Modelling multisera data: The estimation of new joint and conditional epidemiological parameters

HENS, Niel; AERTS, Marc; SHKEDY, Ziv; Theeten, H.; Van Damme, P. & Beutels, P. (2008) Modelling multisera data: The estimation of new joint and conditional epidemiological parameters. In: STATISTICS IN MEDICINE, 27(14). p. 2651-2664.

DOI: 10.1002/sim.3089
Handle: http://hdl.handle.net/1942/8401

Modelling multi-sera data: The estimation of new joint and conditional epidemiological parameters.

N. Hens[1], M. Aerts[1], Z. Shkedy[1], H. Theeten[2], P. Van Damme[2] and Ph. Beutels[2]
[1]Center for Statistics, Hasselt University, B–3590 Diepenbeek, Belgium
[2]Centre for the Evaluation of Vaccination, Epidemiology and Community Medicine, University of Antwerp, Belgium

Abstract

Testing humans for infectious diseases is often done by assessing the presence or absence of diseases-specific antibodies in serum samples. For feasibility and economical reasons, these sera are often tested for more than one antigen. Studying diseases with similar transmission routes can govern new insights for disease dynamics. We use flexible marginal and conditional models to model multi-sera data on the Varicella-Zoster Virus and the Parvo B19-virus in Belgium. Next to the derivation of the age-dependent marginal force of infection, we introduce new epidemiological parameters: the age-dependent joint and conditional force of infection. These parameters allow us to study the association among the occurrence and acquisition of both infections. Furthermore, we show how to test for association and whether the infection-specific age-dependent force of infection curves are proportional and consequently whether separable mixing in the population holds.

1 Introduction

As a part of human epidemiology, mathematical modelling of infectious diseases usually involves describing the flow of individuals between mutually exclusive infection states. For instance, for infections that induce long-lasting immunity, the individuals can be classified into three different stages [1]. In the first stage individuals are susceptible to infection, meaning that they have not been exposed yet. In the second stage, individuals are infected and infectious to others while in the third stage individuals are no longer infectious and have acquired immunity to reinfection. This so-called SIR-model is used to describe the age- and time-dependent transmission of the infection.

Serological data are usually collected in cross-sectional studies. Under the assumptions of lifelong immunity and that the epidemic is in a steady state (i.e. at equilibrium), the prevalence can be estimated from such data. In general, empirical data show that the prevalence is age-dependent. In the literature, several flexible (non)parametric methods, some of them imposing monotonicity, have been proposed to model the age-specific prevalence (see e.g. [2, 3**?**]).

In the presented analyses, focus is on the Varicella-Zoster Virus and the Parvo B19-virus. These viruses are similar in that transmission is by airborne droplets and occurs during close contacts. The contact rate and the infectiousness of the pathogen determine the spread of the infection in a population. It has been shown that the contact rate depends on age through heterogeneity in mixing of individuals from different age-classes.

Except for the work of Farrington et al.[3, 4], Kanaan and Farrington [5] and Sutton et al.[6], who proposed a bivariate model for mumps and rubella and hepatitis B and C, respectively, while ascribing dependency to individual heterogeneity, none of the proposed methods takes account of the association between different infections. In addition, there could be a public health interest in estimating the probability of acquiring a second, altogether different, infection.

In this paper, we introduce the use of marginal and conditional models to study the association among different infectious diseases. We will exploit both flexible parametric and nonparametric methods to achieve the necessary flexibility as a function of age. The essential modelling tool which we use, is the vector generalized additive model methodology of Yee and Wild [7].

In Section 2, we start with introducing the data, while in Section 3 an overview of several existing univariate approaches to model the age-dependent prevalence is given. We introduce the bivariate Dale and the baseline category logits model in Section 4 and apply them to the data in Section 5. Monotonicity constraints for the bivariate models are imposed in Section 6 using the so-called 'pool

adjacent violator algorithm'. We develop marginal, conditional and joint force of infection and a test for proportionality in Section 7. We end with a discussion in Section 8.

2 Data

In a period from November 2001 until March 2003, 2381 serum samples in Belgium were collected and consecutively tested for Varicella-Zoster Virus (VZV) and Parvo B19-virus (B19) (see e.g. [8]). Together with the test result for VZV and B19, the gender and age of the individuals were recorded. In this paper, samples from children under 6 months were omitted because of distortions expected from the presence of maternal antibodies.

The Varicella-Zoster Virus, also known as human herpes virus 3 (HHV-3), is one of eight herpes viruses known to affect humans (and other vertebrates). Primary VZV infection results in chickenpox (varicella), has a two-week incubation period and is highly contagious by air droplets starting two days before symptoms appear. Infectiousness is known to last up to ten days. Therefore, chickenpox spreads quickly through close social contacts.

Parvovirus B19 was the first human Parvovirus to be discovered, in 1975. In clinical terms Parvovirus B19 is best known for causing a childhood exanthem called fifth disease or erythema infectiosum. The virus is primarily spread by infected respiratory droplets. B19 symptoms begin some six days after exposure and last for about a week. After being infected, patients are infectious for five to seven days and usually develop the illness after an incubation period of four to fourteen days.

3 Univariate Modelling

In the context of infectious diseases several flexible modelling techniques have been proposed to model the age-specific prevalence. Often, these models are part of the generalized linear model framework [9]. In this framework, the prevalence π is related to the age at infection $\boldsymbol{x} = (x_1, \ldots, x_n)$ and possibly other covariates (e.g. gender) using the formula $g(\boldsymbol{\pi}|\boldsymbol{x}) = h(\boldsymbol{x})$, where g is a link-function, as e.g. the 'logit'-, 'probit'- or 'complementary log log'-function, and h an assumed systematic component.

Both parametric and nonparametric models have been proposed. Among the parametric models, fractional polynomials (FPs, [10]) offer a wide variety of functional forms for the systematic component h. One attractive property of FPs is the inclusion of more conventional polynomials of the form $h(x) = \beta_0 + \beta_1 x + \beta_2 x^2$.

In contrast to parametric models, local regression methods, splines, etc. provide easy-to-apply nonparametric modelling techniques. In this paper, both flexible parametric (fractional polynomials) and nonparametric models (smoothing splines, [11]) are used to achieve necessary flexibility in modelling multisera data. We defer the application of these techniques to Section 5 and first describe the bivariate approaches.

4 Multivariate Modelling Approaches

4.1 Marginal and Conditional Models

Given bivariate binary dependent data on two infectious diseases $(\boldsymbol{y}_1, \boldsymbol{y}_2)$ from a sample of individuals together with their age $\boldsymbol{x}$, denote the joint probability $\pi_{j_1,j_2} = P(y_1 = j_1, y_2 = j_2)$, where the index $j_k, k = 1, 2$ corresponds to disease 1 and 2, respectively, and $j_k = 1$ (0) indicating past or current infection (susceptibility) for disease $k = 1, 2$. Modelling such multivariate categorical data can be done using conditional or marginal models [12].

A first marginal model that can be considered is the bivariate Dale model (BDM, [13, 14]). The BDM relates the probability of past or current infection for both diseases to the age at infection. The

bivariate Dale model consists of the following three models which are modelled simultaneously

$$\begin{cases} \text{logit}(\pi_{1+}|x) & = h_1(x), \\ \text{logit}(\pi_{+1}|x) & = h_2(x), \\ \log(\text{OR}|x) & = h_3(x). \end{cases} \tag{1}$$

Here OR denotes the age-dependent odds ratio $(\pi_{11}\pi_{00})/(\pi_{10}\pi_{01})$; π_{1+}, π_{+1} the marginal probabilities and h_i, $i = 1, 2, 3$ smooth differentiable functions. Using (1), when OR $\neq 1$, then $\pi_{11} = 1 + (\pi_{1+} + \pi_{+1})(\text{OR} - 1) - \{[1 + (\pi_{1+} + \pi_{+1})(\text{OR} - 1)]^2 + 4\text{OR}(1 - \text{OR})\pi_{1+}\pi_{+1}\}^{1/2}/(2(\text{OR} - 1))$, when OR $= 1$, then $\pi_{11} = \pi_{1+}\pi_{+1}$, it is straightforward to write down the multinomial (log)likelihood in terms of h_i, $i = 1, 2, 3$.

Modelling the OR allows us to describe the association between both diseases. An OR=1 indicates both infectious disease processes to behave independently. Alternatively, the correlation as an association parameter can be modelled in what is called the bivariate probit model (BPM) where probit-link functions are used instead of logit-link functions and a rhobit-link function relates the correlation to a smooth differentiable function of the covariate of interest, i.e. age here [15].

In fully marginal models, the parameters characterize the marginal probabilities and the association is of secondary importance. They also allow other functional forms to be related to the different response variables. Furthermore, the design is reproducible in the sense that the marginal models are consistent with the result obtained from separate univariate analyses.

In contrast, conditional models focus primarily on the association by looking at one variable conditional on the other. Consider the baseline-category logits model (BCL, see e.g. [16]). The BCL is a conditional model where three of the four joint probabilities are modelled proportionally to the remaining joint probability. Taking, e.g., the joint probability of being susceptible to either infection as the reference category, the BCL considers the following three equations simultaneously

$$\begin{cases} \log\left(\frac{\pi_{11}}{\pi_{00}}|x\right) & = h_1(x), \\ \log\left(\frac{\pi_{10}}{\pi_{00}}|x\right) & = h_2(x), \\ \log\left(\frac{\pi_{01}}{\pi_{00}}|x\right) & = h_3(x). \end{cases} \tag{2}$$

Again, using (2), the multinomial (log)likelihood can be expressed in terms of h_i, $i = 1, 2, 3$. For both BDM and BCL, h_i can take different forms. One can opt to use a (orthogonal) quadratic function, a FP or a smoothing spline. By using flexible functionals h_i the difference between marginal and conditional models with respect to their aims diminishes.

A multivariate extension of the smoothing spline approach was provided by the development of vector generalized additive models by Yee and Wild [7], who used vector smoothing [17] to extend the class of generalized additive models to a multivariate setting. In what follows we restrict attention to one covariate and refer to Yee and Wild [7] for the more general additive models.

The multivariate extension of a univariate smoother towards vector smoothers is provided by $\ell(h_1, h_2, h_3; \boldsymbol{y}) - \frac{1}{2}\sum_{i=1}^{3} \lambda_i \int \{h_i''(x)\}^2 dx$, where $\ell(h_1, h_2, h_3; \boldsymbol{y})$ denotes the loglikelihood of the multivariate model, λ_i, $i = 1, 2, 3$ denote component-specific smoothing parameters and $\int \{h_i''(x)\}^2 dx$, $i = 1, 2, 3$ denote component-specific penalties. Determining the optimal values for λ_i, $i = 1, 2, 3$ is done by generalized cross validation.

4.2 Hypotheses Testing

In both the BDM (1) and BCL (2) several hypotheses of interest can be tested using an F-test modification of the likelihood ratio test [7]. For the BDM, potential interest lies in: (1) $H_0 : h_1(x) = c + h_2(x)$ where c is an unknown constant, corresponding to the proportional odds assumption and; (2) $H_0 : h_3(x) = c$ implying that the OR is age-independent. Note that $H_0 : h_3(x) = 0$ corresponds to the hypothesis of independence. In the BCL, proportionality of $\pi_{01}(x)$ and $\pi_{10}(x)$ can be tested using $H_0 : h_2(x) = c + h_3(x)$. Other hypotheses could be tested too, but the interpretation in the context of infectious diseases is less straightforward and mostly not of interest.

Table 1: Parameter estimates and standard errors for the quadratic and best fractional polynomial.

Univariate Model			Bivariate Model		
Parameter	Estimate	s.e.	Parameter	Estimate	s.e.
VZV					
Quadratic Polynomial Model					
Intercept	2.956	0.112	Intercept	2.700	0.101
age	49.381	4.901	age	44.437	3.488
age^2	-39.851	4.334	age^2	-31.167	3.431
Fractional Polynomial Model					
Intercept	5.699	0.295	Intercept	-2.654	0.321
age^{-1}	-4.072	0.987	age	1.048	0.096
$age^{-1}\log(age)$	-13.193	1.480	$age\log(age)$	-0.243	0.026
B19 Virus					
Quadratic Polynomial Model					
Intercept	0.537	0.046	Intercept	0.382	0.047
age	31.254	2.251	age	29.283	1.829
age^2	-27.558	2.298	age^2	-22.001	1.865
Fractional Polynomial Model					
Intercept	-11.084	0.956	Intercept	-11.052	0.955
$age^{0.5}$	6.373	0.582	$age^{0.5}$	6.350	0.581
$age^{0.5}\log(age)$	-1.217	0.121	$age^{0.5}\log(age)$	-1.212	0.121
OR					
Quadratic Polynomial Model					
-	-	-	Intercept	0.792	0.178
Fractional Polynomial Model					
-	-	-	Intercept	0.724	0.186

5 Application to the Data

In Table 1, parameter estimates and standard errors are given for both the univariate model and the BDM with quadratic and best FP, while Table 2 shows the parameter estimates and standard errors for the quadratic polynomial and best FP for the BCL. Note that the quadratic model is a specific FP of degree 2 with powers 1 and 2 and, is, in that sense non-optimal.

There is a moderate difference between the estimates for the BDM-marginal prevalences compared to the corresponding univariate estimates (Table 1). The standard errors of the parameters based on the bivariate model are smaller than the corresponding ones for the univariate models. This corresponds to the increased efficiency associated with bivariate modelling. For VZV, the best FPs differ. While the use of FPs already provides a fair amount of flexibility, using smoothing splines allows more features of the data to be revealed.

A comparison of the AIC-values of the different univariate models showed that the spline-based model provides the best fit for both VZV (1109.17 compared to 1129.81 and 1110.58) and B19 (2739.59 compared to 2796.49 and 2766.12). For the bivariate models, the BDM and BCL with quadratic polynomials resulted in the largest AIC-values (3910.18 and 3887.35). A moderate improvement was found using the FPs (3872.70 and 3872.19), while again the spline-based models resulted in the lowest AIC-values (3848.89 and 3857.26, respectively). Since the AIC-values for both BDM and BCL are comparable due to the multinomial nature of the models, we end up with the spline-based BDM model as the bivariate model with lowest AIC-value.

In Figure 1, the age-dependent prevalences for VZV and B19 together with the OR according to the spline-based BDM are shown. Testing for a constant OR ($H_0 : h_3(x) = c$ in (1)) using the BDM-models and the F-test modification of the likelihood ratio test as mentioned in Section 4.2 resulted

Table 2: Parameter estimates and standard errors for the baseline category logits model using a quadratic and the best fractional polynomial of age.

Parameter	Estimate	s.e.
$\log(\pi_{11}/\pi_{00})$		
Quadratic Polynomial Model		
Intercept	3.082	0.165
age	71.633	5.492
age^2	-55.061	5.040
Fractional Polynomial Model		
Intercept	12.803	0.827
age^{-1}	29.718	4.016
$\text{age}^{-0.5}$	-43.249	3.663
$\log(\pi_{10}/\pi_{00})$		
Quadratic Polynomial Model		
Intercept	2.596	0.167
age	47.479	5.404
age^2	-37.706	4.947
Fractional Polynomial Model		
Intercept	-3.006	0.361
$\text{age}^{0.5}$	1.521	0.141
age^2	-0.001	3e-4
$\log(\pi_{01}/\pi_{00})$		
Quadratic Polynomial Model		
Intercept	0.064	0.210
age	42.739	7.722
age^2	-43.874	7.501
Fractional Polynomial Model		
Intercept	-3.802	0.460
age	0.353	0.059
age^2	-0.005	0.001

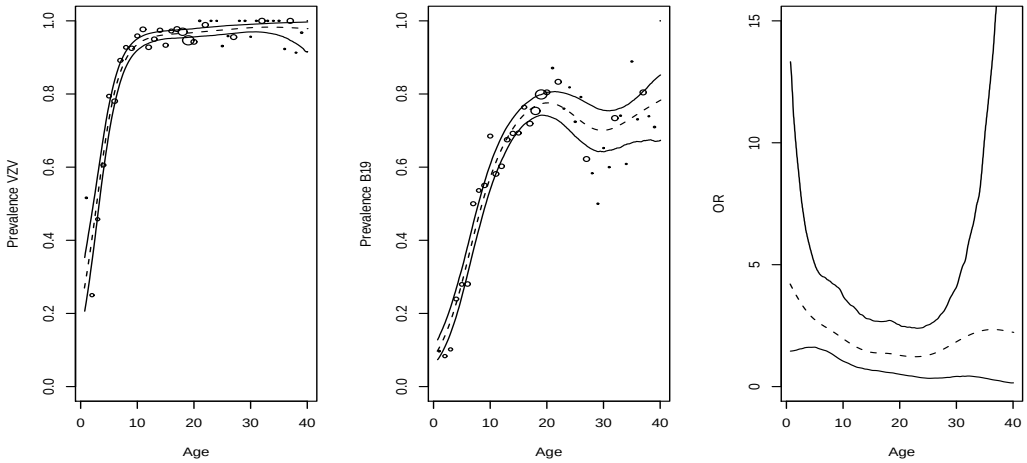

Figure 1: The marginal prevalence-curves for VZV (first panel) and B19 (second panel) together with the OR (third panel) according to the spline-based BDM together with 95% bootstrap-based pointwise confidence intervals.

in a constant OR (a p-value of 0.37 for the spline-based BDM). Bootstrap-based confidence intervals confirmed the latter result (Figure 1).

The odds ratios, according to the different models, are significantly larger than 1, indicating dependency of VZV- and B19-occurrence. Fitting the spline-based BDM model, the estimated OR equals 2.11 with 95% confidence interval (1.45,3.23), meaning that the odds of past or current VZV(B19)-infection among the B19(VZV)-non-susceptible group is 2.11 times larger than the odds of past or current VZV(B19)-infection among the B19(VZV)-susceptible group. Non-susceptibility referring to past or current infection. Similarly, fitting a spline-based bivariate probit model gives an age-independent correlation of 0.21 with 95% confidence interval (0.20,0.61). While the hypothesis of a constant OR was not rejected, the hypotheses corresponding to the proportional odds assumption (BDM), independence (BDM) and proportionality of $\pi_{01}(x)$ and $\pi_{10}(x)$ (BCL) were all rejected at the 5%-significance level.

The left panel of Figure 2 shows the spline-based BDM and BCL estimated joint probabilities together with a barplot of the observed proportions. A deviance of 3043.99 (3046.57) on 5239.33 (5236.43) degrees of freedom for the BDM (BCL) indicates a good fit to the data.

6 Monotonicity

Assuming time homogeneity and the presence of antibodies to be lifelong, the prevalence should be a monotone increasing function of age. In case of heterogeneity, constrained estimation leads to an improved estimation of the (sero)prevalence as a function of age (see e.g. [**?** 18]). [19] showed that constrained smoothing leads to estimates of the form 'smooth then constrain'. This method was applied before to hepatitis A, rubella and measles [18].

For the bivariate modelling this means that π_{11}, π_{1+} and π_{+1} should be monotone increasing and consequently π_{00} should be monotone decreasing, while there is no restriction on π_{10} and π_{01}. Therefore, we apply the Pool Adjacent Violater Algorithm ('PAV', [20]) to monotonize the marginal and joint estimated prevalences. This results in $\hat{\pi}_{10} = \text{PAV}(\hat{\pi}_{1+}) - \text{PAV}(\hat{\pi}_{11})$ and $\hat{\pi}_{01} = \text{PAV}(\hat{\pi}_{+1}) - \text{PAV}(\hat{\pi}_{11})$ as estimates for π_{10} and π_{01}, respectively.

In Figure 2, the non-monotone result (left panel) is contrasted with its monotonized version (right panel) in terms of the joint probabilities, resulting in differences at larger age-values for both $\hat{\pi}_{11}$ and

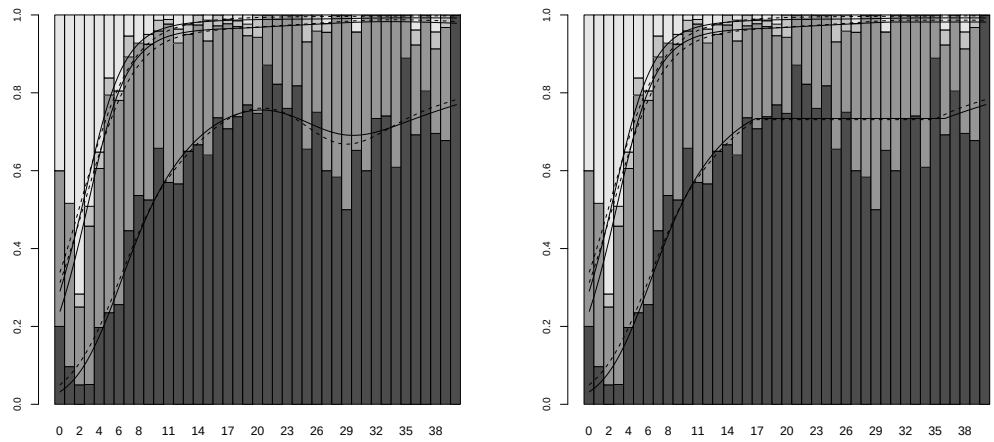

Figure 2: The joint probabilities according to the spline-based BDM (solid lines) and BCL (dashed lines). Observed proportions are shown from dark grey to light grey for p_{11}, p_{10}, p_{01} and p_{00}, respectively. In the left panel: the non-monotone result, in the right panel: the monotonized result.

$\hat{\pi}_{10}$.

Clearly, monotonization is needed for B19. Up to now, the reason for the decrease in seroprevalence around the age of 30 years is unknown. Possible explanations include changes in contact behaviour but this has not been formally assessed.

7 Estimating the Force of Infection

A key parameter in studying infectious disease dynamics is the per capita rate at which a susceptible gets infected, or equivalently the hazard to become infected. The term used in infectious disease modelling for this hazard is the force of infection (FOI). From an assumed functional form for the FOI, the prevalence can be expressed as $\pi(a) = 1 - e^{-\int_0^a \lambda(x)dx}$, and using seroprevalence data and the corresponding binomial loglikelihood, the parameters in the functional form for $\lambda(x)$ can be estimated. [3] for instance parametrized the FOI as a gamma function of age $\lambda(a) = \alpha x^{\beta} exp(-a/\gamma)$ and derived estimators of α, β and γ using maximum likelihood.

Alternatively, the FOI can be estimated from the prevalence $\pi(a)$ using $\lambda(a) = \pi^{'}(a)/(1 - \pi(a))$. This allows us to use existing methodology to model the prevalence and derive the FOI thereof. Especially, the use of (vector) generalized additive models is appealing in terms of their flexibility and the inherited flexibility of the estimated FOI-curve.

Using the marginal prevalences derived from the spline-based BDM and monotonized using the 'PAV'-function, we can derive the FOI. More explicitly, it is straightforward to show that using a logit-link function, the force of infection can be expressed as

$$\lambda_{VZV}(a) = h_1^{'}(a)\pi_{1+}(a), \quad \text{and} \quad \lambda_{B19}(a) = h_2^{'}(a)\pi_{+1}(a), \tag{3}$$

for VZV and B19, respectively. Alternatively, the non-monotonized prevalence can be used to derive the FOI using (3) and the negative values can then be put to zero. The latter estimation method falls again in the 'smooth then constrain' approach. Figure 3 shows the prevalence and FOI curves based on the BDM with application of the 'PAV'-function.

The marginal prevalence for VZV increases rapidly over the first 10 years of life, reaches the 90 percent level already at 8.2 years and an absolute maximum of 98.2 percent at 30.7 years of age. The

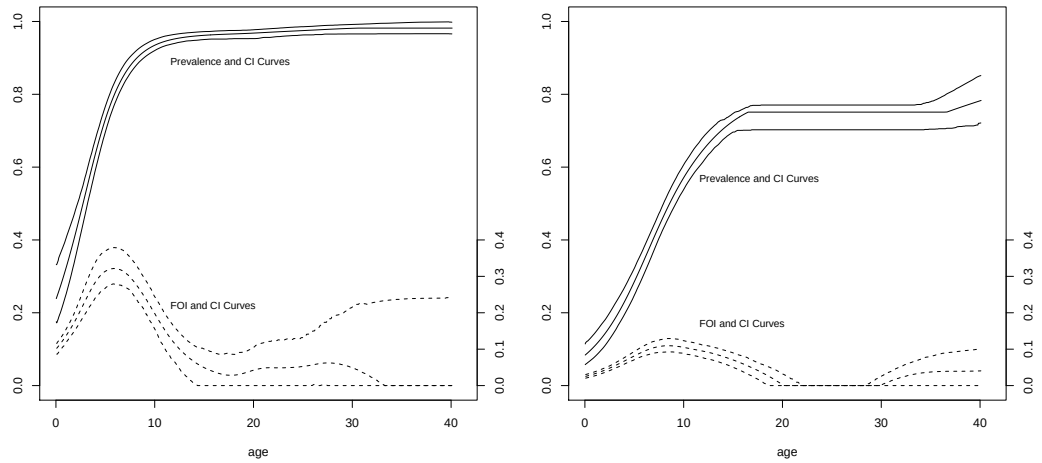

Figure 3: The marginal prevalences and FOI-curves according to the monotonic spline-based BDM together with 95% bootstrap-based pointwise confidence intervals. In the left panel: VZV, in the right panel: B19.

prevalence of B19 knows a more moderate increase over the first 16 years, when it reaches 75 percent and remains constant thereafter. This is translated in the FOI-curves showing the age of maximal FOI at 6.1 years of age (95% CI: (4.9,7.0)) for VZV and 8.3 years of age (95% CI: (7.6,12.0)) for B19. The FOI is nominally much larger for VZV (maximum: 0.32) than for B19 (maximum: 0.11). The confidence intervals show small variability while the application of the 'PAV'-function is invasive for the prevalence for B19 between 16.5 and 36.4 years of age. This is also reflected in the estimated FOI-curve for B19. The latter finding indicates that the assumption of time homogeneity, under which the FOI can be derived from seroprevalence, could be violated. There is however, no indication that this is truly the case.

7.1 Conditional FOI

Using multi-sera data, one can analyse quantities like the prevalence and FOI for one infection conditional on being in a specific state for the second infection (Figure 4). The potential interest in conditional prevalences and FOIs is not only related to quantifying the association between two or more infections (eg to evaluate the impact of combination vaccines) but could also be valuable to analyse chronic co-infections, such as Human Papilomavirus (HPV), Human Immunodeficiency Virus (HIV) and hepatitis B (HBV) or C (HCV) virus where the acquisition of a second related infection (multitype, eg HPV type 16 and HPV type 18, or multi-virus, eg HBV and HCV), could have a dramatic impact on the course of disease and infectiousness to others [21].

Suppose that conditional on B19, one is interested in the rate of acquiring VZV. Thus one looks at the quantities $\lambda_{VZV=1|B19=i} = \pi'_{VZV=1|B19=i}/(1-\pi_{VZV=1|B19=i})$, where $i=1$ ($i=0$) if the state for B19 is (non-)infected (Figure 4). Similarly, one can be interested in looking at the rate of acquiring B19 conditional on VZV and thus $\lambda_{B19=1|VZV=i} = \pi'_{B19=1|VZV=i}/(1-\pi_{B19=1|VZV=i})$, where $i=1$ ($i=0$) if the state for VZV is (non-)infected (Figure 4). In Figure 5, the conditional and marginal prevalence- and FOI-curves are shown.

It is straightforward to verify that $\lambda_{VZV=1|B19=1} = \lambda_{VZV=1|B19=0} = \lambda_{VZV=1}$ ($\lambda_{B19=1|VZV=1} = \lambda_{B19=1|VZV=0} = \lambda_{B19=1}$) if and only if VZV and B19 are independent. In Section 4, VZV-and B19-occurrence were shown to be dependent and as such the difference depicted in Figure 5 is signif-

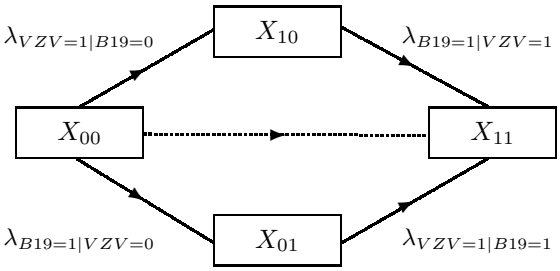

Figure 4: Schematic representation of the flow of individuals among different stages. X_{ij} denotes the number of persons in class $\{i, j\}$, $i, j = 0, 1$ denoting whether infected (1) or not (0) for VZV and B19, respectively.

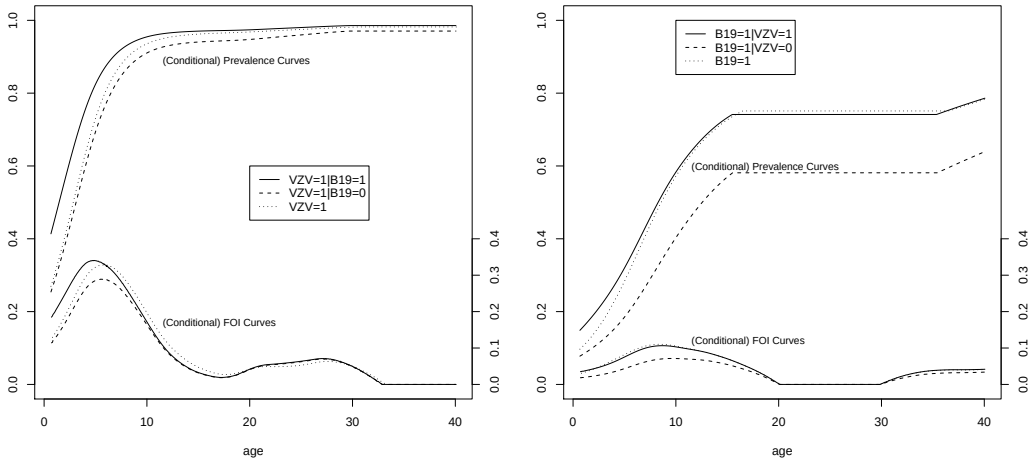

Figure 5: The conditional and marginal prevalences and FOI-curves according to the monotonic spline-based BDM.

icant. Moreover, it is straightforward to verify that a positive dependence (OR>1) is equivalent to $\lambda_{B19=1|VZV=1} > \lambda_{B19=1|VZV=0}$ and $\lambda_{VZV=1|B19=1} > \lambda_{VZV=1|B19=0}$.

Both panels of Figure 5 show that the FOI for VZV (B19) conditional on being infected for B19 (VZV) is higher compared to the FOI for VZV (B19) when conditioning on the non-infected status of B19 (VZV). In other words, a person who is infected for VZV (B19) has a higher probability to become infected for B19 (VZV) compared with persons still susceptible for VZV (B19). Since B19 and VZV are transmitted through close contacts, it is more likely that a person who has already been infected with a first infection has had more close contacts than a person who has not been infected. Such a once infected person is also more likely to continue having more contacts through which a second, similarly transmitted infection, can be acquired, than a person who has not been infected yet. Note that these findings are again in accordance to the positive OR in Section 5.

7.2 Joint Force of Infection

Next to the conditional FOIs, looking at the quantity $\lambda_{VZV=1,B19=1} = \pi'_{11}/(1-\pi_{11})$, could be of interest. However, the interpretation of this joint FOI is tedious. The numerator indicates the proportion that is still susceptible for at least one of the two infections, while the denominator gives us the instantaneous rate at which persons, at least susceptible for one of both infections, go to the state of having (had) both infections. Since VZV and B19 have a short generation interval, simultaneous acquisition is unlikely to occur. Moreover, due to the discrete nature of the data; i.e. age at infection is measured in days; the contribution to this rate from individuals moving from a fully susceptible status to a fully infected status (see dashed arrow in Figure 4) for both diseases should be interpreted as the rate of acquisition of both infections in one day.

In Figure 6, it is shown that the joint FOI coincides almost entirely (especially from 10 years onwards) with the conditional FOI for B19 given one is VZV-infected. This is not surprising given that VZV is the dominant infection and its acquisition is almost complete at 10 years of age. The contribution to the joint FOI thus merely comes from those individuals infected by VZV but still susceptible for B19. Looking more carefully at the expressions of both FOIs in terms of prevalences; $\pi'_{VZV=1,B19=1}/(1-\pi_{VZV=1,B19=1})$ and $\pi'_{B19=1|VZV=1}/(1-\pi_{B19=1|VZV=1})$; near equality holds when $\pi_{VZV=1,B19=1} \cong \pi_{B19=1|VZV=1}$ or equivalently when $\pi_{VZV=1} \cong 1$ which is indeed the case for individuals from 12 years onwards. Moreover, this result indicates that acquiring both diseases on the same day is very unlikely.

In the next section, we go deeper into the relation between the FOI and the way people of different ages mix and how the infections spread.

7.3 Proportional FOI

The average force of infection is the average hazard to acquire infection and, if assumed that we have a short infectious period, can be calculated using the integral equation (see [3])

$$\lambda(x) = \frac{ND}{L} \int_0^\infty \beta(x,y)\lambda(y)S(y)M(y)dy, \tag{4}$$

with population size N, mean duration of the infectious period D, life expectancy L, number of susceptibles at age y: $S(y)$, survivor function $M(y)$ and the contact function $\beta(x,y)$ which denotes the per capita rate at which an individual of age y makes effective (i.e. enabling transfer of infection) contacts with individuals of age x.

Usually, $\beta(x,y)$ is described by a matrix, presenting contact rates between different age-classes, referred to as the contact or mixing matrix. The solution of (4) can often not be determined, since $\beta(x,y)$ is typically unknown. If however, $\beta(x,y)$ is assumed to be of a specific form, Farrington et al.[3] showed how Bayes factors can be used to select the most appropriate form from a set of possible forms. One such form is separable mixing, where it is assumed that there exist functions u and v such that $\beta(x,y) = u(x)v(y)$, in other words the rate for a susceptible person to make contact with an infectious

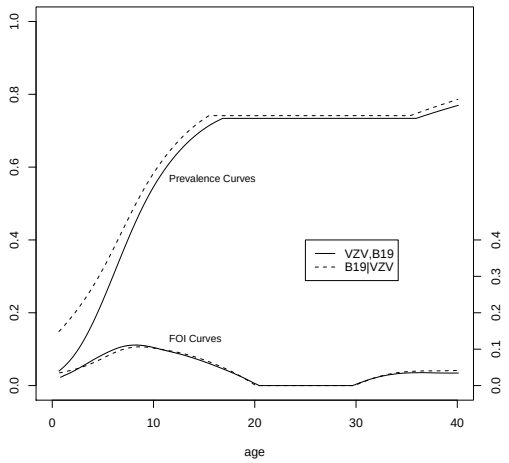

Figure 6: The joint and B19|VZV=1-conditional FOI together with the joint and B19|VZV=1-conditional FOI according to the monotonic spline-based BDM.

person is the product of the density of susceptibles and infected people. Under separable mixing it can be shown that there exists a function $\ell(y)$ such that $\beta(x,y) \propto \lambda(x)\ell(y)$. So if assumed that mixing is separable, and that the infections under consideration are transmitted through the same routes, the forces of infection should be proportional too. If they are not, the separable mixing assumption is not fulfilled. [3] used a likelihood ratio test to test for proportional FOIs while assuming the shape of the FOI to follow a gamma function. By using spline smoothers robustness with respect to possible model misspecification is achieved. Moreover, assuming a gamma function, the FOI-curve is restricted to have only one maximum while a secondary local maximum is likely to be observed because of parent-child transmission.

The proportionality of both marginal FOIs can be expressed as $\lambda_{1+}(x) = \alpha\lambda_{+1}(x)$ and equivalently, in terms of the marginal prevalences, $1-\pi_{1+}(x) = \gamma(1-\pi_{+1}(x))^{\alpha}$, where $\gamma = (1-\pi_{1+}(0))/(1-\pi_{+1}(0))$ and α is an unknown constant. Some further rewriting shows that proportionality of both marginal FOIs can be assessed by looking at $\log(1-\pi_{1+}(x)) - \alpha\log(1-\pi_{+1}(x)) = \gamma^{'}$, where $\gamma' = \log(\gamma)$ and α is the unknown constant. We then consider the bivariate model

$$\begin{cases} \log(1-\pi_{1+}(x)) = h_1(x) \\ \log(1-\pi_{+1}(x)) = h_2(x) \\ \log(\text{OR}) = h_3(x) \end{cases} \tag{5}$$

where $h_i, i = 1,2,3$ are again smooth functions. For the marginal prevalences to be monotone, both $h_i, i = 1,2$ should monotone decreasing. Similar as for the BDM, the corresponding loglikelihood can be derived and monotonicity can be achieved using the 'PAV'-function. Specifying spline functions for $h_i, i = 1,2,3$ in (5) and testing for spline effects resulted in a constant OR = 2.14, while the spline was maintained for both marginal prevalences. Within this model formulated as (5) with $h_1(x) = \beta_0 + s_1(x); h_2(x) = \beta_1 + s_2(x)$; and $h_3(x) = \beta_2$, testing for proportionality reduces to testing the hypothesis: $s_2(x) = s_1(x)/\alpha$. The likelihood ratio test $s_2(x) = s_1(x)/\alpha$ is approximately χ^2-distributed with degrees of freedom equal to the difference in empirical degrees of freedom of the two models. If H_0 is valid, the null model will give us an estimate of α.

The application to VZV and B19 resulted in a p-value of 0.0012 indicating proportionality does not hold and consequently contradicts the separable mixing assumption. Note that using (5) to model bivariate data resulted in an AIC-value of 3857.07 which is comparable to the AIC-value of the spline-based BCL-model but still considerably larger than the AIC-value of the spline-based BDM-model

(Section 5).

8 Discussion

In this paper, based on multisera data, the bivariate Dale model and the baseline category logits model are used to model the marginal prevalence of both Varicella-Zoster Virus and Parvo-virus B19. Using splines to achieve the necessary flexibility through the use vector generalized additive models, the intrinsic difference among marginal and conditional models diminishes.

The use of a bivariate model for this kind of data, improves not only the efficiency but allows us to study the association between infections. It is shown that the acquisition of VZV-and B19 is positively related. Most likely, due to the similarity in transmission by close contacts for both infections. Co-infections, i.e. joint infections caused by more than one pathogen, are an aggravating factor in disease progression for virtually all infections. For VZV and B19, this is unlikely to occur because: (1) acute VZV and B19 are short lived, (2) when people are infected they will change their mixing behaviour (i.e. stay home in bed), and substantially reduce their chance of contacting a person infected by another pathogen unless the first infection is no longer there as is the situation for VZV and B19 since also the infectious periods for both are very short. For HBV, HCV, HPV and HIV chronic or persistent infection may occur, giving rise to a very long infectious period. The methods we expanded here may be most relevant for these potentially chronic infections, as the chance that co-infection occurs during the long asymptomatic period typical of each of these infections is great. The conditional FOI-curves can be regarded as quantifications of how strong the association between the infections is. Individuals already infected for one disease are more likely to become infected for the other disease (with similar transmission routes) at an earlier age as compared to the average acquisition age for that particular disease.

Reparametrizing the Dale model to use 'complementary log'-links for the marginal prevalences to test for proportionality of the FOI and consequently separable mixing showed that proportionality of the FOIs does not hold and therefore, the separability assumption for VZV and B19 is violated. This is due to the fact that contacts are mainly assortative with age.

Extensions towards more than two infections are straightforward using the existing extensions of the bivariate Dale and bivariate category logits model (see [13], [16]). Extensions to include several covariates in the presented model is straightforward. The inclusion of gender in the model even up to its interaction with age resulted in a non-significant contribution to the model (p-value 0.28). Furthermore, the application of the presented methodology to data from several countries is straightforward.

Acknowledgements

We thank both reviewers for their comments leading to an improved presentation of the paper. This work was based on a serum sample collected for the European Commission's ESEN2-project. We are grateful to the Institute of Public Health, Brussels (Dr Robert Vranckx, Dr Veronik Hutse) for assistance with B19 testing. This work has been partly funded by POLYMOD, a European Commission project funded within the Sixth Framework Programme, Contract number: SSP22-CT-2004-502084, by the Fund of Scientific Research (FWO, Research Grant n° G039304) in Flanders, Belgium, by "SIMID", a strategic basic research project funded by the institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), project number 060081 and by the IAP research network nr P5/24 of the Belgian Government (Belgian Science Policy).
We thank Thomas Yee for his technical support concerning the VGAM-library in R.

References

[1] R.M. Anderson and R.M. May. *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, Oxford, 1991.

[2] N. Keiding. Age-specific incidence and prevalence: A statistical perspective (with discussion). *JRSS-A*, 154:371–412, 1991.

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

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

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

[6] A.J. Sutton, N.J. Gay, W.J. Edmunds, V.D. Hope, O.N. Gill, and M. Hickman. Modelling the force of infection for hepatitis b and hepatitis c in injecting drug users in England and Wales. *BMC Infectious Diseases*, 6:93, 2006.

[7] T.W. Yee and C.J. Wild. Vector generalized additive models. *JRSS-B*, 58:481–493, 1996.

[8] A. Nardone and E. Miller. Serological surveillance of rubella in europe: European sero-epidemiology network (ESEN2). *Euro-surveillance*, 9(4):5–7, 2004.

[9] P. McCullagh and J.A. Nelder. *Generalized Linear Models*. Chapman & Hall, 1989.

[10] P. Royston and D.G. Altman. Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling. *Applied Statistics*, 43(3):429–467, 1994.

[11] T.J. Hastie and R.J. Tibshirani. *Generalized Additive Models*. Chapman and Hall, London, 1990.

[12] K.Y. Liang, S.L. Zeger, and B. Qaqish. Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society. Series B*, 54:3–40, 1992.

[13] J. R. Dale. Global cross-ratio models for bivariate, discrete, ordered responses. *Biometrics*, 42:909–917, 1986.

[14] J. Palmgren. Regression models for bivariate binary responses. Technical Report 101, Department of Biostatistics, University of Washington, Seattle, 1989.

[15] J.R. Ashford and R.R. Sowden. Multi-variate probit analysis. *Biometrics*, 26:535–546, 1970.

[16] A. Agresti. *Categorical Data Analysis*. Wiley & Sons, Hoboken, second edition, 2002.

[17] J.A. Fessler. Nonparametric fixed-interval smoothing with vector splines. *IEEE Trans. Signal Process.*, 39:852–859, 1991.

[18] Z. Shkedy, M. Aerts, G. Molenberghs, Ph. Beutels, and P. Van Damme. Modelling forces of infection by using monotone local polynomials. *Applied Statistics*, 52(4):469–485, 2003.

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

[20] R. E. Barlow, D. J. Bartholomew, J. M. Bremner, and H. D. Brunk. *Statistical Inference Under Order Restrictions*. Wiley, New York, 1972.

[21] A. Alberti and others (Jury Panel). Short statement of the first european consensus conference on the treatment of chronic hepatitis b and c in hiv-co-infected patients. *Journal of Hepatology*, 42:615–624, 2005.