Modelling distortions in seroprevalence data using change-point fractional polynomials

DOI: 10.1177/1471082X0801000203
Handle: http://hdl.handle.net/1942/10984

Modelling Distortions in Seroprevalence Data Using Change-point Fractional Polynomials

Hens, N.[1], Kvitkovicova, A.[2], Aerts, M.[1], Hlubinka, D.[2], Beutels, P.[3]

[1] Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Hasselt University and Catholic University of Leuven, Belgium.

[2] Charles University, Prague, Czech Republic.

[3] Center for the Evaluation of Vaccination, Antwerp University, Antwerp, Belgium.

Abstract

This paper shows how to model seroprevalence data using change-point fractional polynomials. The inclusion of a change-point in the fractional polynomial framework allows to detect distortions arising from common (often untestable) assumptions made in the estimation of the age-specific prevalence and force of infection from cross sectional data. The method is motivated using seroprevalence data on the parvovirus B19 and the varicella zoster virus in Belgium.

Key words: Fractional Polynomial; Change-point; Seroprevalence Data; Detecting Distortions; Model Selection Criteria

1 Introduction

Modelling infectious diseases is mostly done using compartmental model that describe the flow of individuals through different disease stages. One of the most important parameters in such a compartmental model describes the per capita rate at which a susceptible person acquires the infection and thus

[1]Corresponding author: Niel Hens, Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Hasselt University, Campus Diepenbeek, Agoralaan - Gebouw D, B–3590 Diepenbeek, BELGIUM, email: niel.hens@uhasselt.be, phone: +32-11-268232 fax: +32-11-268299

moves from the compartment of susceptible to the compartment of infected individuals (Anderson and May, 1991; Diekmann et al., 1990; Capasso, 1993; Thieme, 2003, see e.g.). This per capita rate is the so-called force of infection, i.e. the infection hazard and has been shown to be often age-dependent. Under the steady state assumption and assuming life long immunity once infected, one can estimate the force of infection from seroprevalence data (see e.g. Keiding, 1991).

There exists a vast literature on estimating the force of infection from seroprevalence data, of which most of the methods can be embedded in the generalized linear model framework. Essential in the estimation of the force of infection is expressing the dependency on age using the appropriate functional form. On the one hand, flexible modelling techniques are often a necessity to achieve the appropriate predictive accuracy. On the other hand, they could lead to an overinterpretation of random fluctuations in the data. It is therefore that, while easy-to-apply smoothing methods exist, flexible parametric models are often used in the generalized linear model framework and the like. Among those flexible parametric models, fractional polynomials have gained popularity as evidenced by the high number of citations for the original Royston and Altman (1994) paper.

Shkedy et al. (2006) were the first to use fractional polynomials to model the seroprevalence in infectious disease epidemiology. They showed how, under the steady state assumption and assuming lifelong immunity, the force of infection can be estimated from seroprevalence data using monotone second degree fractional polynomials. The problem of estimating the force of infection from cross-sectional prevalence data has drawn considerable attention (see also Farrington, 1990; Keiding, 1991; Diamond and McDonald, 1992; Capasso, 1993; Farrington et al., 2001). The plausibility of the steady state assumption is however untestable in case of one cross-sectionally collected sample. We refer to Nagelkerke et al. (1999) for a discussion on stationarity (time homogeneity), that is, the assumption that the age-specific force of infection remains constant over time. When observed prevalences increase

monotonically with age, the problem of estimating the force of infection is straightforward. However, unless samples at each age are very large and the steady state assumption is fulfilled, a monotone increase with age of observed prevalences will only rarely occur. As the survival function, one minus the prevalence, is a monotonically decreasing function, one has to estimate the prevalence function under order restrictions. Several solutions, including an isotonic (regression) estimator, the pool adjacent violator algorithm or using a set of monotone candidate models have been proposed to this purpose (see e.g. Shkedy et al., 2006; Hens et al., 2008).

This paper presents an elaboration on the use of fractional polynomials, by including a change-point to detect distortions in age-dependent seroprevalence data. Change-points have been used before in the analysis of epidemiological studies (see e.g. Ulm, 1991; Stasinopoulos and Rigby, 1992; Pastor-Barriuso and Guallar, 1998; Ulm and Küchenhoff, 2000), where the estimation is often used to model threshold effects in biological systems such as for dose-response curves. In a more recent paper, Pastor-Barriuso et al. (2003) elaborate on the use of change-point estimation in logistic regression using transition models. Although their method provides a rigorous modelling approach, we will focus on a somewhat different setting which naturally falls into the fractional polynomial framework. We start with an introduction to the data in Section 2 and present the methodology in Section 3. The proposed method is then applied to the data in Section 4 and inferential measures enabling a more in depth study on the nature of the distortions are introduced in Section 5. We end with a discussion in Section 6.

2 Data

In this section, data on the parvovirus B19 (PVB19) and the varicella zoster virus (VZV) from a serological survey in Belgium are described. These serological data comprise current status data on whether pathogen-specific antibody levels exceed a pre-specified cut-off value in the tested blood samples.

Since after natural infection, antibody presence is assumed to be lifelong, these current status data indicate whether past infection occurred or not. Such data are often analyzed while discarding possible distortions in the observed age-specific profile. The data described here were collected in a period from November 2001 until March 2003 in Belgium. 3374 samples were partly tested for PVB19 (3076 samples) and VZV (2655 samples).

2.1 Varicella Zoster Virus

The Varicella-Zoster Virus 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. In about $10 - 20\%$ of cases, VZV reactivates later in life producing a disease known as herpes zoster or shingles.

Out of the original 3374 samples, 2655 samples from persons with age ranging from 19 days to 40 years have been tested for VZV. The observed age-specific seroprevalence is shown in the upper panel of Figure 1. Previous similar studies on VZV were reported by Thiry et al. (2002) and Nardone et al. (2007).

2.2 Parvovirus B19

PVB19 was the first human parvovirus to be discovered in 1975. PVB19 is best known for causing a childhood exanthem called fifth disease. The virus is primarily spread by infected respiratory droplets. PVB19 symptoms begin some six days after exposure and last about a week. Individuals with PVB19 IgG antibodies are generally considered immune to recurrent infection. About half of adults are PVB19-immune due to a past infection. While the disease is generally mild, most studies have focused on risk factors in pregnant women because of the risk to the fetus (Valeur-Jensen et al.,

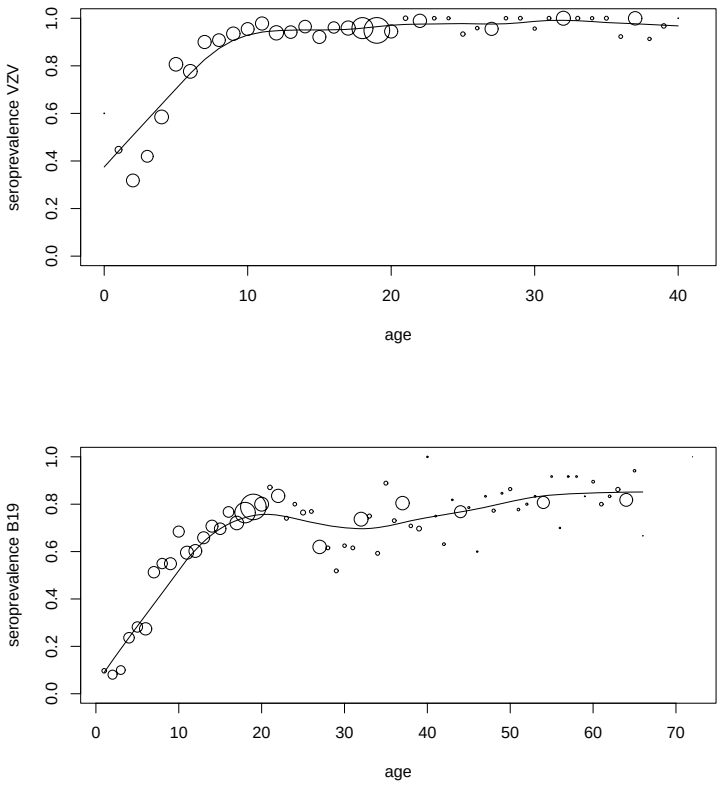

Figure 1: Seroprevalence plot for VZV (upper panel) and PVB19 (lower panel) with lowess curves using a smoother span based on visual inspection. Dots are proportional to the number of samples tested.

1999).

3076 samples from persons with age ranging from 19 days to 71 years, 205 days have been tested for PVB19. A plot of the change in seroprevalence with age is shown in the lower panel of Figure 1. This sample, together with other samples from England & Wales, Finland, Italy and Poland were analyzed before by Mossong et al. (2008) using monotone local polynomials (Shkedy et al., 2003).

Since 2 381 samples were tested for both VZV and PVB19, Hens et al. (2008) studied the association among the two infections by introducing a conditional and joint force of infection. They monotonized their nonparametrically estimated seroprevalence curve using the pool adjacent violator algorithm. However, some distortions with respect to monotonicity are observable in the data. While the data on PVB19 show a distortion around 25 to 35 years of age, the data on VZV show a distortion possibly originating from maternal antibodies for infants aged 19 days to 2 years (Figure 1). Without specific knowledge on the decay of infection-related antibody levels and maternally inherited antibody level decay it is of importance to detect these distortions in the data. These distortions were not explicitly taken into account in the aforementioned studies.

3 Methodology

In this section, change-point FPs are introduced in the generalized linear model framework (GLM, McCullagh and Nelder, 1989). We will first define FPs as proposed by Royston and Altman (1994) in the GLM framework and then extend the model to incorporate one or more change-points.

3.1 Fractional Polynomial Models

A GLM, relating the binary infection status $Y = 0$ (no past infection), $Y = 1$ (past infection) to a predictor variable of interest a (typically age), can be

expressed as

$$g(\pi(a)) = \eta(a), \tag{3.1}$$

where $\pi(a) = P(Y = 1|a)$ denotes the seroprevalence; g is the link-function ('logit', 'cloglog',...) and $\eta(a)$ is the linear predictor expressing the dependency of the seroprevalence on a and which is linear in the coefficients for GLMs. Several choices for η can be made including e.g. polynomial functions of the form $\{1, a, \ldots, a^p\}$ for a certain integer p.

Royston and Altman (1994) proposed to adapt $\eta(a)$ to include FPs, i.e. a set of nonlinear functions using a heuristic procedure as outlined hereafter. Using this procedure their approach falls into the conventional GLM framework with all related advantages. The form of the linear predictor consisting of a FP of degree m is

$$\eta_m(a, \boldsymbol{\beta}, p_1, \ldots, p_m) = \sum_{i=0}^{m} \beta_i H_i(a), \tag{3.2}$$

with m being an integer, $p_1 \leq p_2 \leq \ldots \leq p_m$ the powers, and $H_i(a)$ defined as

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

where $p_0 = 0$ and $H_0 = 1$. Following the suggestions of Royston and Altman (1994), we restrict our attention to models of degree 1 and 2, and choose the powers from the set $\{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$.

3.2 Change-point Model

Discovering distortions in the age-dependent seroprevalence involves searching for an unexpected change in its profile. This change can be governed using a change-point. To achieve the necessary flexibility while retaining the advantage of a parametric model, FPs can be used to model the profile before and after the change-point. Consider a change-point model linking two FPs

$$\eta_{m_1,m_2}(a, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{p}, \boldsymbol{q}, c) = \eta_{m_1}(a, \boldsymbol{\beta}, \boldsymbol{p})\mathrm{I}\{a < c\} + \eta_{m_2}(a, \boldsymbol{\gamma}, \boldsymbol{q})\mathrm{I}\{a \geq c\}, \tag{3.3}$$

where I$\{\cdot\}$ is the indicator function taking value 1 if the specified condition is true and 0 elsewhere; m_1 and m_2 are the degrees; $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ the coefficient vectors; and $\boldsymbol{p}$ and $\boldsymbol{q}$ the vectors of powers for the first and second FP.

There is large variety of potentially interesting submodels of Model (3.3). To define a specific set of models of interest, we rewrite $\eta_{m_1}(a, \boldsymbol{\beta}, \boldsymbol{p})$ (and similarly $\eta_{m_2}(a, \boldsymbol{\gamma}, \boldsymbol{q})$) in (3.3) as

$$\eta_{m_1}(a, \boldsymbol{\beta}, \boldsymbol{p}) = \beta_0 + \sum_{i=1}^{m} \beta_i H_i(a). \tag{3.4}$$

The set of models (C1,C2,C3) of interest can then be defined as

(C1): {(3.3),(3.4)} with change in the non-age-dependent term only:

$$\eta_{m_1}(a, \boldsymbol{\beta}, \boldsymbol{p}) - \eta_{m_2}(a, \boldsymbol{\gamma}, \boldsymbol{q}) = \beta_0 - \gamma_0$$

(C2): {(3.3),(3.4)} with change in the age-dependent term only:

$$\beta_0 = \gamma_0$$

(C3): {(3.3),(3.4)} with continuity constraint in c:

$$\eta_{m_1}(c, \boldsymbol{\beta}, \boldsymbol{p}) = \eta_{m_2}(c, \boldsymbol{\gamma}, \boldsymbol{p})$$

Note that the latter constraint is readily achieved by expressing γ_0 as the analytical solution of $\eta_{m_1}(c, \boldsymbol{\beta}, \boldsymbol{p}) - \eta_{m_2}(c, \boldsymbol{\gamma}, \boldsymbol{p}) = 0$. In addition, Model (3.3) encompasses the conventional FPs by choosing $c > \max(a)$ and e.g. monotonicity constraints can be imposed on both parts either by constrained optimization or using a restrictive set of powers and coefficients. While (C1) and (C2) allow for a discontinuity ((C1) allows for a sudden jump in the seroprofile), (C3) enforces continuity to hold. Other submodels of Model (3.3) could be of interest too but are not considered in the application as presented in this paper.

Selecting the appropriate change-point is done using the same procedure as selecting the appropriate powers for both FPs of the change-point FP by looking over a grid of age-values. While this extends the number of models

considerably, it also shows the high capacity of detecting any distortions in the data. Note that we have 2 061 and 1 973 unique age-values for the PVB19 and VZV sample, respectively. We therefore limited change-points to be selected over the integer grid: 1-41 for VZV and 1-72 for PVB19 (i.e. infection on age was reduced to 'year' instead of 'dd/mm/yyyy').

For a fixed degree m_1 and m_2 in Model (3.3), the optimal set of powers and the change-point are selected using the AIC-criterion (Akaike, 1973; Burnham and Anderson, 2002) given by $-2\ell(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \boldsymbol{p}, \boldsymbol{q}, c|a, y) + 2K$, where ℓ is the loglikelihood corresponding to Model (3.3) and K is the number of parameters in the model. With respect to the selection of powers, this is equivalent with the deviance criterion as proposed by Royston and Altman (1994). Selecting the appropriate degrees and number of change-points (none, one or more; see Section 4.2) is then done using the AIC-and BIC-criterion, respectively. Note that K, i.e. the number of parameters in the model, includes the number of powers and the number of change-points.

4 Results

Since there is no a priori knowledge on the number of change-points, we gradually build up by first considering a one-change-point fractional polynomial in Section 4.1. An extension with a second, different, change-point is then considered in Section 4.2.

4.1 One-change-point Fractional Polynomial

The set of one-change-point FP models used to detect distortions in the seroprevalence profiles of VZV and PVB19 consisted of the change-point

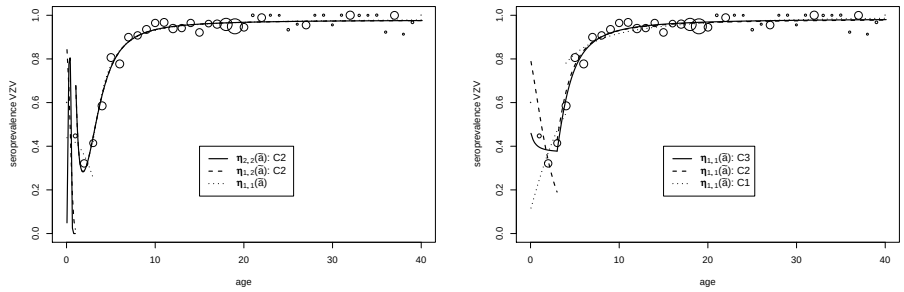

Figure 2: Seroprevalence plot for VZV with the three best change-point fractional polynomials (best: solid line; second best: dashed line; third: dotted line) according to AIC (left panel) and BIC (right panel).

FPs as defined in (4.1).

$$
\begin{aligned}
\eta_1(a) &= \beta_0 + \beta_1 \tilde{a}^{p_1}, \\
\eta_2(a) &= \beta_0 + \beta_1 \tilde{a}^{p_1} + \beta_2 \tilde{a}^{p_2}, \\
\eta_{1,1}(a) &= [\beta_0 + \beta_1 \tilde{a}^{p_1}]\mathrm{I}\{a < c\} + [\gamma_0 + \gamma_1 \tilde{a}^{q_1}]\mathrm{I}\{a \geq c\}, \\
\eta_{1,2}(a) &= [\beta_0 + \beta_1 \tilde{a}^{p_1}]\mathrm{I}\{a < c\} + [\gamma_0 + \gamma_1 \tilde{a}^{q_1} + \gamma_2 \tilde{a}^{q_2}]\mathrm{I}\{a \geq c\}, \\
\eta_{2,1}(a) &= [\beta_0 + \beta_1 \tilde{a}^{p_1} + \beta_2 \tilde{a}^{p_2}]\mathrm{I}\{a < c\} + [\gamma_0 + \gamma_1 \tilde{a}^{q_1}]\mathrm{I}\{a \geq c\}, \\
\eta_{2,2}(a) &= [\beta_0 + \beta_1 \tilde{a}^{p_1} + \beta_2 \tilde{a}^{p_2}]\mathrm{I}\{a < c\} + [\gamma_0 + \gamma_1 \tilde{a}^{q_1} + \gamma_2 \tilde{a}^{q_2}]\mathrm{I}\{a \geq c\}.
\end{aligned}
\tag{4.1}
$$

Here $p_1 \leq p_2$ and $q_1 \leq q_2$, while $\tilde{a} = a + 1$ rather than a is used so that $\log(\tilde{a}) \to 0$ when $a \to 0$. Note that this is similar to what is often done when modelling growth curves. The full range of models, extending the set of models given in (4.1) by the constrained models (C1,C2,C3) together with their AIC-and BIC-value for VZV are shown in Table 1. For each combination the optimal set of powers and the change-point is chosen using AIC.

Both AIC and BIC prefer one-change-point models over the 'original' FPs. While both 'original' FPs have ranks 15 and 16 using AIC, BIC results in ranks 5 and 6. Since it is known that BIC tends to select less complex models (models with fewer parameters) because of the penalization with $(\log(n))$ per extra parameter (n is the sample size), this is no surprise. AIC is known to

Model	Constraint	AIC	Rank	BIC	Rank	$\hat{c}$	Powers
$\eta_1(\tilde{a})$		1 389.2	(16)	1 406.9	(6)	–	0
$\eta_2(\tilde{a})$		1 376.9	(15)	1 406.3	(5)	–	-1, -1
$\eta_{1,1}(\tilde{a})$		1 363.6	(3)	1 404.8	(4)	3	3 ; -1
$\eta_{1,1}(\tilde{a})$	C1	1 373.9	(14)	1 403.3	(3)	4	0
$\eta_{1,1}(\tilde{a})$	C2	1 366.8	(8)	1 402.1	(2)	3	0.5 ; -1
$\eta_{1,1}(\tilde{a})$	C3	1 364.7	(4)	1 400.4	(1)	3	-2 ; -1
$\eta_{1,2}(\tilde{a})$		1 364.7	(5)	1 417.7	(12)	1	3; -2, -2
$\eta_{1,2}(\tilde{a})$	C2	1 363.5	(2)	1 410.5	(7)	1	2; -2, -2
$\eta_{1,2}(\tilde{a})$	C3	1 367.3	(9)	1 414.4	(10)	3	-0.5; -2, 0.5
$\eta_{2,1}(\tilde{a})$		1 367.4	(10)	1 420.3	(13)	3	-2, -2; -1
$\eta_{2,1}(\tilde{a})$	C2	1 365.6	(7)	1 412.7	(9)	3	-0.5, 0.5; -1
$\eta_{2,1}(\tilde{a})$	C3	1 367.6	(11)	1 414.6	(11)	3	3, 3 ; -1
$\eta_{2,2}(\tilde{a})$		1 364.7	(6)	1 429.4	(16)	1	-2, 3 ; -2, -2
$\eta_{2,2}(\tilde{a})$	C1	1 369.6	(12)	1 410.8	(8)	4	-1, -1
$\eta_{2,2}(\tilde{a})$	C2	1 362.8	(1)	1 421.6	(14)	1	0, 3 ;-2, -2
$\eta_{2,2}(\tilde{a})$	C3	1 370.3	(13)	1 429.1	(15)	10	-0.5, 0 ; -2, -2

Table 1: VZV-models with corresponding AIC-and BIC-values, corresponding ranks, selected change-point and power(s).

select more complex models tending to overfit the data. This is reflected in Table 1 where the one-change-point FP models with lower AIC-ranks are found for different degrees (m_1, m_2) while the one-change-point FP models with lower BIC-values are all among the models with degree $(m_1, m_2) = (1, 1)$. Figure 2 shows the best 3 models based on the AIC-criterion (left panel) and BIC-criterion (right panel), respectively. The resulting model fits differ considerably in the segment before 3 years of age, indicating a distortion in the age-dependent VZV seroprofile in that region, presumably due to the presence of maternal antibodies. The most optimal change-points are located at around 3(1) years of age according to the BIC(AIC)-criterion. While AIC-based selection clearly leads to overfitting, BIC-based selection tends to provide more rigorous fits to the data. Note that the model with minimal BIC-value is the continuous one-change-point FP with degrees $(m_1, m_2) = (1, 1)$. With rank 4, this model belongs to the best ones in terms of AIC as well.

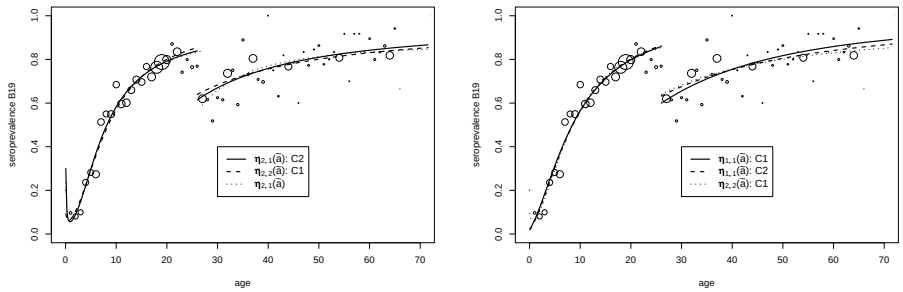

Figure 3: Seroprevalence plot for PVB19 with the three best change-point fractional polynomials (best: solid line; second best: dashed line; third: dotted line) according to AIC (left panel) and BIC (right panel).

Applying the same methodology to the PVB19 data results in the models as listed in Table 2. Again the 'original' FP models have worse ranks than the one-change-point FP models based on AIC, and more clearly still on BIC too. The difference in selecting more complex models using AIC rather than

Model	Constraint	AIC	Rank	BIC	Rank	$\hat{c}$	Powers
$\eta_1(\tilde{a})$		3 485.2	(16)	3 503.3	(13)	–	-1
$\eta_2(\tilde{a})$		3 475.8	(15)	3 505.9	(15)	–	-2, -1
$\eta_{1,1}(\tilde{a})$		3 438.6	(6)	3 480.9	(4)	26	0; -1
$\eta_{1,1}(\tilde{a})$	C1	3 437.3	(4)	3 467.4	(1)	26	0
$\eta_{1,1}(\tilde{a})$	C2	3 437.5	(5)	3 473.6	(2)	26	0; 0
$\eta_{1,1}(\tilde{a})$	C3	3 459.8	(14)	3 496.0	(10)	15	0; 3
$\eta_{1,2}(\tilde{a})$		3 442.5	(11)	3 496.7	(11)	26	0; -2, 3
$\eta_{1,2}(\tilde{a})$	C2	3 440.7	(8)	3 489.0	(6)	26	0; -0.5, -0.5
$\eta_{1,2}(\tilde{a})$	C3	3 445.7	(13)	3 493.9	(9)	21	0; -2, -2
$\eta_{2,1}(\tilde{a})$		3 437.0	(3)	3 491.3	(8)	27	-1, -1; -2
$\eta_{2,1}(\tilde{a})$	C2	3 436.0	(1)	3 484.2	(5)	26	-2, -0.5; -0.5
$\eta_{2,1}(\tilde{a})$	C3	3 441.6	(10)	3 489.8	(7)	28	0.5, 3; -2
$\eta_{2,2}(\tilde{a})$		3 440.9	(9)	3 507.2	(16)	26	-1, -1; -2, 3
$\eta_{2,2}(\tilde{a})$	C1	3 436.1	(2)	3 478.3	(3)	26	-1, -0.5
$\eta_{2,2}(\tilde{a})$	C2	3 439.0	(7)	3 499.3	(12)	26	-1, -1; -1, 1
$\eta_{2,2}(\tilde{a})$	C3	3 445.4	(12)	3 505.7	(14)	29	0.5, 3; -2, -2

Table 2: PVB19-models with corresponding AIC-and BIC-values, corresponding ranks, selected change-point and power(s).

BIC is not so apparent as it is for VZV. Looking at the ranks of the different types of constraints (C1,C2,C3), it is clear that continuity (C3) has an overall worse ranking, which suggests a sudden change in the seroprevalence profile. Figure 3 shows the three best models according to AIC (left panel) and BIC (right panel). While there is an apparent difference in how AIC-and BIC-selected models estimate the seroprevalence at the younger age ranges, there is only a moderate difference among the profiles for higher age-values. The optimal change-point is located at 26-27 years of age.

4.2 Two-change-point Fractional Polynomial

While applying a one-change-point fractional polynomial revealed some apparent distortions in the VZV-and PVB19-seroprofile, it is of interest, if present, to consider more distortions using a second, different, change-point as described by Model (4.2). Note that one-change-point and conventional FPs are special cases of Model (4.2) for $c_2 > \max(a)$ and $c_1 > \max(a)$, respectively. The presence of a second change-point can be assessed by comparing these two-change-point models with the best one-change-point models in terms of AIC and BIC, respectively.

$$\begin{aligned}\eta_{m_1,m_2,m_3}(\tilde{a},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\delta},\boldsymbol{p},\boldsymbol{q},\boldsymbol{r},c_1,c_2) &= \eta_{m_1}(\tilde{a},\boldsymbol{\beta},\boldsymbol{p})\mathrm{I}\{a<c_1\}\\ &+ \eta_{m_2}(\tilde{a},\boldsymbol{\gamma},\boldsymbol{q})\mathrm{I}\{c_1\le a\le c_2\},\\ &+ \eta_{m_3}(\tilde{a},\boldsymbol{\delta},\boldsymbol{r})\mathrm{I}\{c_2<a\}, \qquad (4.2)\end{aligned}$$

Since there exist numerous combinations of first and second degree FPs and change-points, we will limit ourselves to first degree FPs and more specifically to a limited set of submodels of $\eta_{1,1,1}(\tilde{a},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\delta},\boldsymbol{p},\boldsymbol{q},\boldsymbol{r},c_1,c_2)$.

$$\begin{aligned}\eta_{1,1,1}(\tilde{a},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\delta},\boldsymbol{p},\boldsymbol{q},\boldsymbol{r},c_1,c_2) &= [\beta_0+\beta_1\tilde{a}^{p_1}]\mathrm{I}\{a<c_1\}\\ &+ [\gamma_0+\gamma_1\tilde{a}^{q_1}]\mathrm{I}\{c_1\le a\le c_2\}\\ &+ [\delta_0+\delta_1\tilde{a}^{r_1}]\mathrm{I}\{c_2<a\} \qquad (4.3)\end{aligned}$$

Table 3 presents an overview of the resulting fits of the two-change-point FP as defined by Model (4.3) and constraints (C1,C2,C3). In addition, an interesting alternative is (C4) where the middle part changes but the first and last part come from the same model: (4.2) with $\eta_{m_1}(\tilde{a},\boldsymbol{\beta},\boldsymbol{p}) = \eta_{m_3}(\tilde{a},\boldsymbol{\delta},\boldsymbol{r})$. This potentially describes a situation where the prevalence profile changes for a certain age group only. Note that (C1) and (C2) should be interpreted over all FP components now and condition (C3) can be extended into ($C3_1$) denoting continuity in c_1 and ($C3_2$) denoting continuity in c_2, respectively.

VZV	Model	Constraint	AIC	Rank	BIC	Rank	$\hat{c}$	Powers
	$\eta_{1,1,1}(\tilde{a})$	C2	1 363.7	(2)	1 416.7	(4)	3, 12	-0.5; 1; 0
	$\eta_{1,1,1}(\tilde{a})$	C1 & C4	1 369.6	(4)	1 404.9	(1)	4, 38	0.5
	$\eta_{1,1,1}(\tilde{a})$	C2 & C4	1 365.1	(3)	1 406.3	(2)	4, 38	0.5; 0
	$\eta_{1,1,1}(\tilde{a})$	$C3_1$ & C4	1 363.6	(1)	1 416.6	(3)	3, 30	3; -1; -2
PVB19	Model	Constraint	AIC	Rank	BIC	Rank	$\hat{c}$	Powers
	$\eta_{1,1,1}(\tilde{a})$	C2	3 435.9	(1)	3 490.2	(3)	3, 26	0.5; -0.5; -0.5
	$\eta_{1,1,1}(\tilde{a})$	C1 & C4	3 436.0	(3)	3 472.2	(1)	6, 26	0
	$\eta_{1,1,1}(\tilde{a})$	C2 & C4	3 436.7	(4)	3 478.9	(2)	6, 26	0; 0
	$\eta_{1,1,1}(\tilde{a})$	$C3_1$ & C4	3 436.0	(2)	3 490.3	(4)	2, 26	0.5; -0.5; -2

Table 3: Selected VZV-and PVB19-models with corresponding AIC-and BIC-value and ranks, selected change-points and power(s).

Figure 4 shows the resulting two best two-change-point FP fits on VZV (upper row) and PVB19 (lower row) using AIC (left column) and BIC (right column), respectively. The seroprofiles for VZV suggest that one change-point is sufficient to capture the distortions in the data. Note indeed that none of the two-change-point FP models for VZV result in lower AIC-nor lower BIC-value as compared with the best three one-change-point FP. The PVB19-models differ from the previous results in how they estimate seroprevalence for infants and young children (up to 8 years of age). The models, better in terms of BIC, fit an increased prevalence for the age group 6 to 26 years of age, whereas those better in terms of AIC extend the one-change-point models by a decrease in the prevalence for infants, resembling the maternal antibodies influence observed for VZV. The two-change-point FP models for PVB19 are comparable to the best one-change-point models in terms of AIC. They are worse when turning to BIC, which is expected due to their complexity. Let us now synthesize these finding for both the VZV and B19 sample.

Both AIC and BIC have been used to select the most optimal model from the candidate set of models. There was a clear indication that using AIC,

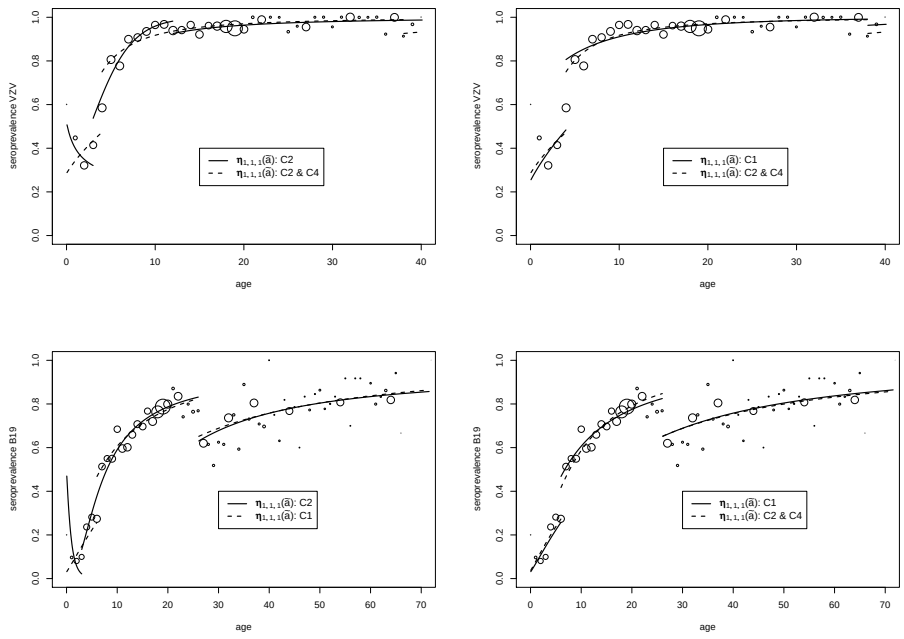

Figure 4: Seroprevalence plot for VZV (upper row) and PVB19 (lower row) with the two best two-change-point fractional polynomials (best: solid line; and second best: dashed line) according to AIC (left column) and BIC (right column).

models tend to overfit the data while using BIC less complex models are chosen. It is therefore that we limit ourselves to the BIC-based best models to synthesize our findings.

Figure 5 shows the best FP, one-change-point FP and two-change-point FP for VZV (left panel) and PVB19 (right panel) based on the BIC-criterion. For both VZV and PVB19, in terms of BIC, the one-change-point FP models provide the best fit (Tables 1, 2, 3) with change-point respectively at 3 and 26 years of age.

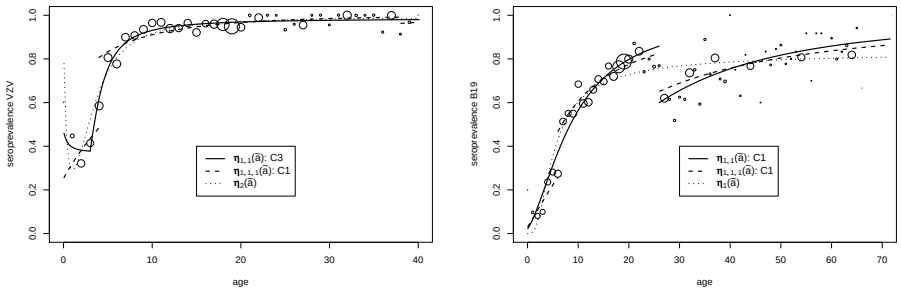

Figure 5: Seroprevalence plot for VZV (left panel) and PVB19 (right panel) with best no-changepoint, one-changepoint and two-changepoint models according to BIC.

5 Inferential Measures

Once a change-point has been detected and estimated, it is of importance to estimate the associated uncertainty. In a Bayesian setting this is often done by using an (un)informative prior distribution for the change-point and by looking at the 95% credibility interval of the corresponding posterior distribution. In a likelihood setting, one can use the BIC-criterion to calculate approximate posterior probabilities for each of the models (see e.g. Burnham and Anderson, 2002).

Therefore, given the BIC-based best model for either infection, i.e. $\eta_{1,1}(\tilde{a}) : (C3)$ and $\eta_{1,1}(\tilde{a}) : (C1)$ for VZV and PVB19, respectively, consider a series of models defined by a grid of change-points c_j, $j = 1, \ldots, J$. The prior distribution on this set of models is taken as uninformative (uniform distribution). For each model (for each c_j), BIC_j is calculated and the posterior probability is approximated by

$$\frac{\exp\{-\frac{1}{2}\text{BIC}_j\}}{\exp\{-\frac{1}{2}\sum_{j=1}^{J}\text{BIC}_j\}}.$$

Thus, we can estimate the posterior density of c for both $\eta_{1,1}(\tilde{a}) : (C3)$ (VZV) and $\eta_{1,1}(\tilde{a}) : (C1)$ (PVB19). Figure 6 depicts a smoothed density

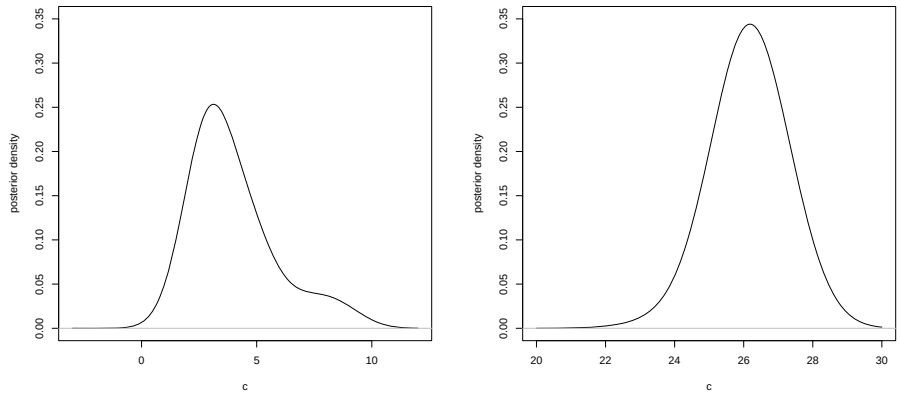

Figure 6: Posterior density plots for the change-points of models $\eta_{1,1}(\tilde{a})$: $(C3)$ (VZV, left panel) and $\eta_{1,1}(\tilde{a}) : (C1)$ (PVB19, right panel).

estimate of this posterior density for both VZV (left panel) and B19 (right panel). The resulting posterior densities show a rather high uncertainty associated with the corresponding change-point. For VZV, the location situates itself at 'pre-primary school'-ages, i.e. the approximate 95% credibility interval equals (0.22,8.88), strengthening the idea that maternal antibodies could be causing this change in profile. For B19, the approximate 95% credibility interval equals (21.61,25.89). Although there is no clear cause for this distortion, we will formulate some hypotheses worthwhile to further investigate in Section 6.

The BIC-based best models for VZV and B19 allow to derive quantities that can be used to gain more insight in the nature of these distortions. While the resulting model VZV is continuous in the change-point, there is an abrupt change on the derivative scale. The fraction of the left- and right-derivative in c has a natural interpretation as the relative change of the force of infection in the change-point. Indeed, denote $\eta_{1,1}(c)^L$ and $\eta_{1,1}(c)^R$, the left and right limit of the linear predictor in the change-point c and $\pi_{1,1}(c)^L$ and $\pi_{1,1}(c)^R$ the corresponding prevalences. Using a 'logit'-link function the force of infection $\lambda_{1,1}^{L,R}(c)$ can be expressed as $\eta_{1,1}'^{L,R}(c) \times \pi_{1,1}^{L,R}(c)$ (Shkedy et al.,

2006). Since $\pi^L_{1,1}(c) = \pi^R_{1,1}(c)$, it follows that $\lambda^L_{1,1}(c)/\lambda^R_{1,1}(c) = \eta'^L_{1,1}(c)/\eta'^R_{1,1}(c)$. The estimated ratio equals -0.005 while the confidence interval for this ratio was calculated using the delta method: $(-0.039, 0.029)$, reflecting a non-significant change in sign and a significant change in magnitude.

When looking at the model for B19, the change in intercept at the change-point results in a 'jump' at the level of the prevalence. The estimated jump equals 0.259 with confidence interval $(0.210, 0.309)$ (delta method). Alternatively, the OR in the change-point can be calculated as 4.062 (95% CI: $(2.937, 5.186)$), reflecting a fourfold increase in odds at the change-point.

6 Discussion

Estimating the force of infection from seroprevalence data involves making untestable assumptions about time homogeneity. An assumption often violated due to changes in the mixing behavior of humans, in the virulence of the pathogen, in cross-reactions with other related emerging diseases, epidemics or any other unknown causes. Current practice shows that distortions with respect to monotonicity in seroprevalence are often discarded; maternal antibodies are not explicitly taken into account and monotonicity is often imposed. In this paper, we propose to use a change-point fractional polynomial model to explicitly search for these distortions. The use of fractional polynomials retains the features of a parametric model while incorporating considerable flexibility. The proposed methodology was applied to data on VZV and B19 in Belgium and change-points at the age of 3 and 26 years of age, respectively, were identified for both data sets using the BIC-criterion. A posterior density of the change-point was obtained using that same BIC-criterion and its link to Bayesian methodology (Burnham and Anderson, 2002).

While the distortion in the seroprofile of VZV is presumably caused by the presence of maternal antibodies (although it is believed that maternal antibodies have disappeared around 6 months of age), different hypotheses for

the distortion in the seroprofile of B19 can be formulated. A first hypothesis is a new type of the virus which emerged in the period of 1985-1988 mainly affecting children aged 10-14 years (known to be the ages with highest force of infection). Although this hypothesis is untestable and no empirical data provide evidence for this, its plausibility can be based on the findings of new emerging types of PVB19 by Ekman et al. (2007). A second hypothesis is given by waning antibodies, i.e. once infected, a boost in the amount of antibodies is observed after which this amount gradually decreases possibly below the cut-off value which is used to classify individuals as seronegative or seropositive. Since the force of infection is highest among 10-14 years of age and infection afterwards is less likely until parenthood, it is potentially possible that the individuals around the age of 26 years are wrongly classified as negatives. The rise in the seroprofile after 26 years of age is associated with the likely occurrence of child-parent transmission, which could boost the parents' antibody level. Note that the average age of an adult becoming a parent was 26-28 years of age in Belgium (source: EUROSTAT). Some simulations omitted from the text, support the plausibility of this hypothesis, however empirical data, such as incidence data, is lacking. A third hypothesis is given by the occurrence of an epidemic. More specifically, it is known that an epidemic occurred in the Netherlands and the UK in 1998 (source: eurosurveillance 1998). However it is unclear whether this or any epidemic like this could fully explain the elevated seroprevalence for the younger age groups. Specific simulation models could show whether this is a reasonable hypothesis, but are beyond the scope of the paper.

This paper does not aim to present an all encompassing methodology to model seroprevalence data and derive infectious disease parameters such as the force of infection and the basic reproduction number. The proposed methodology, possibly of interest in other research domains, has been developed with the sole purpose of providing a rigorous way of detecting distortions in seroprofiles, which due to the ability of capturing abrupt changes in the seroprofile surpasses the use of ordinary scatterplot smoothers for this

purpose. Once these distortions are detected and whenever antibody level data is available, one can use mixture models and the like to take into account the presence of maternal antibodies and/or waning antibodies and/or other plausible effects. The application of these techniques is however not entirely straightforward (Gay, 1996; Bollaerts et al., 2008), as is using these results in mathematical models of infectious diseases. Clearly, care is required when observing distortions. Different scenarios can be implemented and contrasted to the observed seroprofile to validate the appropriateness of the mathematical model.

While the proposed approach searches for a change-point over a suitable grid on the age range using the AIC-criterion in resemblance with the selection of the appropriate fractional polynomial, another option is to embed the change-point methodology in a Bayesian framework where an uninformative prior distribution on the change-point reflects the uncertainty around the threshold (see e.g. Carlin et al., 1992).

Elaborating on the use of a transition model for change-point estimation in logistic regression as proposed by Pastor-Barriuso et al. (2003) is not straightforward in combination with the use of fractional polynomials. Several adjustments need to be made since the predictor in a FP needs to be positive, which is not necessarily the case for the models considered by Pastor-Barriuso et al. (2003). This however is a topic of further research.

Acknowledgements

We gratefully acknowledge two referees and an associate editor for provoking thoughts that have lead to an improved version of the manuscript. 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 PVB19 testing. This work is part of the research project MSM 0021620839, has been funded 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, by the Fund of Scientific Research (FWO, Research Grant G039304) in Flanders, Belgium, by the IAP research network nr P6/03 of the Belgian Government (Belgian Science Policy) and by the Grant Agency of Charles University, project number 252387/2007. This work benefited from discussions held in POLYMOD, a European Commission project funded within the Sixth Framework Programme, Contract number: SSP22-CT-2004-502084

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. Petrov and F. Csaki (Eds.), *2nd International Symposium on Information Theory*, pp. 267–281. Budapest: Akademia Kiado.

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

Bollaerts, K., M. Aerts, N. Hens, Z. Shkedy, C. Faes, P. Van Damme, and P. Beutels (2008). Estimating the force of infection directly from antibody levels. Technical report, Center for Statistics, Hasselt University.

Burnham, K. and D. Anderson (2002). *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach*. Springer-Verlag New York Inc.

Capasso, V. (1993). *Mathematical Structures of Epidemic Systems*. Springer.

Carlin, B., A. Gelfand, and A. Smith (1992). Hierarchical bayesian analysis of changepoint problems. *Applied Statistics 41*, 389–405.

Diamond, L. D. and J. M. McDonald (1992). *Demographic Application of*

Event History Analysis., Chapter Analysis of current-status data. Oxford University Press.

Diekmann, O., J. Heesterbeek, and J. Metz (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*, 65–382.

Ekman, A., K. Hokynar, L. Kakkola, K. Kantola, L. Hedman, H. Bondn, M. Gessner, C. Aberham, P. Norja, S. Miettinen, K. Hedman, and M. Sderlund-Venermo (2007). Biological and immunological relations among human parvovirus B19 genotypes 1 to 3. *Journal of Virology 81*, 6927–6935.

Farrington, C. P. (1990). Modeling forces of infection for measles, mumps and rubella. *Statistics in Medicine 9*, 953–967.

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

Gay, N. (1996). Analysis of serological surveys using mixture models: application to a survey of parvovirus B19. *Statistics in Medicine 15*, 1567–1573.

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

Keiding, N. (1991). Age-specific incidence and prevalence: A statistical perspective (with discussion). *Journal of the Royal Statistical Society, Series A 154*, 371–412.

McCullagh, P. and J. Nelder (1989). *Generalized Linear Models*. Chapman & Hall.

Mossong, J., N. Hens, V. Friederichs, I. Davidkin, M. Broman, B. Litwinska, J. Siennicka, V. P. Trzcinska, A., P. Beutels, A. Vyse, Z. Shkedy, M. Aerts, M. Massari, and G. Gabutti (2008). Parvovirus B19 infection in five european countries: seroepidemiology, force of infection and maternal risk of infection. *Epidemiology and Infection*, In press.

Nagelkerke, N., S. Heisterkamp, M. Borgdorff, J. Broekmans, and H. Van Houwelingen (1999). Semi-parametric estimation of age-time specific infection incidence from serial prevalence data. *Statistics in Medicine 18*, 307–320.

Nardone, A., F. de Ory, M. Carton, D. Cohen, P. van Damme, I. Davidkin, M. Rota, H. de Melker, J. Mossong, M. Slacikova, A. Tischer, N. Andrews, G. Berbers, G. Gabutti, N. Gay, L. Jones, S. Jokinen, G. Kafatos, M. Martnez de Aragn, F. Schneider, Z. Smetana, B. Vargova, R. Vranckx, and E. Miller (2007). The comparative sero-epidemiology of varicella zoster virus in 11 countries in the European region. *Vaccine 25*, 7866–7872.

Pastor-Barriuso, R. and E. Guallar (1998). Use of two-segmented logistic regression to estimate change-points in epidemiologic studies. *American Journal of Epidemiology 148*, 631–642.

Pastor-Barriuso, R., E. Guallar, and J. Coresh (2003). Transition models for change-point estimation in logistic regression. *Statistics in Medicine 22*, 1141–1162.

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

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

Shkedy, Z., M. Aerts, G. Molenberghs, P. Beutels, and P. Van Damme (2006). Modeling age dependent force of infection from prevalence data using fractional polynomials. *Statistics in Medicine 5:9*, 1577–1591.

Stasinopoulos, D. and R. Rigby (1992). Detecting break points in generalized linear models. *Computational Statistics and Data Analysis 13*, 461–471.

Thieme, H. (2003). *Mathematics in Population Biology*. Princeton University Press.

Thiry, N., P. Beutels, Z. Shkedy, R. Vranckx, C. Vandermeulen, M. Van Der Wielen, and P. Van Damme (2002). The seroepidemiology of primary varicella-zoster virus infection in Flanders (Belgium). *European Journal of Pediatrics 161*, 588–593.

Ulm, K. (1991). A statistical method for assessing a threshold in epidemiological studies. *Statistics in Medicine 10*, 341–349.

Ulm, K. and H. Küchenhoff (2000). Re: Use of two-segmented logistic regression to estimate change-points in epidemiologic studies." (letter). *Am. J. Epidemiology 152*, 289.

Valeur-Jensen, A., C. Pedersen, T. Westergaard, I. Jensen, M. Lebech, P. Andersen, P. Aaby, B. Pedersen, and M. Melbye (1999). Risk factors for parvovirus B19 infection in pregnancy. *Journal of the American Medical Association 281*, 1099–1105.