

## Modeling forces of infection using monotone local polynomials

Peer-reviewed author version

SHKEDY, Ziv; AERTS, Marc; MOLENBERGHS, Geert; Beutels, Phillipe & Van Damme, Pierre (2003) Modeling forces of infection using monotone local polynomials. In: Journal of the Royal Statistical Society: series C: applied statistics, 52(4). p. 469-485.

DOI: 10.1111/1467-9876.00418

Handle: <http://hdl.handle.net/1942/431>

# Modeling Forces of Infection

## Using Monotone Local Polynomials

Ziv Shkedy, Marc Aerts, Geert Molenberghs

Limburgs Universitair Centrum, Center for Statistics, Biostatistics, Universitaire Campus,  
B3590 Diepenbeek, Belgium

Philippe Beutels, Pierre Van Damme

University of Antwerp, Epidemiology and Community Medicine, Center for Evaluation of  
Vaccination, B2610 Antwerp, Belgium

**Summary.** Based on serological data from prevalence studies of rubella, mumps and hepatitis A, this paper describes a flexible local maximum likelihood method for the estimation of the rate at which susceptible individuals acquire infection at different ages. In contrast to parametric models used before in the literature, the local polynomial likelihood method allows one to model this age-dependent force of infection without making any assumptions about the parametric structure. Moreover, this method allows for simultaneous nonparametric estimation of age-specific incidence and prevalence. Unconstrained models may lead to negative estimates for the force of infection at certain ages. To overcome this problem and to guarantee maximal flexibility, the local smoother can be constrained to be monotone. It turns out that different parametric and nonparametric estimates of the force of infection can exhibit considerable different qualitative features like location and number of maxima, emphasizing the importance of a well-chosen flexible statistical model.

**Keywords:** Prevalence data; Incidence; Force of Infection; Local Polynomial; Smoothing

## 1. Introduction

Mathematical models are often used to describe the process of infectious diseases at population level (Anderson and May, 1991). 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 induced by infection is assumed to be lifelong. We further assume that the mortality caused by the infection is negligible. Let  $q(a, t)$  be the fraction of susceptible individuals (not yet infected) 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

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

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

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

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

The force of infection can be estimated from an age-specific cross-sectional prevalence sample, which is a sample taken at a certain time point and for each one of the individuals in the sample the observed information consists of whether the individual has been infected

or not before his/her age at the test. If only one prevalence sample is available, then age and time dependence cannot be separated and one has to assume that the force of infection is time independent in order to model age dependent force of infection. Hence, if it is assumed that the disease is in steady state, then the age dependent force of infection can be modeled according to (2).

Muench (1959) 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  $\beta > 0$ . In this approach, the force of infection  $\ell(a)$  is age independent and the solution to (2) is given by  $q(a) = \exp(-\beta a)$ . Griffiths (1974) considered a catalytic infection model for measles in which the force of infection increases linearly in the age range 0–10 years. Grenfell and Anderson (1985) extended the model and used functions  $q(a) = \exp(-\sum_i \beta_i a^i)$  leading to a polynomial form for the force of infection. Other parametric models were proposed by Becker (1989), who assumed that the time spent in the susceptible class follows the Weibull distribution with monotone force of infection  $\ell(a) = \mu \alpha a^{\alpha-1}$  for  $\mu > 0$  and  $\alpha > 0$ . He also considered a model with piecewise constant force of infection,  $\ell(a) = \beta_i$  for  $a_{i-1} \leq a < a_i$ , implying that  $q(a) = \exp(-[\sum_{j=1}^{i-1} \beta_j (a_j - a_{j-1}) + \beta_i (a - a_{i-1})])$ , for  $a_{i-1} \leq a < a_i$ .

A cross-sectional sample can be seen as a particular case of current-status data. This type of data consists of information about individual age and whether or not a specific event occurred before the individual's age at the time of test (events such as first marriage, first birth, infection by a disease, etc.). The analysis of current-status data was discussed by Diamond and McDonald (1992), who proposed several parametric models for the incidence and pointed out that in a current-status sample the time to event for all individuals is censored: individuals who had experienced an event before their age at test are left censored and individuals who did not experience an event before their age at test are right censored.

As a nonparametric approach, Keiding (1991) proposed to use isotonic regression (Barlow *et al.*, 1972) to estimate the prevalence,  $\pi(a) = 1 - q(a)$ , which is, in this case, a step function  $\hat{\pi}(a)$ . He suggested to apply a kernel smoothed estimate  $\int K((a-u)/h)/(h(1-\hat{\pi}(u_-)))d\hat{\pi}(u)$ , where  $K$  is a kernel and  $h$  a bandwidth, as an estimate for  $\ell(a)$ . In case that other covariates are included in the model (such as gender, nationality, etc.), Shiboski (1998) proposed, in the context of current status data, to use generalized additive models in order to estimate the prevalence. A semi-parametric approach to model age-time dependent force of infection was discussed by Nagelkerke *et al.* (1999).

As a nonparametric method, higher order local polynomial estimation is known to have several desirable properties, as compared to kernel estimates (Fan and Gijbels 1996). In this paper we use local polynomials to estimate age-specific prevalence and force of infection of rubella and mumps in the UK and hepatitis A in Belgium. The first two datasets were discussed in Farrington *et al.* (2001), who used nonlinear models for the prevalence and for the force of infection. These datasets consist of 4230 and 8179 individuals for rubella and mumps, respectively, with age range between 1 to 44 years old (Figures 1a and 1c). While rubella and mumps are common airborne childhood infectious diseases, the hepatitis A virus (HAV) is mainly (> 95 %) transmitted by the feco-oral route (e.g. through food and water polluted by faeces containing the virus). Transmission is facilitated by poor hygienic living and housing conditions, and is particularly common in developing countries (Hadler, 1991, Beutels *et al.*, 1997). In these countries HAV is mainly a childhood infection, whereas in industrial countries HAV infection occurs during adulthood as well as childhood. In the poorest developing countries, the pattern of high endemicity is characterized by rapid infection at a very young age; over 90% of the children become infected by the age of 5. In 1993 and early 1994, a study of the prevalence of HAV antibodies was conducted in

the Flemish Community of Belgium. The purpose of this study was to obtain data on the prevalence of hepatitis A in Flanders and to analyze the epidemiological pattern of HAV. During the study period serum samples were collected from hospitals (non-infectious disease wards) in the Flemish Community. The dataset contains the serological results of 3161 Belgian individuals together with their age in years, ranging from 0.5 to 85 years (Figure 1e). The study group was similar in composition to the Flemish population in terms of age.

#### FIGURE 1 ABOUT HERE

Estimates of the force of infection are negative whenever the estimated probability to be infected before age  $a$  is a nonmonotone function. One way to circumvent this problem is to define a nonnegative force of infection,  $\ell(a) \geq 0$  for all  $a$ , and to estimate a model with  $\eta(a) = \int_0^a \ell(u)du$ . Farrington (1990) applied this method to measles, mumps and rubella. However, such an approach is still parametric in nature and hence less flexible. It inevitably puts constraints on other qualitative characteristics of the functional form of the predictor  $\eta(a)$ , which in turn restricts the flexibility of the estimate of the true force of infection. A nonparametric approach however only assumes some form of smoothness for  $\pi(a)$  or  $\ell(a)$ . Figure 1 also shows the approach of Keiding (1991), a step function estimate for  $\pi(a)$  and the smooth force of infection estimates using two different bandwidths. The values chosen are  $h = 6$  and  $h = 8$  for rubella,  $h = 4.5$  and  $h = 6$  for mumps and  $h = 15$  and  $h = 20$  for HAV; values were chosen by visual inspection. Although the nonparametric estimate of the probability  $\pi(a)$  is consistent under very general conditions, it has the disadvantage of being nonsmooth. The three estimates for the force of infection in the right hand panels have quite different shapes. The parametric models in Figure 1 (indicated by FP) lead a model with a unique maximum for the force of infection. These models were estimated

using fractional polynomials and will be discussed in detail in Sections 2 and 3. For rubella and mumps, the nonparametric models predict multi peaks models and for HAV, according to different choices for the bandwidth  $h$  the smooth nonparametric estimates have one maximum (at age 44.2) or two maxima (at ages 36.3 and 50.2). This illustrates two important issues: i) nonparametric versus parametric models, ii) the critical choice of the bandwidth and the need for an optimal and data-driven bandwidth choice. In this paper, we propose to estimate the force of infection by local polynomials. Compared to the method of Keiding (1991), this approach allows simultaneous estimation of prevalence and force of infection. As a consequence, the estimated probability curve is also smooth. Moreover, local polynomials are known to have several desirable properties like automatic boundary correction (Fan and Gijbels 1996). Whereas Keiding (1991) chose his bandwidth  $h$  by visual inspection, we will select an optimal data driven bandwidth, minimizing the mean squared error of the estimated force of infection. According to the principle “smooth then constrain” (Mammen *et al.* 2001), the fitted probability curve is constrained to be monotone leading to a nonnegative estimated force of infection. While the focus is on local polynomial models, we will also use fractional polynomials (Royston and Altman 1994). Fractional polynomials extend classical polynomials to multiple noninteger powers, leading to very flexible parametric models. Although these fractional polynomials offer an interesting alternative on its own, in this paper we will mainly use them to determine the optimal bandwidth for the local polynomial method. An elaborate discussion of the use of fractional polynomial as a parametric alternative for the estimation of the prevalence and the force of infection is presented in Shkedy *et al.* (2002).

This paper is organized as follows. In Section 2 we describe the use of local polynomials to estimate  $\pi(a)$  and  $\ell(a)$ , including bandwidth selection. Section 3 applies the method on

the three datasets mentioned above. A small simulation study, comparing the performance of our method with Keiding's approach, is discussed in Section 4 and Section 5 concludes with some final remarks and ideas for further research.

## 2. Modeling Age-Dependent Force of Infection with Local Polynomials

Consider an age-specific cross-sectional prevalence sample of size  $N$  and let  $a_i$  be the age of the  $i$ th subject. Instead of observing the age of infection, we observe a binary response indicator  $Y_i$  taking the value 1 if subject  $i$  had experienced infection before age  $a_i$  and 0 otherwise. Let  $\pi(a_i)$  denote the probability to be infected before age  $a_i$ , so  $\pi(a_i) = 1 - q(a_i)$ . The log-likelihood is given by  $L(\boldsymbol{\beta}) = \sum_{i=1}^N Q_i \{Y_i, g^{-1}(\eta(a_i))\}$  where  $Q_i$  is the contribution of the  $i$ th subject to the Bernoulli log-likelihood with success probability  $\pi(a_i) = g^{-1}\{\eta(a_i)\}$ ,  $\eta(a)$  is the linear predictor and  $g$  the link function. The functional form describing how the force of infection changes with age is determined by the link function and the parametric structure of the linear predictor. Using a model with a log link as in Muench (1959), Griffiths (1974) and Grenfell and Anderson (1985),  $\eta(a)$  is the cumulative hazard and therefore the force of infection is simply the first derivative of the linear predictor. Indeed, under the catalytic model  $\pi(a) = 1 - e^{-\eta(a)}$ , and using the definition for the hazard rate, we get  $\ell(a) = \pi'(a)/\{1 - \pi(a)\} = \eta'(a) \exp\{-\eta(a)\}/\exp\{-\eta(a)\} = \eta'(a)$ . Note however that the log link suffers from the structural defect that the estimated probabilities can be exceed unity. In the general case, when the link function is not restricted to be the log link, the force of infection can still be expressed as a product of two functions

$$\ell(a) = \varphi\{\eta(a), \eta'(a)\} = \eta'(a)\delta\{\eta(a)\} \quad (3)$$

where the form of  $\delta$  is determined by the link function  $g$ . Table 1 shows four typical link functions with their corresponding structure for the force of infection.



TABLE 1 ABOUT HERE

The choice of a link function together with a specification of the functional form of  $\eta(a)$  as a function of the age  $a$  determines a fully parametric model. When turning to a local polynomial likelihood method in which no specific form for the predictor is assumed, the choice of a particular link function is less important (Fan, Heckman and Wand 1995). The local polynomial likelihood method provides consistent estimates for  $\eta(a)$  and  $\eta'(a)$ , without any parametric restriction on the functional form. They only have to satisfy some smoothness condition (Fan and Gijbels 1996, Chapter 3). Therefore, for a given link function, the local force of infection can be estimated according to (3).

Using a kernel  $K$ , assigning higher weights to data points in the neighborhood of some fixed age  $a$ , and a bandwidth parameter  $h$ , the local likelihood estimation is based on maximization of  $\sum_{i=1}^n Q_i \{Y_i, g^{-1}(\eta(a_i - a))\} K((a_i - a)/h)$ . The linear predictor is locally approximated by a polynomial of order  $p$ , e.g. for  $p = 1$  by a linear function  $\eta(a_i - a) \approx \eta(a) + \eta'(a)(a_i - a) = \beta_0(a) + \beta_1(a)(a_i - a)$ . The local estimate for  $\eta(a)$  is the local intercept,  $\hat{\beta}_0(a)$ , and the local slope,  $\hat{\beta}_1(a)$ , is the estimate for the first derivative  $\eta'(a)$ . Higher order polynomials can be considered as well. The estimation of  $\beta_0(a)$  and  $\beta_1(a)$  has to be repeated for each value of  $a$ . The choice of kernel is less important; typical choices are the symmetrical beta family,  $K(u) = (1 - u^2)^\gamma / \text{Beta}(0.5, \gamma + 1)$  for  $|u| \leq 1$ ,  $\gamma = 0, 1, 2, \dots$ , and the Gaussian kernel given by  $K(u) = \exp(-u^2/2) / \sqrt{2\pi}$ . Throughout the next sections, we will always use the latter Gaussian kernel function. The choice of the smoothing parameter  $h$  however is crucial and will be discussed in more detail in what follows. As explained later in this section, it will turn out that in our setting a local quadratic model is of special importance. For a given link function, the local force of infection can be estimated by

$$\hat{\ell}(a) = \hat{\eta}'(a) \delta \{ \hat{\eta}(a) \} = \hat{\beta}_1(a) \delta \{ \hat{\beta}_0(a) \} = \varphi \{ \hat{\beta}_0(a), \hat{\beta}_1(a) \}. \quad (4)$$

The last column in Table 1 presents the local estimates for the force of infection corresponding to different link functions.

The local polynomial estimation procedure requires a data driven value for the bandwidth  $h$ . Several methods to choose the value of  $h$  are discussed in Chapter 4 of Fan and Gijbels (1996). The minimization of the asymptotic mean squared error (*AMSE*) of  $\hat{\ell}(a)$  as a function of  $h$  leads to an optimal local bandwidth  $h(a)$ . The asymptotic normality of  $\hat{\ell}(a)$  follows from the asymptotic joint normality of  $(\hat{\beta}_0(a), \hat{\beta}_1(a))$  and the delta method can be used to derive expressions for the asymptotic bias and variance of  $\hat{\ell}(a)$ . Indeed, from the main theorem in Fan, Heckman and Wand (1995), it follows that for  $p = 2$  (local quadratic) and  $h = cn^{-1/7}$  ( $c$  some constant)

$$\begin{bmatrix} n^{\frac{3}{7}}(\hat{\beta}_0(a) - \eta(a)) \\ n^{\frac{2}{7}}(\hat{\beta}_1(a) - \eta'(a)) \end{bmatrix} \xrightarrow{D} N(\mathbf{b}(c), V(c)). \quad (5)$$

where  $\mathbf{b}(c) = (b_1(c), b_2(c))$  is the asymptotic bias and  $V(c)$  the asymptotic covariance matrix. Consider the function  $\varphi$  in (3). Using the delta method (some details in the Appendix), we get the following asymptotic normality result for the estimated force of infection:

$$n^{\frac{2}{7}} \left( \varphi(\hat{\beta}_0(a), \hat{\beta}_1(a)) - \varphi(\eta(a), \eta'(a)) \right) \xrightarrow{D} N(\gamma, \tau^2), \quad (6)$$

where

$$\gamma = \delta \{ \eta(a) \} b_2(c), \quad (7)$$

and

$$\tau^2 = \delta^2 \{ \eta(a) \} V_{22}(c). \quad (8)$$

The results in (6)–(8) indicate that  $\eta(a)$  influences the asymptotic bias and variance of the estimated force of infection only by the term  $\delta(\eta(a))$ . The asymptotic mean square error of

$\hat{\ell}(a)$  is given by

$$AMSE = \delta^2 \{ \eta(a) \} \{ b_2^2(c) + V_{22}(c) \}. \quad (9)$$

The optimal choice for the constant  $c$  of the optimal bandwidth  $h = cn^{-1/7}$  is the solution  $c_{opt}$  to

$$\frac{\partial b_2^2}{\partial c}(c_{opt}) + \frac{\partial V_{22}}{\partial c}(c_{opt}) = 0. \quad (10)$$

It follows from (10) that  $\eta(a)$  is not directly involved in the determination of the optimal bandwidth which can be obtained by just minimizing the  $AMSE$  of  $\hat{\beta}_1(a)$  as an estimator for  $\eta'(a)$ . Fan and Gijbels (1996) explain in their Section 3.3 why the choice  $p = 2$  is optimal for estimation of the first derivative  $\eta'(a)$ . Other odd choices for  $p - 1$  are also appropriate but for most applications the choice  $p = 2$  suffices. In the context of a binary response, Fan, Heckman and Wand (1995) showed that for  $p = 2$ , the optimal constant  $c_{opt}$  is given by

$$c_{opt}(a) = \left\{ 27 \frac{\int K^{*2}(z) dz}{\left( \int z^3 K^*(z) dz \right)^2} \frac{\pi(a)(1 - \pi(a))g'(\pi(a))^2}{f_A(a)\eta^{(3)}(a)^2} \right\}^{1/7} \quad (11)$$

where  $f_A(a)$  is the (unknown) density of the age distribution. The factors in (11) depending on the so-called equivalent kernel  $K^*$  are known and integrate to a constant for a given kernel  $K$  (see Section 3.2.2 in Fan and Gijbels 1996). The unknown quantities  $\eta^{(3)}(a)$ ,  $\pi(a)$  can be estimated using initial estimators resulting from a global fractional polynomial (or any other flexible parametric) model and the density  $f_A(a)$  can be estimated by a kernel estimator.

Note that the choice of the optimal bandwidth as discussed here minimizes the  $AMSE$  for the local estimate of the force of infection. Another option is to minimize the  $AMSE$  of  $\hat{\pi}(a)$  which would optimally lead to a local linear ( $p = 1$ ) instead of a local quadratic ( $p = 2$ ) approach. In that case the optimal bandwidth is  $c_{opt}n^{-1/5}$  with for  $c_{opt}$  an expression similar to (11) with  $\eta^{(2)}(a)$  instead of  $\eta^{(3)}(a)$ .

As mentioned above, we need initial estimators for  $\eta^{(3)}(a)$  (or  $\eta^{(2)}(a)$  in case of a local linear model) and  $\pi(a)$ . Using smoothers again would require new bandwidth choices and would make the procedure unnecessary complicated. At this stage, it is typically sufficient to estimate these unknown quantities based on a flexible parametric model. Here, fractional polynomials (Royston and Altman 1994) will be used. High order conventional polynomials offer a wide range of curve shapes but often fit the data badly at the extremes of the observed age range. Moreover, conventional polynomials do not have asymptotes and fit the data poorly whenever asymptotic behavior of the infection process is expected. To improve conventional polynomial models on these shortcomings, Royston and Altman (1994) introduced the family of fractional polynomials as a generalization of the conventional polynomial class of functions. In the context of binary responses, a fractional polynomial of degree  $m$  for the linear predictor is defined as

$$\eta_m(a, p_1, p_2 \dots p_m) = \sum_{i=0}^m \beta_i H_i(a), \quad (12)$$

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

$$H_i(a) = \begin{cases} a^{p_i} & \text{if } p_i \neq p_{i-1} , \\ H_{i-1} \times \log(a) & \text{if } p_i = p_{i-1} , \end{cases} \quad (13)$$

with  $p_0 = 0$  and  $H_0 = 1$ . Royston and Altman (1994) argued that, in practice, fractional polynomials of order higher than 2 are rarely needed. Note that, for models with log link function, the model proposed by Muench (1959) is defined as  $\eta_1(a, p = 1)$  and  $\eta_2(a, p_1 = 1, p_2 = 2)$  corresponds to the model proposed by Griffiths (1974). The models considered by Grenfell and Anderson (1985) have the general form of  $\eta_m(a, p_1, p_2 \dots, p_m)$  with  $p_i = i$  for  $i = 1, 2, \dots, m$ .

Keiding’s method (1991) to estimate  $\pi(a)$  is based on isotonic regression of the observed prevalence on age and results in a step function for  $\hat{\pi}(a)$ . In practice, the *pool adjacent violator algorithm* (PAV) (Barlow, 1972) can be used to calculate  $\hat{\pi}(a)$ , which is monotone by construction. Our local polynomial smooth estimate  $\hat{\pi}(a) = g^{-1}(\hat{\beta}_0(a))$  can be non-monotone as a function of age  $a$  and therefore result in negative estimates for the force of infection. Following Friedman and Tibshirani (1984) and Mammen *et al.* (2001), we suggest to estimate  $\pi(a)$  and  $\ell(a)$  (using the optimal bandwidth) and then, if necessary, to “isotonize” the estimates by the PAV algorithm. This is in line with the findings of Mammen *et al.* (2001). They showed that constrained smoothing leads to estimates of the form “smooth then constrain”. One could also try estimates based on the idea “constrain then smooth” (as in Keiding 1991). For local polynomials this idea does not work: smoothing by polynomials is not monotonicity preserving.

### 3. Application to the Data

Throughout the analysis, the logit link function is used. We start with selecting the best fractional polynomial, by an extensive grid search over powers  $\mathbf{p} = (p_1, p_2, \dots, p_m)$ . The deviance of the linear model  $\eta_1(a, 1)$  is taken as a baseline and improvement by other models is measured by  $G(m, \mathbf{p}) = D(1, 1) - D(m, \mathbf{p})$ , where  $D(m, \mathbf{p})$  is the deviance of the model with fractional polynomial of order  $m$  and sequence of powers  $\mathbf{p}$ . Note that a large value of  $G$  indicates a better fit. Although fractional polynomials provide a wide range of curve shapes, there is no guarantee that the corresponding estimated curve  $\pi(a)$  will be monotone in age. Therefore, the model with the largest value of  $G$  among all monotone models is chosen as the best fractional polynomial model. To decide whether a model of degree  $m$  is adequate or a degree  $m + 1$  model is needed, Royston and Altman (1994) recommend to use

the criterion  $D(m, \tilde{\mathbf{p}}) - D(m + 1, \tilde{\mathbf{p}}) > \chi_{2,0.9}^2$  where  $\tilde{\mathbf{p}}$  is the power sequence for the model that has the best goodness-of-fit. This model is of interest in its own as a flexible model for estimating the force of infection and will be used subsequently to provide a data-driven bandwidth for the local likelihood approach.

## TABLE 2 ABOUT HERE

Table 2 presents deviances for the best monotone fractional polynomials that were fitted to the data. Estimated models are shown in Figure 1. Clearly, for all three examples, first order fractional polynomials are not adequate and second order fractional polynomials are required.

As mentioned in the previous section, the computation of the bandwidth of the local polynomial estimator requires estimates for the density  $f_A(a)$ , initial estimates for  $\pi(a)$  and for the second (for  $p = 1$ ) and third derivative (for  $p = 2$ ) of the linear predictor  $\eta(a)$ . The age density was estimated with a kernel estimator, shown in Figure 2. The estimate for the linear predictor of the optimal fractional polynomial model is given by  $\hat{\eta}(a) = \hat{\beta}_0 + \hat{\beta}_1 a^{\hat{p}_1} + \hat{\beta}_2 a^{\hat{p}_2}$  such that  $\hat{\eta}^{(k)}(a) = \hat{\beta}_1 a^{\hat{p}_1 - k} \prod_{i=1}^k (\hat{p}_1 - i + 1) + \hat{\beta}_2 a^{\hat{p}_2 - k} \prod_{i=1}^k (\hat{p}_2 - i + 1)$ . As an initial estimate for  $\hat{\pi}(a)$  we took the probability estimated by the optimal fractional polynomial. The local optimal bandwidth was then estimated according to (11). Figure 2 shows estimates for the different unknowns (for hepatitis A) in (11) and indicates that the local bandwidth which minimizes the  $AMSE$  of  $\hat{\eta}'(a)$  (with  $p = 2$ ) is higher than the optimal bandwidth that minimizes the  $AMSE$  of  $\hat{\eta}(a)$  (with  $p = 1$ ) (which is expected and reflects that more data are needed to fit locally a more complicated model).

## FIGURES 2 AND 3 ABOUT HERE

Figures 3a and 3b show the local linear and local quadratic fit for rubella. Up to age 20 both models predict the same patterns for the force of infection, except at age 1.5 where the local quadratic predicts higher values for the force of infection. From age 20 onwards, the local quadratic model predicts steadily decreasing trend with force of infection equal to zero from age 38 onwards. The local linear model indicates that the force of infection flattens around the value of 0.06 and even slightly increases from age 40 onwards. The estimated models for mumps are shown in Figure 3c and 3d. Both models predict a maximum around age 5 (4.97 and 5.4 for the local linear and local quadratic models, respectively), while the local linear model predicts a secondary peak at age 10. Due to the larger bandwidth, this peak is smoothed out by the local quadratic model. Similar to rubella, the local quadratic model predicts a steady decrease of the force of infection becoming zero from age 30 onwards while the local linear model indicates that the force of infection slightly increases as from age 30. Note that for both rubella and mumps the local quadratic estimate for the force of infection is zero at older age groups due to initially negative estimates for the force of infection. The PAV algorithm was applied to the estimated prevalence at these age groups which leads to a nondecreasing prevalence curve and force of infection equal to zero at these age groups. For HAV, the local linear model has a bimodal form with maxima at 28 and 55 years (Figure 1f). Note the close resemblance with Keiding's smoothed estimate with bandwidth  $h = 15$ , except for the smallest and largest ages where the local linear estimate seems to flatten out more nicely (less boundary effects). The local quadratic polynomial however produces negative estimated forces of infection from age 70.4 onwards. This is due to the larger values (at higher age groups) of the optimal bandwidth that was used to fit the quadratic local polynomial model. Again, the PAV algorithm was applied in order to “monotonize” the probability estimates and as a result the force of infection is estimated

to be zero after age 70.4. The force of infection estimated with the local quadratic model shows a unimodal form with a maximum at age 40 and is quite similar to Keiding's smoothed estimate with bandwidth  $h = 20$  (except for ages above 75). Since the optimal bandwidth for the local quadratic model is chosen to minimize the local *AMSE* of the force of infection we recommend the local quadratic method.

To assess the local variability of  $\hat{\ell}(a)$ , a bootstrap procedure (Davison and Hinkley, 1997) was applied to calculate pointwise confidence intervals for  $\hat{\ell}(a)$ . Specifically,  $B$  bootstrap samples were generated by resampling the original data (with replacement, each sample containing  $N$  pairs  $(a_i^*, Y_i^*)$ ) and  $(1 - 2\alpha) \times 100\%$  percentile confidence intervals  $(\hat{\ell}^*(a)_{[(B+1)\alpha]}, \hat{\ell}^*(a)_{[(B+1)(1-\alpha)]})$  were calculated, where  $\hat{\ell}^*(a)_{[(B+1)\alpha]}$  is the  $(B+1)\alpha$ th order statistic of the bootstrap replicated local forces of infection  $\hat{\ell}_1^*(a), \dots, \hat{\ell}_B^*(a)$ . The same optimal local bandwidth as shown in Figure 2 was used (a data driven local bandwidth within each bootstrap run was computationally not feasible). Since the estimation procedure was not constrained for the bootstrap samples, estimates for the force of infection at higher ages might become negative, for both linear and quadratic models. Equivalently to the PAV algorithm, one can define the lower and upper confidence limits to be  $\max\{0, \hat{\ell}^*(a)_{[(B+1)\alpha]}\}$  and  $\max\{0, \hat{\ell}^*(a)_{[(B+1)(1-\alpha)]}\}$  respectively. Figure 4 shows local estimates for  $\pi(a)$  and  $\ell(a)$  for rubella and mumps together with their bootstrap confidence intervals. The variability of  $\hat{\pi}(a)$  increases at older age groups, which can be explained by the smaller sample sizes at these age groups. For rubella, at younger age groups, the lower and upper confidence limits range between 0 to 0.15. Note that for mumps, from age 30 and onwards, the lower and upper confidence limits for the constrained force of infection are both zero.

FIGURES 4 AND 5 ABOUT HERE



Figure 5 shows the bootstrap estimates for the prevalence and force of infection for HAV. The right hand panels in Figure 5 display the corrected pointwise confidence interval for HAV. The confidence intervals obtained from the linear polynomial turned out to be wider than those obtained from the quadratic model. This is due to the larger value of the optimal bandwidth that was used to estimate the quadratic model. Based on its optimal theoretical properties (being optimal for estimating the force of infection) and because of the observed superior accuracy characteristics, we recommend the use of the local quadratic smoother with optimal data-driven bandwidth to estimate the force of infection based on an age-specific prevalence sample.

#### 4. Simulation study

We performed a small simulation study to investigate the performance of the monotonized local polynomial models compared to the isotonic regression approach. The test function considered is  $\pi(a) = \exp \{-4.98 - 0.0258a^{1.3958} + 0.3081a^{0.9375}\}$ , which is the estimated fractional polynomial from the previous section. With the same sample size and age values as in the HAV dataset,  $M = 150$  new datasets were generated with the number of infected individuals at age  $a$  drawn from the binomial distribution with probability  $\pi(a)$ . In each simulation run,  $\pi(a)$  was estimated by isotonic regression and by monotonized local linear and local quadratic polynomials. We used fixed global bandwidths, approximately the global average of the optimal local bandwidths as depicted in Figure 2, leading to  $h = 7$  for the local linear and  $h = 13$  for the local quadratic fits. For Keiding's (1991) smoothed estimate we took  $h = 10$ . Using an optimal data-driven bandwidth for each method within each simulation run was computationally not feasible. Let  $\hat{\pi}_j(a)$  be the estimated probability at age  $a$  in the  $j$ th simulation,  $j = 1, 2, \dots, M$ . The local squared bias is estimated by

$\hat{b}^2(a) = \{\bar{\hat{\pi}}(a) - \pi(a)\}^2$ , with  $\bar{\hat{\pi}}(a) = \sum_{j=1}^M \hat{\pi}_j(a)/M$  and the local variance is estimated by  $\hat{v}(a) = \sum_{j=1}^M \{\hat{\pi}_j(a) - \bar{\hat{\pi}}(a)\}^2/M$ , leading to the simulation estimate for the local mean squared error  $MSE$ , given by  $\widehat{MSE}(a) = \hat{b}^2(a) + \hat{v}(a)$ .

#### FIGURE 6 AND 7 ABOUT HERE

Figure 6 shows that the three different mean curves  $\bar{\hat{\pi}}(a)$  can hardly be distinguished, except in the last 5 age groups (80-85) where the probabilities estimated by the isotonic regression increase. Between age 1 to 80, the local squared bias of the three models is essentially the same. The local variance of the isotonic regression is however much higher than the local variance of monotonized local polynomial models. Since the local variance is the dominant term in  $\widehat{MSE}(a)$ , the isotonic regression model has also higher values for  $\widehat{MSE}(a)$ . This pattern can also be seen in panel b, which shows the 5% and the 95% quantiles of the isotonic regression and the local polynomial models. The variability in the isotonic regression model is clearly higher than in the local polynomial models. Table 3 shows global simulated squared bias, variance and MSE, averaging over all age groups. The global MSE of the isotonic regression is 3.2 times higher than the global MSE of the local linear model and 4.1 times higher than the global MSE of the local quadratic mode. The results remain essentially the same for the trimmed (5%) means.

At each simulation the force of infection was estimated according to (4) for the local polynomial models and using a kernel smoother for the isotonic regression. Let  $\hat{\ell}_j(a)$  be the estimated force of infection at age  $a$  in the  $j$ th simulation,  $j = 1, 2, \dots, M$  and  $\bar{\hat{\ell}}(a) = \sum_{j=1}^M \hat{\ell}_j(a)/M$ . The local squared bias, variance and mean square error were calculated using  $\hat{b}^2(a) = \{\bar{\hat{\ell}}(a) - \ell(a)\}^2$ ,  $\hat{v}(a) = \sum_{j=1}^M \{\hat{\ell}_j(a) - \bar{\hat{\ell}}(a)\}^2/M$ .

The last row in Table 3 shows that the local quadratic model has the smallest global MSE  $,0.84 \times 10^{-4}$  compared to  $1.21 \times 10^{-4}$  and  $3.53 \times 10^{-4}$  for the local linear modal

and the isotonic regression respectively. Figure 7 displays the simulation results for  $\ell(a)$ . The variability of  $\ell(a)$  increases with age in all models but the local polynomial models have smaller square bias and variance than the isotonic regression model (locally and globally). Note that the pattern of increasing variability in panel b was already observed in the nonparametric bootstrap estimate for the confidence intervals of the force of infection (Figure 5).

### TABLE 3 ABOUT HERE

Although these results should be interpreted with some caution (no optimal bandwidths were used), there is a clear preference for the local polynomial models with some advantage for the local quadratic model.

## 5. Conclusion

We have suggested to model the force of infection for rubella, mumps and hepatitis A using the nonparametric technique of local polynomial estimation. Specification of a fully parametric model for the linear predictor will inevitably restrict the shape of the estimated force of infection. This is not always recognized as being possibly too restrictive. It is here where nonparametric methods can contribute to the analysis and, because they are fully unconstrained and highly data-driven, they may reveal aspects of the data which are ignored or hidden by parametric models. Local polynomial estimators are consistent without model assumptions (only require sufficient smoothness) and are known to have many desirable properties. This approach also allows simultaneous estimation of prevalence and force of infection. Asymptotic results for the local estimate of the force of infection were derived leading to a data-driven bandwidth selector. According to the principle “smooth

then constrain”, the fitted probability curve can be constrained to be monotone leading to a nonnegative estimated force of infection. Results from a small simulation study show that estimates obtained from monotonized local polynomials are less variable than those based on isotonic regression. As an overall conclusion we recommend, based on theoretical considerations and our findings in the data analysis and the small simulation study, the use of the local quadratic model to estimate the force of infection.

In further research it is examined how such a nonparametric smoothing method can be extended to estimate the force of infection of hepatitis A allowing time heterogeneity. Indeed, in contrast with rubella and mumps, prior to widespread vaccination an upward shift in the age at hepatitis A infection has been observed in industrialized countries following overall improvements in hygienic conditions in the second half of the 20th century. Also in Flanders a decrease in prevalence in the youngest age groups (0-14 years) has been observed by comparing the sample from 1993-94 with previous small samples obtained from Belgian first time blood donors in 1979 and 1989 (Beutels *et al.*, 1998). As a consequence of the age shift the assumption of time independence is likely to be violated in relation to HAV. We shall further study this issue when we analyze a new sample to be taken in 2001-2.

For rubella and mumps, the interpretation of the estimated force of infection might be more straightforward. Both are airborne infections which are usually acquired in childhood and the effects of time dependence are likely to be limited. Mumps is generally more infectious than rubella (Christie 1980), resulting in a greater force of infection at most ages. This very general observation can be made with all of the methods under discussion. However, a closer examination of the results using nonparametric methods, reveals that even for these airborne infections with simple general characteristics the parametric methods may

lead to overly rigid (and simplified) estimates for the force of infection. Indeed, the local polynomials yield a higher maximum and more fluctuation in the force of infection with regards to age, which we intuitively ascribe to different intensities of mixing in and between children and their parents. As indicated by figure 3, transmission in the first years of life occurs more frequently for rubella than for mumps, when the force of infection is also estimated higher. In contrast, mumps is known to infect susceptibles of all ages, but only very exceptionally during the first year of life. This last feature is also reflected in our nonparametric estimates for rubella (which has a single peak, but starts with a non-zero rate at the first age) and mumps (which starts at zero, but shows multiple peaks).

### Acknowledgments

We thank the associate editor and two referees for their valuable comments, which improved presentation of the paper substantially. The first three authors gratefully acknowledge support from the Belgian IUAP/PAI network “Statistical Techniques and Modeling for Complex Substantive Questions with Complex Data”. We gratefully acknowledge financial support from the Flemish Fund for Scientific Research (FWO-nr G.0023.01N).

### Appendix

The following modification of the delta method is used to derive the asymptotic normality result (6).

**Lemma** *Let  $\{\mathbf{T}_n\} = \{(T_{1n}, T_{2n})\}$  be a sequence of bivariate estimators for  $\boldsymbol{\theta} = (\theta_1, \theta_2)$  such*

that, as  $n \rightarrow \infty$

$$\begin{bmatrix} \sqrt{a_n}(T_{1n} - \theta_1) \\ \sqrt{b_n}(T_{2n} - \theta_2) \end{bmatrix} \xrightarrow{D} N(\mathbf{b}, V),$$

where  $a_n$  and  $b_n$  are sequences of constants tending to infinity,  $\mathbf{b} = (b_1, b_2)$  represents the asymptotic bias and  $V$  the asymptotic covariance matrix.

Consider a real-valued function  $\varphi(\mathbf{t}) = \varphi(t_1, t_2)$  such that  $(\partial\varphi/\partial t_1, \partial\varphi/\partial t_2)$  is non-null at  $\mathbf{t} = \boldsymbol{\theta}$  and continuous in a neighborhood of  $\boldsymbol{\theta}$ . If  $\delta_n = a_n/b_n \rightarrow \infty$ , then as  $n \rightarrow \infty$

$$\sqrt{b_n}\{\varphi(\mathbf{T}_n) - \varphi(\boldsymbol{\theta})\} \xrightarrow{D} N(\gamma, \tau^2), \quad (14)$$

where  $\gamma = \frac{\partial\varphi}{\partial t_2}(\boldsymbol{\theta})b_2$  and  $\tau^2 = \left(\frac{\partial\varphi}{\partial t_2}(\boldsymbol{\theta})\right)^2 V_{22}$ .

### Proof

We have that

$$\begin{aligned} & \sqrt{b_n}\{\varphi(\mathbf{T}_n) - \varphi(\boldsymbol{\theta})\} \\ &= \sqrt{\frac{a_n}{\delta_n}}(T_{1n} - \theta_1)\tilde{\varphi}_1(\mathbf{T}_n, \boldsymbol{\theta}) + \sqrt{b_n}(T_{2n} - \theta_2)\tilde{\varphi}_2(\mathbf{T}_n, \boldsymbol{\theta}) \end{aligned}$$

where

$$\begin{aligned} \tilde{\varphi}_1(\mathbf{T}_n, \boldsymbol{\theta}) &= \frac{\varphi(T_{1n}, \theta_2) - \varphi(\theta_1, \theta_2)}{(T_{1n} - \theta_1)}, \\ \tilde{\varphi}_2(\mathbf{T}_n, \boldsymbol{\theta}) &= \frac{\varphi(T_{1n}, T_{2n}) - \varphi(T_{1n}, \theta_2)}{(T_{2n} - \theta_2)}. \end{aligned}$$

Since  $\mathbf{T}_n \xrightarrow{P} \boldsymbol{\theta}$ , it follows that for  $i = 1, 2$

$$\tilde{\varphi}_i(\mathbf{T}_n, \boldsymbol{\theta}) \xrightarrow{P} \frac{\partial\varphi}{\partial t_i}(\boldsymbol{\theta}).$$

Applying Slutsky's theorem and the fact that  $\delta_n \rightarrow \infty$  together with  $\sqrt{a_n}(T_{1n} - \theta_1) = O_P(1)$ ,

we get

$$\sqrt{b_n}\{\varphi(\mathbf{T}_n) - \varphi(\boldsymbol{\theta})\} \xrightarrow{D} N(\gamma, \tau^2).$$

## References

- [1] Anderson, R.M. and May, R.M. (1991) *Infectious diseases of humans, dynamics and control*. New York: Oxford University Press Inc.
- [2] Barlow, R.E., Bartholomew, D.J., Bremner, M.J. and Brunk, H.D. (1972) *Statistical inference under order restriction*, New York: Wiley.
- [3] Becker, N.G. (1989) *Analysis of infectious disease data*. London: Chapman and Hall.
- [4] Beutels, M., Van Damme, P., Aelvoet, W., Desmyter, J., Dondeyne, F., Goilav, C., Mak, R., Muylle, L., Pierard, D., Stroobant, A., Van Loock, F., Waumans, P. and Vranckx, R. (1997) Prevalence of Hepatitis A, B and C in the Flemish Population. *Eur. J. Epidem.* , **13**, 275–280 .
- [5] Beutels, M., Van Damme, P., Vranckx, R. and Meheus, A. (1998) The shift in prevalence of hepatitis A immunity in Flanders, Belgium. *Acta Gastro-enterologica Belgica*, **61**, 4–7.
- [6] Christie A.B. (1980), *Infectious Diseases: Epidemiology and Clinical Practice*. Edinburgh London Melbourne New York: Churchill Livingstone.
- [7] Davison, A.C. and Hinkley, D.V. (1997) *Bootstrap Methods and Their Application*. Cambridge University Press.
- [8] Diamond, I.D. and McDonald, J.M. (1992) Analysis of current-status data. In *Demographic Application of Event History Analysis* (eds. J. Trussel, R. Hankinson and J. Tiltan), Ch. 12. Oxford University Press.
- [9] Fan, J. and Gijbels, I. (1996) *Local polynomial modeling and its application*. London: Chapman and Hall.

- [10] Fan, J., Heckman, N.E. and Wand, M.P. (1995) Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *J. Am. Statist. Assoc.*, **90**, 141–150.
- [11] Farrington, C.P. (1990) Modeling Forces of infection for measles, mumps and rubella. *Statist. Med.*, **9**, 953–967.
- [12] 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.
- [13] Friedman, J. and Tibshirani, R. (1984) The monotone smoothing of scatterplots. *Technometrics*, **26**, 243–247.
- [14] Grenfell, B.T and Anderson, R.M (1985) The estimation of age-related rates of infection from case notifications and serological data. *J. Hygiene*, **95**,(2), 419–436.
- [15] Griffiths, D. (1974) A catalytic model of infection for measles. *Appl. Statist.*, **23**, 330–339.
- [16] Hadler, S.C. (1991) Global impact of hepatitis A virus infection: changing patterns. In *Viral hepatitis and Liver Disease* (eds F.B. Hollinger, S.M. Lemon, H.S. Margolis), pp. 14–20. Baltimore: Williams & Wilkins.
- [17] Keiding, N. (1991) Age-specific incidence and prevalence: a statistical perspective. *J. R. Statist. Soc. A*, **154**, 371–412.
- [18] Mammen, E., Marron, J.S., Turlach, B.A. and Wand, M.P. (2001) A general framework for constrained smoothing. *Statistical Science*, **16**, 232–248.



- [19] Muench, H. (1959) *Catalytic models in epidemiology*. Boston: Harvard University Press.
- [20] Nagelkerke, N., Heisterkamp, S., Borgdorff, M., Broekmans, J. and Van Houwelingen, H. (1999) Semi-parametric estimation of age-time specific infection incidence from serial prevalence data. *Statist. Med.*, **18**, 307–320.
- [21] Shiboski, S.C. (1998) Generalized additive models for current status data. *Lifetime Data Analysis*, **4**, 29–50.
- [22] Shkedy, Z., Aerts, M., Molenberghs, G., Beutels, Ph. and Van Damme, P. (2002), Modeling Age Dependent Force of Infection From Prevalence Data Using Fractional Polynomials. Submitted.
- [23] Royston, P. and Altman, D.G. (1994) Regression using fractional polynomials of continuous covariates : parsimonious parametric modeling. *Appl. Statist.*, **43**, 429–467.

Table 1. General expressions for the force of infection according to different link functions.  $\Phi$  denotes the cumulative distribution function and  $\phi$  the density function of the standard normal distribution

Link function	$\pi(a)$	$\delta(\eta(a))$	local estimate for $\ell(a)$
log	$1 - e^{-\eta(a)}$	1	$\hat{\beta}_1(a)$
Complementary log-log	$1 - e^{-e^{\eta(a)}}$	$e^{\eta(a)}$	$\hat{\beta}_1(a)e^{\hat{\beta}_0(a)}$
logit	$\frac{e^{\eta(a)}}{1 + e^{\eta(a)}}$	$\frac{e^{\eta(a)}}{1 + e^{\eta(a)}}$	$\hat{\beta}_1(a) \frac{e^{\hat{\beta}_0(a)}}{1 + e^{\hat{\beta}_0(a)}}$
probit	$\Phi(\eta(a))$	$\frac{\phi(\eta(a))}{1 - \Phi(\eta(a))}$	$\hat{\beta}_1(a) \frac{\phi(\hat{\beta}_0(a))}{1 - \Phi(\hat{\beta}_0(a))}$

Table 2. Deviance and Gain values for first and second order monotone fractional polynomials with logit link function.

First order ( <b>m=1</b> )					Second order ( <b>m=2</b> )		
Dataset	df	Deviance	$p$	$G(1, p)$	df	Deviance	$p_1, p_2$
Hepatitis A	83	115.34	0.32	34.21	81	97.62	0.9, 1.4
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 3. Simulation results: global simulated squared bias, variance and mean squared error ( $\times 10^4$ ) for isotonic regression, monotonized local linear and local quadratic fits. The numbers in parenthesis are the trimmed means ( $trim = 5\%$ ).

		local linear	local quadratic	isotonic regression
$\pi(a)$	$\bar{b}^2$	0.40 (0.39)	0.02 (0.02)	0.72 (0.18)
	$\bar{v}$	2.63 (2.5)	2.34 (2.22)	9.01 (8.85)
	$\overline{MSE}$	3.04 (2.91)	2.36 (2.24)	9.73 (9.03)
$\ell(a)$	$\bar{b}^2$	0.28 (0.21)	0.06 (0.03)	1.16 (0.92)
	$\bar{v}$	0.94 (0.77)	0.77 (0.63)	2.37 (1.87)
	$\overline{MSE}$	1.21 (0.98)	0.84 (0.66)	3.53 (2.79)

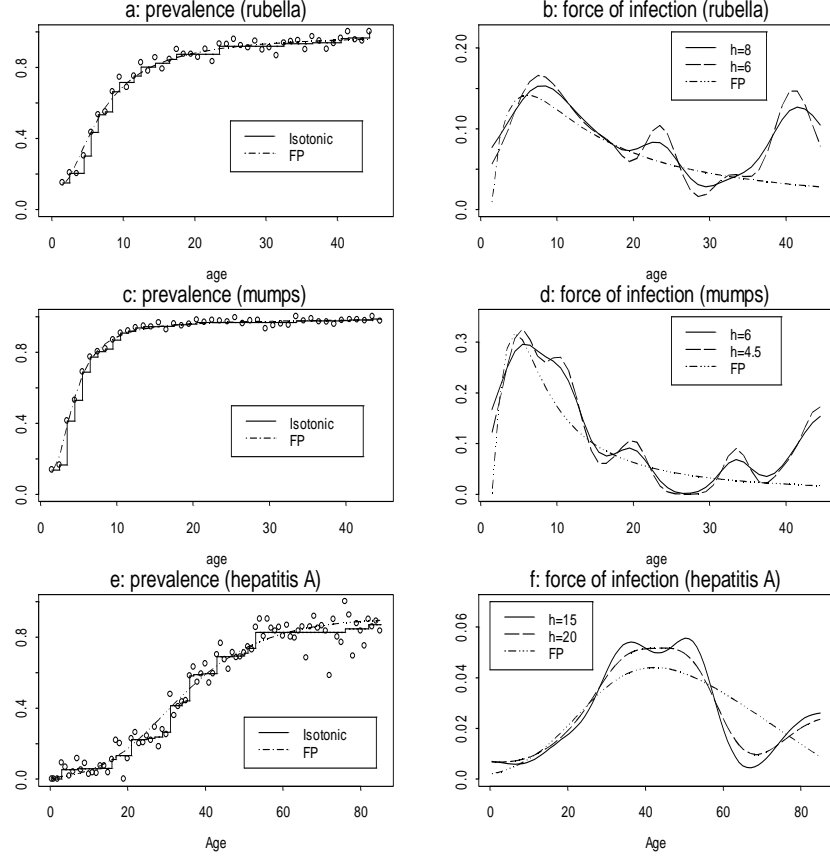


Fig. 1. Panels a, c and e. Rubella, mumps and HAV, data with estimated probability curve  $\hat{\pi}(a)$ : isotonic regression (solid line), optimal fractional polynomial (threedot-dash). Panels b, d and f. Estimated force-of-infection curve  $\hat{\ell}(a)$ : estimate based on optimal fractional polynomial (threedot-dash). Keiding's (1991) smoothed estimate using the standard normal density function as the kernel function and two bandwidths. The bandwidths for rubella are equal to 8 (solid line) and 6 (longdash line), for mumps  $h = 6$  (solid line) and  $h = 4.5$  (longdash line), and for HAV  $h = 15$  (solid line) and  $h = 20$  (longdash line).

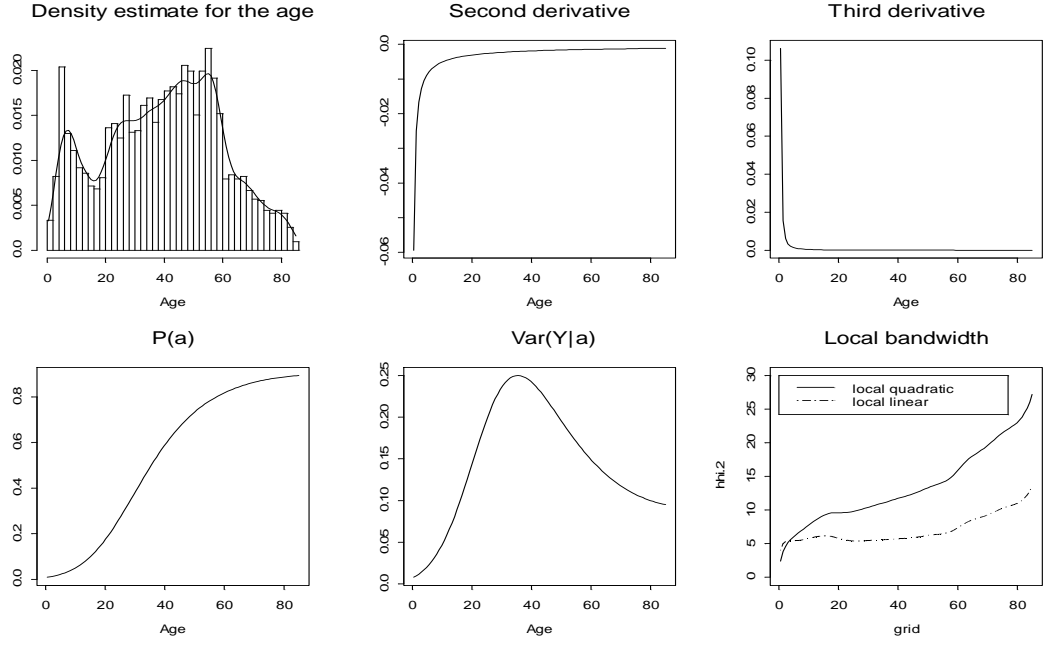


Fig. 2. Optimal local bandwidth for hepatitis A. From top left to bottom right: kernel density estimate for  $f_A(a)$ , estimates for  $\hat{\eta}^{(2)}(a), \hat{\eta}^{(3)}(a)$ ,  $\hat{\pi}(a)$ ,  $\widehat{\text{Var}}(Y|a) = \hat{\pi}(a)(1 - \hat{\pi}(a))$  based on the optimal fractional polynomial of order 2 and the corresponding optimal local linear (dashed line) and local quadratic (solid line) bandwidth estimates.

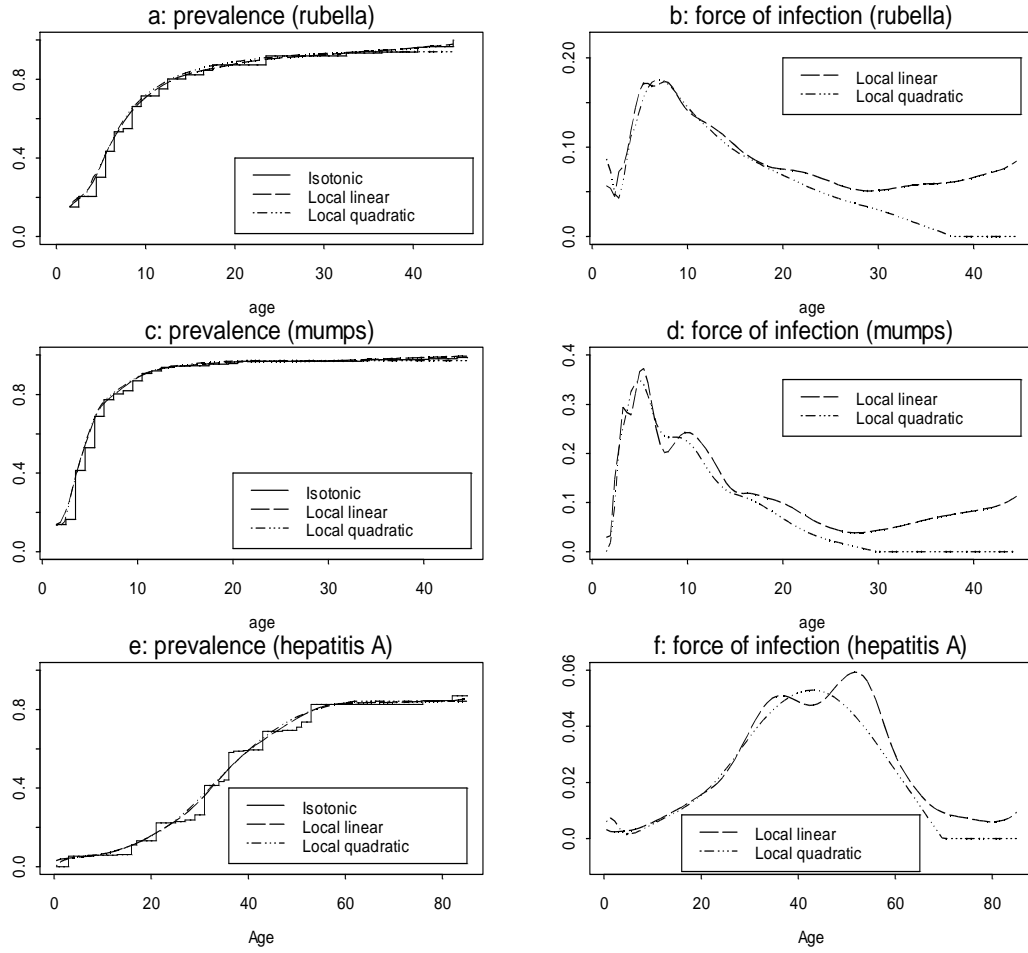


Fig. 3. Panels a, c and e. Estimated probability curve  $\hat{\pi}(a)$  for rubella, mumps and HAV: isotonic regression (solid line), local linear estimate (long dashed), local quadratic (threedot-dash). Panels b, d and f. Estimated force-of-infection curve  $\hat{\ell}(a)$ : constrained local quadratic estimate (threedot-dash) and local linear estimate (long dashed).

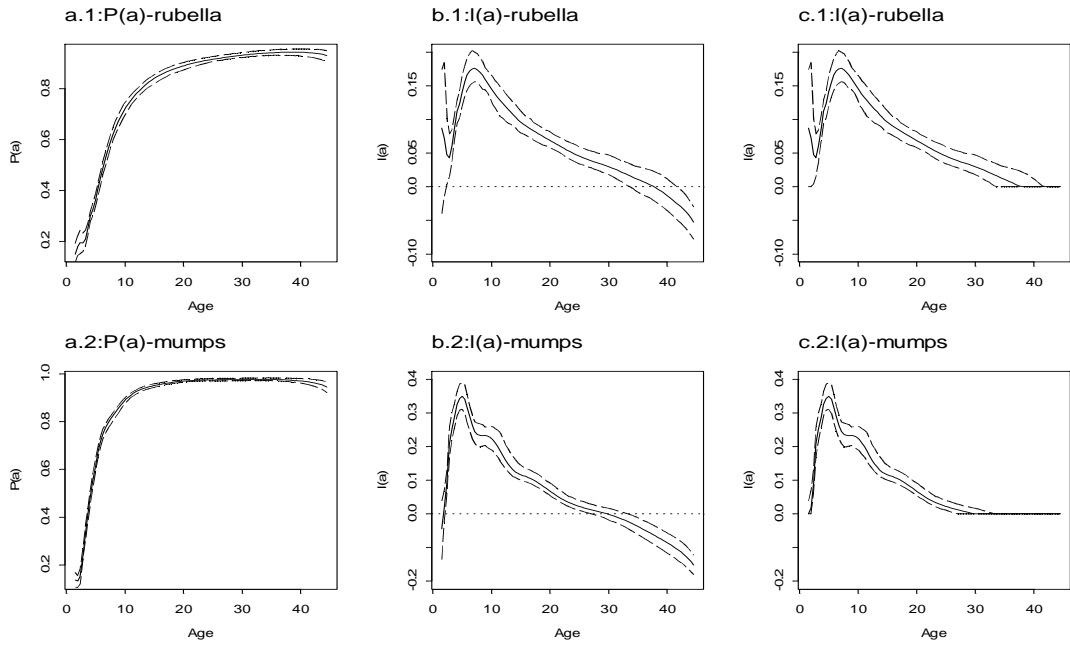


Fig. 4. Bootstrap confidence intervals for rubella (upper panels a1, b1, c1) and mumps (lower panels a2, b2, c2). From left to right: local quadratic estimates with confidence intervals for  $\pi(a)$ ,  $l(a)$  and constrained  $l(a)$ .



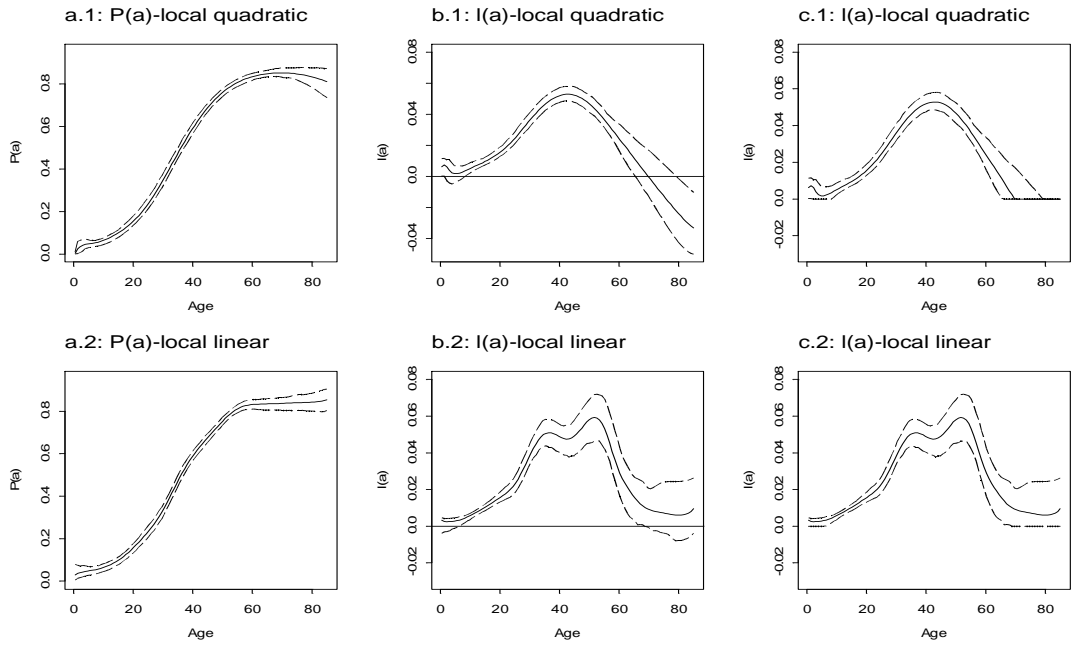


Fig. 5. Bootstrap confidence intervals for HAV. From left to right: local estimates with confidence intervals for  $\pi(a)$ ,  $\ell(a)$  and constrained  $\ell(a)$ . Upper panels a1, b1, c1: local quadratic polynomials. Lower panels a2, b2, c2: local linear polynomials.

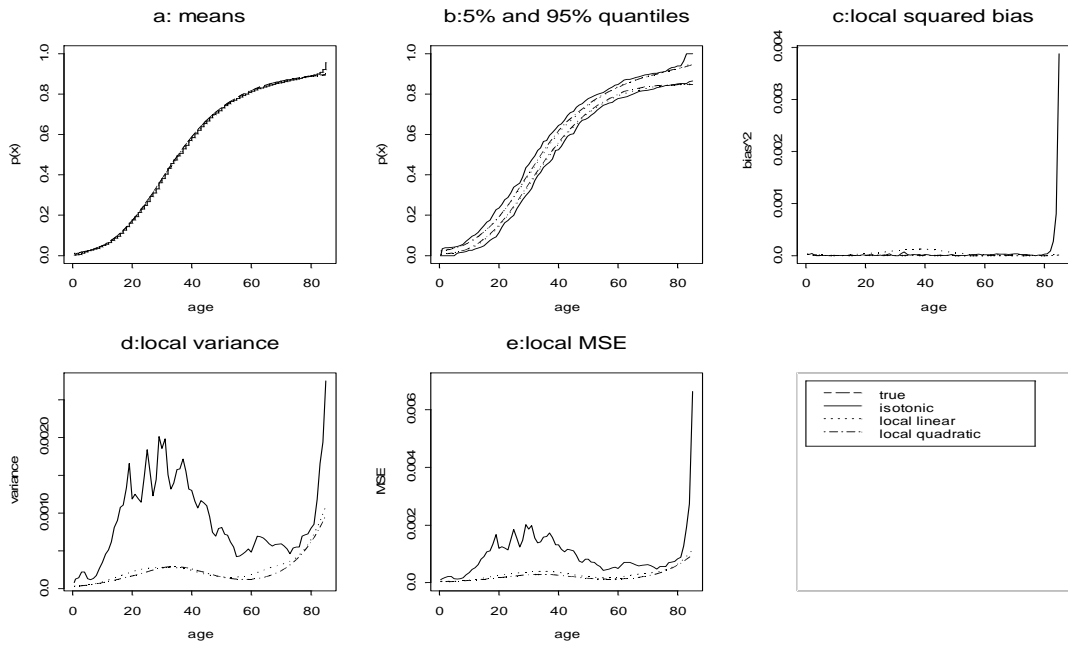


Fig. 6. Simulation results for test function  $\pi_A(a)$ , including isotonic regression (solid lines), mono-tonized local linear (dotted lines) and local quadratic (dot-dashed lines) estimates. From top left to bottom right: average probability estimates  $\hat{\pi}(a)$ , 5% and 95% quantiles, simulated squared bias  $b^2(a)$ , variance  $v(a)$  and mean squared error  $b^2(a) + v(a)$ .

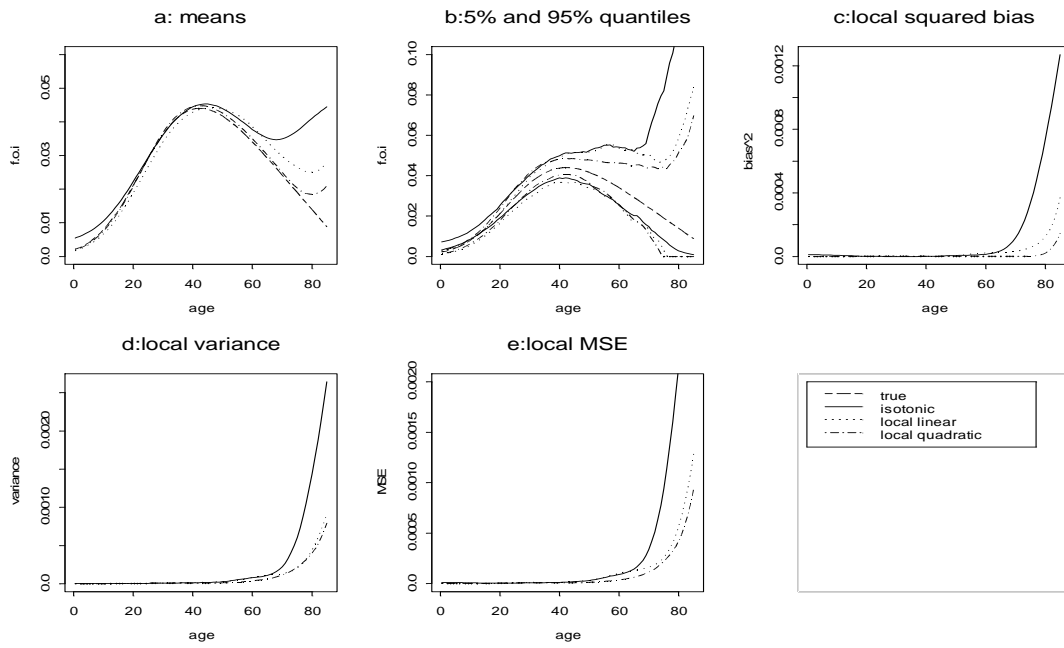


Fig. 7. Simulation results for force of infection  $\ell(a)$ , including estimates based on isotonic regression (solid lines), monotonized local linear (dotted lines) and local quadratic (dot-dashed lines) models. From top left to bottom right: average probability estimates  $\bar{\ell}(a)$ , 5% and 95% quantiles, simulated squared bias  $b^2(a)$ , variance  $v(a)$  and mean squared error  $b^2(a) + v(a)$ . Panels a and b also show the true force of infection  $\ell(a)$  (dashed lines).