TECHNICAL
REPORT

0233

Modelling age dependent force of infection from prevalence data using fractional polynomials

Z. Shkedy, M. Aerts, G. Molenberghs, P. Beutels, and P. Van Damme

IAP STATISTICS
NETWORK

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Modeling Age Dependent Force of Infection From Prevalence Data Using Fractional Polynomials

Z. Shkedy[1,*], M. Aerts[1], G. Molenberghs[1], Ph. Beutels[2] and P. Van Damme[2]

[1] *Center for Statistics, Limburgs Universitair Centrum, Universitaire Campus - gebouw D,*
B–3590 Diepenbeek, Belgium

[2] *University of Antwerp, Epidemiology and Community Medicine, Center for Evaluation of Vaccination*
B 2610 Antwerp, Belgium

Summary

The force of infection is one of the primary epidemiological parameters of infectious diseases. For many infectious diseases it is assumed that the force of infection is age dependent. Although the force of infection can be estimated directly from a follow up study, it is much more common to have cross-sectional seroprevalence data from which the prevalence and the force of infection can be estimated. In this paper we propose to model the force of infection within the framework of fractional polynomials. We discuss several parametric examples from the literature and show that all of these examples can be expressed as special cases of fractional polynomial models. We illustrate the method on five seroprevalence samples, two of Hepatitis A, and one of Rubella, Mumps and Varicella.

Keywords: Seroprevalence; Force of Infection; Conventional Polynomials; Fractional

[*] Correspondence to: Ziv Shkedy, Center for Statistics, Limburgs Universitair Centrum, Universitaire Campus - gebouw D, B–3590 Diepenbeek, Belgium, email: ziv.shkedy.@luc.ac.be

Polynomials; Generalized Linear Models.

1. Introduction

Mathematical models are often used to describe the process of infectious diseases at population level (Anderson and May, [1]). Such compartmental models consist of a set of differential equations which aim to describe the flow of individuals from one disease stage to the other. In this paper, we assume the disease is irreversible, meaning that the immunity is assumed to be lifelong. We further assume that the mortality caused by the infection is negligible and can be ignored. Let $q(a,t)$ be the fraction of susceptible individuals at age a and time t. Under the assumptions stated above the partial differential equation which describes the change in the susceptible fraction at age a and time t is given by :

$$\frac{\partial}{\partial a}q(a,t)+\frac{\partial}{\partial t}q(a,t)=-\ell(a,t)q(a,t). \tag{1}$$

Here $\ell(a,t)$ is the rate at which susceptible individuals become infected and is called the hazard or the force of infection. Note that (1) assumes that the natural death rate is zero up to the life expectancy and thereafter infinity. In a steady state, the time homogeneous form of the model, $\frac{\partial}{\partial t}q(a,t)=0$ and (1) reduces to

$$\frac{d}{da}q(a)=-\ell(a)q(a). \tag{2}$$

Differential equation (2) describes the change in the susceptible fraction with the host age. This representation of the model is called the static model.

In practice, the force of infection can be estimated from a seroprevalence cross-sectional sample. Figure 1 shows five datasets that will be discussed in this paper. The Rubella and the Mumps datasets were used by Farrington [2] to illustrate the use of nonlinear models for estimating the force of infection. Keiding [3] used the Hepatitis A dataset from Bulgaria to illustrate the use of isotonic regression as an nonparametric approach to estimate the prevalence and the force of infection. Shkedy *et al.* [4] used the Hepatitis A dataset from Belgium to illustrate the use of local polynomials as a nonparametric method to estimate both the prevalence and the force of infection. This dataset consists of 3161 individuals with age range between 1 and 86 years old, sampled in 1993 in Belgium. For more details about the sample we refer to Beutels *et al.* [5]. Lastly, the Varicella dataset consists of 1673 individuals with age range from 1 to 44 years old that was sampled in Belgium between October 1999 to April 2000. For more details about this study we refer to Thiry *et al.* [6].

FIGURE 1, ABOUT HERE.

Muench [7] suggested to model the infection process with a catalytic model, in which the distribution of the time spent in the susceptible class is exponential with rate β. The force of infection in this case, β, is age independent. Under the catalytic model $q(a) = e^{-\int_0^a \beta ds} = e^{-\beta a}$ and $\frac{d}{da}q(a) = -\beta e^{-\beta a}$. Griffiths [8] proposed a model for measles in which the force of infection increases linearly in the age range 0–10. Grenfell and Anderson [9] extended the model further and used polynomial functions to model the force of infection. Their model assumes that $q(a) = e^{-\Sigma \beta_i a^i}$ which implies that the force of infection is $\ell(a) = \sum \beta_i i a^{i-1}$. For the general case the solution for (2) under the catalytic model is $q(a) = e^{-\gamma(a)}$, where $\gamma(a) = \int_0^a \ell(s)ds$ is the cumulative hazard.

One problem that arises when a higher order polynomial model is fitted is that the estimate for the force of infection can get negative. In fact, a force of infection estimate turns negative whenever the estimated probability to be infected before age a is a nonmonotone function. One solution to this problem is to define a nonnegative force of infection, $\ell(a, \boldsymbol{\beta}) \geq 0$ for all a, and to estimate $\pi(a)$ under these constrains. Farrington [2], Farrington *et al.* [10] and Edmunds *et al.* [11] applied this method for measles, mumps and rubella, using a nonlinear model for $\pi(a)$. However, Farrington's method requires prior knowledge about the dependence of the force of infection on age. Other parametric models, fitted within the framework of generalized linear models (GLM) with binomial error (McCullagh and Nelder [12]), were discussed by Becker [13], Diamond and McDonald [14] and Keiding *et al.* [15] who used models with complementary log-log link function in order to parameterize the prevalence and the force of infection as a Weibull model. Becker [13] suggested to model a piecewise constant force of infection by fitting a model with log link. For the case that other covariates, in addition to age, are included in the model, Jewell and Van Der Laan [16] proposed, in the context of current status data, a proportional hazard model with constant force of infection which can be fitted as a GLM with complementary log-log link. Grummer-Strawn [17] discussed two parametric models, the first being a Weibull proportional hazard model with complementary log-log link and the second being a log-logistic model with logit link function. For the latter, the proportionality in the model is interpreted as proportional odds.

A nonparametric method was discussed by Keiding [3] who used isotonic regression to estimate the prevalence and applied kernel smoothers to estimate the force of infection. Keiding *et al.* [15] proposed to model the force of infection using natural cubic splines. Recently Shkedy *et al.* [4] proposed to use local polynomials to estimate both the prevalence, $1 - q(a)$, and the force of infection.

Shiboski [18] proposed a semiparametric model, based on generalized additive models (Hastie and Tibshirani [19]), in which the dependency of the force of infection and age is modeled nonparametrically and the covariate effect is the parametric component of the model. Depending on the link function, the model proposed by Shiboski [18] assumes proportionality; proportional hazard (complementary log-log link) or proportional odds (logit and probit links). Other semiparametric models, assuming a logit link, were proposed by Rossini and Tsiatis [20]. In this paper we restrict the discussion to parametric models for which the only covariate in the model is the host age. In Section 2 we describe a general age-dependent model for the force of infection, based on prevalence data. Section 3 discusses fractional polynomials as a flexible parametric approach to model the force of infection. The method is applied within the framework of generalized linear models for binary response. In Section 4 we apply the method to the datasets mentioned above. The models in Section 4 assume a logistic form of $q(a)$ and were fitted with the logit link function. In Section 5 we modify this assumption and model the force of infection with fractional polynomial for which $q(a) = \exp(-\gamma(a))$.

2. Age-Dependent Force of Infection

Consider an age-specific cross-sectional prevalence sample of size N and let a_i be the age of the ith subject. Instead of observing the age at infection we observe a binary variable Y_i such that

$$Y_i = \begin{cases} 1 & \text{if subject } i \text{ had experienced infection before age } a_i, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

With $\pi(a_i)$ be the probability to be infected before age a_i, $\pi(a_i) = 1 - q(a_i)$, the log likelihood is given by

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{N} Y_i \log\left\{\pi(a_i)\right\} + (1 - Y_i)\log\left\{1 - \pi(a_i)\right\}. \tag{4}$$

Here, $\pi(a) = g^{-1}(\eta(a))$, where $\eta(a)$ is the linear predictor and g is the link function. For binary responses, g is often taken to be a logit link function, $\log(\pi/(1-\pi))$, but other link functions such as the complementary log-log link, $\log(-\log(1-\pi))$, and log link, $-\log(1-\pi)$, can be used as well. The models proposed by Muench [7], Griffiths [8] and Grenfell and Anderson [9] assume g to be the log link function (for $1-\pi$) and $\eta(a) = \sum_{i=0}^{k} \beta_i a^i$, where k is equal to 1 (Muench), 2 (Griffiths) and K (Grenfell and Anderson). Using a model with log link function leads to a simple interpretation of the first derivative of the linear predictor. Indeed, $\eta(a)$ is the cumulative hazard and therefore the force of infection is simply the first derivative of the linear predictor. Under the catalytic model $\pi(a) = 1 - e^{-\eta(a)}$, using the definition for the hazard rate, we get

$$\ell(a) = \frac{\pi'(a)}{1-\pi(a)} = \frac{\eta'(a)e^{-\eta(a)}}{e^{-\eta(a)}} = \eta'(a). \tag{5}$$

In the general case, when the link function is not restricted to be the log link, the force of infection can still be derived according to (5). It is easy to see that for the binomial distribution, the force of infection can be expressed as a product of two functions:

$$\ell(a) = \eta'(a)\delta(\eta(a)), \tag{6}$$

where the form of $\delta(\cdot)$ is determined by the link function g. Table 1 shows three possible link functions with their corresponding structure for the force of infection.

TABLE 1, About Here.

3. Fractional Polynomial Models for Binomial Data

3.1. Motivating Example

Viral hepatitis is a serious problem throughout the world. In Belgium, the most common form of viral hepatitis infection is caused by the hepatitis A virus. We consider a cross-sectional prevalence sample ($N = 3161$), taken in 1993 and at the beginning of 1994 from 11 hospitals in Belgium. We consider two generalized linear models with logit and complementary log-log link functions. For the logit model the linear predictor is $\eta(a) = \beta_0 + \beta_1 a + \beta_2 a^3$. This model has a deviance of 82.74 on 83 degrees of freedom. For the complementary log-log model $\eta(a) = log(\beta_0) + \beta_1 a^2 + \beta_2 a^3$. The deviance of this model is 81.41 on 83 degrees of freedom. The force of infection of these models can be derived from Table 1. Although both models fit the data well, Figure 2 shows that both models predict negative forces of infection at the higher age groups.

FIGURE 2, ABOUT HERE.

3.2. Fractional Polynomials

The motivation to model the force of infection with fractional polynomials is to allow for flexible changes in the force of infection over the age of the host. Indeed, high order conventional polynomials offer a wide range of curve shapes but often fit the data badly at the extremes of the observed age. Moreover, conventional polynomials do not have asymptotes and fit the data poorly whenever asymptotic behavior of the infection process is expected. Royston and Altman [21] introduced the family of fractional polynomials as a generalization of the conventional polynomials class of functions. In the context of binary responses, a fractional polynomial of degree m for the linear predictor is defined as

$$\eta_m(a,\ \boldsymbol{\beta},\ p_1, p_2 \ldots p_m) = \sum_{i=0}^{m} \beta_i H_i(a), \tag{7}$$

where m is an integer, $p_1 \leq p_2 \leq \cdots \leq p_m$ is a sequence of powers and $H_i(a)$ is a transformation function given by

$$H_i(a) = \begin{cases} a^{p_i} & \text{if } p_i \neq p_{i-1} \\ \\ H_{i-1}(a) \times \log(a) & \text{if } p_i = p_{i-1} \end{cases} \tag{8}$$

with $p_0 = 0$ and $H_0 = 1$. Royston and Altman [21] argued that, in practice, fractional polynomials of order higher than 2 are rarely needed and suggested to choose the value of the powers from the set $\{-2, -1, -0.5, 0, 0.5, 1, 2, \max(3, m)\}$. We note that for models with log link function $\eta_1(a, \boldsymbol{\beta}, p = 1)$ is Muench's model, $\eta_2(a, \boldsymbol{\beta}, p_1 = 1, p_2 = 2)$ corresponds to the model proposed by Griffiths [8] and the models considered by Grenfell and Anderson [9] have the general form of $\eta_m(a, \boldsymbol{\beta}, p_1, p_2 \ldots, p_m)$ with $p_i = i$ for $i = 1, 2, \ldots, m$.

Table 2, About Here.

Table 2 shows a selection of parametric models discussed in the literature and their representation as fractional polynomials. For example, the model proposed by Keiding [15] is a first order fractional ploynomial with $\mathbf{p} = 0$. The model with linear force of infection can be parameterized as first order fractional polynomial with complementary log-log link for which $\mathbf{p} = 0$ with the constrained that $\beta_1 = 2$. In this case $\ell(a) = 2\beta_0 a$ which implies that the force of infection is zero at birth and increases linearly thereafter.

3.3. Model Selection

Within the fractional polynomials framework the deviance of the model with $\eta_1(a, \boldsymbol{\beta}, 1)$ is taken to be the baseline deviance and improvement by other models is measured by

$$G(m, \mathbf{p}) = D(1, 1) - D(m, \mathbf{p}), \tag{9}$$

where $D(m, \mathbf{p})$ is the deviance of the model with fractional polynomial of order m and a sequence of powers, $\mathbf{p} = (p_1, p_2, \dots p_m)$. Note that a large value of G indicates a better fit. Fitting models within the framework of fractional polynomials requires to start the modeling procedure from first order fractional polynomials. To decide whether a model of first degree is adequate or a second degree model is needed, Royston and Altman [21] recommend to use the criterion $D(1, \tilde{\mathbf{p}}) - D(2, \tilde{\mathbf{p}}) > \chi^2_{2,0.9}$ where $\tilde{\mathbf{p}}$ is the power sequence for the model that has the best goodness-to-fit (hence, the model with the highest likelihood or, equivalently, the smallest deviance).

3.4. Constrained Fractional Polynomials

Although fractional polynomials provide a wide range of curve shapes, there is no guarantee that $\pi(a)$ will be a monotone function of age and therefore fractional polynomials can still result in a negative estimate for the force of infection. It is clear from Table 1 that the estimate for the force of infection is negative whenever $\eta'_m(a, \hat{\boldsymbol{\beta}}, \mathbf{p}) < 0$ (since $\delta(\eta_m(a, \hat{\boldsymbol{\beta}}, \mathbf{p}))$ is strictly positive). Therefore, one should fit model (7) subject to the constraint that $\eta'_m(a, \hat{\boldsymbol{\beta}}, \mathbf{p}) \geq 0$, for all ages a in the predefined range. In the framework of fractional polynomials this cannot be done analytically. But in practice, one can fit a large number of fractional polynomials, over a grid of powers, and check for each fitted model if $\eta'_m(a, \hat{\boldsymbol{\beta}}, \mathbf{p}) \geq 0$, for all ages a. In case that a given sequence of powers leads to a negative derivative of the linear predictor, the model is not considered an appropriate model. This means that we choose the model with the best goodness-to-fit among all fractional polynomials for which $\eta'_m(a, \hat{\boldsymbol{\beta}}, \mathbf{p}) \geq 0$.

4. Application to the Data

In this section, we apply our method to the datasets mentioned above. For each dataset, first and second order fractional polynomials were fitted and the criterion proposed by Royston and Altman [21] was used to deicide whether the second order model is needed or not. Table 3 presents the deviance and gain values for the best first order fractional polynomials. Clearly, for all datasets except the Bulgarian dataset, first order fractional polynomials are not adequate and second order fractional polynomials are required. For the first order models, the gain values in Table 3 also indicate that, for all datasets except the Bulgarian dataset, the first order fractional polynomials with $\mathbf{p} = 1$ are not adequate and other powers are needed.

TABLE 3, About Here.

4.1. Hepatitis A

The upper two panels in Figure 3 show the estimated models for the prevalence and the force of infection for Hepatitis A in Belgium. The model with the best goodness-of-fit has a gain value of 51.94 and $\mathbf{p} = (1, 1.3)$. For this model the deviance is 97.61 on 81 degrees of freedom. The estimated force of infection reaches a peak at age 40 ($\ell(40) = 0.04159$) and drops down thereafter. Figure 4 shows the unrestricted profile likelihood surface, $L(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, p_1, p_2)$, for this dataset. The point a represents the likelihood's value of the best unrestricted fractional polynomial for which $\mathbf{p} = (1.9, 1.9)$ and the deviance is 79.60 on 81 degrees of freedom. However, this model cannot be retained since it predicts a negative force of infection at older age groups. Point b represents the likelihood's value of the conventional polynomial ($\mathbf{p} = (1, 3)$, deviance equal to 82.74 on 83 degrees of freedom) which was discussed in Section 3 and will not be considered either. The point c represents the likelihood's value of the best constrained fractional polynomial. Hence, the fractional polynomial presented in Figure 3 can be seen as the model that has the best goodness-of-fit among all fractional polynomials satisfying $\eta'(a, \hat{\boldsymbol{\beta}}, \mathbf{p}) \geq 0$.

For the Bulgarian dataset, the second order fractional polynomial with $\mathbf{p} = (1.9, 1.9)$ has a deviance of 77.77 on 78 degrees of freedom. This model suggests that the force of infection is maximal at age 41.5 ($\ell(41.5) = 0.0815$). However, the first order fractional polynomial is to be preferred since $D(1, \mathbf{p}) - D(2, \mathbf{p}) = 1.74$. Interestingly, the first order fractional polynomial with $\mathbf{p} = 1$ and logit link is just a simple linear logistic regression model. For this model $\ell(a) = \beta_1 \pi(a)$ such that it predicts an upward trend for the force of infection.

Figures 3 and 4, About Here.

4.2. Varicella

The upper two panels in Figure 5 show the estimated model for both prevalence and force of infection for the Varicella dataset. The deviance of the model is 43.90 on 39 degrees of freedom and $\mathbf{p} = (-0.6, -0.7)$. For varicella, the force of infection reaches a maximum at age 3 with value $\ell(3) = 0.324$ and drops down thereafter. At age 44 the force of infection is estimated to be 0.0214.

4.3. Rubella and Mumps

For Rubella, the fractional polynomial model with $\mathbf{p} = (-0.9, -0.4,)$ has the best goodness-of-fit with a deviance of 42.34 on 39 degrees of freedom. For Mumps, the model with the best goodness-to-fit uses power $\mathbf{p}$ =(-1.2,-0.9,). For this model, the deviance is 47.94 on 39 degrees of freedom. Figure 5 (middle panels) show that for Rubella the force of infection rises to a peak at age 6.5 ($\ell(6.5) = 0.1415$). For Mumps, the force of infection reaches a maximum value at age 4.5, $\ell(4.5) = 0.317$.

FIGURE 5, ABOUT HERE.

5. Influence of the Link Function

In the previous section, all models were fitted with the logit link function. In this section, we consider models of the general form $\pi(a) = 1 - \exp(-\gamma(a))$. More precisely, for the first order fractional polynomials we specify

$$\pi(a) = \begin{cases} 1 - \exp\left(-\beta_0 e^{\beta_1 H(a)}\right) & p \neq 0, \\ \\ 1 - \exp\left(-\beta_0 a^{\beta_1}\right) & p = 0. \end{cases} \tag{10}$$

For the second order fractional polynomials, we consider the following specification

$$\pi(a) = 1 - \exp\left(-\beta_0 e^{\beta_1 H_1(a) + \beta_2 H_2(a)}\right), \tag{11}$$

with corresponding linear predictor

$$\begin{cases} \eta_2(a, \boldsymbol{\beta}, p_1, p_2) = \log(\beta_0) + \beta_1 a^{p_1} + \beta_2 a^{p_2} & \text{if } \; p_1 \neq p_2, \\ \\ \eta_2(a, \boldsymbol{\beta}, p_1, p_2) = \log(\beta_0) + \beta_1 a^{p_1} + \beta_2 a^{p_1} \log(a) & \text{if } \; p_1 = p_2. \end{cases} \tag{12}$$

We note that the models specified in (10) and (11) are GLM with a complementary log-log link function. The first order model specified in (10) with $p = 0$ implies a Weibull distribution for the time spent in the susceptible class. Such a Weibull model was used by Keiding [15] to model the force of infection for Rubella from an Austrian seroprevalence sample. A model with a constant force of infection is a special case of a first order fractional polynomial with complementary log-log link function with β_1 fixed at value 1; in that case $\eta(a, \boldsymbol{\beta}) = \log(\beta_0) + \log(a)$. Such a model was used recently by Farrington [10] to model the force of infection for Hepatitis A in Bulgaria. Furthermore, a model with linear force of infection is a first order fractional polynomial with $\mathbf{p} = 0$ and $\beta = 2$.

Figure 6 shows the estimated forces of infection for all datasets when the optimal fractional polynomials were fitted with logit (solid lines) and complementary log-log link functions (dashed lines). We note that although the power sequence had changed, the change of the link function has only little influence on the estimated forces of infection. For example, the deviance of the model for Varicella is 44.04 on 39 degrees of freedom and $\mathbf{p} = (-1.3, -0.9)$ but the estimated force of infection is the same for the logit and complementary log-log models.

FIGURE 6, ABOUT HERE.

Figure 7 shows the estimated prevalence and force of infection for Hepatitis A in Bulgaria. For the first order models, the best fractional polynomial has a deviance of 82.75 on 80 degrees of freedom and $\mathbf{p} = 0.5$. The force of infection for this model steeply increases with age. Similar to the models with logit link, a second order fractional polynomial is not needed. Since models with different link function are not nested, we use the Akaike's information criterion (AIC) for model selection (Akaike [22]. The smallest value of AIC, 382.83, is obtained for the first order logit model (see Table 4). We note that the upward trend of the force of infection estimated by the first order logit model was already observed by Groeneboom [23] in his discussion of Keiding's paper.

Table 4 and Figure 7, About Here.

6. Discussion

We have shown that modeling the prevalence and the force of infection with fractional polynomials is a very flexible method, allowing a variety of different types of relationships between the force of infection and age. The method can compete with nonparametric smoothers while keeping the attractive features of parametric models. Furthermore, we have shown that well known parametric models for the distribution of the age at infection, such as exponential, Weibull and log-logistic distributions, can be expressed as a special case of fractional polynomials. For models with complementary log-log link function, the curve shape of the force of infection depends on the slope of the first order fractional polynomial with $\mathbf{p} = 0$. Therefore, we need to fit the model $\eta_1(a, \boldsymbol{\beta}, \mathbf{p} = 0)$ and to check the parameter estimate for β_1. The force of infection is constant if $\beta_1 = 1$, linear if $\beta_1 = 2$ and monotone if $\beta_1 \neq 1$. Thus, by fitting a large number of fractional polynomials with logit and complementary log-log link function we account for the possibility of constant, linear, monotone or flexible curve shapes for force of infection. However, we do not require the force of infection to have a specific curve shape in advance, the choice is data-driven.

In case that other covariates, in addition to age, are included, the following semiparametric additive model parameterizes the prevalence as

$$\text{link}(\pi(a)) = \phi(a) + Z\alpha, \tag{13}$$

where Z is the additional categorical covariate(s). The nonparametric component of the model, $\phi(a)$, is used to model the dependency of $\pi(a)$ on age while $Z\alpha$, the parametric component of the model, is used to model the covariate effects. In order to ensure a nonnegative estimate for the force of infection, one needs to estimate $\pi(a)$ with a nondecreasing function. This can be done by applying the pool adjacent violators algorithm (Barlow *et al.*[24] and Robertson [25]) to the data. This approach has been followed by Grummer-Strawn [17] and Shiboski [18]. Within the framework of fractional polynomial we can replace the nonparametric component of the model with a fractional polynomial

$$\text{link}(\pi(a)) = \eta_m(a, \mathbf{p}, \beta) + Z\alpha,$$

where $\eta_m(a, \mathbf{p}, \beta)$ is the fractional polynomial modeling the dependence on age. Similar to the semiparametric model in (13), depending on the link function, this model implies proportionality. For example, suppose that Z is binary variable, then for models with complementary log-log link we get $\ell(a|Z = 1) = \exp(\alpha)\ell(a|Z = 0)$ and for models with logit link we obtain $\ell(a|Z = 1)/\ell(a|Z = 0) = \alpha q(a|Z = 0)/q(a|Z = 1)$.

All models discussed above are generalized linear models which imply that standard software, such as PROC GENMOD in SAS or the function `glm()` in Splus, can be used. Although our method requires to fit a large number of fractional polynomials and to choose the one with the best goodness-of-fit, the modeling procedure is not time consuming. In fact, the optimal fractional polynomial for each dataset was found in less than 3 minutes.

The models reported in this paper were fitted with a sequence of powers from -2 to 3 with an increment of 0.1. Of course, when a more sensitive grid is used the final powers of the best model will be slightly different. For example, for Hepatitis A (Belgium) the best second order fractional polynomial, fitted using a grid with increment of 0.02, has powers 1.132653 and 1.153061 with deviance 97.44. However, the force of infection is the same as for the model with $\mathbf{p} = (1, 1.3)$ (maximal of absolute difference between the forces of infection is 5.68×10^{-5}). The problem of estimating a negative force of infection was addressed by fitting constrained fractional polynomials by excluding models that lead to negative force of infection as appropriate models. In our opinion, blind use of conventional linear predictors to model the force of infection can yield misleading results. Flexible models should be consider and the family of fractional polynomials offers an interesting choice. They can also be used as an exploratory tool or to perform a sensitivity analysis of a particular parametric model that, for instance, reflects prior information about the force of infection.

REFERENCES

1. Anderson, R.M. and May, R.M. (1991), *Infectious diseases of humans, dynamic and control.* Oxford University Press Inc. New York.
2. Farrington, C.P. (1990), Modeling Forces of infection for measles, mumps and rubella, *Statistics in Medicine* **9**, 953–967.
3. Keiding, N. (1991), Age-specific incidence and prevalence: a statistical perspective. *J. R. Statist. Soc.* A, **154**, 371–412.
4. Shkedy, Z., Aerts, M., Molenberghs, G., Beutels, Ph. and Van Damme, P. (2002), Modeling Hepatitis A force of infection using monotone local polynomials. Submitted.
5. Beutels, M., Van Damme, P. and Aelvoet, W. (1997), Prevalence of Hepatitis A, B and C in the Flemish Population, *European Journal of Epidemiology*, **13**, 275–280 .
6. Thiry, N., Beutels, Ph., Van Damme, P. and Vranckx, R (2002) The seroepidemiology of primary varicella-zoster virus (vzv) infection in Flanders (Belgium), submitted.
7. Muench, H. (1959), *Catalytic models in epidemiology.* Harvard University Press, Boston.
8. Griffiths, D. (1974), A catalytic model of infection for measles, *Applied Statistics*, **23**, 330–339.
9. Grenfell, B.T and Anderson, R.M (1985), The estimation of age-related rates of infection from case notifications and serological data, *Journal of Hygiene* **95**,(2), 419–436.
10. Farrington, C.P., Kanaan, M.N., Gay, N.J. (2001)., Estimation of the basic reproduction number for infectious diseases from age-stratified serological survey data (with discussion). *Appl. Statist.*, **50**, 251–292.
11. Edmunds, W.J., Gay, N.J., Kretzschmar , M., Pebody, R.G and Wachmann, H. (2000), The pre vaccination epidemiology of measles, mumps and rubella in Europe: implications for modeling studies, *Epidemiol. infect.*, **125**, 635-650.
12. McCullagh, P. and Nelder, J.A (1989), *Generalized Linear Models.* Chapman and Hall. New York.
13. Becker, N.G. (1989), *Analysis of infectious disease data.* Chapman and Hall.
14. Diamond, I.D. amd McDonald, J.M. (1992), Analysis of current-status data, In *Demographic Application of Event History Analysis* (eds. Trussel J., Hankinson R. and Tiltan J.) Ch. 12, Oxford, Oxford University Press.
15. Keiding, N., Begtrup, K., Scheike, T.H., and Hasibeder, G. (1996), Estimation from current status data in continuous time, *Lifetime data analysis*, **2**, 119–129.

16. Jewell, N.P, and Van Der Leen, M, (1995), Generalizations of current status data with applications, *Lifetime data analysis*, **1**, 101–109.

17. Grummer-Strawn, L.M. , (1993), Regression analysis of current status data: an application to breast feeding, *Journal of the American statistical association* **88**, 758-765.

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

19. Hastie, T.J. and Tibshirani, R.J. (1990), *Generalized additive models.* Chapman and Hall.

20. Rossini, A.J. and Tsiatis, A.A. (1996), A semiparametric proportional odds regression model for the analysis of current status data, *Journal of the American statistical association* **91**, 423, 713-721

21. Royston, P. and Altman, D.G. (1994), Regression using fractional polynomials of continuous covariates : parsimonious parametric modeling, *Applied Statistics* **43**, 429–467.

22. Akaike, H. (1974), A new look at the statistical identification model, *IEEE transactions on automatic control*, **19**, 716–723.

23. Groeneboom (1991), Discussion on : Age-specific incidence and prevalence: a statistical perspective, by Keiding N. (1991), *J. R. Statist. Soc.* A, **154**, 396–398.

24. Barlow, R.E, Bartholomew, D.J, Bremner, M.J and Brunk, H.D (1972), *Statistical inference under order restriction.* Wiley.

25. Robertson, T., Wright, F.T. and Dykstra, R.L. (1988), *Order restricted statistical inference*, Wiley.

26. Diamond, I.D (1991), Discussion on : Age-specific incidence and prevalence: a statistical perspective, by Keiding N. (1991), *J. R. Statist. Soc.* A, **154**, 396–398.

Table I. *General forms for the force of infection.*

Link function	$\pi(a)$	$\ell(a)$	$\delta(\eta(a))$
log	$1-e^{-\eta(a)}$	$\eta'(a)$	1
Complementary log-log	$1-e^{-e^{\eta(a)}}$	$\eta'(a)e^{\eta(a)}$	$e^{\eta(a)}$
logit	$\frac{e^{\eta(a)}}{1+e^{\eta(a)}}$	$\eta'(a)\frac{e^{\eta(a)}}{1+e^{\eta(a)}}$	$\frac{e^{\eta(a)}}{1+e^{\eta(a)}}$

Table II. *Parametric models presented in the literature. The models presented by Grummer-Strawn (1993) and Jewell and Van Der Laan (1995) included other covariate in addition to age. For these models $\eta(m,p,\beta)$ is the fractional polynomial that use to model the dependency of the force of infection on age. For the models discussed in Grummer-Strawn, we do not include the adjusted parameter in our analysis since it is assumed that susceptibility is* 100% *at birth.*

Publication	Force of infection	Fractional polynomial	Link function
Munch (1959),Farrington (2001), Jewell and Van der laan (1995)	constant	$\eta(m=1,p=0,\beta=0)$	cloglog
Griffiths(1974)	linear	$\eta(m=1,p=0,\beta=2)$	cloglog
Grenfell and Anderson (1985)	polynomial	$\eta(m=k,p_i=i)$	log
Keiding (1996),Becker (1989), Diamond and McDonald (1992), Grummer-Strawn (1993)	monotote	$\eta(m=1,p=0,\beta)$	cloglog
Grummer-Strawn (1993)	flexible	$\eta(m=1,p=0,\beta)$	logit

Table III. *Deviance and Gain values for first and second order fractional polynomials with logit link function.*

	First order (**m=1**)				Second order (**m=2**)		
Dataset	df	Deviance	p	$G(1,p)$	df	Deviance	p_1, p_2
Hepatitis A (Be)	83	115.34	0.32	34.21	81	97.61	1.0,1.3
Hepatitis A (Bul)	80	79.51	1	0	78	77.77	1.9,1.9
Varicella	41	50.94	0.07	69.59	39	43.90	-0.7,-0.6
Rubella	41	56.28	0.03	165.13	39	42.34	-0.9,-0.4
Mumps	41	82.31	-0.2	516.88	39	47.94	-1.2,-0.9

Table IV. *Deviance summaries of the fitted models for the Bulgarian Hepatitis A dataset. The Weibull model has 81 degrees of freedom since in this case* $p = 0$*, the exponential model with constant force of infection has 82 degrees of freedom since we fixed both* p *and* β*.*

Model (link)	df	Deviance	**p**	Likelihood	AIC	
Second order(logit)	78	77.77	1.9,1.9	375.967	385.96	
First order (logit)	80	79.51	1	376.83	382.83	
Second order(cloglog)	78	79.21	1.3,1.3	376.68	386.68	
First order (cloglog)	80	82.75	0.5	378.45	384.44	
First order (cloglog)	81	94.40	0 ($\beta_1 \neq 1$)	384.27	388.27	Weibull
First order (cloglog)	82	94.67	0 ($\beta_1 = 1$)	384.41	386.41	Constant force of infection

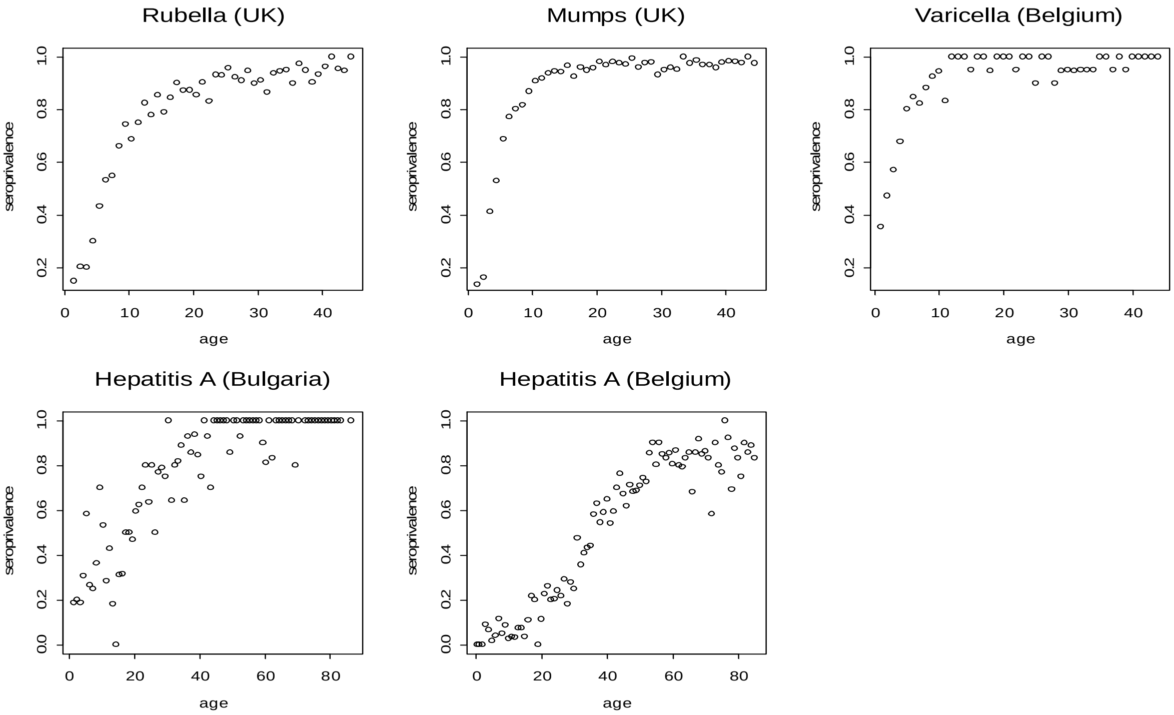

Figure 1. *Five cross sectional seroprevalence datasets.*

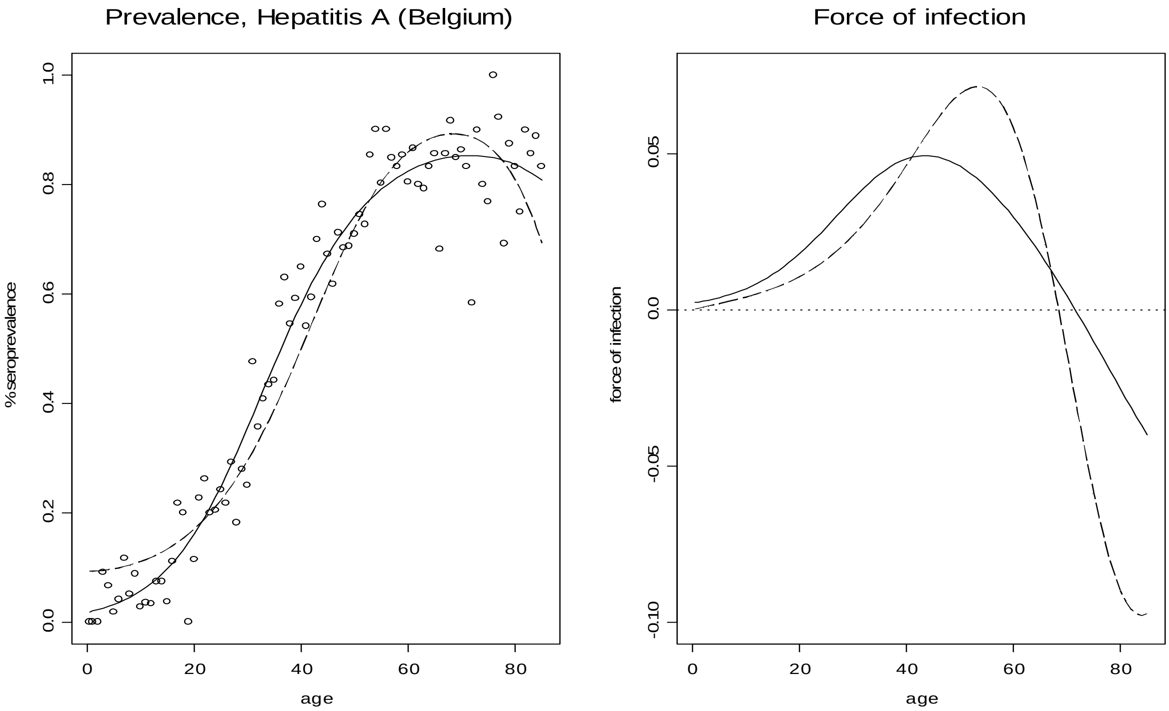

Figure 2. *Hepatitis A in Belgium. Left panel: data and estimated models for the prevalence. Right panel: estimated forces of infection. Solid line: model with logit link function. Dashed line: model with complementary log-log link function.*

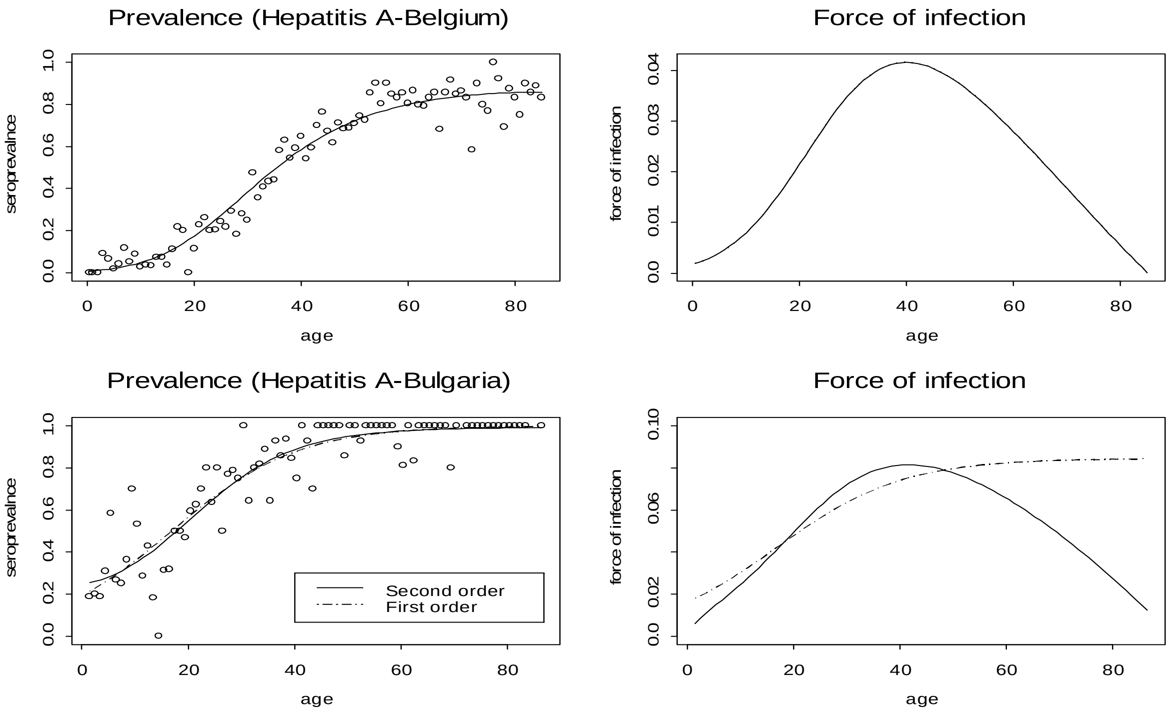

Figure 3. *Hepatitis A in Belgium (upper panels) and in Bulgaria (lower panels).*

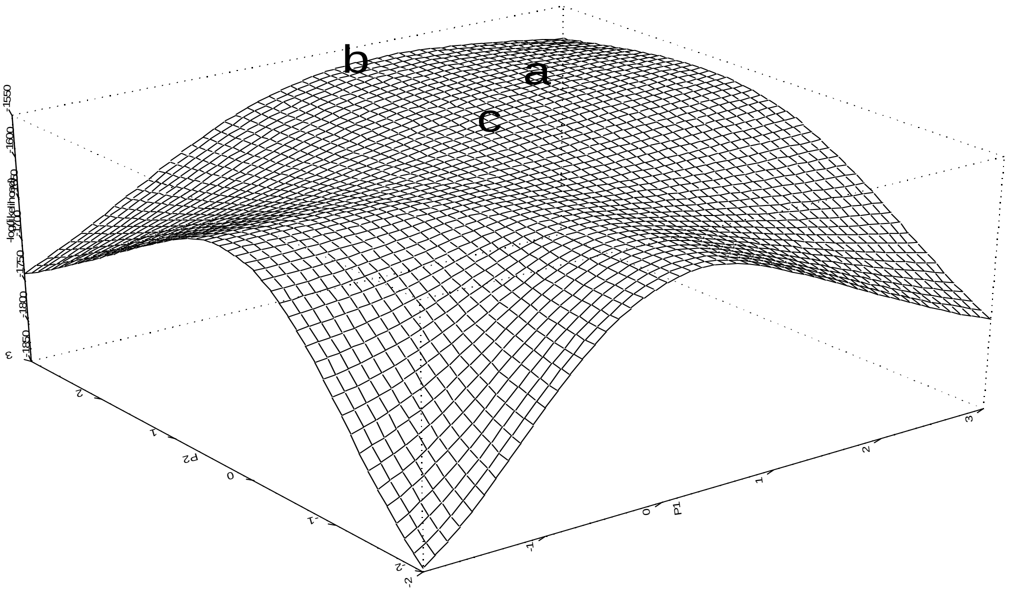

Figure 4. *Non restricted likelihood surface for Hepatitis A in Belgium. The points on the surface: (a) the best second order fractional polynomial* $\mathbf{p} = (1.9, 1.9)$*, (b) the conventional polynomial* $\mathbf{p} = (1, 3)$ *and (c) the best restricted fractional polynomial* $\mathbf{p} = (1, 1.3)$.

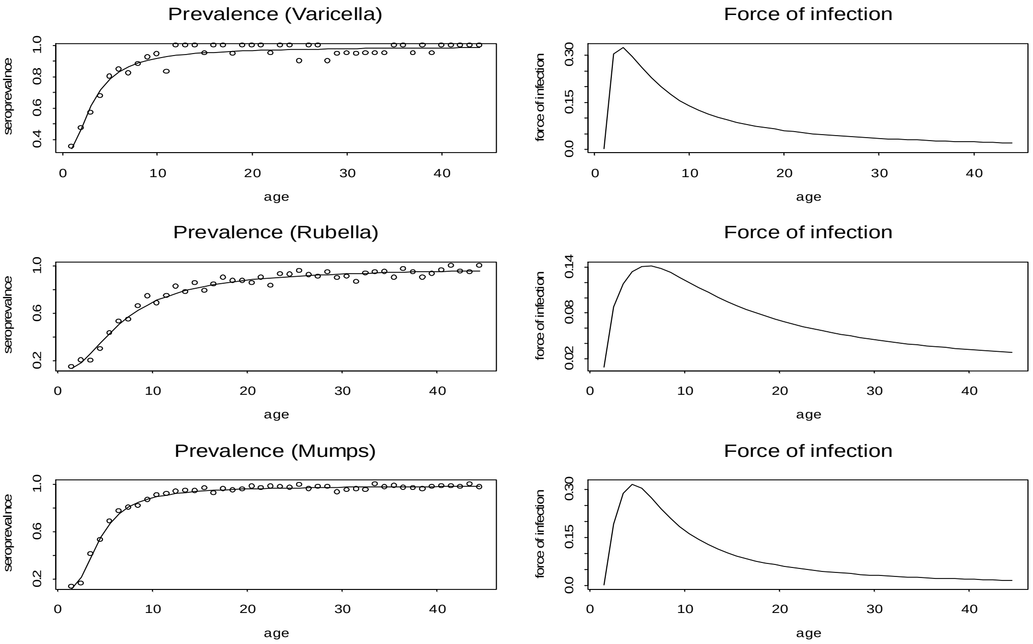

Figure 5. *Varicella,Rubella and Mumps.*

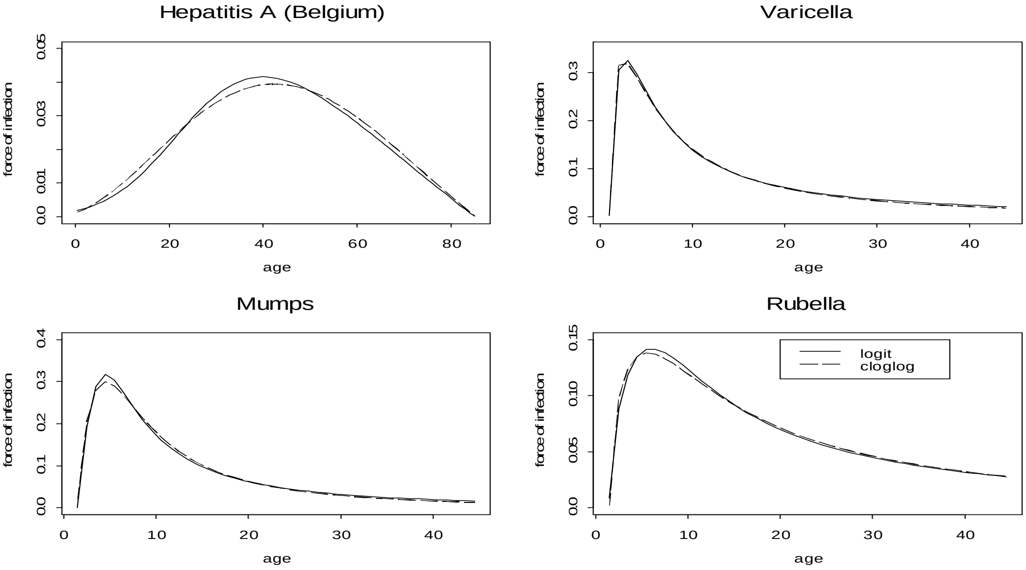

Figure 6. *Force of infection for second order fractional polynomial with logit (solid line) and complementary log-log link functions (dashed line). The power sequence are p=(0.5,0.9), p=(-1.3,-0.9), p=(-1.6,-1.5) and p=(-1.2,-0.9) for Hepatitis, Varicella, Mumps and Rubella respectively.*

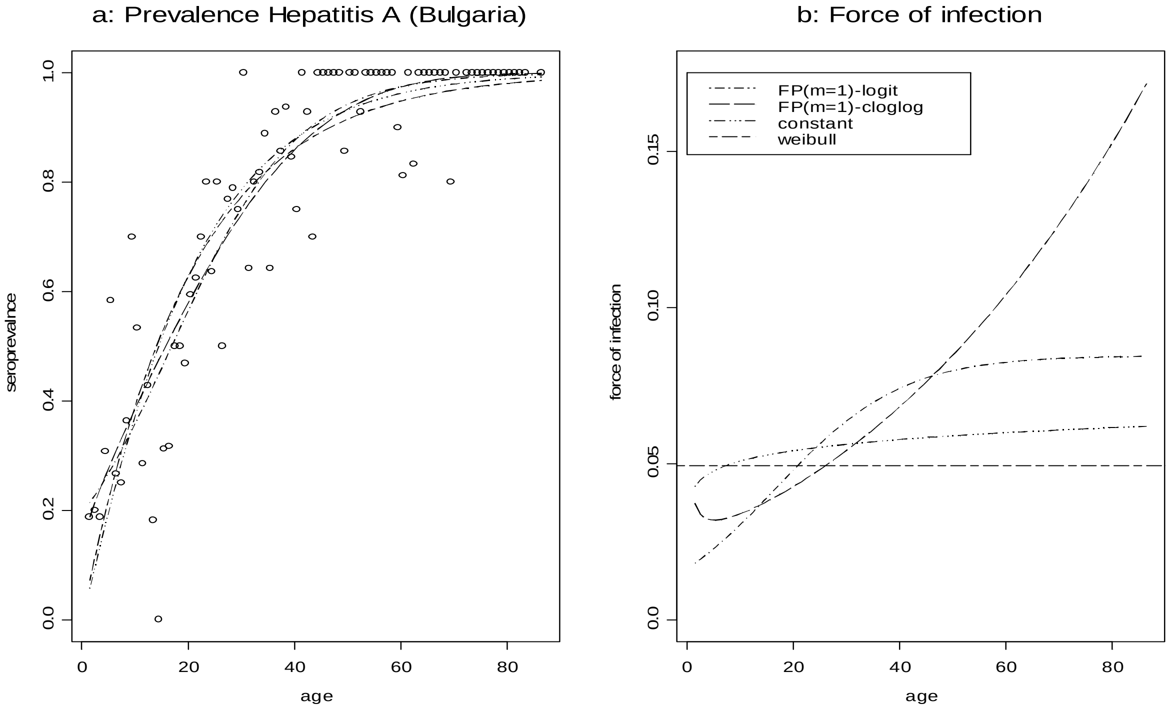

Figure 7. *Hepatitis A in Bulgaria, models with complementary log-log link function. First order fractional polynomials with logit and complementary log-log link function (FP(m=1)-logit and FP(m=1)-cloglog respectively). The models with constant and monotone force of infection were both fitted with complementary log-log link function and* $p = 0$.