Statistical Modelling 2012; **12**(**5**): 441–462

Estimating the population prevalence and force of infection directly from antibody titres

K Bollaerts[1], M Aerts[2], Z Shkedy[2], C Faes[2], Y Van der Stede[4], P Beutels[3] and N Hens[2,3]

[1]Scientific Institute of Public Health, Brussels, Belgium

[2]Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Hasselt University & Katholieke Universiteit Leuven, Diepenbeek, Belgium

[3]Centre for Health Economics Research and Modeling Infectious Diseases, Centre for the Evaluation of Vaccination, Vaccine & Infectious Disease Institute, University of Antwerp, Antwerp, Belgium

[4]Veterinary and Agrochemical Research Centre, Brussels, Belgium

Abstract: The use of threshold values in order to diagnose individual subjects as being 'susceptible' or 'infected or recovered/immune' for a specific infection is virtually always prone to false positive, false negative or inconclusive classifications. Such misclassifications might lead to biased estimates for epidemiological parameters, such as the prevalence and the force of infection. In this article, we propose to estimate these epidemiological parameters directly from antibody titres, using an underlying mixture model. The method is applied to estimate the *Salmonella* serological prevalence in pigs and the age-dependent force of infection using serological data on the Varicella-Zoster virus (VZV) in humans. The threshold and direct method are compared through a simulation study.

Key words: force of infection; mixture model; serological data; test misclassification; prevalence

Received December 2010; **revised** October 2011 & April 2012; **accepted** May 2012

1 Introduction

Epidemiology deals with the study of the occurrence of a disease and its determinants within a given population of humans or animals. A key characteristic of disease occurrence is the prevalence π, being the proportion of infected individuals within a population (here and throughout this article 'infected' refers to past or current infection). The dependency of the prevalence π on covariates is often of interest. For instance, to control and prevent an infectious disease, the dependency of the prevalence $\pi(t)$ on time t is crucial. In our first application we will estimate the time-dependent prevalence $\pi(t)$ of *Salmonella*, based on data from the Belgian *Salmonella* surveillance programme.

Address for correspondence: Marc Aerts, Interuniversity Institute for Statistics and Statistical Bioinformatics, Hasselt University, Agoralaan 1, 3590 Diepenbeek, Belgium. E-mail: marc.aerts@uhasselt.be

 10.1177/1471082X12457495

Another important epidemiological parameter is the force of infection λ (FOI, or infection hazard), which is the instantaneous rate at which susceptible individuals become infected, being typically investigated as a function of age and/or time (Anderson and May, 1991). More formally, denote the fraction of the individuals susceptible to a specific infectious disease at age a and time t as $q(a, t)$. Assuming that maternal antibodies are absent and that the disease is irreversible, meaning that immunity is lifelong and that mortality caused by infection is negligible, the partial differential equation which describes the change in the susceptible fraction at age a and time t is given by $\frac{\partial}{\partial a}q(a, t) + \frac{\partial}{\partial t}q(a, t) = -\lambda(a, t)q(a, t)$, with $\lambda(a, t)$ being the age- and time-specific FOI. In a steady state, assuming time homogeneity or $\frac{\partial}{\partial t}q(a, t) = 0$, this differential equation simplifies to $q'(a) = -\lambda(a)q(a)$, with now $q'(a)$ denoting the derivative and $\lambda(a)$ denoting the age-specific FOI. As the prevalence $\pi(a) = 1 - q(a)$, the following simple equation relates the FOI to the prevalence:

$$\lambda(a) = \frac{\pi'(a)}{1 - \pi(a)}. \tag{1.1}$$

So, having an estimator for the age-dependent prevalence $\pi(a)$, equation (1.1) can be easily applied to get an estimate for the FOI $\lambda(a)$. It is also important to mention that the prevalence $\pi(a)$ is non-decreasing as a function of the age a and that the estimator should be as well. Of course, the proportion susceptible $q(a)$ will decrease with age, and hence the prevalence $\pi(a)$ cannot decrease. In our second application we estimate the age-dependent prevalence $\pi(a)$ and FOI $\lambda(a)$ for the Varicella-Zoster virus (VZV), based on a Belgian serological study.

Although not equal, the prevalence $\pi(t)$ or $\pi(a)$ is commonly replaced by the so-called 'seroprevalence', i.e., the proportion of subjects at time t or of age a who test positive on a serological test. More precisely, a test is considered positive if the antibody titre measurement exceeds a threshold value. In the one-threshold case (Figure 1: left panel), individuals are diagnosed as infected if their test result exceeds a certain threshold value ζ and as being still susceptible otherwise. In case two threshold values are used (Figure 1: right panel), individuals having test results higher than the highest threshold value ζ_2 are diagnosed as infected, individuals having test results smaller than the lowest threshold value ζ_1 are diagnosed as susceptible, whereas all remaining individuals are labelled inconclusive.

However, the use of threshold value(s) in order to diagnose individual subjects is virtually always prone to misclassification, encompassing false negative results (infected subjects testing negative, 1-sensitivity), false positive results (susceptible subjects testing positive, 1-specificity) and inconclusive classifications (in case two thresholds are used), yielding biased estimates of the epidemiological parameters. In case the sensitivity and specificity characteristics of the applied test are known, the seroprevalence can be corrected for these misclassifications (see, e.g., Rogan and Gladen, 1978). But, in practice, such corrections are often not applied. Moreover, it is common practice just to discard the inconclusive individuals from any analysis.

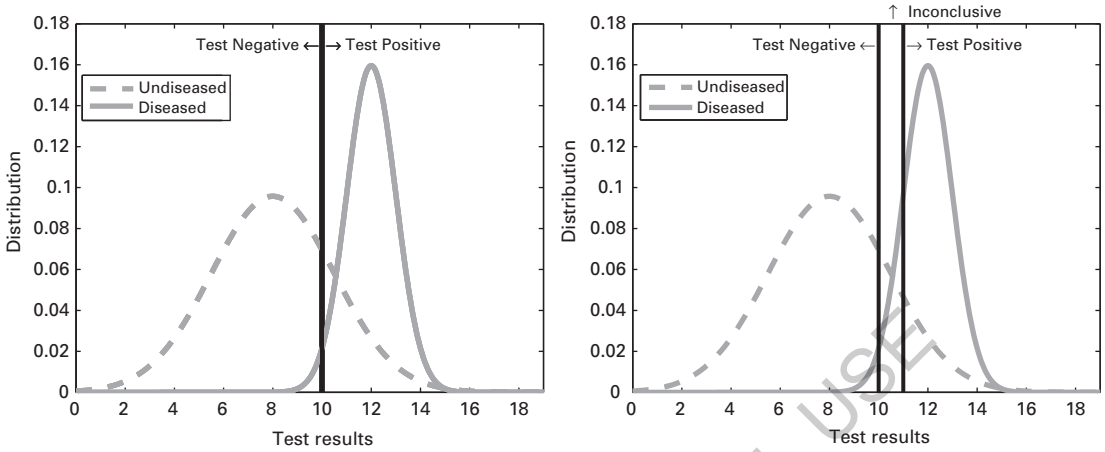

Figure 1 Distribution of fictive test results from a population of susceptible and infected subjects with subjects classified using one threshold value (left panel) and two threshold values (right panel). The assumed distribution of the susceptible subjects is $N(8, 2.5^2)$ and of the infected subjects is $N(12, 1)$.

In this article, we investigate the effect of test misclassification also on the estimation of the FOI and show that the optimal threshold minimizing the (asymptotic) bias appears to be different for both parameters, prevalence and FOI. To avoid the need for correcting for bias and for selecting different optimal thresholds, we propose to estimate the prevalence and the FOI 'directly' (without the use of any threshold) from the serological data. We will call these estimators the 'direct' estimators, and although these new estimators are also based on serological data (on the continuous scale), the term seroprevalence is reserved for the popular approach using the dichotomized data (on the binary scale, after comparing with a threshold). These latter estimators will be referred to as the 'threshold estimators'.

The direct estimators for the prevalence and the force of infection are derived from an underlying mixture model. Mixture models are well-established models nowadays, based on an extensive literature showing its applicability in a wide range of applications (see, e.g., Titterington *et al.*, 1985; Böhning *et al.*, 1992; Schlattmann, 2009). In our definition and application of the direct estimators, the mixture model itself is not of main interest, but its (latent) structure allows a direct definition of prevalence and FOI. All components of the direct estimators can be implemented fully parametrically, semi-parametrically or even fully non-parametrically.

The relation between the direct estimator and the threshold estimator will be elucidated in Section 2. The methodology is illustrated in Section 3 by estimating the time-dependent prevalence $\pi(t)$ of *Salmonella* in pigs and by estimating the age-dependent prevalence $\pi(a)$ and FOI $\lambda(a)$ of VZV in humans. In both applications the estimation is based on serological test results using ELISA. A Monte Carlo simulation study comparing both methods (direct and threshold estimators) is conducted as well in Section 4.

2 Methodology

In this methodological section, we first briefly review the threshold approach. Next, the direct approach is introduced showing how to estimate the prevalence and the FOI directly from the serological test results on the continuous scale (such as the antibody titre measurements), avoiding any use of thresholds. Finally, the connection between both approaches is discussed. For simplicity all parameters of interest are formulated as age dependent in this section. But everything also applies on time dependent parameters as illustrated in the application in Section 3.1.

2.1 Threshold approach

Consider a serological sample of n subjects and let y_i be the (continuous) test result of subject i of age a_i ($i = 1, 2, \ldots, n$) and let ζ be a single threshold value. Then, assuming infected individuals have test results on average at the higher end of the scale, y_i is dichotomized as

$$z_i = \begin{cases} 0, & y_i \leq \zeta \\ 1, & y_i > \zeta, \end{cases} \tag{2.1}$$

where '0' denotes test-negative and '1' test-positive. The accuracy of a diagnostic test with threshold value ζ is typically quantified by the test sensitivity SE (the probability of a positive test result given the individual has been infected) and the test specificity SP (the probability of a negative test result given the individual is still susceptible). Note that, of course, SE and SP depend on the threshold: a higher value for ζ implies a lower value for SE and a higher value for SP. For a test to be valid, it is required that $\text{SE} > 1 - \text{SP}$. This condition comes down to requiring that the probability of a test-positive result is larger for an infected individual than for a susceptible individual.

Unless the test result is perfect (SE = SP = 1) the age-dependent seroprevalence $\pi_z(a) = P(z = 1|a)$ is not equal to the unknown age-dependent prevalence $\pi(a)$. Indeed the following relationship holds

$$\pi_z(a) - \pi(a) = (1 - \pi(a))(1 - \text{SP}) - \pi(a)(1 - \text{SE}), \tag{2.2}$$

and consequently an (asymptotically) unbiased estimator for $\pi_z(a)$ is not an (asymptotically) unbiased estimate for the true prevalence $\pi(a)$. The threshold estimator $\widehat{\pi}_z(a)$ is typically obtained from a logistic regression model.

Clearly, the bias depends on both the threshold ζ (determining SE and SP) and $\pi(a)$ and can be positive as well as negative. Replacing $\pi(a)$ by $\pi_z(a)$ in identity (1.1),

$$\lambda_z(a) = \frac{\pi_z'(a)}{1 - \pi_z(a)}, \tag{2.3}$$

and plugging in an unbiased estimator for $\pi_z(a)$, will of course also lead to a biased threshold estimator for the true FOI $\lambda(a)$, with (asymptotic) bias equal to

$$\lambda_z(a) - \lambda(a) = -\left(\frac{1 - \text{SE}}{(1 - \pi(a))\text{SP} + \pi(a)(1 - \text{SE})}\right)\lambda(a). \tag{2.4}$$

It is interesting to note that this bias is always negative, and increasing in magnitude as a function of the true FOI $\lambda(a)$ and decreasing as a function of the sensitivity SE.

To illustrate the magnitude and direction of the (age-dependent) bias given in (2.2) and (2.4), we calculated $\pi_z(a)$ and $\lambda_z(a)$ for different choices $\{2, 3, \ldots, 18\}$ for the threshold ζ, assuming that the non-dichotomized test results y of the susceptible individuals are distributed as $N(8, 2.5^2)$ and of the infected individuals as $N(12, 1^2)$ (see also Figure 1) and for a true age-dependent prevalence $\pi(a)$ as shown by the solid black curve in the left upper panel of Figure 2.

The results for $\pi_z(a)$ (respectively, $\lambda_z(a)$) are given in the left upper (respectively, lower) panel of Figure 2. To visualize the bias, $\pi(a)$ and $\lambda(a)$ are graphically presented as well. In addition, mean absolute errors (MAE) are calculated for several threshold choices ζ. In particular, $\text{MAE}(\pi_z) = \frac{1}{G}\sum_{g=1}^{G}|\pi_z(a_g) - \pi(a_g)|$ and $\text{MAE}(\lambda_z) = \frac{1}{G}\sum_{g=1}^{G}|\lambda_z(a_g) - \lambda(a_g)|$ are calculated for an equally spaced grid on age between minimum and maximum age with $G = 500$ points on the grid. The results for $\pi_z(a)$ (respectively, $\lambda_z(a)$) are given in the right upper (respectively, lower) panel of Figure 2 with the vertical dashed lines representing the means of the susceptible and infected population. As can be seen, the $\text{MAE}(\lambda_z)$ is minimized for threshold values much smaller than the one minimizing $\text{MAE}(\pi_z)$, clearly illustrating that the choice of threshold might be problematic when interest is in estimating (different) epidemiological parameters.

A bias correction to obtain an (asymptotically) unbiased estimate of $\pi(a)$ when using seroprevalence data was provided by Rogan and Gladen (1978). Starting from rewriting (2.2) as

$$\pi(a) = \frac{\pi_z(a) + \text{SP} - 1}{\text{SE} + \text{SP} - 1}, \tag{2.5}$$

it immediately follows that an asymptotically unbiased estimator of $\pi(a)$ is given by the so-called Rogan–Gladen estimator or

$$\widehat{\pi}_{RG}(a) = \frac{\widehat{\pi}_z(a) + \widehat{\text{SP}} - 1}{\widehat{\text{SE}} + \widehat{\text{SP}} - 1}, \tag{2.6}$$

given asymptotically unbiased estimates $\widehat{\pi}_z(a)$, $\widehat{\text{SP}}$ and $\widehat{\text{SE}}$ for $\pi_z(a)$, SP and SE. To make sure that the Rogan–Gladen is well defined with values between 0 and 1, definition (2.6) can be modified as $\widehat{\pi}_{RG}(a) = \max\{0, \min\{(\widehat{\pi}_z(a) + \widehat{\text{SP}} - 1)/(\widehat{\text{SE}} + \widehat{\text{SP}} - 1), 1\}\}$.

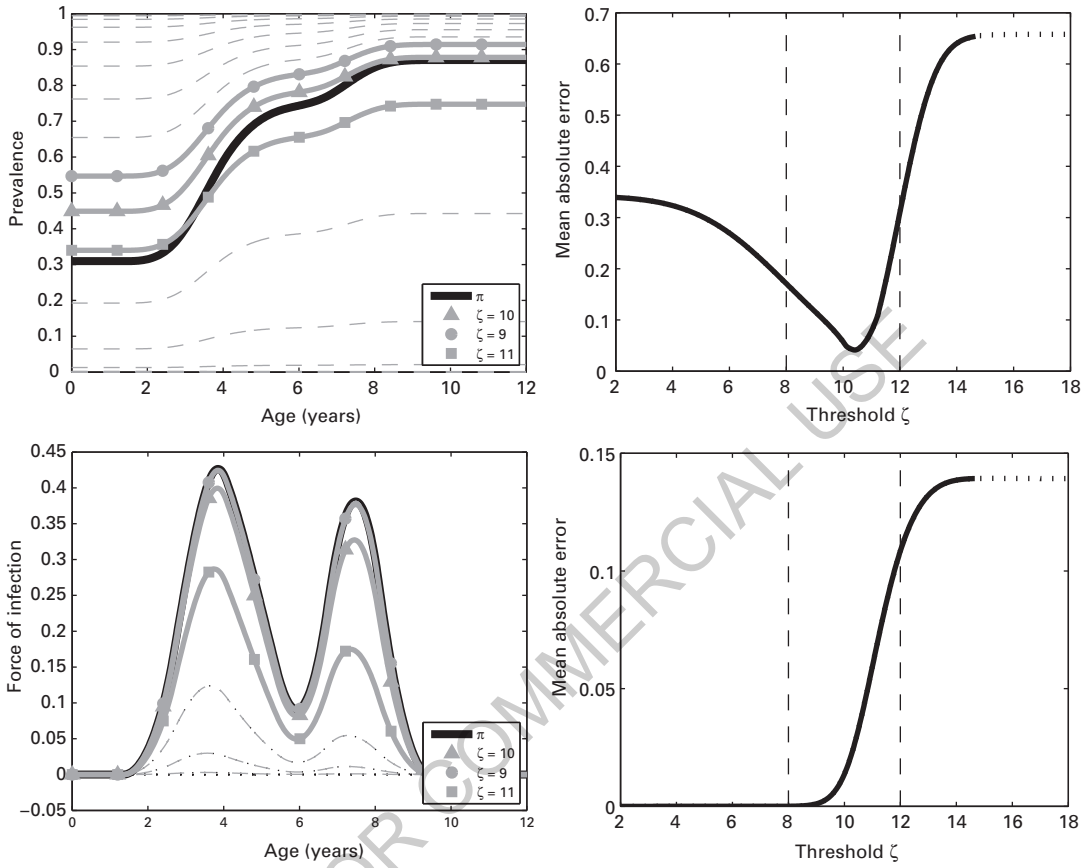

Figure 2 Age-dependent prevalence (upper left panel) and FOI (lower left panel) analytically derived for different choices of threshold values ζ and the corresponding mean absolute errors (right panels). In the right panels, results for tests that are not valid are indicated with dotted lines. Assumed distribution susceptible subjects: $N(8, 2.5^2)$; infected subjects: $N(12, 1)$.

An expression of the FOI $\lambda(a)$ corrected for test misclassification can be obtained by substituting (2.5) in (1.1), leading to

$$\lambda(a) = \frac{\pi_z^{'}(a)}{\mathrm{SE} - \pi_z(a)}. \tag{2.7}$$

A nice feature of expression (2.7) is its independence of the test specificity SP. Furthermore, since $\mathrm{SE} \leq 1$, expression (2.7) readily explains the direction of the asymptotic bias given in (2.4). Indeed, if $\mathrm{SE} < 1$, then $\lambda_z(a) < \lambda(a)$. Only in case $\mathrm{SE} = 1$, we have $\lambda_z(a) = \lambda(a)$. Starting from (2.7), an estimator $\widehat{\lambda}_{RG}(a)$ of the FOI corrected for test

misclassification is naturally obtained by plugging in the estimators $\widehat{\pi}_z(a)$, $\widehat{\pi}_z'(a)$ and $\widehat{\text{SE}}$.

Analogously to the one-threshold case, we also investigated the two-threshold case. In particular, we investigated the effect of test misclassification and discarded inconclusive classifications on the estimation of $\pi(a)$ and $\lambda(a)$. The results (not shown) are in line with the ones obtained in the one-threshold case, again indicating that the choice of thresholds is problematic when interest is in estimating (different) epidemiological parameters. Furthermore, the bias corrections are very complicated expressions and require the estimation of the probability of a test result to be discarded given the individual is infected (respectively, susceptible). Whereas estimates of SE and SP are sometimes provided with the test kit or obtained from previous studies or expert opinion, estimates for the probabilities to be discarded are generally lacking.

2.2 Direct approach

As an alternative to the seroprevalence based estimators reviewed in the previous section, one might consider estimating epidemiological parameters directly from the non-dichotomized serological data y. Such an approach has the major advantage that no threshold is needed, nor estimates for SE and SP in order to correct for bias. Also no inconclusive observations are discarded.

In particular, $\pi(a)$ and $\lambda(a)$ can be derived directly from a two-component mixture model assuming an age-dependent mixing probability. Formally, the two-component mixture model for a non-dichotomized test result y for an individual of age a can be represented as

$$f(y|a) = (1 - \pi(a))\ f_S(y;\theta_S) + \pi(a)\, f_I(y;\theta_I), \tag{2.8}$$

with f_S and f_I the densities referring to the susceptible and infected subpopulations, possibly depending on parameters θ_S and θ_I, respectively, and with $\pi(a)$ being the age-dependent mixing probability (i.e., the true prevalence).

The mixture (2.8) and the estimation of both components f_S and f_I is of no direct interest, but the underlying mixture structure allows us to relate the mean $E(y|a)$ to the prevalence $\pi(a)$. Indeed, it immediately follows that the mean $E(y|a)$ of y given age a equals

$$\mu(a) = (1 - \pi(a))\mu_S + \pi(a)\mu_I, \tag{2.9}$$

with μ_S and μ_I the mean test results for susceptible and infected subpopulations respectively. To guarantee identifiability, we assume $\mu_I > \mu_S$ (which corresponds to the requirement that SE $>$ 1 $-$ SP for a test to be valid).

An expression for the true prevalence $\pi(a)$ is naturally obtained by rewriting (2.9) as

$$\pi(a) = \frac{\mu(a) - \mu_S}{\mu_I - \mu_S}, \tag{2.10}$$

showing that $\pi(a)$ is the excess of the mixture mean $\mu(a)$ to the mean of the susceptible population μ_S relative to the difference in means of both subpopulations. It follows immediately that $\pi(a) = 0$ if $\mu(a) = \mu_S$ and that $\pi(a) = 1$ if $\mu(a) = \mu_I$.

Given asymptotically unbiased estimates $\widehat{\mu}(a)$, $\widehat{\mu}_S$ and $\widehat{\mu}_I$, an asymptotically unbiased direct estimator for $\pi(a)$ is given by

$$\widehat{\pi}(a) = \frac{\widehat{\mu}(a) - \widehat{\mu}_S}{\widehat{\mu}_I - \widehat{\mu}_S}. \tag{2.11}$$

To estimate μ_S and μ_I (with $\mu_S < \mu_I$) all data are collapsed over the age dimension, and a two-component mixture model can be applied. The estimator $\widehat{\mu}(a)$ is just a simple regression estimator, with the constraint that the estimator is monotone non-decreasing as a function of a and such that that $\widehat{\mu}_S \leq \widehat{\mu}(a) \leq \widehat{\mu}_I$.

An expression for the age-dependent FOI $\lambda(a)$ is obtained by substituting (2.10) in (1.1) and subsequently simplifying the expression, yielding

$$\lambda(a) = \frac{\mu'(a)}{\mu_I - \mu(a)}. \tag{2.12}$$

Expression (2.12) shows that the force of infection equals the derivative of the mixture mean $\mu'(a)$ relative to the difference between the mean of the infected population μ_I and the mixture mean $\mu(a)$. Note that $\mu(a) \leq \mu_I$ with $\mu(a) = \mu_I$ if the whole population is infected at age a (from which point on the FOI is no longer useful and not defined). As can be seen in (2.12), the force of infection depends on the mean of the infected population μ_I only and not on the mean of the susceptible population μ_S. Using identity (2.12), an asymptotically unbiased direct estimator for $\lambda(a)$ is given by

$$\widehat{\lambda}(a) = \frac{\widehat{\mu}'(a)}{\widehat{\mu}_I - \widehat{\mu}(a)}, \tag{2.13}$$

using (asymptotically) unbiased estimates $\widehat{\mu}(a)$, $\widehat{\mu}'(a)$ and $\widehat{\mu}_I$.

2.3 Distributional properties of the direct estimator

The regression function $\mu(a)$ can be estimated in a fully parametric, semi-parametric or fully non-parametric way. In Sections 3 and 4 we will use monotone penalized splines to estimate $\mu(a)$ in a flexible way (see, e.g., Eilers and Marx, 1996). In principle it should be possible to derive the asymptotic variance and asymptotic distribution of the spline-based estimators (2.11) and (2.13). But, although penalized splines have gained much popularity over the last decade, their asymptotic properties have been little explored until recently. Claeskens *et al.* (2009) present the first general treatment of the asymptotic properties of penalized splines. Depending on an assumption on the number of knots, sample size and penalty, they show that the theoretical properties of penalized spline estimators are either similar to those of regression splines or to those of smoothing splines, with a clear breakpoint distinguishing the cases. They

also obtain expressions for bias and variance. But in our case we fit penalized splines constrained to be monotone, a condition which further complicates the derivation of the asymptotic distribution.

Next, we need the distributional properties of the estimator $\widehat{\pi}(a)$ which equals $(\widehat{\mu}(a) - \mu_S)/(\mu_I - \mu_S)$ with μ_S and μ_I replaced by their respective estimators. These estimates are obtained in a separate step, based on a mixture model, but using the same data as for estimating $\widehat{\mu}(a)$. This brings us to the theory of, e.g., Randles (1982), who describes in which way the limiting distribution is affected when unknown parameters (such as μ_S and μ_I) are replaced by their respective estimators. As the FOI combines the estimator for $\mu(a)$ (in the denominator) with the estimator for the derivative $\mu'(a)$ (in the numerator), results of Claeskens *et al.* (2009) have also to be extended to the joint distribution of $(\widehat{\mu}(a), \widehat{\mu}'(a))$.

Deriving asymptotic results taking all these necessary extensions of existing theory into account would be very interesting, but is beyond the scope of this article. Moreover we believe that even if the asymptotic results would be available, they would require the non-trivial estimation of unknown parameters. We think that the bootstrap as presented in Section 3 is a very practical alternative. Although it is quite computationally intensive, its application and implementation is rather straightforward. Nowadays, in situations as the current one, the application of the bootstrap is forming less a burden than the application of asymptotic analytic expressions.

2.4 Monotone transformations and connection between threshold and direct approaches

The seroprevalence $\pi_z(a) = P(z = 1|a)$ is invariant for non-decreasing transformations of the quantitative test result y and corresponding threshold ζ. Indeed $\{z = 1\} = \{y \leq \zeta\} = \{g(y) \leq g(\zeta)\}$ for any non-decreasing function g. The application of a transformation g, however, might affect the direct approach (2.10) as it depends on means which are not invariant (e.g., $E(g(y)|a) \neq g(E(y|a))$). This gives rise to the natural question which transformation is optimal for the direct method. This optimization problem can be addressed from a theoretical point of view or rather from practical considerations. The theoretical optimization in terms of efficiency seems far from trivial and is beyond the scope of this article. A practical solution, however, is to select g in order to optimize the estimation of the mixture model and consequently of the means μ_S and μ_I. A typical transformation, for instance, would be the log-transformation as the quantitative test results are often skewed to the right.

One particular non-decreasing but non-smooth transformation g reveals the relation between the threshold and the direct method. Indeed, the 'threshold' transformation defined as

$$g(y) = I(y > \zeta) = \begin{cases} 0 \text{ if } & y \leq \zeta \\ 1 \text{ if } & y > \zeta \end{cases} \tag{2.14}$$

turns the mixture in a mixture of two binary components with age-dependent mean $\mu_g(a) = E[g(y)|a]$ equal to $\pi_z(a)$ and mixture identity (2.9) translates to

$$\begin{aligned}\pi_z(a) &= (1 - \pi(a)) \int_\zeta^\infty f_S(y;\theta_S)dy + \pi(a) \int_\zeta^\infty f_I(y;\theta_I)dy \\ &= (1 - \pi(a))\,(1 - \mathrm{SP}) + \pi(a)\mathrm{SE}\end{aligned}$$

and consequently for this specific transformation

$$\pi(a) = \frac{\pi_z(a) + \mathrm{SP} - 1}{\mathrm{SE} + \mathrm{SP} - 1}, \tag{2.15}$$

which is exactly leading to the Rogan–Gladen estimator (2.6). In a similar way, the FOI corresponding to this particular transformation equals the corrected threshold formula (2.7). Thus an interesting conclusion is that the misclassification-corrected expressions (2.5) and (2.7) are in fact a special case of the expressions based on the direct approach (2.10) and (2.12).

2.5 Estimation and inference

In the direct approach, the prevalence and FOI are estimated as a combination of two separate estimation steps, both using the data $\{(a_i, y_i), i = 1, \ldots, n\}$: estimation of the mean $\mu(a) = E(y|a)$ and its derivative $\mu'(a)$, on the one hand, and estimation of μ_I and μ_S, on the other hand. The estimates $\widehat{\mu}(a)$, $\widehat{\mu}'(a)$, $\widehat{\mu}_I$ and $\widehat{\mu}_S$ are then combined by identities (2.11) and (2.13) to get the estimates $\widehat{\pi}(a)$ and $\widehat{\lambda}(a)$.

Estimation of $\mu(a)$ and $\mu'(a)$

This part refers to a typical regression model with one covariate, as well as its derivative. As is the case in any regression model, there are several options depending on the application at hand. One can use a classical parametric linear model or a more flexible parametric model, such as a fractional polynomial or a nonlinear model. But one could also prefer a non- or semiparametric regression model, such as a local polynomial model or a spline model. In the applications and in the simulations we used penalized splines as we prefer a very flexible model for our applications.

Estimation of μ_I and μ_S

In the first application (*Salmonella*) a parametric gamma mixture model is fitted, and for computational reasons the estimation is based on MCMC rather than on the EM-algorithm (Dempster *et al.*, 1977). In the second application (Varicella-Zoster) a normal mixture is fitted with the EM-algorithm.

Inference

As explained in Section 2.3 the derivation of the distributional properties is not straightforward because of the separate estimation of regression mean and mixture means, because of the use of P-splines with data driven smoothness penalty and other complexities such as a complex survey design in the first application and monotonicity in the second application. Therefore bootstrap standard errors and bootstrap percentile intervals are reported, based on a bootstrap algorithm which allows us to take all complexities into account as well as all sources of variability (such as the data driven smoothness penalty). In principle one could also compute improved percentile intervals such as the bias-corrected and accelerated intervals, but the implementation of these improved intervals dealing with the above-mentioned issues is not straightforward.

3 Applications

The first application addresses the estimation of the time-dependent prevalence $\pi(t)$ of *Salmonella*-infected pigs at primary production in Belgium using both the threshold based and the direct approach. To allow flexible time trends, the regression mean $\mu(t)$ is estimated using splines. The mixture component means μ_S and μ_I are estimated by fitting a marginal two-component gamma mixture density.

The second application addresses the estimation of the age-dependent prevalence $\pi(a)$ and FOI $\lambda(a)$ using serological data on the VZV in humans. In this case a normal mixture was used to estimate μ_S and μ_I.

All methods were implemented in MatLab R2009a (code is available upon request).

3.1 *Salmonella* in pigs

Consumption of food that is contaminated with the *Salmonella* bacteria can cause human salmonellosis, which is a common gastrointestinal zoonotic disease worldwide. Mostly, salmonellosis is self-limiting, but it can evolve into serious illness or cause death especially within enhanced susceptible persons (e.g., children, elderly, pregnant women and immuno-compromised persons). The data come from the Belgian *Salmonella* surveillance programme launched in January 2005 (Van der Stede *et al.*, 2008). Following this programme, all professional pig herds are obliged to collect 10 to 12 blood samples from pigs every 3 to 4 months a year. The blood samples are tested for *Salmonella*-specific antibody levels using indirect ELISA following the test manufacturer's guidelines. The test results are reported as sample-to-positive ratios (SP-ratios), which are extremely positively skewed and normally range between 0 and 4. For the current application, we use only data on 314 herds for the year 2005, previously used to develop methods to identify *Salmonella* risk farms (Bollaerts *et al.*, 2008; Cortiñas Abrahantes *et al.*, 2009). In this first example, the age dependency of the prevalence is replaced by time dependency.

We first focus on the estimation of the regression mean $\mu(t)$, based on the transformed SP-ratios, log(SP+1). It is standard to estimate $\mu(t)$ using least-squares techniques, parametrically or non-parametrically, depending on the application of interest. To flexibly model seasonal trends, we opt to use linear P-splines regression (Eilers and Marx, 1996), which is essentially regression using an excessive number of equally spaced B-splines (Dierckx, 1993) and an additional discrete penalty to correct for overfitting. In particular, we use a regression basis of 15 equally-spaced B-splines between minimum and maximum sampling time. The degree of B-splines is chosen to be $d = 3$ because of its good trade-off between model flexibility, model smoothness and complexity. A second order smoothness penalty is used and the optimal values for the smoothness parameter are determined using cross-validation (CV) with the candidate values chosen from an approximate geometric grid $\{0.001, 0.01, 0.1, 0.5, 1, 5, 10, 50, 100, 500\}$. The optimal smoothness parameter is found to be 1. The obtained estimate $\widehat{\mu}(t)$ is graphically presented in Figure 3a. To estimate the mixing component means μ_S and μ_I, a two-component gamma mixture model is fitted to log(SP+1). For the current application gamma mixtures are an attractive choice since they allow skewly distributed mixing components. The gamma mixture is fitted using the MCMC-algorithm as outlined by Wiper *et al.* (2001) and using the following non-informative priors: a beta prior beta(1,1) for the mixing proportion, an inverted gamma distribution GI(1,1) for the (gamma) means and an exponential distribution E(0.01) for the (gamma) scales. Other mixtures, e.g., of log-normals, are natural candidates too, but the gamma mixture is fitting best. Three chains of each 50 000 samples are run, of which the first 30 000 are discarded as burn-in. The obtained gamma mixture density estimate is given in Figure 3b and shows a very good fit to the data. The estimates $\widehat{\mu}(t)$, $\widehat{\mu}_S$ and $\widehat{\mu}_I$ are combined as in (2.11) to obtain the direct estimate $\widehat{\pi}(t)$, which is graphically represented by means of the solid line in Figure 3d.

To compare, we also estimate $\pi(t)$ from SP-ratios dichotomized using the commonly used threshold values $\zeta = 0.50$ and $\zeta = 1.00$. Logistic regression is used to estimate the seroprevalences $\pi_z(t)$ for both threshold values. Similar as before, we used P-splines regression (Eilers and Marx, 1996), using a basis of 15 B-splines of degree $d = 3$ and a second order smoothness penalty of which the smoothness parameter is determined using cross-validation. The obtained optimal smoothness parameters are 0.01 for $\zeta = 0.50$ and 0.1 for $\zeta = 1.00$ and a graphical representation of $\widehat{\pi}_z(t)$ is given in Figure 3c for both threshold values. Clearly, the choice of threshold affects the results with lower thresholds yielding higher apparent prevalences. Estimates of the test characteristics (subscripts denoting the threshold) $SE_{0.50}$, $SP_{0.50}$, $SE_{1.00}$ and $SP_{1.00}$ can be derived from the obtained gamma mixture density estimate given in Figure 3b. The misclassification-corrected estimates (2.6) are graphically represented by the dashed and dotted line in Figure 3d as well. As can be seen, the three different estimates of $\pi(t)$ are almost perfectly overlapping and seasonal effects are observed with peaks during the summer months.

Finally, to estimate the variability in the estimated curves, generically denoted as $f(t, \widehat{\theta})$, the two-stage bootstrap procedure involving resampling 'populations' and

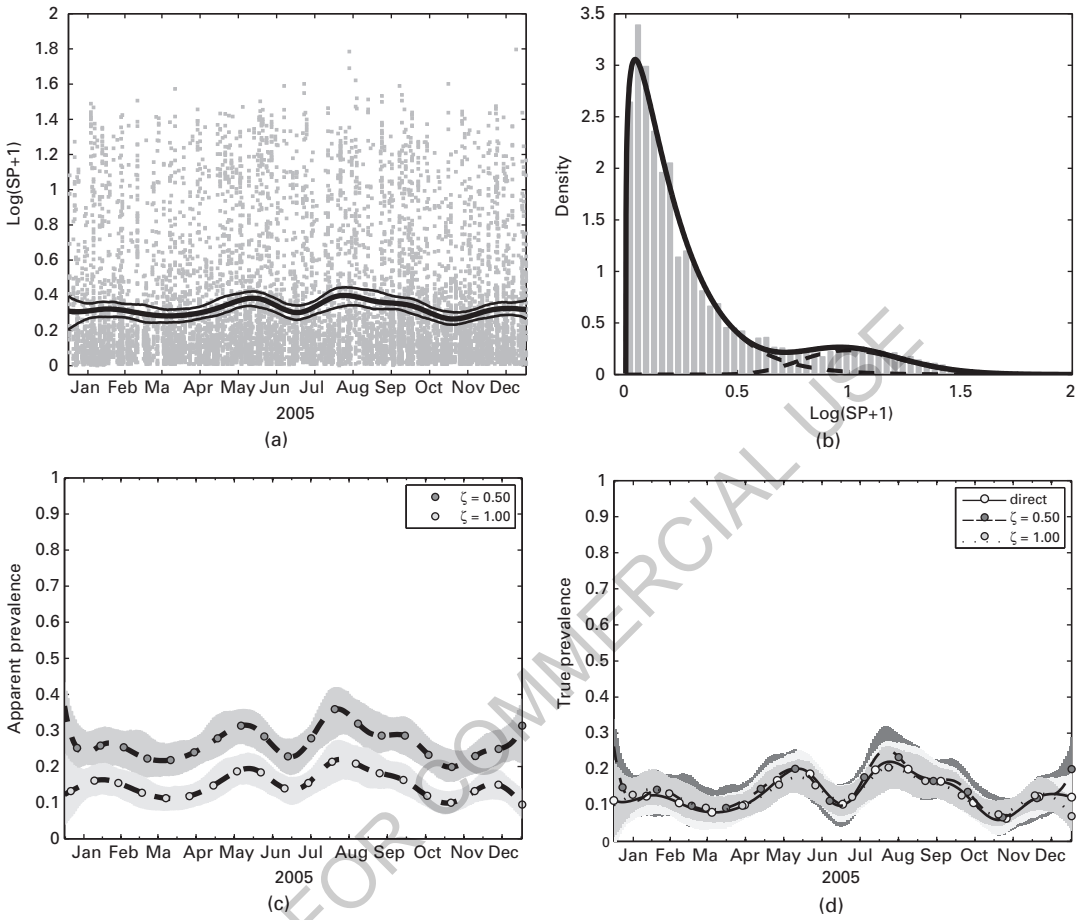

Figure 3 *Salmonella* data: (a) mixture mean of log(SP+1) as function of sampling time, (b) two-component gamma mixture density estimate of log(SP+1), (c) estimated seroprevalence based on $\zeta = 0.50$ and $\zeta = 1.00$ and (d) estimated prevalence using direct estimation, and using the misclassification-corrected seroprevalence estimators for $\zeta = 0.50$ and $\zeta = 1.00$.

then resampling observations within populations (Rao and Wu, 1988) is used. As explained earlier, the *Salmonella* data set is hierarchically structured, with (bottom-up) animal $i = 1, \ldots, n_{jk}$ sampled at a particular sampling time $j = 1, \ldots, r_k$ within a particular herd $k = 1, \ldots, K$. To reflect this hierarchical structure of the data, bootstrap confidence intervals are calculated by resampling with replacement K herds from the observed set of herds and, when the k^*th herd is selected, resampling with replacement for every $j = 1, \ldots, r_{k^*}$, n_{jk^*} observations $y^*_{ijk^*}$ corresponding to the jth sampling time within the selected herd k^*. Then, to assess the variability in $f(t, \widehat{\theta})$, the same model is repeatedly fitted to the bootstrap samples leading to B

bootstrap estimates $f^*(t, \widehat{\theta}^*)$. We chose $B = 500$. In order to obtain the $100(1-2\alpha)\%$ pointwise confidence interval for $f(t)$, percentile intervals are calculated conditional on time t. The latter are defined as $[f^*(t, \widehat{\theta}^*)[(B+1)\alpha]; f^*(t, \widehat{\theta}^*)[(B+1)(1-\alpha)]]$ with $f^*(t, \widehat{\theta}^*)[(B+1)\alpha]$ being the $[(B+1)\alpha]^{\text{th}}$ order statistic of the $B = 500$ bootstrap estimates $f^*(t, \widehat{\theta}^*)$. Of course, in order to estimate the variability of $\widehat{\pi}_z(t)$, each bootstrap sample is dichotomized first.

3.2 Varicella-Zoster virus

Varicella-Zoster virus is one of the eight herpes viruses known to affect humans. Primary infection with VZV results in chickenpox, being a highly infectious disease mostly affecting young children. Following primary infection there is usually lifelong protective immunity from further episodes of chickenpox. The data came from a Belgian study, in which 2381 serum samples were collected from November 2001 until March 2003 and tested for VZV-specific antibody levels (AL) using ELISA (Nardone and Miller, 2004).

The age-dependent FOI is derived directly from the log transformed antibody-levels log(AL+1). For VZV, the presence of antibody levels is assumed to be lifelong. This implies, given the additional assumption of time homogeneity and ignoring maternal antibody levels, that the prevalence $\pi(a)$ is a non-decreasing function of age a. The fact that $\pi(a)$ is non-decreasing in a implies together with $\mu_S < \mu_I$ that the mixture mean $\mu(a)$ is also non-decreasing in a (see identity (2.9)). To avoid the estimation of decreasing age-trends in antibody level mean $\mu(a)$, but still allow flexible estimation, isotone constrained linear P-splines regression (Bollaerts *et al.*, 2008) is used. Constrained P-splines regression, which is a non-parametric smoothing technique by which different types of shape constraints can be imposed, extends P-splines regression as introduced by Eilers and Marx (1996) with an additional asymmetric discrete penalty enforcing the constraints. For the current application, the antibody level mean $\mu(a)$ is estimated as a function of age using a basis of 30 equally-spaced B-splines of $d = 3$ between minimum and maximum age. Monotonicity is imposed using an asymmetric first-order penalty with its weight chosen as high as 10^6 to ensure that violations of the monotonicity assumption are negligible. A second-order smoothness penalty is used with the smoothness weight being optimally chosen using cross-validation with candidate values selected from the approximate geometric grid, $\{0.001, 0.01, 0.1, 0.5, 1, 5, 10, 50, 100, 500\}$. The optimal smoothness parameter is found to be 0.01. The estimated log(VZV+1) mixture mean $\widehat{\mu}(a)$ as a function of age is graphically displayed in Figure 4a, on top of a scatter plot of the data by age. A histogram of the data is given in Figure 4b.

The mixing component means μ_S and μ_I can be obtained in various ways, using EM or NPMLE (see, e.g., Böhning *et al.*, 1992, 1998; McLachlan and Peel, 2000; Schlattmann, 2009) or MCMC-sampling (Gilks *et al.*, 1996). The use of the EM algorithm is illustrated here by estimating the mixing component means based on a

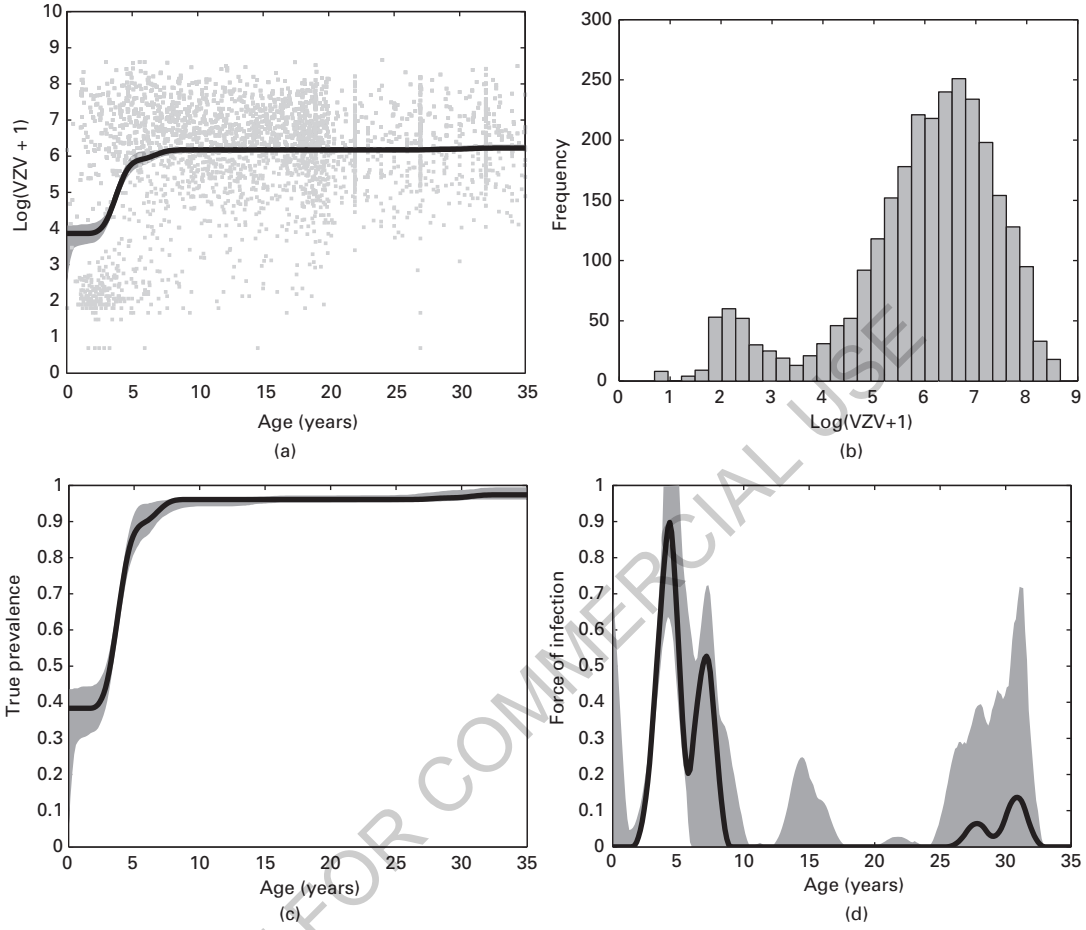

Figure 4 VZV data: (a) mixture mean of log(VZV+1) as function of age, (b) histogram of log(VZV+1), (c) age-dependent true prevalence and (d) age-dependent FOI.

mixture with two normally distributed components ($\widehat{\mu}_S = 2.3324$ and $\widehat{\mu}_I = 6.3279$). The estimates $\widehat{\mu}(a)$, $\widehat{\mu}_S$ and $\widehat{\mu}_I$ are combined as in (2.11) to obtain an estimate of the prevalence $\widehat{\pi}(a)$, given Figure 4c. To obtain an estimate of the FOI, $\widehat{\mu}'(a)$, $\widehat{\mu_S}$ and $\widehat{\mu_I}$ are combined as in (2.13) with $\widehat{\mu}'(a)$, the first derivative of $\widehat{\mu}(a)$, being analytically derived from $\widehat{\mu}(a)$ (Dierckx, 1993). The estimated curve $\widehat{\lambda}(a)$ is given in Figure 4d. The FOI curve shows peaks at about 4 and 7 years, which can be explained by the Belgian school system, with the typical age of joining nursery school (not compulsory but commonly done) being 3 years and the one of joining primary school (compulsory primary education) being 6 years.

The variability of the estimated curves is again assessed by means of pointwise bootstrap confidence intervals. In total, $B = 500$ bootstrap samples are generated by resampling pairs of the original data (a_i, y_i) with replacement. For each bootstrap sample, which contains $N = 1295$ observations (a_i^*, y_i^*), $\mu(a)$, μ_S and μ_I are estimated as before using $\log(y_i^* + 1)$, leading to $B = 500$ different estimates $\widehat{\mu}^*(a)$, $\widehat{\mu}_S^*$ and $\widehat{\mu}_I^*$. In order to estimate the $100(1 - 2\alpha)\%$ pointwise percentile intervals are calculated as before, now conditional on age.

4 Simulation study

In Section 2.4, we showed the equivalence between the corrected threshold based approach and the direct approach. However, when it comes to estimation, both methods may still differ in the way they can be implemented and in their distributional properties. A Monte Carlo simulation study was conducted to get further insights in the performance of both methods for a range of sample sizes and parameter settings.

To simulate data, we start from a two-component Gaussian mixture model given by $(1 - \pi(a))\, N(y|\mu_S, \sigma_S^2) + \pi(a) N(y|\mu_I, \sigma_I^2)$. Without loss of generality we put $\mu_S = 0$ throughout the simulation study such that $\mu_I = \mu_I - \mu_S$. Data are generated for different 'true' values for five parameters: the mixture model parameters μ_I, σ_S^2 and σ_I^2, the threshold parameter ζ (implying a certain sensitivity and specificity) and the sample size n. To ensure a good spread, age values a (in years) are generated using Latin hyper-cube sampling (Mckay *et al.*, 1979) from a uniform U(0,12) distribution for each simulation setting. In Latin hyper-cube sampling, the sampling domain is divided in n segments having equal sampling probability and exactly one sample is taken within each segment. The age-dependent (true) prevalence $\pi(a)$ is kept fixed throughout the whole simulation study, equal to the estimated curve from the VZV example, as graphically represented in Figure 4c, but limited to the maximal age of 12 years.

Instead of taking some fixed values (the number of which is always limited) for the above-mentioned five parameters, random values are generated from 'input' distributions. This approach allows to explore a large variety of settings and consequently has the potential to lead to more general results and insights. The mixture component variances σ_S^2 and σ_I^2 are randomly and (independently) sampled from $U(0.5, 1.5)$ and the threshold ζ is randomly sampled from $U(\mu_S = 0, \mu_I)$ where μ_I is randomly sampled from $U(0.5, 6)$. Finally, sample sizes n are randomly sampled from the set $\{100, 101, \ldots, 2500\}$.

In total, $K = 1000$ simulation settings $\mathbf{simset}_i = (\sigma_{Si}^2, \sigma_{Ii}^2, \mu_{Ii}, \zeta_i, n_i)$, with $i = 1, 2, \ldots, K$, are generated. Then, for each simulation setting, 100 data sets $\mathbf{y}_{ij}$, $j = 1, 2, \ldots, 100$, are generated. In particular, given a simulation setting $\mathbf{simset}_i$, a data set $\mathbf{y}_{ij}$ is generated containing n_i pairs of observations (a_{ijk}, y_{ijk}) of ages (a) and test results (y) and with $k = 1, \ldots, n_i$. Then, for a given a_{ijk}, the corresponding $\pi(a_{ijk})$ is

calculated and y_{ijk} is obtained as $N(0, \sigma_{Si}^2)$ or as $N(\mu_{Ii}, \sigma_{Ii}^2)$ depending on whether the value generated for the indicator $I_{ijk} \sim Bernoulli(\pi(a_{ijk}))$ equals 0 or 1 respectively.

Then, for each data set, the age-dependent prevalence $\pi(a)$ and FOI $\lambda(a)$ are estimated from $\mathbf{y}_{ij}$ using the direct approach and from $\mathbf{z}_{\zeta ij} = I(\mathbf{y}_{ij} > \zeta_i)$ using the threshold approach with correction for test misclassification. The age-dependent mixture mean (respectively, seroprevalence) is estimated using isotone constrained linear (respectively, logistic) P-splines regression (Bollaerts *et al.*, 2006) as in Section 3.2. The smoothness parameter is optimally selected for each simulated data set separately. The mixture component parameters are estimated using the two-component Gaussian EM-algorithm.

We use the mean squared error (MSE, combining bias and variance) to quantify the quality of the estimates for the prevalence $\pi(a)$ and the FOI $\lambda(a)$ and the relative MSE-difference to compare the performance of the two types of estimators (threshold based versus direct approach). Let us first focus on the estimates for the prevalence using the direct approach with $\widehat{\pi}_{Dij}$ denoting the estimated prevalence for data set j within simulation setting i. Then, for an equally spaced grid on age between minimum and maximum age with $G = 500$ points a_g on the grid, the age-dependent MSEs are calculated as $\text{MSE}\left(\hat{\pi}_{Di}(a_g)\right) = \frac{1}{100}\sum_{j=1}^{100}\left(\hat{\pi}_{Dij}(a_g) - \pi(a_g)\right)^2$, for $g = 1, 2, \ldots, G$. The corresponding overall MSE is then simply given by $\text{MSE}\left(\hat{\pi}_{Di}\right) = G^{-1}\sum_{g=1}^{G}\text{MSE}\left(\hat{\pi}_{Di}(a_g)\right)$. In a similar way, the overall MSEs for the threshold based estimation of the prevalence and for the direct and threshold based estimation of the FOI are calculated, being denoted as $\text{MSE}(\hat{\pi}_{\zeta i})$, $\text{MSE}(\hat{\lambda}_{Di})$ and $\text{MSE}(\hat{\lambda}_{\zeta i})$, respectively. The relative differences in MSEs are then calculated as $\Delta_{\pi i} = (\text{MSE}(\hat{\pi}_{\zeta i}) - \text{MSE}(\hat{\pi}_{Di}))/\text{MSE}(\hat{\pi}_{\zeta i})$ and as $\Delta_{\lambda i} = (\text{MSE}(\hat{\lambda}_{\zeta i}) - \text{MSE}(\hat{\lambda}_{Di}))/\text{MSE}(\hat{\lambda}_{\zeta i})$.

To compare the performance of both methods under different simulation settings, we examine how the relative MSE-differences depend on the simulation settings. This can be done by analyzing the relative MSE-differences as the 'responses' in a regression model with the simulation settings as 'explanatory' variables. More precisely we use Generalized Additive Models (GAMs) for this purpose, being introduced by Hastie and Tibshirani (1990). GAMs extend the framework of Generalized Linear Models (GLMs; McCullagh and Nelder, 1989) by allowing the explanatory variables X_j for $j = 1, \ldots, p$, to enter the linear predictor η as smooth functions $f_j(\cdot)$, as such, keeping the generality of a GLM, but relaxing its polynomial structure of the additive effects. We adopt the GAM-approach by Marx and Eilers (1998), who proposed to fit all smooth components $f_j(\cdot)$ simultaneously using penalized likelihood with every smooth component $f_j(\cdot)$ being a B-splines function. In particular, to the response variables Δ_π and Δ_λ, penalized GAM models are fitted assuming a normal response and using the identity link. In total, 6 explanatory variables X_j for $j = 1, \ldots, 6$ are considered. In particular, the explanatory variables considered are (transformations of) the simulation settings, i.e., the mixture component variances σ_S^2 and σ_I^2, the mean of the second mixture component μ_I, test sensitivity $\text{SE} = 1 - \Phi(\zeta, \mu_I, \sigma_I^2)$, test

specificity SP = $\Phi(\zeta, \mu_S, \sigma_S^2)$ and sample size n. Every smooth component $f_j(X_j)$, for $j = 1, \ldots, 6$, is taken to be a B-splines function of 15 equally spaced B-splines of third degree. A second order smoothness penalty is chosen and the smoothness parameter is optimally determined using Akaike's Information Criterion from a 6-dimensional geometric grid Λ^6 where $\Lambda = 10^{\mathcal{P}}$ with $\mathcal{P} = \{-1, 0, 1, 2, 3\}$. The obtained optimal smoothness parameters are (on $\log_10$ scale) $\{3, 3, 3, 3, 3, 2\}$ and $\{3, 3, 3, 3, 3, 3\}$. To allow a direct comparison of both approaches on the original scale, the marginal additive smooth effects

$$E(Y|X_j = x_j) = \int \cdots \int g^{-1}\left(\eta(X_1, \ldots, X_p)\right) f(X_1, \ldots, X_p)^{-j} \prod_{k \neq j} dx_k,$$

are obtained through numerical integration by simulation as

$$\widehat{E}(Y|X_j = x_j) = \frac{\sum_{r=1}^{R} g^{-1}(\widehat{\eta}(X_1 = x_1^{(r)}, \ldots, X_p = x_p^{(r)}) \phi\left(\frac{x_j^{(r)} - x_j}{\sigma}\right)}{\sum_{r=1}^{R} \phi\left(\frac{x_j^{(r)} - x_j}{\sigma}\right)},$$

where R is a (large) number of simulation settings, generated by randomly sampling from the input distributions as specified before. We will take $R = 10^3$. This expression represents the Gaussian-kernel weighted average with ϕ being the Gaussian density given by $\phi(z) = (2\pi)^{-1/2} e^{-\frac{z^2}{2}}$. The variance σ^2 is of fundamental importance since it controls the trade-off between bias and variances and is optimally selected using ordinary least squares cross-validation (Silverman, 1986). Percentile bootstrap confidence intervals of the marginal effects are calculated as well using $B = 500$ bootstrap samples. Each bootstrap sample is obtained by resampling $K = 1000$ quintuples of 'observations' (**simset**$_i$, MSE($\hat{\pi}_{Di}$), MSE($\hat{\pi}_{\zeta i}$), MSE($\hat{\lambda}_{Di}$), MSE($\hat{\lambda}_{\zeta i}$)), based on which the simulation based marginal effects for the relative differences in MSEs are calculated, yielding $B = 500$ replicated estimates of the marginal effects $\widehat{E}^*(Y|X_j = x_j)$ for $j = 1, \ldots, 6$. In order to estimate the $100(1-2\alpha)\%$ pointwise confidence interval for $\widehat{E}(Y|X_j = x_j)$, percentile intervals are calculated conditional on X_j as in Section 3.

The results for the relative MSE-differences between the direct and threshold based approach on the true prevalence scale and FOI scale are graphically displayed in Figure 5. Clearly, for most plots the marginal smooth additive effects are larger than 0, indicating that the direct approach performs better than the threshold based approach. For the prevalence, the degree in which the direct method outperforms the threshold method improves with higher values for μ_I and of σ_I^2, with lower values for σ_S^2, with smaller values for the specificity SP and with smaller or larger values for the sensitivity SE. It does not seem to depend very much on the sample size. For the FOI the differences are less pronounced. Sensitivity seems to have the largest impact and it improves with sample size. Only on the FOI scale, we observe that the

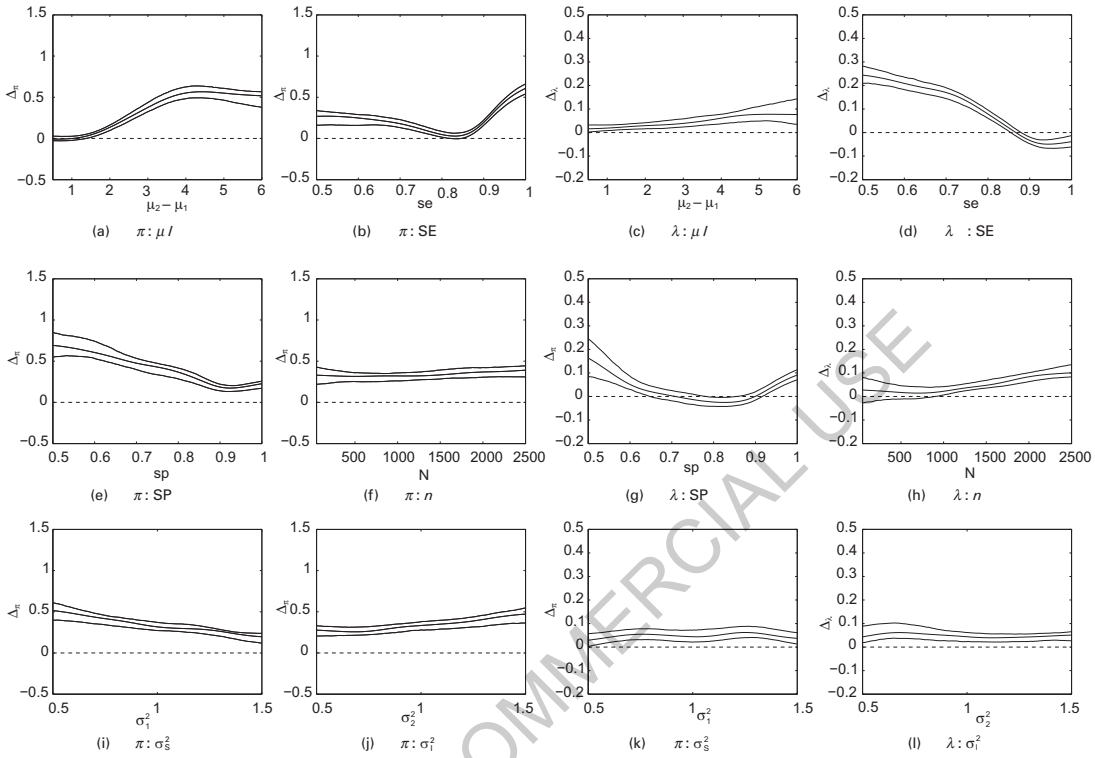

Figure 5 Simulation study: marginal additive smooth effects for the relative differences in MSEs on the true prevalence scale (left panels) and FOI scale (right panels) together with 95% bootstrap confidence intervals.

threshold based approach performs slightly better than the direct method for high values of the sensitivity.

In summary, the main conclusion is that, in case both approaches use equivalent model specifications (being the same mixture model, and the same spline model for the dependency on age), the new direct estimator is generally performing better than the corrected threshold method.

5 Discussion

In this article, we illustrated that the 'optimal' threshold minimizing the asymptotic bias is different for the true prevalence and the age-dependent FOI. Realizing that the choice of threshold is troublesome, we propose estimating epidemiological parameters directly from serological data, avoiding the use of thresholds. The close connection between the common threshold based method and the direct method is revealed and the versatility of the methodology is illustrated by its application in two

diverse settings. In addition, a Monte Carlo simulation study comparing both methods is conducted, based on which it is concluded that the direct approach performs better.

The proposed estimators are derived from an underlying two-component mixture model for (serological) test results assuming a covariate-dependent mixing probability and covariate-independent mixing components. Mixture models are a very natural choice to analyze serological data and several applications are reported in the literature. Hardelid *et al.* (2008) use finite mixture regression models to investigate the effect of maternal age and maternal country of birth on rubella antibody distribution of newborns. Gay (1996) modelled age-stratified serological data using two-component mixture models with age-dependent mixing probabilities and age-dependent mixture components. The model is parameterized at a limited number of age groups with intermediate values obtained by linear interpolation. Furthermore, order restrictions are imposed assuming that the mixing probability, mixing component means and variances increase by age. Results indicate that changes in mixing probability are the dominant age effect. In later work when using age-stratified three-component mixture models (reflecting strong positive, weak positive and negative components) to analyze oral fluid test results, Gay *et al.* (2003) omit the age-dependency of the parameters of the mixture components. Also Vyse *et al.* (2004, 2006) assume age-independent mixture components to estimate age-stratified disease prevalences using two-component mixture models within age groups. In all these applications, the (covariate dependent) mixture model is estimated, whereas we proposed estimators of epidemiological parameters that are derived from an underlying mixture model, but for which the mixture model as such does not need to be estimated. Indeed, the direct approach only requires the estimation of first-order moments. This is an important practical advantage because it is often difficult to find a mixture model that adequately fits the data. Observe that also the threshold approach requires a full specification of the mixture density, in order to obtain estimates for SE and SP.

Finally it is important to emphasize that the threshold estimator cannot be corrected for misclassifications in settings with two thresholds, and moreover that it deletes the inconclusive cases laying in between these two thresholds. The direct method, on the other hand, does not discard any inconclusive data, and does not need to be corrected for misclassifications.

The proposed estimators rely on the assumption of a covariate-dependent mixing probability and covariate-independent mixing components, which is equivalent to the use of constant thresholds when adopting the threshold approach. Although this assumption is a common one (see Gay, 1996; Gay *et al.*, 2003; Vyse *et al.*, 2004, 2006), it would be interesting to relax this assumption to model, e.g., the mechanism of decaying antibody levels. Finally, note that since the mixture mean $\mu(a)$ and the mixture component means μ_S and μ_I are estimated separately, the natural order restriction, $\mu_S < \mu(a) < \mu_I$, might be violated, yielding direct estimates of the epidemiological parameters outside the valid range. Observe that a similar order restriction holds for the threshold approach, i.e., $1 - \text{SP} < \pi_z(a) < \text{SE}$. Upon

occurrence, violations against these order restrictions might be prevented by using constrained optimization.

Acknowledgements

The authors thank an associate editor and two reviewers for their valuable comments and suggestions, which led to a substantially improved version of the manuscript. This study was partly funded and benefited from discussions as part of SIMID, a strategic basic research project funded by the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), project number 060081; and of POLYMOD, a European Commission project funded within the Sixth Framework Programme, Contract Number: SSP22-CT-2004-502084. The authors gratefully acknowledge the financial support from the IAP research Network P6/03 of the Belgian Government (Belgian Science Policy). Kaatje Bollaerts acknowledges support from the Belgian Federal Public Service of Health, Food Chain Safety, and Environment research programme (R-04/003-Metzoon) 'Development of a Methodology for Quantitative Assessment of Zoonotic Risks in Belgium Applied to the "Salmonella in Pork" Model'.

References

Anderson RM and May RM (1991) *Infectious diseases of humans, dynamic and control.* New York. Oxford University Press Inc.

Böhning D, Schlattmann P and Lindsay B (1992) Computer-assisted analysis of mixtures (c.a.man): Statistical algorithms. *Biometrics*, **48**, 283–304.

Böhning D, Dietz E and Schlattmann P (1998) Recent developments in c.a.man (computer assisted analysis of mixtures). *Biometrics*, **54**, 525–36.

Bollaerts K, Eilers PHC and Van Mechelen I (2006) Simple and multiple p-splines regression with shape constraints. *Brit. J. of Math. and Stat. Psych.*, **59**, 451–69.

Bollaerts K, Aerts M, Ribbens S, Boone I, Van der Stede Y and Mintiens K (2008) Identification of *Salmonella* high risk pig-herds in belgium by using semiparametric quantile regression. *Journal of Royal Statistical Society, Series A*, **171**, 449–64.

Claeskens G, Krivobokova T and Opsomer JD (2009) Asymptotic properties of penalized spline estimator. *Biometrika*, **96**, 529–44.

Cortiñas Abrahantes J, Bollaerts K, Aerts M, Van der Stede Y, Ogunsanya V and Mintiens K (2009) *Salmonella* serosurveillance: different statistical methods to categorize pig herds based on serological data. *Preventive Veterinary Medicine*, **89**, 59–66.

Dempster AP, Laird NM and Rubin DB (1977) Maximum likelihood from incomplete data via the em algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

Dierckx P (1993) *Curve and surface fitting with splines*. Oxford: Clarendon.

Eilers PHC and Marx B (1996) Flexible smoothing using b-splines and penalized likelihood (with comments and rejoinder). *Statistical Science*, **11**, 89–121.

Gay NJ (1996) Analysis of serological surveys using mixture models: application to a

survey of parvovirus B19. *Statistics in Medicine*, **15**, 1567–73.

Gay NJ, Vyse AJ, Enquselassie F, Nigatu W and Nokes DJ (2003) Improving sensitivity of oral fluid testing in IgG prevalence studies: application of mixture models to a rubella antibody survey. *Epidemiology and Infection*, **130**, 258–91.

Gilks WR, Richardson S and Spiegelhalter DJ (1996) *Markov chain Monte Carlo in practice*. London: Chapman and Hall.

Hardelid P, Williams D, Dezateux C, Tookey PA, Peckham CS, Cubitt WD and Cortina-Borja M (2008) Analysis of rubella antibody distribution from newborn dried blood spots using finite mixture models. *Epidemiological Infection*, **136**, 1698–706.

Hastie T and Tibshirani R (1990) *Generalized additive models*. New York: Chapman and Hall.

Marx B and Eilers PHC (1998) Direct generalized additive modeling with penalized likelihood. *Computational Statistics & Data Analysis*, **28**, 193–209.

McCullagh P and Nelder JA (1989) *Generalized Linear Models*. London: Chapman & Hall.

Mckay MD, Conover WJ and Beckman RJ (1979) A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, **21**, 239–45.

McLachlan GJ and Peel D (2000) *Finite mixture models*. New York: John Wiley & Sons.

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

Randles RH (1982) On the asymptotic normality of statistics with estimated parameters. *Annals of Statistics*, **10**, 462–74.

Rao JNK and Wu CFJ (1988) Resampling inference with complex survey data. *Journal of American Statistical Association*, **83**, 231–41.

Rogan WMJ and Gladen B (1978) Estimating prevalence from the results of a screening test. *American Journal of Epidemiology*, **107**, 71–76.

Schlattmann P (2009) *Medical applications of finite mixture models*. Statistics for Biology and Health, Springer-Verlag Berlin Heidelberg.

Silverman BW (1986) *Density estimation for statistics and data analysis*. London: Chapman and Hall.

Titterington DM, Smith AFM and Makov UE (1985) *Statistical analysis of finite mixture distributions*. New York: Wiley.

Van der Y, Stede Bollaerts K, Cortiñas Abrahantes J, Imberechts H and Mintiens K (2008) Identification of *Salmonella* risk farms by serological surveillance at pre-harvest level: mission impossible!? *Proceedings of the fourth lustrum of the Dutch and Flemish society for Veterinary Epidemiology and Economics*, Wageningen, The Netherlands.

Vyse AJ, Gay NJ, Hesketh LM, Morgan-Capner P and Miller E (2004) Seroprevalence of antibody to varicella zoster virus in England and Wales in children and young adults. *Epidemiology and Infection*, **132**, 1129–34.

Vyse AJ, Gay NJ, Hesketh LM, Pebody R, Morgan-Capner P and Miller E (2006) Interpreting serological surveys using mixture models: the seroepidemiology of measles, mumps and rubella in England and Wales at the beginning of the 21st century. *Epidemiology and Infection*, **134**, 1303–12.

Wiper M, Insua DR and Ruggeri F (2001) Mixtures of gamma distributions with applications. *Journal of Computational and Graphical Statistics*, **10**, 440–54.