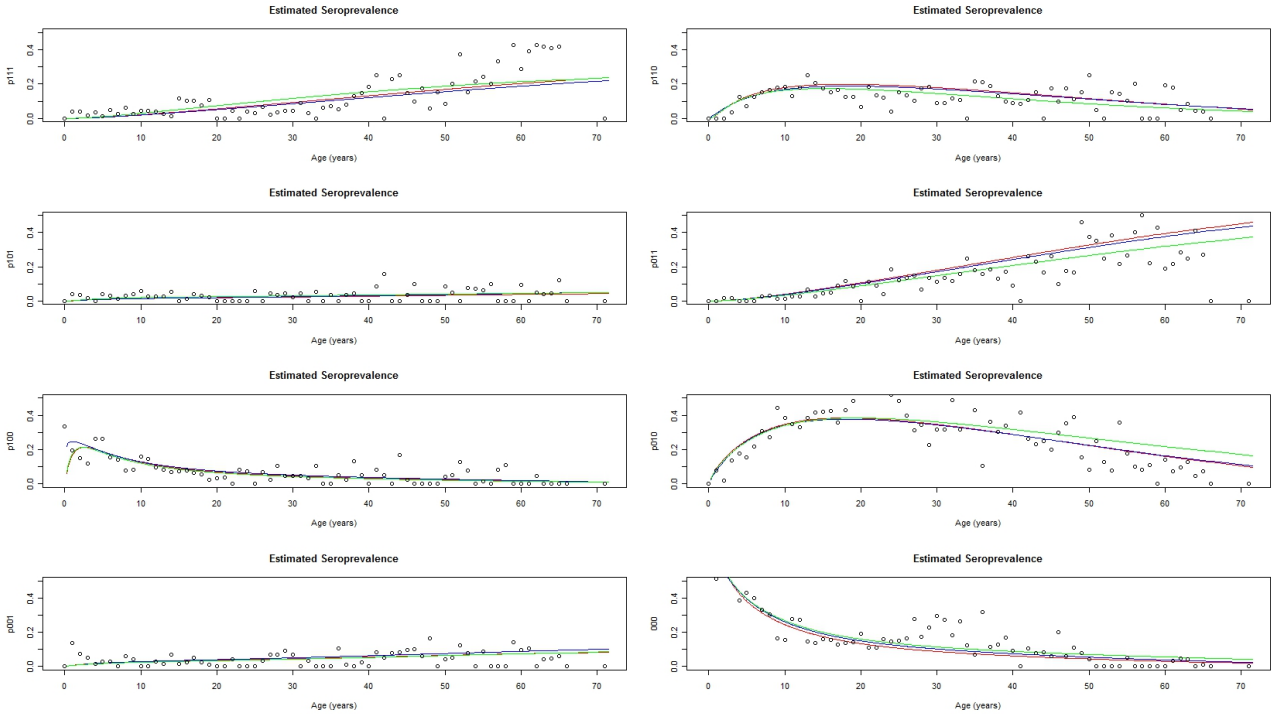

Figure 6.1: Estimated seroprevalences based on the gamma frailty models with Gompertz baseline hazard function.

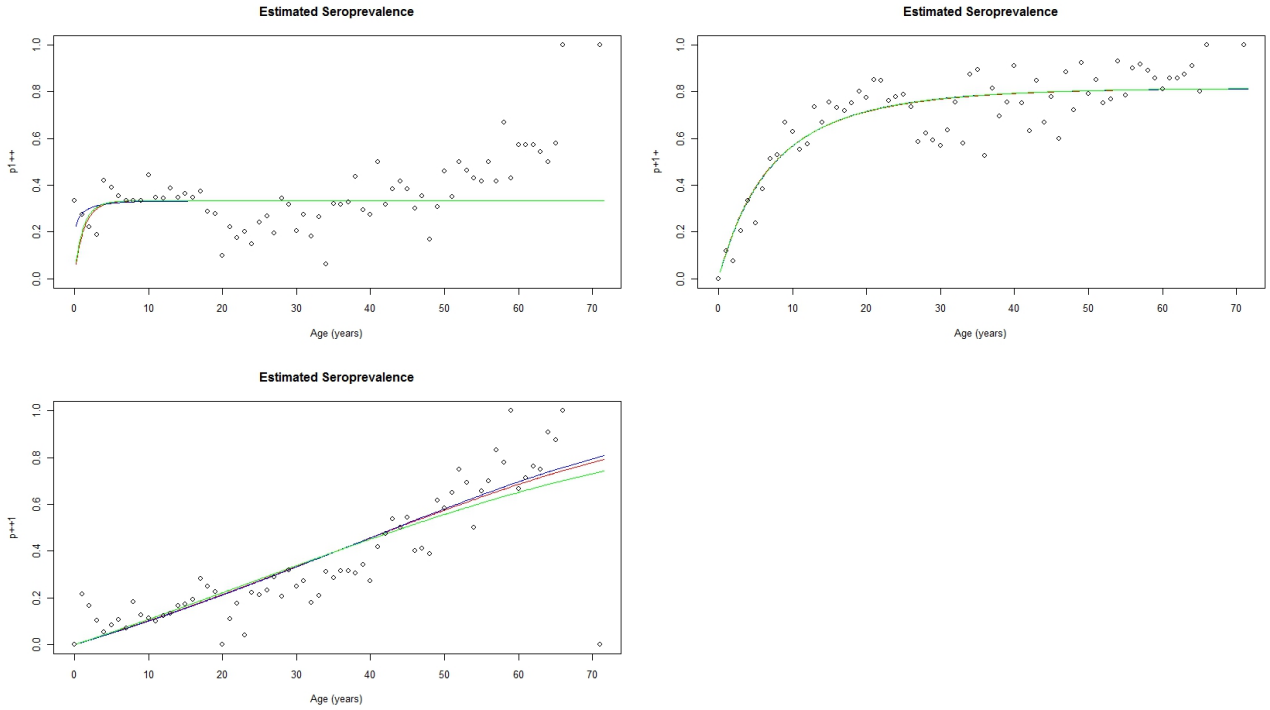

Figure 6.2: Estimated marginal seroprevalences based on the gamma frailty models with Gompertz baseline hazard function. CMV (upper left), PVB19 (upper right) and HAV (lower).

serological data at hand. The results of the trivariate correlated PVF frailty model with Gompertz baseline force of infection is included in Table 6.3. One observes that the correlated PVF frailty model outperforms the correlated gamma frailty model based on the Akaike Information Criterion. The estimated parameters of the Gompertz baseline force of infection are close to the estimates obtained in the unrestricted correlated gamma frailty model. Furthermore, the inverse Gaussian frailty distribution is derived when restricting the parameter value of η to 0.5 (see Subsection 4.4.3). Results of the correlated inverse Gaussian frailty distribution are included as well in Table 6.3. Among all presented frailty models, the smallest AIC-value is observed for the correlated inverse Gaussian frailty model.

Table 6.3: Correlated PVF and inverse Gaussian frailty models with Gompertz baseline hazard functions. Parameter estimates and corresponding standard error estimates of the Gompertz baseline hazard function, estimated frailty variances, estimated correlation coefficients and AIC-values of the frailty models.

	PVF	Inverse Gaussian
Parameter	Estimate (s.e.)	Estimate (s.e.)
a_1	0.3332 (0.1322)	0.4241 (0.1529)
b_1	-0.2879 (0.0828)	-0.3370 (0.0960)
a_2	0.1280 (0.0096)	0.1146 (0.0075)
b_2	-0.0768 (0.0093)	-0.0669 (0.0075)
a_3	0.0530 (0.0442)	0.0427 (0.0180)
b_3	0.0911 (0.0221)	0.0759 (0.0052)
σ_1^2	6.3646 (3.3299)	10.1570 (4.7892)
σ_2^2	0.0629 (0.0636)	0.0107 (0.0613)
σ_3^2	38.9558 (24.7869)	65.4141 (33.7265)
ρ_{12}	0.0991 (0.0587)	0.0020 (0.0085)
ρ_{13}	0.3411 (0.0458)	0.3940 (0.0380)
ρ_{23}	0.0401 (0.0250)	0.0041 (0.0140)
η	0.3605 (0.0677)	0.5000 (-)
AIC	9833.7	9830.6

Although the correlated frailty models have the advantage of being more flexible as compared to the shared frailty models, the implied correlation structure for the infection-specific frailties is constrained as mentioned in Section 4.3. For our application, one can easily observe that the correlation coefficients are at the implicit boundaries resulting from the additive decomposition of the frailty variables in the correlated frailty models. Therefore, one can argue that the proposed correlated frailty models are not flexible enough to model the true underlying correlation structure. The solution for this shortcoming of the presented correlated frailty models will probably involve a reformulation of the models without additive decomposition of the frailty terms. The latter reformulation in order to overcome the boundary constraints on the correlation coefficients is beyond the scope of this master thesis.

6.2 Simulation Study

In order to assess the performance of the unconstrained correlated gamma frailty model, we conducted a simulation study in which different parameter settings are considered. Furthermore, the aim is also to evaluate the amount of information that is lost when turning from time-to-event (TTE) data which is seldomly observed in practice to right censored (RC) data, and finally to current status (CS) data. Without loss of generality $n_s = 150$ serological datasets of size $n = 2890$ are generated for which the motivation lies in the sample size of the original serological dataset. First of all, we consider as a starting point the parameter estimates corresponding to the trivariate correlated gamma frailty model with Gompertz baseline hazard functions included in Table 6.1. Consequently, trivariate TTE data is generated using the following stepwise procedure. In the first step, the correlated gamma frailties (Z_1,Z_2,Z_3) are obtained via its additive component variables Y_l ($l = 0, 1, ..., 6$) which are assumed to follow a gamma distribution with parameters k_l and 1 as discussed in Section 4.3. Given a value z_i for the frailty variables Z_i, the event times t_i^* (independent values generated from the random variable T_i^*) ($i = 1, 2, 3$) are calculated using the transformation formula:

$$t_i^* = \frac{1}{b_i} \log(1 - \frac{b_i \log(u_i)}{a_i z_i}),$$

where u_i is sampled at random from a uniform distribution on the unit interval and a_i and b_i are to be replaced by the corresponding parameter estimates in Table 6.1. However, in the extreme parameter setting with respect to the baseline hazard of infection for CMV, the sampling of u_1 is restricted to the possible range of the survival function. Indeed, sampling values too close to zero makes no sense since these values can never been achieved given the selected parameter configuration. In addition, values for the censoring time variable C_i are generated from a uniform distribution on the interval ranging from zero to 72. Therefore, the censoring indicators δ_i are derived from the comparison of the true event times t_i^* and the censoring times c_i. Especially, δ_i equals 1 whenever $c_i > t_i^*$ and zero otherwise ($i = 1, 2, 3$).

The correlated gamma frailty model with Gompertz baseline hazard function is fitted for the simulated TTE, RC and CS data derived from each of the generated serological datasets. The likelihood functions are specified using the likelihood contributions included in Section 4.5. In Table 6.4, the simulation results are summarized for the original parameter setting obtained from the correlated gamma frailty model. These include the averaged parameter estimates and empirical standard error estimates for the frailty model parameters. One can observe that the averaged parameter estimates are in general quite far from the true values for all three data types. This observation is surprising especially for TTE data since TTE data consists of the most detailed information with respect to the parameters of the underlying simulation process and therefore is believed to yield estimates very close to the true parameter values. Moreover, the difficulty lies in the estimation of the parameters a_1 and b_1, which was already highlighted in the previous section. The value of a_1 is consistently overestimated whereas the averaged estimates of b_1 have the opposite sign of the true parameter value. A more interesting parameter configuration consists of less extreme parameter values for the parameters describing the first baseline force of infection. In Table 6.5, averaged parameter estimates and empirical standard error estimates are presented for the correlated gamma frailty model with Gompertz baseline hazard function. Furthermore, the simulated datasets used in the simulation study are generated under the parameter setting in which the true value for b_1 is changed to -0.0226 in contrast to the first parameter setting.

The averaged parameter estimates turn out to be much closer to the true parameter values as compared to the previous results. Although the averaged parameter estimates of the Gompertz baseline hazard functions can be considered stable, the empirical standard errors are increasing

Table 6.4: Simulation results for $n_s = 150$ simulated datasets of size $n = 2890$ with respect to time to event (TTE) data, right censored (RC) data and current status (CS) data.

		TTE	RC	CS
Parameter	True Value	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)
a_1	0.3286	0.8615 (0.0257)	0.8614 (0.0241)	0.9239 (1.0980)
b_1	-0.7226	0.0188 (0.0947)	0.0071 (0.0429)	0.5593 (0.9543)
a_2	0.1173	0.1281 (0.0040)	0.1324 (0.0039)	0.1272 (0.0086)
b_2	-0.0665	-0.0135 (0.0051)	-0.0198 (0.0032)	-0.0125 (0.0069)
a_3	0.0104	0.0107 (0.0009)	0.0107 (0.0006)	0.0101 (0.0015)
b_3	0.0258	0.0236 (0.0062)	0.0167 (0.0024)	0.0337 (0.0352)
σ_1^2	0.5505	0.0829 (0.1228)	0.0680 (0.0513)	0.7050 (1.1712)
σ_2^2	0.0507	0.0346 (0.0603)	0.0174 (0.0247)	0.0478 (0.0616)
σ_3^2	0.6251	0.5412 (0.2449)	0.0772 (0.1103)	1.0686 (2.0888)
ρ_{12}	0.2981	0.5758 (0.3348)	0.4585 (0.3270)	0.3057 (0.2937)
ρ_{13}	0.9363	0.3755 (0.2583)	0.9429 (0.1441)	0.8230 (0.2550)
ρ_{23}	0.2842	0.2132 (0.1713)	0.4562 (0.3270)	0.3021 (0.2909)

when going from TTE data to RC data, and turning finally to CS data. The latter consideration is important when comparing the performance of the correlated frailty model since TTE is often not available. An investigation of the effects with respect to the parameter estimates obtained from the restrictive CS data is therefore crucial. In addition, the variance estimates are in general not close to the true values as observed in Table 6.5. Although it seems difficult to estimate the variance and correlation parameters given the specified parameter configuration, we already pointed out a major limitation of the correlated gamma frailty model in the sense that the correlation coefficients are constrained. For the parameter setting under investigation, the correlation coefficients with respect to the frailty variables are at the boundary as demonstrated in Section 6.1.

Table 6.5: Simulation results for $n_s = 150$ simulated datasets of size $n = 2890$ with respect to time to event (TTE) data, right censored (RC) data and current status (CS) data. Less extreme parameter values for a_1 and b_1 are considered.

		TTE	RC	CS
Parameter	True Value	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)
a_1	0.3286	0.3316 (0.0133)	0.3285 (0.0119)	0.3208 (0.0487)
b_1	-0.0226	0.0087 (0.0058)	0.0006 (0.0094)	-0.0025 (0.0188)
a_2	0.1173	0.1286 (0.0035)	0.1332 (0.0038)	0.1262 (0.0068)
b_2	-0.0665	-0.0131 (0.0023)	-0.0183 (0.0034)	-0.0134 (0.0051)
a_3	0.0104	0.0102 (0.0005)	0.0102 (0.0006)	0.0102 (0.0011)
b_3	0.0258	0.0281 (0.0023)	0.0283 (0.0043)	0.0306 (0.0141)
σ_1^2	0.5505	0.6751 (0.0675)	0.6214 (0.0770)	0.5746 (0.2043)
σ_2^2	0.0507	0.0427 (0.0240)	0.0365 (0.0243)	0.0365 (0.0406)
σ_3^2	0.6251	0.7196 (0.0879)	0.7563 (0.2238)	0.8967 (0.7835)
ρ_{12}	0.2981	0.2229 (0.0774)	0.2051 (0.0852)	0.1986 (0.1355)
ρ_{13}	0.9363	0.8175 (0.0460)	0.8527 (0.0921)	0.8230 (0.1527)
ρ_{23}	0.2842	0.1955 (0.0697)	0.1901 (0.0790)	0.1817 (0.1253)

The simulation results concerning the same parameter setting and an increased sample size of

$n = 1000$ is documented in Appendix C. One final extension of the simulation study presented in this thesis covers the situation in which the correlation coefficients are away from the implied boundary. The averaged parameter estimates and empirical standard error estimates are included in Table 6.6. The parameter estimates are in line with those obtained in Table 6.5 and the correlation coefficients are quite closer to the true values in all three cases.

Table 6.6: Simulation results for $n_s = 150$ simulated datasets of size $n = 2890$ with respect to time to event (TTE) data, right censored (RC) data and current status (CS) data. Correlation coefficients away from the boundary contraints.

		TTE	RC	CS
Parameter	True Value	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)
a_1	0.3286	0.3384 (0.0129)	0.3349 (0.0136)	0.3510 (0.0851)
b_1	-0.0226	0.0136 (0.0045)	0.0090 (0.0110)	0.0141 (0.0285)
a_2	0.1173	0.1286 (0.0034)	0.1328 (0.0037)	0.1270 (0.0086)
b_2	-0.0665	-0.0132 (0.0023)	-0.0178 (0.0033)	-0.0134 (0.0055)
a_3	0.0104	0.0105 (0.0005)	0.0103 (0.0006)	0.0101 (0.0013)
b_3	0.0258	0.0252 (0.0031)	0.0270 (0.0078)	0.0344 (0.0261)
σ_1^2	0.5505	0.7235 (0.0544)	0.6835 (0.0887)	0.7417 (0.3249)
σ_2^2	0.0507	0.0413 (0.0242)	0.0389 (0.0262)	0.0382 (0.0422)
σ_3^2	0.6251	0.5984 (0.1241)	0.6844 (0.4342)	1.0909 (1.4430)
ρ_{12}	0.2981	0.2105 (0.0666)	0.2080 (0.0784)	0.1885 (0.1407)
ρ_{13}	0.3063	0.2849 (0.0567)	0.2999 (0.0802)	0.2807 (0.1720)
ρ_{23}	0.2842	0.2463 (0.0796)	0.2460 (0.1143)	0.2233 (0.1815)

6.3 SIRS Infections

In Chapter 5, the univariate frailty model is extended to capture the disease dynamics for infections that do not confer lifelong immunity. In Section 6.1, we pointed out that the estimated seroprevalence did not fit the observed data well, especially for CMV and PVB19. However, these infections are known to be nonimmunizing infections for which individuals can be reinfected during their lives. Therefore, we investigate the performance of the univariate frailty model under the assumption of an SIRS transmission process for both diseases. In order to illustrate the limitations of the traditional univariate frailty model (see Section 4.1) in this context, univariate current status data is generated under the assumption of an underlying SIRS infection process. Afterwards, the univariate frailty model for immunizing infections is fitted to the simulated data. For convenience, a Gompertz baseline force of infection with parameters $a_1 = 0.3286$ and $b_1 = -0.0226$ is assumed in addition to a constant replenishment rate $\sigma = 0.05$. The gamma frailty distribution is considered with a frailty variance σ_f^2 equal to 0.2. In the left panel of Figure 6.3, the simulated current status data is plotted together with the estimated seroprevalence using a univariate SIR frailty model (black solid line) and a univariate SIRS frailty model (red dashed line). One can easily observe that the univariate frailty model derived under the assumption of an SIR infection mechanism is not able to describe the simulated data well. In addition, the estimated frailty variance using the SIR model is substantially smaller than 0.2 which supports the previous conclusion. Alternatively, it is also very useful to check whether the univariate SIRS model yields a proper fit to generated current status data on immunizing infections. The results of this analysis are included in the right panel of Figure 6.3. The SIRS model (red dashed line) turned out to model the observed prevalence quite good while estimating the replenishment rate to be nearly equal to zero (3.5e-9). Furthermore, the estimated model parameters are almost equal for both the

SIR as SIRS models (see Appendix C.2.). Therefore, one can conclude that it is worthwhile to consider the univariate SIRS frailty model when the observed prevalence lacks a monotone pattern.

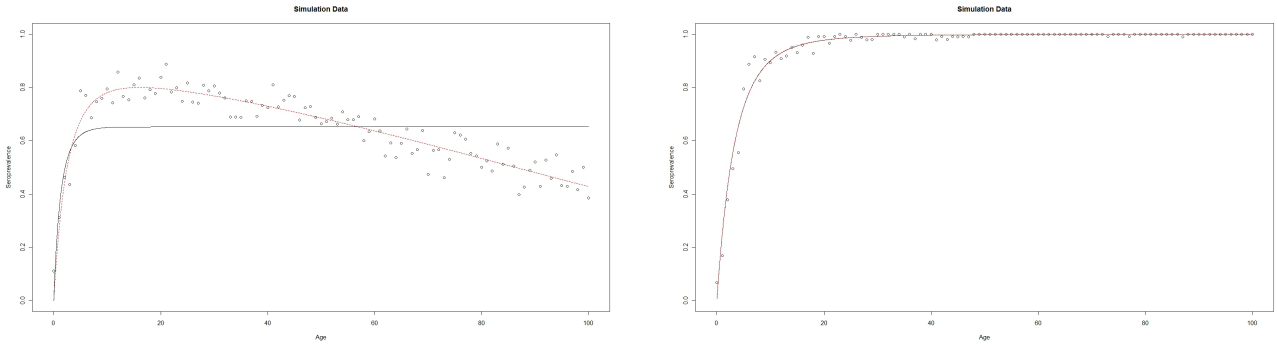

Figure 6.3: Estimated seroprevalence based on the univariate SIR (black solid line) and SIRS (red dashed line) gamma frailty models. The data is generated based on an SIRS infection process with Gompertz baseline force of infection (a_1 = 0.3286 and b_1 = -0.0226) and constant replenishment rate $\sigma = 0.05$ (left panel), and based on an SIR infection with Gompertz baseline force of infection (a_1 = 0.3286 and b_1 = -0.0226) (right panel).

Secondly, a parametric baseline force of infection $\lambda_0(a)$ and parametric age-dependent replenishment rate $\sigma(a)$ are assumed in the univariate SIRS model as stated in Section 5.2. Therefore, the Gompertz baseline force of infection (with parameters a_i and b_i, $i = 1, 2$) is applied together with a Gaussian replenishment rate, i.e.

$$\sigma(a) = \frac{d}{\sqrt{2\pi s^2}} e^{\frac{1}{2s^2}(a-m)^2}, \qquad m > 0, s > 0, d \geq 0,$$

where m represents the mean, s represents the standard deviation and d is an amplifying factor. In addition, the convenient gamma frailty distribution is assumed again. The parameter estimates and standard error estimates for the univariate gamma frailty models with respect to CMV and PVB19 infection are presented in Table 6.7. Furthermore, in Figure 6.4, the estimated seroprevalence for CMV and PVB19 is depicted based on the univariate frailty models with Gompertz baseline force of infection and Gaussian replenishment rate $\sigma(a)$. The estimated seroprevalence describes the data better as compared to the frailty models in Section 6.1 (see Figure 6.2). The AIC-values corresponding to these models are equal to 3578.9 and 3297.3 for CMV and PVB19, respectively. For comparison, the univariate frailty models derived under the assumption of immunizing infections have AIC-values equal to 3667.9 and 3304.2, respectively. Therefore, an improvement with respect to the model fit is observed by turning towards the extended univariate frailty models allowing for reinfections with the pathogens.

Although the presented univariate frailty models outperform the frailty models in Section 6.1, the parametric assumptions made to model the seroprevalence are based on visual examination of the data pattern. However, a Gompertz baseline force of infection seems to contradict the natural history of the infections and induces a somewhat strange assumption with respect to the parametric form of $\sigma(a)$. An alternative approach to model the seroprevalence of CMV and PVB19 and thereby to avoid the ad hoc method considered previously, is the use of social contact data to augment the serological data. In other words, contact rates supplement the serological data as described in Chapter 5.

The univariate SIRS gamma frailty model is considered and the procedure is illustrated for serological data on PVB19 and Belgian social contact data obtained from the POLYMOD survey.

Table 6.7: Parameter estimates and standard error estimates for the parametric Gompertz baseline force of infection $\lambda_0(a) = a_i e^{b_i a}$ and Gaussian replenishment rate $\sigma(a)$ in the SIRS univariate frailty model.

	CMV	PVB19
Parameter	Estimate (s.e.)	Estimate (s.e.)
a_1	1.5948 (0.0480)	-
b_1	0.0449 (0.0066)	-
a_2	-	0.1080 (0.0074)
b_2	-	-0.0202 (0.0105)
d	40.8586 (1.2264)	0.5629 (0.2695)
m	28.7919 (0.0537)	23.6923 (1.4679)
s	6.1796 (0.0715)	14.1951 (0.6810)
σ_1^2	11.6325 (0.3237)	-
σ_2^2	-	0.1933 (0.1001)
AIC	3578.9	3297.3

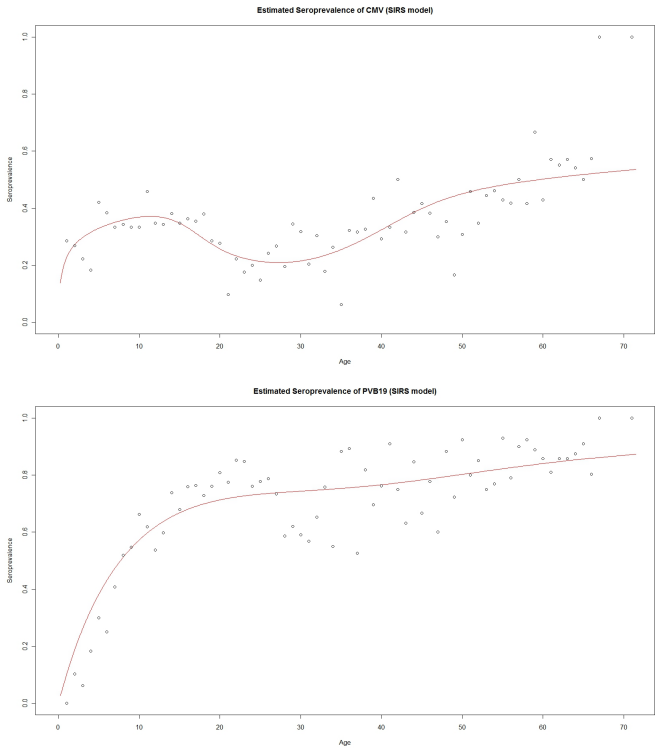

Figure 6.4: Estimated seroprevalence of CMV (upper graph) and PVB19 (lower graph) based on the univariate gamma frailty models for SIRS infections with Gompertz baseline force of infection and Gaussian replenishment rate.

Furthermore, the constant proportionality assumption (i.e. $q(a, a'|c) = q_0$) is supplemented with the assumption of a constant replenishment rate σ. The iterative procedure suggested by Farrington *et al.* (2001) and Kanaan and Farrington (2005) is applied to estimate the piecewise constant force of infection from contact data (see Section 5.3). Consequently, the model parameters q_0, σ and σ_f^2 are estimated based on the serology and the estimated piecewise force of infection. In Figure 6.5,

the estimated age-dependent seroprevalence of PVB19 is plotted (left panel) in combination with the estimated piecewise constant force of infection (right panel) driving this infection process. The AIC-value of the described model equals 3291.6 which is smaller than the values corresponding to the previous models for the seroprevalence of PVB19. The estimated model parameters are equal to $\hat{q}_0 = 0.2785$, $\hat{\sigma} = 0.0241$ and $\hat{\sigma}_f^2 = 0.1104$. After an initial increase in the estimated seroprevalence, the seroprevalence seems to stabilize after which it slightly decreases with age. Furthermore, the general trend in the estimated piecewise force of infection is a decrease with age wich seems reasonably due to the natural history of PVB19 infection. Despite the improved fit to the observed current status data, further refinements of the univariate SIRS model need to be considered and a more thorough examination of the derived results is essential. One of the first extensions of the basic SIRS model should allow for a nonconstant replenishment rate. Therefore, a piecewise constant replenishment rate could be useful instead of the constant one proposed earlier. Moreover, the constant proportionality assumption can be relaxed by assuming a parametric form for $q(a,a'|c)$.

These frailty models are quite complex and require a considerable amount of time to run. Consequently, further refinements of the specified model such as relaxing the constant proportionality assumption or the assumption of a nonconstant replenishment rate are not included in this master thesis. Since the univariate SIRS gamma frailty model shows to have advantages over the univariate SIR model, multivariate extensions are important to consider as well. These shared and correlated frailty models were not touched upon in this master thesis and are an interesting topic for further research. Although a deeper investigation with respect to frailty models in the context of nonimmunizing infections is required, the idea was to present some exploratory results in this section in order to illustrate the usefulness of such scientific research.

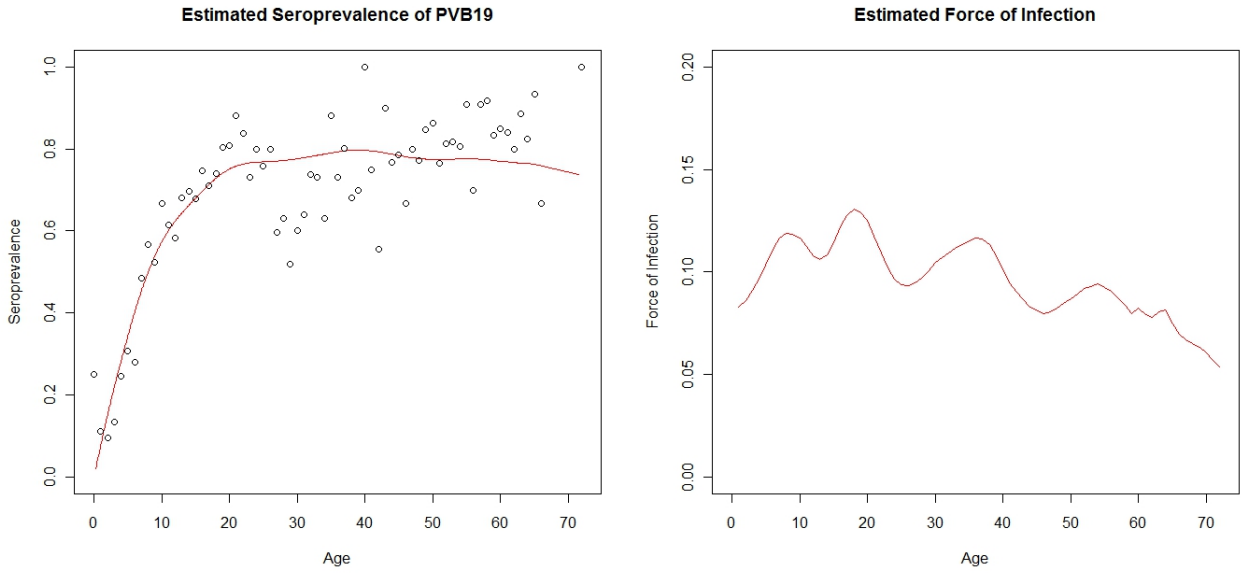

Figure 6.5: Estimated seroprevalence of PVB19 (left panel) based on the univariate SIRS gamma frailty model with Gompertz baseline force of infection and constant replenishment rate. The piecewise constant force of infection (right panel) is derived from social contact data augmenting the serology.

Chapter 7

Conclusion

Individuals within a population differ greatly in their responses towards specific events such as treatment, drugs or disease. Ordinary methods in survival analysis assume implicitly that the population under study is homogeneous and all individuals are having the same risk of experiencing these events. Consequently, more advanced techniques need to be considered in order to account for individual heterogeneity within a population. Frailty models constitute a specific area in survival analysis and provide a convenient way of introducing individual heterogeneity into models for survival data. In fact, in its most simple form, a frailty represents an unobserved random proportionality factor modifying the hazard function of an individual or related subjects. The concept of univariate frailty modelling goes back to the work of Beard (1959) considering different mortality models whereas the term *frailty* was suggested by Vaupel *et al.* (1979) in the univariate setting. Later, extensions of the univariate frailty models were made to analyse multivariate survival data. Shared frailty models and correlated frailty models were found very useful in describing the dependence of event times in a natural way. As frailty models have shown their strength in modelling multivariate survival data and individuals also differ in the acquisition of infections, frailty modelling has been introduced in infectious disease modelling as well. After the seminal work of Coutinho *et al.* (1999), many authors have studied the performance of the univariate, shared and correlated frailty models to describe infection dynamics within human populations. Hens *et al.* (2009) suggested the use of the shared and correlated gamma frailty model to model the seroprevalence of hepatitis A and B in the Belgian population. As time-to-event data is very hard to collect, they relied on cross-sectional serological data with respect to the presented infections. In this master thesis, we focused on the extension of the existing bivariate frailty models proposed by Hens *et al.* (2009) to trivariate models useful in modelling current status data on CMV, PVB19 and HAV. In addition, misspecifications of these trivariate frailty models in case of nonimmunizing infections were considered in more detail in the thesis. The derivations in the context of frailty misspecifications are illustrated by augmenting the serological data at hand with data on the frequency of close social contacts between individuals in the Belgian population.

Following the standard assumption of a gamma frailty distribution and selecting the frequently used Gompertz baseline hazard function, the trivariate correlated gamma frailty model was found to have an advantage over the shared and univariate models when applying these models to describe the serological data. Due to its flexibility, the correlated frailty model elevates the restrictive assumption of a shared frailty while allowing infections with different diseases to be dependent. The latter refers to the classical assumption of independence between events in univariate frailty models. Furthermore, relaxing the assumption of a gamma frailty distribution when using inverse Gaussian and PVF frailty distributions supported the idea that the correlated frailty model outperforms the shared and univariate models in an obvious way. Although some credit has been given to the correlated frailty models, one also needs to be aware of the limitations regarding these models. First of all,

the simplicity of the additive decomposition of the frailties comes with the price of an implicit constraint on the correlation coefficients. For the data application included in this thesis, it was shown that the estimated correlations are precisely at the boundary which leads to the conclusion that the correlated frailty models are not able to fully capture the disease mechanisms. Additional work needs to be done to avoid these restrictions on the correlation coefficients. The latter implies that the additive decomposition of the frailty terms needs to be reconsidered. Secondly, these models were fitted under the assumption of a Gompertz baseline force of infection for the three diseases. As illustrated within the thesis, the Gompertz parametric form is a questionable choice for the infections at hand. However, the choice was inspired by the comparison of the performance of models using other popular parametric assumptions for the baseline hazard. A simulation study was performed to evaluate the usefulness of the trivariate correlated frailty model. Therefore, three different parameter settings were selected to generate datasets which are then analyzed using the trivariate correlated gamma frailty model. The simulation results pointed out that the variance components and correlation coefficients are difficult to estimate, especially when the correlation coefficients are at the implicit boundaries. In general, the usefulness of the correlated model in case of trivariate current status data is underlined as information loss when turning from TTE data to CS data is rather limited.

The traditional trivariate frailty models were implemented assuming that the infections confer lifelong immunity. However, CMV and PVB19 clearly do not fulfill this requirement since reinfections with the pathogens are possible. Consequently, some refinements of the traditional frailty models are made. In this thesis, we limited ourselves to the univariate frailty model for infections with dynamics that can be successfully described using a mathematical SIRS transmission model. Nevertheless, multivariate models allowing for reinfections are an interesting topic for further research. The survival function for an SIRS infection was found to have an additional integral term. We turned to numerical integration techniques in order to calculate the unconditional survival function after inclusion of the frailty terms. Both a parametric approach as well as a more realistic alternative based on the combination of serology and social contact data are illustrated within this thesis. Although the parametric version in which a Gompertz baseline force of infection and Gaussian replenishment rate were assumed, performed quite well in modelling the observed seroprevalence data for CMV and PVB19, a more realistic approach is to estimate a piecewise constant force of infection based on social contact data. The mass action principle is thereby of crucial importance while relating the force of infection to an augmented effective contact function. In the present monologue, the augmented effective contact function is assumed to factorize into individual frailties acting multiplicatively on the baseline effective contact function. In addition, the baseline effective contact function reduces to the product of a constant proportionality factor times the age-dependent contact rates. Under the assumption of a constant replenishment rate, a univariate frailty model is fitted to the serology of PVB19 and social contact data. The univariate SIRS model outperforms both the univariate SIR frailty model and the parametric SIRS model. However, the aim in this master thesis was primarily to sketch the approach required to tackle the problem of modelling the seroprevalence for nonimmunizing infections.

Multivariate models have shown to be important in modelling the associations among infections such that additional extensions of the univariate SIRS models are essential. Moreover, the restrictive assumptions with respect to the constant replenishment rate and constant proportionality factor need to be relaxed. In addition to SIRS infections, numerous other mathematical transmission models are available in literature to describe infections without lifelong immunity and derivations of the survival functions will give rise to an extensive amount of frailty models. In conclusion, despite the promising results with respect to the SIRS frailty models, there exists some remaining work to be done in the future.

Appendix A

A.1. Mass Action Principle

One of the most important parameters in infectious disease modelling is the so-called force of infection. The force of infection, also called the hazard of infection, is the per-capita rate at which a susceptible individual acquires the infection from an infectious individual. The force of infection resembles the hazard function in survival analysis as described in Section 3.1 and is considered to be time-varying and age-dependent. Often, one assumes time homogeneity which implies that the infection under investigation is in endemic steady state at the population level, thereby allowing the disease incidence to undergo cyclical epidemics and fluctuating around a stationary average over time (Farrington *et al.*, 2001). Under the assumption of time homogeneity the force of infection can be estimated from serological data. An excellent overview of the available methods to estimate the force of infection from serological data is given in Hens *et al.* (2010). In general, let $\lambda(a)$ represent the age-dependent force of infection for an infection in steady state. Anderson and May (1991) stated that the age-dependent force of infection $\lambda(a)$ can be written as:

$$\lambda(a) = \int_0^\infty \beta(a, a')I^*(a')da', \tag{1}$$

where $\beta(a,a')$ represents the effective contact function or transmission rate. Farrington *et al.* (2001) argued that the infection process is driven by the effective contact function $\beta(a,a')$, representing the per capita rate at which an individual of age a' makes effective contacts with individuals of age a. An effective contact is an event such that an individual of age a' infects a person of age a, given that the person of age a' is infectious and the other is susceptible for the infection under study. Equation (1) reflects the so-called *mass action principle*. The mass action principle implicitly assumes that the susceptible and infectious individuals in the population mix homogeneously with each other. However, this signals immediately one of the major drawbacks of definition (1) since contacts are often directed and clustered which is not taken into account in the formula (Goeyvaerts, 2011).

The effective contact function $\beta(a,a')$ can be factorized in terms of the contact rate between individuals of age a' and age a, denoted by $c(a,a')$ and a proportionality factor related to the degree of infectiousness and susceptibility of the two individuals, given a contact between those, say $q(a,a'|c)$ (Farrington *et al.*, 2001):

$$\beta(a, a') = q(a, a'|c)c(a, a'). \tag{2}$$

The contact rate $c(a,a')$ is characterized by a definition of a contact relevant for the disease under consideration. In fact, the contact rate resembles the mixing pattern in the population with respect to the relevant route of transmission, as pointed out by Farrington *et al.* (2001). The proportionality factor $q(a,a'|c)$ determines an age-related transmission probability for the infectious antigen, given that a contact between a susceptible and infectious individual has occurred. Hence, the effective contact function is a joint feature of population characteristics and the organism. Furthermore, due to the nature of contacts between individuals, the contact rate $c(a,a')$ needs to be symmetric, as

discussed by Wallinga *et al.* (2006) and Goeyvaerts *et al.* (2010).

For the purposes of this thesis, infection- or disease specific mortality is ignored. The assumption of no excess mortality associated with disease holds for most of the common childhood diseases, at least in developed countries. As mentioned earlier, the infection is considered to be in endemic steady state and the present discussion is only valid for large populations. Assuming the population to be large makes it possible to rely on deterministic epidemiological theory. In addition, the population size N is taken fixed and births and deaths are equally balanced (demographic equilibrium). In that way, one can easily derive the following expression for the life expectancy L^* if the natural mortality rate is represented by $\mu(a)$:

$$L^* = \int_0^\infty e^{-\int_0^a \mu(u)du} da. \tag{3}$$

The mean duration of infectiousness is denoted by D and the age-dependent recovery rate for the infection equals $\gamma(a)$. Therefore, the mean infectious period can be written as (Goeyvaerts, 2011):

$$D = \int_0^\infty e^{-\int_0^a \gamma(u)du} da,$$

analogue to the results for the life expectancy in equation (3). Consider now the situation in which an infectious individual of age a' is introduced into a population of size N with a fraction $n(a)$ of individuals of age a from which a proportion of $S(a)$ is susceptible for the specific infection. The latter notation refers to the fact that $S(a)$ resembles the survival function in Section 3.1 given that the event of interest here is infection with the pathogen. Under these circumstances, the average number of newly infected individuals of age a which are infected by the single individual of age a' during his or her entire infectious period equals

$$Nn(a)S(a)\int_0^\infty \beta(a, a'+u)e^{-\int_0^u \gamma(v)dv} e^{-\int_{a'}^{a'+u} \mu(v)dv} du. \tag{4}$$

If the infectious period is short compared to the timescale on which the effective contact rates and natural mortality rates vary, equation (4) simplifies to

$$NDn(a)S(a)\beta(a, a'). \tag{5}$$

Often, one writes the density of the population age distribution $n(a)$ in terms of the natural mortality rate $\mu(a)$ and the life expectancy L^*. Since we assume demographic equilibrium and a constant population size N, one obtains:

$$n(a) = \frac{1}{L^*} e^{-\int_0^a \mu(u)du}. \tag{6}$$

The next generation operator G acting on a single instantaneous infected individual with density $i(a')$ gives rise to the next generation of infected individuals of age a, i.e.

$$G[i](a) = NDn(a)S(a)\int_0^\infty \beta(a, a')i(a')da'. \tag{7}$$

From the next generation operator, one can determine the total number of cases infected by the distributed individual during its infectious period on average, $\int_0^\infty G[i](a)da$. Another key parameter in describing an epidemic is the reproduction number R. The reproduction number R is defined as the average number of secondary infections produced by a single infectious individual during his or her infectious period. Consequently, the reproduction number equals the spectral radius of the next generation operator G (Diekmann *et al.*, 1990). In a fully susceptible population, i.e. $S(a) = 1$, the reproduction number is called the basic reproduction number R_0. Knowledge of the (basic)

reproduction number is important since it summarizes in a very simple and convenient way the amount of effort required to eradicate an epidemic within a population. This statement results from the fact that an infection with reproduction number larger than one in a large population can become epidemic whereas reducing the reproduction number to a level lower than one induces the infection to disappear (Farrington *et al.*, 2001).

The mass action principle relates the age-dependent force of infection $\lambda(a)$ to the effective contact function $\beta(a,a')$ as in equation (1). Farrington *et al.* (2001) reformulates this formula as follows:

$$\lambda(a) = \frac{N}{L^*} \int_0^\infty \left(\int_0^\infty \beta(a, a'+u) e^{-\int_0^u \gamma(v)dv} e^{-\int_{a'}^{a'+u} \mu(v)dv} du \right) \lambda(a') S(a') e^{-\int_0^{a'} \mu(u)du} da'. \tag{8}$$

Again under the assumption that the mean duration of infectiousness is short compared to the timescale on which the transmission and mortality rates change, equation (8) can be approximated by:

$$\lambda(a) = \frac{ND}{L^*} \int_0^\infty \beta(a, a') \lambda(a') S(a') e^{-\int_0^{a'} \mu(u)du} da'. \tag{9}$$

For infections conferring lifelong immunity, the proportion of susceptible individuals $S(a)$ of age a equals:

$$S(a) = e^{-\int_0^a \lambda(u)du}, \tag{10}$$

where $\lambda(u)$ is the age-dependent force of infection. As demonstrated in the master thesis, the survival function (fraction of seronegatives) has a different expression for nonimmunizing infections following an SIRS transmission model.

Coutinho *et al.* (1999) were pioneers in accounting for heterogeneity in the acquisition of infections. Following the work of Farrington *et al.* (2001), the mass action principle as explained previously, can incorporate individual heterogeneity in a straightforward way. The authors focus on an augmented effective contact function denoted by $\beta(a,u;a',v)$ which represents the per capita rate of acquisition of the infection by an individual of age a and frailty u from an individual of age a' and frailty v (see Section 5.3). The frailties are assumed to follow a continuous nonnegative distribution with density function $f(x)$, x in $[0,\infty)$. As before, the mortality rate is assumed to be independent of the infection and the frailties. For infections conferring lifelong immunity and a short mean infectious period, the mass action principle in (9) reduces to:

$$\lambda(a,u) = \frac{ND}{L^*} \int_0^\infty \int_0^\infty \beta(a,u;a',v) \lambda(a',v) e^{-\int_0^{a'} \lambda(r,v)dr} e^{-\int_0^{a'} \mu(r)dr} f(v) da' dv. \tag{11}$$

Under a simple multiplicative decomposition of the augmented effective contact function, i.e. $uv\beta_0(a,a')$, as suggested by Farrington *et al.* (2001) and a standard assumption of expectation one for the frailty distribution, the force of infection factorizes into $\lambda(a,u) = u\lambda_0(a)$. $\lambda_0(a)$ is called the baseline hazard function and $\beta_0(a,a')$ is equal to the baseline effective contact function. This simple multiplicative decomposition of the augmented effective contact function justifies the use of the term frailty for the activity levels u and v as each individual has a specific frailty with respect to the acquisition of infections. The mass action principle in equation (11) equals then:

$$\lambda_0(a) = \frac{ND}{L^*} \int_0^\infty \int_0^\infty v^2 \beta_0(a,a') \lambda_0(a') e^{-\int_0^{a'} v\lambda(r)dr} e^{-\int_0^{a'} \mu(r)dr} f(v) da' dv. \tag{12}$$

A.2. Optimization Algorithms

Optimization is an important tool in many areas such as decision theory and the analysis of physical systems. In statistics, one is also frequently faced with numerical optimization problems. One of the most important examples is the maximization of the likelihood function in the likelihood inferential framework. In general, optimization problems are defined using an objective function which needs to be optimized. The objective function depends on several parameters, variables or unknows and quantifies the performance of a system by means of a single number. The process of identifying the objective function is called modelling. The construction of a model is a first step in the optimization problem and is often the most important one. Indeed, too simple models are not able to capture the particularities associated with a certain physical system whereas too complex models are often very difficult too solve. After the identification of the model, optimization algorithms can be used to select plausible values for the unknown variables of the objective function for which the objective is optimized. Although there exists no universal optimization algorithm, numerous possibilities are available to solve different problems. Often it is the responsibility of the user to select carefully the optimization algorithm which is best suited for the problem at hand. After the application of an optimization algorithm, one needs to verify whether the algorithm converged to a valid solution for the optimization problem. In many cases, nice mathematical expressions called optimality conditions are derived based on which one determines whether the algorithm successfully solved the optimization problem of interest (Nocedal and Wright, 1999).

In the present elaboration, we summarize the most important optimization algorithms which are used to solve the optimization problems throughout this master thesis. We initiate out discussion with the basic mathematical formulation of an optimization problem. Mathematically speaking, optimization is the maximization or minimization of a function subject to constraints on its variables. Let us denote x as the vector of variables, also called parameters or unknowns. The objective function f is a function of x which is to be maximized or minimized. In the applications included in this thesis, the objective function equals the likelihood function L. Finally, c equals the constraints which are to be satisfied by the variables x. In full generality, one can formulate the optimization problem as follows (Nocedal and Wright, 1999):

$$\min_x \ f(x) \quad \text{subject to } c(x),$$

where $c(x)$ is a matrix of scalar-valued functions of the variables in the vector x. In maximization problems, one can change f into -f in the above problem formulation.

Optimization algorithms are iterative procedures. This implies that they start with an initial guess, often called starting values, for the unknown parameters and consequently generate a sequence of improved estimates until hopefully convergence towards the optimal values has been achieved. The strategy applied to move from one iteration point to the next distinguishes one algorithm from another. Most of the approaches make use of the value of the objective function f and possibly the first and second derivative of f. In order to evaluate an optimization algorithm, concepts such as accuracy, robustness and efficiency are essential. However, these notions are out of the scope of this master thesis and are therefore not handled in this part. For an extensive presentation with respect to the key items in optimization problems, the reader is referred to Nocedal and Wright (1999), Nocedal (1992), and Bulirsch and Stoer (1980).

Overview of Algorithms

The last fourty years has seen the development of many powerful optimization algorithms for unconstrained problems of smooth functions. All algorithms require the user to specify a starting point, denoted by x_0. Beginning at this starting point, the algorithms generate a sequence of points $\{x_k\}_{k=0}^{\infty}$ which is terminated when either no more progress is made or when a solution point is approximated with a specified accuracy. In order to decide in which direction one moves from x_k to x_{k+1}, the algorithm uses information on the function f at x_k, and possibly some other information regarding the other iterates and the derivatives of f in x_k. There exist two fundamental strategies to move from one iterate to the next.

In the line search strategy, the algorithm chooses a direction p_k and searches along the direction from the current iterate x_k for a new iterate x_{k+1} with a smaller function value. The distance to move along the direction p_k can be found by approximately solving the following one-dimensional minimization problem to find a step length α:

$$\min_{\alpha} \ f(x_k + \alpha p_k), \quad \alpha \geq 0.$$

The equation is solved approximately by consideration of a limited number of trial step lengths until it finds one that loosely approximates the minimum of the equation. The process is repeated over and over again.

The second algorithmic strategy is known as trust region and consists of the construction of a model function m_k using the information gathered about f. The behaviour of the model function around x_k is similar to that of the actual function f. Because the model function m_k can approximate the function f poorly when x is far from x_k, one minimizes m_k in a small region of x_k. Therefore, the candidate step p is obtained by solving:

$$\min_{p} \ m_k(x_k + p),$$

where $x_k + p$ lies within the trust region. Usually the trust region is a ball defined by $||p||_2 \leq \Delta$, where $\Delta > 0$ represents the trust region radius. The model function is often defined as a quadratic function of the form:

$$m_k(x_k + p) = f_k + p^T \nabla f_k + \frac{1}{2} p^T B_k p,$$

where f_k, ∇f_k and B_k are a scalar for the function value, vector for the gradient and matrix representing the Hessian $\nabla^2 f_k$ or some approximation of it, respectively with the indices referring to the evaluation of these functions in x_k. The equation is the Taylor series expansion of f around x_k. Therefore, the functions m_k and f are in agreement upto order one in iterate x_k.

In fact, the line search and trust region strategies differ in the order in which they choose the direction and distance of the move to the next iterate. Indeed, line search starts by fixing the direction p_k and then identifying the appropriate distance, i.e. the step length α_k. However, in trust region optimization, first the maximum distance is chosen, i.e. the trust region radius Δ_k, after which the direction and step are selected which attain the best improvement.

Line Search Methods

The most straightforward and obvious direction to move from x_k to x_{k+1} in a line search approach is the steepest-descent direction -∇f_k. From all possible directions, this is the one in which f decreases most rapidly. The latter conclusion can be verified using the Taylor series expansion of f around x_k. The steepest descent method is a line search method that moves along p_k = -∇f_k in each iteration step. One of the advantages of the steepest descent direction is that it only requires the calculation of the gradient ∇f_k, but not of the Hessian matrix B_k. The disadvantage however is that the steepest descent method can be excruciatingly slow for complex optimization problems.

Another important search direction in the line search approach is the Newton direction. This direction is derived from the second order Taylor series approximation of $f(x_k + p)$, which is equal to:

$$f(x_k + p) \approx f_k + p^T \nabla f_k + \frac{1}{2} p^T \nabla^2 f_k p.$$

Assuming that the Hessian matrix $\nabla^2 f_k$ is positive definite, we obtain the Newton direction by minimizing the approximation with respect to p. Setting the derivative of the approximation equal to zero, one easily obtains:

$$p_k^N = -\nabla^2 f_k^{-1} \nabla f_k.$$

The Newton direction is reliable when the difference between the true function value $f(x_k + p)$ and the quadratic model is not too large. The Newton direction can be used in a line search approach when -$\nabla^2 f_k$ is positive definite. There is in contrast to the steepest descent direction, a step length of size 1 associated with the Newton direction. The main drawback of the Newton direction is the requirement of the Hessian matrix $\nabla^2 f_k$ for which explicit calculation is sometimes an error-prone and expensive process.

Quasi-Newton search directions are therefore a quite useful alternative to the basic Newton direction. They do not require the explicit calculation of the Hessian matrix and use an approximation B_k instead which is updated after each step in order to account for the additional information gained after each iteration step. These updates make use of the fact that changes in the gradient provide knowledge about the second derivative of f along the search direction of interest. One can proof that (Nocedal and Wright, 1999):

$$\nabla f(x + p) = \nabla f(x) + \nabla^2 f(x) p + \int_0^1 \left[\nabla^2 f(x + tp) - \nabla^2 f(x)\right] p dt.$$

Because $\nabla f(.)$ is a continuous function, the size of the integral term is equal to $o(||p||)$. Setting $x = x_k$ and $p = x_{k+1} - x_k$, the equation simplifies to:

$$\nabla f_{k+1} = \nabla f_k + \nabla^2 f_{k+1}(x_{k+1} - x_k) + o(||x_{k+1} - x_k||).$$

If x_k and x_{k+1} lie in a region near the solution x^*, i.e. the true minimum of the objective function f, within which ∇f is positive definite, the final terms is eventually dominated by the term $\nabla^2 f_{k+1}(x_{k+1} - x_k)$. Therefore, one can write:

$$\nabla^2 f_{k+1}(x_{k+1} - x_k) \approx \nabla f_{k+1} - \nabla f_k.$$

Consequently, the new Hessian approximation B_{k+1} such that the previous property of the true Hessian is satisfied. Therefore, B_{k+1} needs to satisfy the following condition, known as the secant equation:

$$B_{k+1} s_k = y_k,$$

where $s_k = x_{k+1} - x_k$ and $y_k = \nabla f_{k+1} - \nabla f_k$. Typically, additional requirements on B_{k+1} are imposed such as symmetry. The initial approximation B_0 needs to be specified by the user.

Two of the most popular formulae for updating the Hessian approximation B_k are the symmetric-rank-one (SR1) formula and the BFGS formula. The SR1 formula is defined by:

$$B_{k+1} = B_k + \frac{(y_k - B_k s_k)(y_k - B_k s_k)^T}{(y_k - B_k s_k)^T s_k}.$$

The BFGS formula is named after its inventors, Broyden, Fletcher, Goldfarb, and Shanno, which is defined by:

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k}.$$

The Quasi-Newton search direction equals the basic Newton direction except for the replacement of the true Hessian matrix by the Hessian approximation B_k, i.e.

$$p_k^{QN} = -B_k^{-1} \nabla f_k.$$

A last class of search directions is generated by nonlinear conjugate gradient methods. They have the following form:

$$p_k = -\nabla f_k + \beta_k p_{k-1},$$

where β_k ensures that the directions p_k and p_{k-1} are conjugate. Further details regarding this approach are omitted here.

Nelder-Mead Simplex Method

The Nelder-Mead Method is a simplex method for finding a local minimum of a function of severable variables devised by Nelder and Mead (1965). For two variables, a simplex is a triangle, and the Nelder-Mead method is a pattern search that compares function values at the three vertices of a triangle. The worst vertex (at which the function evaluation is largest), is rejected and replaced with a newly selected vertex. Consequently, a new triangle is formed and the search procedure is continued. Therefore, a sequence of triangles is generated by the algorithm in which the size of the triangles is reduced such that eventually the coordinates of the minimum point are found. The algorithm can be generalized to triangles in N dimensions in order to find the minimum of a function in N variables. The Nelder-Mead simplex method is effective and computationally compact.

We formulate the Nelder-Mead simplex method for a two-dimensional optimization problem. Let $f(x,y)$ be the objective function to be minimized (Note that the notation is slightly different from the one used in the preceeding sections). The starting values constitute the three vertices of a triangle: $V_k = (x_k,y_k)$, $k = 1, 2, 3$. The objective function is then evaluated in the three vertices, $z_k = f(x_k,y_k)$, $k = 1, 2, 3$. The ordered function values are given by $z_{(1)} \leq z_{(2)} \leq z_{(3)}$. We now introduce the notation (Mathews and Fink, 2004):

$$B^* = (x_{(1)}, y_{(1)}) \qquad G^* = (x_{(2)}, y_{(2)}) \qquad W^* = (x_{(3)}, y_{(3)}),$$

to distinguish between the best vertex, the second best and the worst vertex respectively. The construction process uses the midpoint of the line segment between B^* and G^* defined by:

$$M = \frac{B^* + G^*}{2} = \left(\frac{x_{(1)} + x_{(2)}}{2}, \frac{y_{(1)} + y_{(2)}}{2} \right).$$

The function decreases as we move along the side of the triangle from W^* to B^*, and if we move along the side from W^* to G^*. Therefore it seems feasible that $f(x,y)$ takes on smaller values at points away from W^* on the opposite side of the line connecting B^* and G^*. One chooses a test point R^* by reflecting the triangle through the side B^*G^*. In order to determine the test point R^*, the midpoint M is connected with the worst point W^* and its length equals d. The vector point R^* is defined by:

$$R^* = M + (M - W^*).$$

If the function value at R^* is smaller compared to the evaluation of the objective function at W^*, then we have moved in the correct direction towards the minimum value. As the minimum can be located just a bit further than the point R^*, one extends the line segment through M and R^* to the point E^*. In such a way, one obtains an expanded triangle $B^*G^*E^*$. The point E^* is found by moving an additional distance d along the line segment connecting M and R^*. The point E^* is a better vertex than R^*, if the function value is decreases. The formula for E^* equals:

$$E^* = R^* + (R^* - M) = 2R^* - M.$$

If the function values at R^* and W^* are the same, another test point must be selected. Perhaps the function is smaller at M, but we cannot replace W^* with M since a traingle needs to be constructed. Consider the two midpoints C_1 and C_2 of the line segments W^*M and MR^*, respectively. The point with the smaller function value is called C^*, and the new triangle is $B^*G^*C^*$. If the function value at C^* is not less as compared to the function value at W^*, one shrinks the points G^* and W^* towards B^*. Hence, G^* is replaced by M whereas W^* is substituted with P^*, i.e. the midpoint of the line segment joining B^* and W^*. In each iteration step, a new vertex point is found, which replaces the worst vertex W^* at the previous iteration. More details concerning a wide variaty of numerical optimization algorithms are found in Nocedal and Wright (1999), and Mathews and Fink (2004).

The Nelder-Mead Simplex Method is implemented as default optimization procedure in the R-function optim. In the thesis, this algorithm is used due to its good general performance. Different optimization algorithms are implemented in the SAS procedure PROC NLMIXED. For the purposes of this master thesis, the Quasi-Newton methods are used in order to maximize the loglikelihood functions for the trivariate frailty models. However, the trust region method, Newton-Raphson method with line search, Nelder-Mead Simplex Method and conjugate gradient mehods can be requested as well in the NLMIXED procedure among several others.

A.3. Numerical Integration

Most of the applications of integration within and outside of mathematics, involve the evaluation of the definite integral:

$$I = \int_a^b f(x)dx.$$

The Fundamental Theorem of Calculus enables us to evaluate such definite integrals by first finding an antiderivative of f. Therefore, one has spent considerable time developing several integration techniques. However, there are in general two obstacles in the calculation of the definite integral I.

- Finding the antiderivative of f in terms of familiar functions may be impossible, or at least extremely difficult.
- We may not have a formula for $f(x)$ in terms of the unknown x, e.g. $f(x)$ may be an unknown function whose values at certain points in the interval $[a, b]$ have been determined by means of experimental research.

We now investigate the problem of approximating the value of the definite integral I using only the values of $f(x)$ at finitely many points of the interval $[a, b]$. Obtaining such an approximation of the definite integral I is called numerical integration. Upper and lower sums, and in fact any Riemann sum, can be used for these purposes, but they usually require much more calculations to yield a desired precision then the methods that are most often used for numerical integration. Special attention is directed towards the Trapezoid Rule, the Midpoint Rule and Simpson's Rule for numerical integration. All the techniques require the calculation of the values of $f(x)$ at a set of equally spaced points in $[a, b]$. The computational *expense* involved in approximating the definite integral I is roughly proportional to the number of function values required. Therefore, the fewer function evaluations needed to achieve a certain degree of accuracy, the better the method that is used to approximate the integral I (Adams, 2006).

Trapezoid Rule

Consider the function $f(x)$ and assume that the function is continuous on the interval $[a, b]$. In addition, subdivide $[a, b]$ into n subintervals of equal length $h = (b - a)/n$ using the $n + 1$ points:

$$x_0 = a, \quad x_1 = a + h, \quad x_2 = a + 2h, \quad ..., \quad x_n = a + nh = b.$$

Moreover, we assume that the value of $f(x)$ at each of these points is known:

$$y_0 = f(x_0), \quad y_1 = f(x_1), \quad y_2 = f(x_2), \quad ..., \quad y_n = f(x_n).$$

The Trapezoid Rule approximates the definite integral I by using straight line segments between the points (x_{j-1},y_{j-1}) and (x_j,y_j) $(1 \leq j \leq n)$ and consequently summing the areas of the resulting n trapezoids. A trapezoid is defined as a four-sided polygon with one pair of parallel sides.

The first trapezoid has vertices $(x_0,0)$, (x_0,y_0), (x_1,y_1) and $(x_1,0)$. The two parallel sides are vertical and have lengths y_0 and y_1. The perpendicular distance between them equals $h = x_1 - x_0$. The area of the trapezoid is h times the average of the parallel sides:

$$h\frac{y_0 + y_1}{2} \text{ square units.}$$

We can then approximate the integral of f over any subinterval as follows:

$$\int_{x_{j-1}}^{x_j} f(x)dx \approx h\frac{y_{j-1}+y_j}{2}, \quad (1 \leq j \leq n).$$

The original definite integral I can be approximated by the sum of these trapezoidal areas:

$$\begin{aligned}\int_a^b f(x)dx &\approx h\left(\frac{y_0+y_1}{2}+\frac{y_1+y_2}{2}+...+\frac{y_{n-1}+y_n}{2}\right)\\ &= h\left(\frac{1}{2}y_0+y_1+y_2+...+y_{n-1}+\frac{1}{2}y_n\right).\end{aligned}$$

Midpoint Rule

A somewhat simpler approximation of the definite integral I is based on the partition of $[a,b]$ into n equal subintervals and involves forming a Riemann sum of the areas of rectangles whose height are taken at the midpoints of the n subintervals. The Midpoint Rule is defined by:

$$h\left(f(m_1)+f(m_2)+...+f(m_n)\right) = h\sum_{j=1}^{n} f(m_j),$$

where $m_j = a + (j-\frac{1}{2})h$, $1 \leq j \leq n$. The Midpoint Rule can be used in combination with the Trapezoid Rule in the approximation of the integral I. More details regarding this approach are included in Adams (2006).

Simpson's Rule

The Trapezoid Rule approximation to I results from approximating the graph of f by straight line segments through adjacent pairs of data points on the graph. Intuitively, one would expect to do even better if we approximate the graph by more general curves. Since straight lines are the graphs of linear functions, the simplest obvious generalization is to use the class of quadratic functions to approximate the graph of f by segments of parabolas. This constitutes the basis for Simpson's Rule.

Suppose that three points are given in the plane, one on each of three equally spaced vertical lines, spaced, say h units apart. If one selects the middle line as being the y-axis, then the coordinates of the three points are equal to (-h,y_L), (0,y_M), and (h,y_R). Constants A, B, and C can be chosen so that the parabola $y = A + Bx + Cx^2$ passes through these points. By substituting the coordinates of the three points into the equation of the parabola, we get (Adams, 2006):

$$\begin{aligned}y_L &= A - Bh + Ch^2\\ y_M &= A\\ y_R &= A + Bh + Ch^2\end{aligned}$$

Therefore, one obtains $A = y_M$ and $2Ch^2 = y_L - 2y_M + y_R$. Now one can easily derive that:

$$\begin{aligned}\int_{-h}^{h}\left(A+Bx+Cx^2\right)dx &= \left(Ax+\frac{B}{2}x^2+\frac{C}{3}x^3\right)\Big|_{-h}^{h}\\ &= 2Ah+\frac{2}{3}Ch^3\\ &= h\left(2y_M+\frac{1}{3}(y_L-2y_M+y_R)\right)\\ &= \frac{h}{3}\left(y_L+4y_M+y_R\right).\end{aligned}$$

Hence, the area of the plane region bounded by the parabolic curve, the interval of length $2h$ on the x-axis, and the left and right vertical lines is equal to $(h/3)$ times the sum of the heights of the region at the left and right edges, and four times the height at the middle. In order to approximate the definite interval I based on the subdivision of the interval $[a, b]$ into an even number n of subintervals of equal length h, we have:

$$\int_a^b f(x)dx \approx \frac{h}{3}y_0 + 4y_1 + 2y_2 + 4y_3 + 2y_4 + ... + 2y_{n-2} + 4y_{n-1} + y_n$$
$$= \frac{h}{3}\left(\sum y_{ends} + 4\sum y_{odds} + 2\sum y_{evens}\right).$$

The Simpson's Rule approximation for I requires no more data than does the Trapezoid Rule approximation, but it requires the values of $f(x)$ at $n+1$ equally spaced points. However, in contrast to the Trapezoid Rule, Simpson's Rule treats the data differently by weighting successive values differently. In general, the latter can produce a much better approximation of the definite integral I of interest (Adams, 2006).

The methods presented above are all based on an equal subdivision of the interval $[a, b]$. This restriction can be avoided using for example Gaussian approximations. Gaussian approximations involve selecting evaluation points and weights in an optimal way such that the most accurate results are obtained for well-behaved functions (Adams, 2006).

Adaptive Gaussian Quadrature Approximation

The principle underlying the most state-of-the-art deterministic approximations of I is Gaussian quadrature. A quadrature rule is an approximation of the definite integral of a function, e.g. the definite integral I as introduced earlier. A quadrature rule is most often stated as a weighted sum of function values at specified points within the domain of integration. An n-point Gaussian quadrature rule, named after Carl Friedrich Gauss, is a quadrature rule constructed to give rise to an exact result for polynomials of degree $2n-1$ or less by suitable choices for evaluation points x_i and weights ω_i, i = 1, 2, ..., n. The domain for such integration is conventionally taken to be equal to $[-1, 1]$, so that the rule is stated as:

$$\int_{-1}^{1} f(x)dx \approx \sum_{i=1}^{n} \omega_i f(x_i).$$

Gaussian quadrature will produce only accurate results for the integration problem if the function $f(x)$ is well approximated by a polynomial function on the interval $[-1, 1]$. However, if the integrated function can be written as $f(x) = p(x)w(x)$, where $p(x)$ is approximately polynomial of degree $2n-1$ or lower and $w(x)$ is a known basis function, then there exist points $x_i \in [-1, 1]$ and alternative weights $\omega_i^{'}$ associated with each point x_i such that:

$$\int_{-1}^{1} f(x)dx = \int_{-1}^{1} p(x)w(x)dx \approx \sum_{i=1}^{n} \omega_i^{'} p(x_i).$$

These points and weights only depend on a, b and the basis function $w(x)$. Infinite intervals and semi-infinite intervals can be treated through appropriate transformation of the variable to a finite interval. Different choices for the basis function or weighting function $w(x)$ can be made such that one obtains Gauss-Legendre quadrature, Chebyshev-Gauss quadrature, Gauss-Hermite quadrature, Guass-Jacobi quadrature, etc. For many purposes Gauss-Legendre quadrature, with $w(x) = 1$, is adequate to evaluate the integral of interest. Moreover, it can be shown (Press *et al.*, 2007) that

the evaluation points are just the roots of a polynomial belonging to a specific class of orthogonal polynomials. Gaussian quadrature is preferred over the traditional methods presented earlier since they have fewer function evaluations for a given order. With Gaussian quadrature the weights and evaluation points are determined such that the integration rule is exact to as high an order as possible. As mentioned in the preceeding paragraph, the evaluation points x_i, $i = 1, 2, ..., n$, are the roots of the Legendre polynomial of degree n in case of Gauss-Legendre quadrature. The weights $\omega_i^{'}$, $i = 1, 2, ..., n$, are called the Christoffel weights. By transforming the function variable, one can derive an estimate for the definite integral I on a more general interval $[a, b]$ by means of the Gauss-Legendre integration formula presented previously.

In applied mathematics, adaptive quadrature is the process in which the definite integral of a function $f(x)$ is approximated using the static quadrature rules introduced above on adaptively refined subintervals of the integration domain. Generally, adaptive algorithms are just as efficient and effective as traditional algorithms for well-behaved integrands, but they are also effective for badly-behaved integrands for which traditional algorithms fail. In general, an approximation Q of the integral I by means of a static quadrature rule gives rise to an error estimate. If the error estimate is larger than the tolerance, one subdivides the interval in two parts and the quadrature rule is applied on the two intervals. Either the initial estimate or the sum of recursively computed halves is returned. A more thorough explanation on the fundamental aspects of numerical integration is incorporated in Press *et al.* (2007). The Adaptive Gaussian Quadrature Approximation is used for the approximation of the definite integrals within this thesis and implemented using the integrate function in R.

Appendix B

B.1. Derivation of the probability density function for right-censored survival data

Let H denote the cumulative distribution function of the observation times $T_j = \min\{T_j^*, C_j\}$. In survival analysis, one often relies on the assumption of independence between the event times and the censoring times. Therefore, under the independence assumption, the cumulative distribution function H is equal to:

$$\begin{aligned} H(t_j) &= P(\min\{T_j^*, C_j\} \leq t_j) = 1 - P(\min\{T_j^*, C_j\} > t_j) \\ &= 1 - P(T_j^* > t_j, C_j > t_j) = 1 - (1 - F(t_j))(1 - G(t_j)) \end{aligned}$$

Furthermore, let H_0 and H_1 denote subdistribution functions of the observation time T_j defined by:

$$H_k(t_j) = P(T_j \leq t_j, \Delta_j = k), \qquad k \ = 1,\ 2.$$

The following relationship holds for the subdistribution function H_1 under the assumption of independence between event and censoring times:

$$\begin{aligned} H_1(t_j) &= P(T_j \leq t_j, \Delta_j = 1) = P(T_j^* \leq t_j, T_j^* \leq C_j) \\ &= \int_{t_j^* \leq t_j, t_j^* \leq c_j} \int f(t_j^*) g(c_j) dt_j^* dc_j \\ &= \int_{t_j^* \leq t_j} f(t_j^*) \left(\int_{t_j^* \leq c_j} g(c_j) dc_j \right) dt_j^* \\ &= \int_{t_j^* \leq t_j} f(t_j^*)(1 - G(t_j^*)) dt_j^* \end{aligned}$$

An analogue result can be obtained for H_0, namely:

$$H_0(t_j) = \int_{c_j \leq t_j} g(c_j)(1 - F(c_j)) dc_j.$$

Consequently, for the corresponding density function h_0 and h_1, we have $h_0(t_j) = dH_0(t_j)/dt_j = g(t_j)(1\text{-}F(t_j))$ and $h_1(t_j) = dH_1(t_j)/dt_j = f(t_j)(1\text{-}G(t_j))$. The joint probability density function of the right-censored survival data, represented by the random vector (T_j, Δ_j) is given by:

$$\begin{aligned} f(t_j, \delta_j) &= \delta_j h_1(t_j) + (1 - \delta_j) h_0(t_j) \\ &= h_1(t_j)^{\delta_j} h_0(t_j)^{(1-\delta_j} \\ &= (f(t_j)(1 - G(t_j)))^{\delta_j} (g(t_j)(1 - F(t_j)))^{1-\delta_j}. \end{aligned}$$

B.2. Laplace transform of a gamma distributed random variable

The Laplace transform is defined in equation (3.4) for a continuous random variable X. Let X follow a gamma distribution with shape parameter k and inverse scale parameter ψ. In that case, one has:

$$\mathbf{L}(u) = E(e^{-uX}) = \int_0^\infty e^{-ux} f(x) dx,$$

where $f(x)$ equals the probability density function of the random variable X (see Section 3.4.4). Using partial integration and induction, one can calculate the previous definite integral as follows:

$$\begin{aligned}
\mathbf{L}(u) &= \int_0^\infty e^{-ux} \frac{\psi^k}{\Gamma(k)} x^{k-1} e^{-\psi x} dx \\
&= \frac{\psi^k}{\Gamma(k)} \int_0^\infty x^{k-1} e^{-(\psi+u)x} dx \\
&= \frac{\psi^k}{\Gamma(k)} \left(\frac{-1}{(\psi+u)} x^{k-1} e^{-(\psi+u)x} \,|_0^\infty + \frac{k-1}{(\psi+u)} \int_0^\infty x^{k-2} e^{-(\psi+u)x} dx \right) \\
&= \frac{\psi^k}{\Gamma(k)} \left(\frac{k-1}{\psi+u} \int_0^\infty x^{k-2} e^{-(\psi+u)x} dx \right) \\
&= ... \\
&= \frac{\psi^k}{\Gamma(k)} \left(\frac{(k-1)(k-2)...2.1}{(\psi+u)^{k-1}} \int_0^\infty e^{-(\psi+u)x} dx \right) \\
&= \frac{\psi^k}{\Gamma(k)} \left(\frac{(k-1)!}{(\psi+u)^k} (-e^{-(\psi+u)x} \,|_0^\infty) \right) \\
&= \frac{\psi^k}{(\psi+u)^k}
\end{aligned}$$

Therefore, the above expression reduces to:

$$\mathbf{L}(u) = \left(\frac{\psi+u}{\psi} \right)^{-k} = \left(1 + \frac{u}{\psi} \right)^{-k}.$$

The latter expression corresponds to the one presented in equation (3.10).

B.3. Laplace transform of an inverse Gaussian distributed random variable

The Laplace transform of an inverse Gaussian distributed random variable X is derived using the probability density function $f(x)$ in equation (4.20). The Laplace transform is defined as:

$$\mathbf{L}(u) = E(e^{-uX}) = \int_0^\infty e^{-ux} f(x) dx.$$

The Laplace transform reduces to (Wienke, 2010):

$$\begin{aligned}
\mathbf{L}(u) &= \int_0^\infty e^{-ux} \frac{\sqrt{\zeta}}{\sqrt{2\pi x^3}} e^{-\frac{\zeta}{2\phi^2 x}(x-\phi)^2} dx \\
&= \int_0^\infty \frac{\sqrt{\zeta}}{\sqrt{2\pi z^3}} e^{-\frac{(\zeta+2\phi^2 u)x^2 - 2\phi\zeta x + \zeta\phi^2}{2\phi^2 x}} dx \\
&= \int_0^\infty \frac{\sqrt{\zeta}}{\sqrt{2\pi z^3}} e^{-\frac{x}{2}\frac{(\zeta+2\phi^2 u)}{\phi^2} + \frac{\zeta}{\phi} - \frac{\zeta}{2x}} dx \\
&= e^{-\frac{\zeta\sqrt{1+\frac{2\phi^2 u}{\zeta}}}{\phi} + \frac{\zeta}{\phi}} \int_0^\infty \frac{\sqrt{\zeta}}{\sqrt{2\pi z^3}} e^{-\frac{\zeta x}{2}\frac{1+\frac{2\phi^2 u}{\zeta}}{\phi^2} + \frac{\zeta\sqrt{1+\frac{2\phi^2 u}{\zeta}}}{\phi} - \frac{\zeta}{2x}} dx
\end{aligned}$$

The following relationship is used to simplify the above expression:

$$-\frac{\zeta x}{2}\frac{1+\frac{2\phi^2 u}{\zeta}}{\phi^2} + \frac{\zeta\sqrt{1+\frac{2\phi^2 u}{\zeta}}}{\phi} - \frac{\zeta}{2x} = -\frac{\zeta}{2\frac{\phi^2}{1+\frac{2\phi^2 u}{\zeta}}x}\left(x - \frac{\phi}{\sqrt{1+\frac{2\phi^2 u}{\zeta}}}\right)^2.$$

Indeed,

$$\frac{\sqrt{\zeta}}{\sqrt{2\pi z^3}} e^{-\frac{\zeta x}{2}\frac{1+\frac{2\phi^2 u}{\zeta}}{\phi^2} + \frac{\zeta\sqrt{1+\frac{2\phi^2 u}{\zeta}}}{\phi} - \frac{\zeta}{2x}} = \frac{\sqrt{\zeta}}{\sqrt{2\pi z^3}} e^{-\frac{\zeta}{2\frac{\phi^2}{1+\frac{2\phi^2 u}{\zeta}}x}\left(x - \frac{\phi}{\sqrt{1+\frac{2\phi^2 u}{\zeta}}}\right)^2}$$

is the probability density function of an inverse Gaussian distributed random variable such that the definite integral from zero to infinity of this expression equals one. Therefore, the expression for the Laplace transform simplifies to:

$$\mathbf{L}(u) = e^{-\frac{\zeta\sqrt{1+\frac{2\phi^2 u}{\zeta}}}{\phi} + \frac{\zeta}{\phi}}$$

which corresponds to the expression in equation (4.21).

Appendix C

C.1. Additional Simulation Results

Additional simulation results are provided here with respect to the same parameter setting as described in Section 6.2 while increasing the sample size of the simulated data sets to 10000 instead of 2890. The idea is to compare the averaged estimates and emperical standard error estimates with those obtained in the original situation. The number of simulated data sets equals $n_s = 50$.

		TTE	RC	CS
Parameter	True Value	Estimate (s.e.)	Estimate (s.e.)	Estimate (s.e.)
a_1	0.3286	0.3324 (0.0096)	0.3290 (0.0083)	0.3128 (0.0279)
b_1	-0.0226	0.0086 (0.0043)	0.0013 (0.0067)	-0.0059 (0.0128)
a_2	0.1173	0.1282 (0.0023)	0.1323 (0.0023)	0.1264 (0.0038)
b_2	-0.0665	-0.0126 (0.0019)	-0.0182 (0.0022)	-0.0143 (0.0030)
a_3	0.0104	0.0102 (0.0003)	0.0104 (0.0003)	0.0104 (0.0006)
b_3	0.0258	0.0278 (0.0014)	0.0275 (0.0035)	0.0283 (0.0078)
σ_1^2	0.5505	0.6775 (0.0490)	0.6268 (0.0555)	0.5365 (0.1407)
σ_2^2	0.0507	0.0472 (0.0190)	0.0346 (0.0167)	0.0283 (0.0235)
σ_3^2	0.6251	0.7117 (0.0589)	0.7289 (0.1935)	0.7928 (0.4472)
ρ_{12}	0.2981	0.2419 (0.0577)	0.2114 (0.0651)	0.1940 (0.1134)
ρ_{13}	0.9363	0.8167 (0.0353)	0.8585 (0.0774)	0.8302 (0.1421)
ρ_{23}	0.2842	0.2097 (0.0520)	0.1961 (0.0622)	0.1767 (0.1009)

In general, the same conclusions hold as those stated in Chapter 6 with respect to the amount of information lost by turning from TTE data to RC data, and consequently to CS data. In addition, the empirical standard error estimates are found to be somewhat smaller as those included in Table 6.5. Of course, this is what we would expect if the sample size of the simulated datasets is increased.

C.2. Comparison between SIR and SIRS Models

The parameter estimates and estimated standard error estimates obtained from fitting the univariate SIR and SIRS gamma frailty models to simulated current status data (see Section 6.3) are included in the following table. The current status data is generated under the assumption of an SIRS and SIR infection process, respectively for the disease under investigation. The Gompertz baseline force of infection with parameters $a_1 = 0.3286$ and $b_1 = -0.0226$ is used in both cases. The frailty variance is assumed to be equal to 0.2. Under an SIRS infection process, the replenishment rate σ is taken equal to 0.05 for both simulation sets. The simulated datasets have a sample size equal to $N = 10000$.

	SIRS CS data			SIR CS data		
Para-meter	True Value	SIRS model Estimate (s.e.)	SIR model Estimate (s.e.)	True Value	SIRS model Estimate (s.e.)	SIR model Estimate (s.e.)
a_1	0.3286	0.3447 (0.0449)	0.5328 (0.0579)	0.3286	0.2936 (0.0220)	0.2943 (0.0203)
b_1	-0.0226	-0.0264 (0.0032)	-0.5051 (0.0558)	-0.0226	-0.0245 (0.0176)	-0.0241 (0.0153)
σ	0.0500	0.0413 (0.0075)	-	0.0000	3.5e-9 (5.5e-7)	-
σ_f^2	0.2000	0.3180 (0.2051)	0.0003 (0.0108)	0.2000	0.1040 (0.1210)	0.1078 (0.1049)
AIC		12347.1	12913.2		1720.5	1718.5

As stated earlier, the SIR model describes the dynamics of an SIRS infection rather poorly such that the more complex SIRS model seems worthwhile to consider. Moreover, the SIRS model is also capable of fitting the univariate SIR simulation data while estimating the replenishment rate almost equal to zero. These observations are in favor of the SIRS model in the situation of nonimmunizing infections following an SIRS transmission model.

Bibliography

Aalen, O.O. (1992). Modelling heterogeneity in survival analysis by the compound Poisson distribution. *Annals of Applied Probability*, **4**, 951-972.

Adams, R.A. (2006). *Calculus: A Complete Course (6th ed.)*. Pearson Education Canada Inc.: Toronto.

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

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

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

Beard, R.E. (1959). Note on some mathematical mortality models. In: *The Lifespan of Animals*. G.E.W. Wolstenholme, M.O'Conner (eds.), Ciba Foundation Colloquium on Ageing, Little, Brown, Boston, 302-211.

Bernoulli, D. (1760). An attemt at a new analysis of the mortality caused by smallpox and of the advantages of inoculation to prevent it. *Histoire et Mémoires de l' Académie des Sciences*, **2**, 1-79.

Bulirsch, R. and Stoer, J. (1980). *Introduction to Numerical Analysis*. Springer-Verlag: New York.

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

Coutinho, F.A.B., Massad, E., Lopez, L.F., Burattini, M.N., Struchiner, C. and Azevedo-Neto, R. (1999). Modelling heterogeneities in individual frailties in epidemic models. *Mathematical and Computer Modelling*, **30**, 97-115.

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

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

Feller, W. (1971). *An introduction to Probability Theory and its Applications*. John Wiley and Sons: New York.

Fiore, A.E. (2004). Hepatitis A transmitted by food. *Food Safety*, **38**, 705-715.

Goeyvaerts, N., Hens, N., Ogunjimi, B., Aerts, M., Shkedy, Z., Van Damme, P. and Beutels, P. (2010). Estimating transmission parameters and the basic reproduction number using social contact data and serological data on varicella zoster virus in belgium. *Journal of the Royal Statistical Society Series C*, In Press.

Goeyvaerts, N. (2011). *Statistical and Mathematical Models to Estimate the Transmission of Airborne Infections from Current Status Data*. Doctoraatsproefschrift.

Gompertz, B. (1825). On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. *Philosophical Transactions of the Royal Society of London*, **115**, 513-585.

Hens, N., Aerts, M., Shkedy, Z., Theeten, H., Van Damme, P. and Beutels, Ph. (2008). Modelling multi-sera data: The estimation of new joint and conditional epidemiological parameters. *Statistics in Medicine*, **27**(14), 2651-2664.

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

Hens, N., Aerts, M. Faes, C., Shkedy, Z., Lejeune, O., Van Damme, P. and Beutels, P. (2010). Seventy-five years of estimating the force of infection from current status data. *Epidemiology and Infection*, **138**(6), 802-812.

Ho, M. (1990). Epidemiology of cytomegalovirus infections. *Reviews of Infectious Diseases*, **12**, Sup7, S701-S710.

Hougaard, P. (1999). Fundamentals of survival data. *Biometrics*, **55**, 13-22.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. Springer-Verlag: New York.

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

Keiding, N., Andersen, P. (eds.) (2006). *Survival and Event History Analysis*. Wiley: New York.

Lancaster, T. (1979). Econometric methods for the duration of unemployment. *Econometrica*, **47**, 939-956.

Mathews, J.H. and Fink, K.K. (2004). *Numerical Methods Using Matlab (4th ed.)*. Prentice Hall Inc.: New Jersey.

Melnick, J.L. (1995). History and epidemiology of hepatitis A virus. *The Journal of Infectious Diseases*, **171**, Sup1, S2-S8.

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

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

Muench, H. (1959). *Catalytic Models in Epidemiology*. Harvard University Press: Boston.

Nelder, J.A. and Mead, R. (1965). A simplex method for function minimization. *Computer Journal*, **7**, 308-313.

Nocedal, J. and Wright, S.J. (1999). *Numerical Optimization*. Springer-Verlag: New York.

Nocedal, J. (1992). Theory of algorithms for unconstrained optimization. *Acta Numerica*, **1**, 199-242.

Pawitan, Y. (2001). *In all likelihood: statistical modelling and inference using likelihood*. Oxford University Press: Oxford.

Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P. (2007). *Numerical Recipes: The Art of Scientific Computing (3rd ed.)*. Cambridge University Press: New York.

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

Taylor, G.H. (2003). Cytomegalovirus. *American Family Physician*, **67**(3), 519-524.

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

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

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

Weibull, W. (1939). A statistical theory of the strength of materials. Ingenior Ventenskaps Akademien Handlinger 151. Generalstabens Litografiska Anstalts Förlag, Stockholm.

Wienke, A., Arbeev, K., Locatelli, I. and Yashin, A.I. (2005). A comparison of different bivariate correlated frailty models and estimation strategies. *Mathematical Biosciences*, **198**, 1-13.

Wienke, A. (2010). *Frailty Models in Survival Analysis*. Chapman & Hall/CRC: Boca Raton.

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

Young, N.S. and Brown, K.E. (2004). Mechanisms of disease: Parvovirus B19. *The New Engeland Journal of Medicine*, **350**, 586-597.

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
The trivariate correlated gamma frailty model for current status data

Richting: **master of Statistics-Epidemiology & Public Health Methodology**
Jaar: **2011**

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,

Abrams, Steven

Datum: **12/09/2011**