

# 4

## Estimating seroprevalence of human papillomavirus type 16 using a mixture model with smoothed age-dependent mixing proportions.

MA Vink<sup>1,2</sup>

J van de Kasstele<sup>3</sup>

J Wallinga<sup>1</sup>

PFM Teunis<sup>1,4</sup>

JA Bogaards<sup>1,2</sup>

1. Centre for Infectious Disease Control, National Institute for Public Health and the Environment, Bilthoven, the Netherlands
2. Department of Epidemiology and Biostatistics, VU University Medical Center, Amsterdam, the Netherlands
3. Department of Statistics, Mathematical Modelling and Data Logistics, National Institute for Public Health and the Environment, Bilthoven, the Netherlands
4. Hubert Department of Global Health, Rollins School of Public Health, Emory University, Atlanta (GA), USA

Epidemiology, 2015;26(1):8–16

## **ABSTRACT**

### **Background**

The presence of antibodies to viral antigens in serum is generally considered a well-defined marker of past infection or vaccination. However, analyses of serological data that use a cut-off value to classify individuals as seropositive are prone to misclassification bias, in particular when studying infections with a weak serological response, such as the sexually transmitted human papillomavirus (HPV).

### **Methods**

We analyzed the serological concentrations of HPV type 16 (HPV16) antibodies in the general Dutch population in 2006-07, prior to the introduction of mass vaccination against HPV. We used a two-component mixture model to represent persons who were seronegative or seropositive for HPV16. Component densities were assumed to be log-normally distributed, with parameters possibly dependent on sex. The age-dependent mixing proportions were smoothed using penalized splines to obtain a flexible seroprevalence profile.

### **Results**

Our results suggest that HPV16 seropositivity is associated with higher antibody concentrations in women. Seroprevalence shows an increase starting from adolescence in men and women alike, coinciding with the age of sexual debut. Seroprevalence stabilizes in men around age 40 years, whereas it has a decreasing trend from age 50 years onwards in women. Analyses that rely on a cut-off value to classify persons as seropositive yield substantially different seroprevalence profiles, leading to a qualitatively different interpretation of HPV16 infection dynamics.

### **Conclusions**

Our results provide a benchmark for examining the effect of HPV16 vaccination in future serological surveys. Our method may prove useful for estimating seroprevalence of other infections with a weak serological response.

## INTRODUCTION

Human papillomavirus (HPV) is one of the most prevalent sexually transmitted infections and the necessary cause for the development of cervical cancer, the second most common cancer among women worldwide.<sup>1</sup> Many countries have implemented vaccination programs for females, specifically targeting HPV types 16 and 18, which are most commonly associated with cervical cancer.<sup>2</sup> In 2009, vaccination of 12-year-old girls against HPV16 and HPV18 was added to the Dutch national immunization program. As development of cervical cancer after HPV infection takes more than 25 years on average,<sup>3</sup> it will be a couple of decades before the effects of vaccination become apparent in a reduction of cervical cancer cases. Monitoring surrogate endpoints is required to infer effectiveness of vaccination sooner, and data from serological surveys might be a useful tool for observing changes in HPV infection dynamics. In the Netherlands, two population-based serological surveys have been carried out (in 1995-96 and in 2006-07) and a third survey is being scheduled for 2016. Knowledge of the pre-vaccine seroprevalence of HPV is necessary as it will serve as a benchmark for examining changes in infection dynamics from post-vaccine serological surveys.

The presence of antibodies to viral antigens in serum is generally considered a well-defined marker of past infection or vaccination. Serological cross-sectional studies have been used to estimate the prevalence and the force of infection,<sup>4</sup> and they provide a tool to monitor the impact of mass vaccination.<sup>5</sup> The analysis of serological data often relies on methods that use a fixed cut-off value to discriminate between seropositive and seronegative individuals. However, when test sensitivity and specificity are not known, such methods are prone to misclassification bias, especially when antibody concentrations of likely negative and likely positive persons overlap. In such cases, it is common practice to set a relatively high cut-off value to accurately identify true antibody-negative individuals, thus safeguarding test specificity.<sup>6</sup>

The usefulness of serosurveillance for monitoring HPV infection dynamics is presently hampered by a limited understanding of the serological response to infection with HPV.<sup>7</sup> This is partly because anogenital infection is exclusively intraepithelial with little or no viraemia, resulting in low antibody concentrations to the major capsid protein L1.<sup>8</sup> Consequently, assays that measure L1 antibody concentrations are relatively insensitive, with a low signal to noise ratio, and there is no clear threshold to discriminate between seronegative and naturally infected seropositive individuals, complicating the analysis of HPV serological data. It is commonly held that not everyone who is infected with HPV will seroconvert; seroconversion rates based on a fixed cut-off value are estimated at around 50%-60% in women.<sup>9</sup> Antibody concentrations may wane over time,<sup>10</sup> making HPV seroconversion an imperfect marker of past infection and hence complicating the a priori choice for the shape of the age-specific seroprevalence.

HPV seroprevalence figures in the Dutch population prior to vaccination were estimated using a fixed cut-off value to classify persons as seropositive or seronegative. As a result, estimates of seroprevalence among young children were relatively high, contradicting the sexual mode of HPV transmission, and differences between men and women were reported—in particular, a delay of

10 years in step-up of male seroprevalence— that cannot be explained by sexual behavior only.<sup>11</sup>

<sup>12</sup> To improve our understanding of the sero-epidemiology of HPV, it is worthwhile to compare these estimates obtained by prior classification to an alternative that uses full information of the data and that adjusts for covariates such as sex, if necessary.

In this paper, we present novel methodology for estimating the age-specific seroprevalence directly from antibody titers or concentrations. We apply our method to estimate the pre-vaccine seroprevalence of HPV16, the most prevalent of all HPV types. Briefly, we model the concentration of HPV16 antibodies in the Dutch general population as a combination of concentrations from seropositive and seronegative individuals by a mixture model. We combine the mixture model with a flexible approach to model the age-dependent mixing proportions using penalized splines (P-splines). Because not much is known of HPV serology, this allows us to estimate the seroprevalence without the need for a predefined shape. Possible differences in the serological response between men and women are investigated within this model framework, and seroprevalence profiles from the mixture model are compared with a method that uses a cut-off value to denote seropositivity.

## METHODS

### Serological data

HPV-specific antibodies were measured in a cross-sectional population-based serological survey (Pienter 2),<sup>11</sup> performed in the Netherlands in 2006-2007 before HPV vaccination was included in the national immunization program.<sup>13</sup> A total of 7,179 randomly sampled participants, of whom 3,304 men and 3,875 women between 0 and 79 years of age, provided a blood sample. HPV type-specific IgG serum antibodies against L1 virus-like-particles were tested with a virus-like-particle-based multiplex immunoassay. This non-neutralizing assay measures the antibody concentration of seven high-risk HPV types. The assay has a lower limit of detection for HPV16 of 0.08 Luminex units per milliliter (LU/ml). A cut-off value of 9 LU/ml for seropositivity was determined by a one-sided 99% prediction interval using antibody concentrations from persons who are believed to be seronegative for HPV16, here taken to be children between 0 and 10 years of age.<sup>11</sup>

### Mixture model

For each participant we used age, sex, and concentration of HPV-16 specific IgG antibodies. We modeled the log-transformed concentration of HPV16 antibodies by a mixture model with two components, representing people seronegative and seropositive for HPV16. Each individual  $i$  ( $i = 1 \dots n$ ) with observed log-concentration  $y_i$ , contributes to the likelihood:

$$f(y_i) = (1 - I_i)f_0(y_i|\mu_0, \sigma_0) + I_if_1(y_i|\mu_1, \sigma_1),$$

with  $l_i = 0$  if the person was assigned to the seronegative mixture component and  $l_i = 1$  if the person belongs to the seropositive component. We assumed

$$I_i \sim \text{Bern}(p_i)$$

with

$$\text{logit}(p_i) = s^M(a_i)1_{[\text{sex}_i=M]} + s^F(a_i)1_{[\text{sex}_i=F]}.$$

Here 1 denotes the indicator function which equals one if the statement in the subscript applies and zero otherwise. The component densities  $f_0$  and  $f_1$  are assumed to be normally distributed, independent of age but possibly dependent on the covariate sex:

$$\begin{aligned} f_0(y_i) &\sim N(\mu_0^M 1_{[\text{sex}_i=M]} + \mu_0^F 1_{[\text{sex}_i=F]}, \sigma_0^M 1_{[\text{sex}_i=M]} + \sigma_0^F 1_{[\text{sex}_i=F]}), \\ f_1(y_i) &\sim N(\mu_1^M 1_{[\text{sex}_i=M]} + \mu_1^F 1_{[\text{sex}_i=F]}, \sigma_1^M 1_{[\text{sex}_i=M]} + \sigma_1^F 1_{[\text{sex}_i=F]}). \end{aligned}$$

In case the observed log-concentration was left censored (below the detection limit of the serological assay), the contribution to the likelihood was assumed to be a cumulative Normal distribution. Because a survey on sexual health of youth in the Netherlands showed that sexual intercourse below age 12 years is rare,<sup>14</sup> we assumed the concentration of 0-10 year olds to contribute to the seronegative component density only.

The terms  $s^M(a)$  and  $s^F(a)$  denote log-odds of seropositivity of men and women by age, respectively. To obtain smooth functions of age, they were modeled by the mixed model representation of P-splines:<sup>15,16</sup>

$$\begin{aligned} s^M(a) &= BX\beta^M + BZb^M, \\ s^F(a) &= BX\beta^F + BZb^F. \end{aligned}$$

Here  $B$  is a  $n \times k$  cubic B-spline basis with  $k$  equally spaced knots,  $X$  is a  $k \times d$  matrix such that  $X\beta$  is a polynomial of degree  $d-1$ , and  $Z = D^T(DD^T)^{-1}$  a  $k \times (k-d)$  matrix, where  $D$  is a  $(k-d) \times k$  difference matrix of order  $d$ .  $\beta$  is a vector of length  $d$  and  $b$  is a vector of length  $k-d$ . Within this setting, the degree of B-splines, the number of knots and the  $d$ -order difference penalties on coefficients of adjacent B-spline basis functions, are to be chosen. Given the setting of the P-spline model, the smoothing parameter that controls large values of  $d$ -order differences is automatically estimated. This parameter regulates the number of knots to an effective number needed to fit the model properly. Therefore,  $k$  can be any number, but it is preferred to overstate it in order not to underfit the model. The parameter  $d$  has some influence on the curvature of the spline, with a larger  $d$  resulting in a less curved line. In this study, we put knots on 15 equally distributed age groups and penalize second-order differences, i.e.  $k=15$  and  $d=2$ .

## Estimation procedure

Parameters were estimated in a Bayesian framework through Gibbs sampling, using JAGS.<sup>17</sup> To avoid label switching, the  $\mu$ 's were re-parameterized:  $\mu_1^M = \mu_0^M + \Delta\mu^M$  and  $\mu_1^F = \mu_0^F + \Delta\mu^F$ . As JAGS

works with precision  $\tau$  instead of variance  $\sigma^2$ , the former was estimated. To minimize the correlation between  $\beta_0$  and  $\beta_1$ , we subtracted 35 from all observed ages (roughly corresponding to the mean age of the people in our data set).

Non-informative normal priors were set on parameters  $\mu_0$  and  $\beta$ . Non-informative gamma priors were set on the precision parameters. A normal prior was set on  $b$  with a gamma hyperprior on the precision of  $b$ . To ensure that  $\Delta\mu$  is positive, non-informative half-normal distributions were set on the  $\Delta\mu$ 's. We ran four parallel Markov chain Monte Carlo (MCMC) chains. For each chain 500 iterations were taken as burn-in time, and 12,500 iterations for sampling from the posterior distributions. To reduce autocorrelation we retained every 10<sup>th</sup> iteration, resulting in 5,000 samples in total. Convergence of MCMC chains was inspected visually. The model and estimation method were tested on a simulated data set on which the model parameters were correctly estimated (not shown). The code is provided in the Appendix.

### Sex-specific serological response

We investigated whether the component densities of the mixture model differed between men and women, by evaluating the following models:

1. Both component densities different for men and women;
2. Negative component density the same, positive component density different;
3. Negative component density different, positive component density the same;
4. Both component densities the same for men and women.

The models were evaluated by the deviance information criterion (DIC) which can be viewed as the Bayesian analogue of the Akaike information criterion (AIC).<sup>18</sup> Models are penalized to have a higher DIC both by the deviance (the larger this is, the worse the fit) and by the effective number of model parameters (so favoring parsimony). The scenario with the lowest DIC is preferred. Note that there is not yet a satisfactory answer to the question what a noteworthy difference between two models is.<sup>19</sup> As a rule of thumb, a difference of more than 7-10 points favors the model with the smallest DIC. A difference less than 2-5 points suggests little or no evidence to favor a particular model.<sup>18,20</sup>

### Model validation

To confirm whether the model provides a reasonable fit of the data, we plotted the empirical cumulative distribution function (CDF) with 95% confidence intervals of the HPV16 log-concentration, together with the model-predicted CDF, per sex, per 5-year age group. For the model-predicted CDF per age group, we used the estimated parameters  $\mu_0, \mu_1, \sigma_0, \sigma_1$  and an average seroprevalence, calculated as the weighted mean of the age-specific seroprevalence.

### Cut-off method to estimate seroprevalence

The mixture model relaxes the assumption of a fixed cut-off value to distinguish between seronegative and seropositive individuals. The effect of this relaxation was also evaluated by modeling the seroprevalence with a cut-off value of 9 LU/ml for individuals above 10 years of age. Individu-

als were classified as seropositive if their HPV16 antibody concentration exceeded the cut-off value. The probability that a person  $i$  ( $i = 1, \dots, n$ ) was seropositive is Bernoulli distributed;

$$I_i \sim \text{Bern}(p_i)$$

with

$$\text{logit}(p_i) = s^M(a_i)\mathbf{1}_{[\text{sex}_i=M]} + s^F(a_i)\mathbf{1}_{[\text{sex}_i=F]}.$$

This is similar to the mixture model setting, except that no sex-specific cut-off values were considered. The smoothed prevalences  $s^M(a)$  and  $s^F(a)$  were again represented by P-splines taking 15 equally spaced knots and penalizing second order differences. Estimation was done within the statistical software R version 3.0.1 using the gam function of the mgcv package.<sup>21, 22</sup>

### Sensitivity analysis

By means of sensitivity analysis we tested the influence of the assumption that 0-10 year olds are strictly seronegative. First, we re-estimated the model parameters in case only the 0-5 year olds are assumed strictly seronegative, and second, we ran the model without any prior assumptions on seronegativity. As infants might have acquired increased antibody levels from their mothers,<sup>23</sup> we also tested the influence of this age group by re-estimating the model parameters after excluding the youngest infants (age 0-0.5 years).

## RESULTS

The data provide substantial evidence that the seropositive component density differs between the sexes, as demonstrated by a clearly smaller DIC of models 1 and 2 compared with models 3 and 4 (Table 1). The preferred models specify a sex-specific location of the seropositive component density, whereas the variance of the seropositive component density is similar for men and women. Apparently, detection of HPV16-specific antibodies is performed with similar noise for men and women alike, but the humoral immune response to an HPV16 infection leads to a

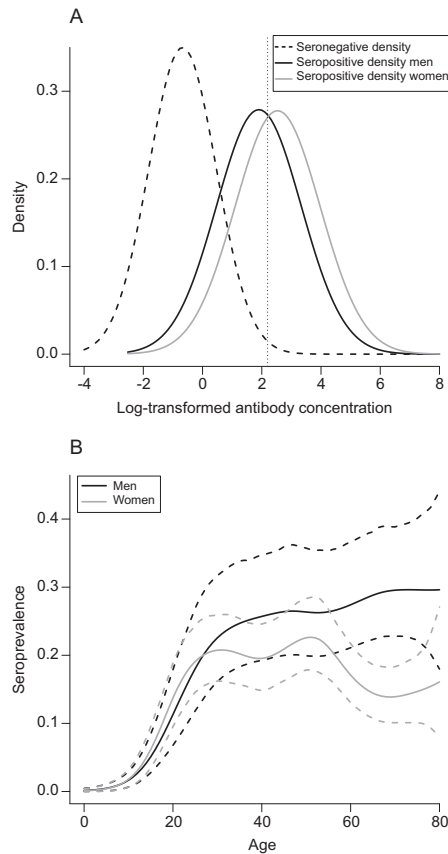
**Table 1.** Estimates of the seropositive and seronegative component densities of antibody concentrations to human papillomavirus type 16, according to four mixture models assessing the sex-specific serological response.

Model	Seronegative component		Seropositive component		DIC	pD
	Men	Women	Men	Women		
1	N(-0.73,1.11 <sup>2</sup> )	N(-0.63,1.17 <sup>2</sup> )	N(1.74,1.47 <sup>2</sup> )	N(2.70,1.38 <sup>2</sup> )	25618.8	22.4
2	N(-0.68,1.14 <sup>2</sup> )		N(1.90,1.43 <sup>2</sup> )		25620.8	20.4
3	N(-0.68,1.13 <sup>2</sup> )	N(-0.68,1.14 <sup>2</sup> )	N(-2.19,1.47 <sup>2</sup> )		25640.0	20.0
4	N(-0.68,1.14 <sup>2</sup> )		N(-2.19,1.48 <sup>2</sup> )		25635.7	17.9

DIC indicates deviance information criterion pD, effective number of parameters

stronger signal in women than in men. There is some evidence that the seronegative component density should be sex-specific as well, but the difference between the DIC of model 1 and model 2 is small. Moreover, model 1 and model 2 provide similar results, both in terms of parameter estimates (Table 1) as well as in estimated seroprevalence profiles (see Appendix Figure 1). We choose to continue with the results provided by model 2, as from a biological point of view we expect that the measurement error in seronegative individuals is independent of sex. In Appendix Table 1, 95% credible intervals of its parameter estimates are presented.

Figure 1A shows the mixture component densities for men and women according to model 2, together with the cut-off value used to denote seropositivity. Test specificity of the cut-off value of 9 LU/ml is calculated at 99% for both men and women, but test sensitivity is low – only 42% for men and 59% for women. Figure 1B shows the resulting HPV16 seroprevalence; a comparable

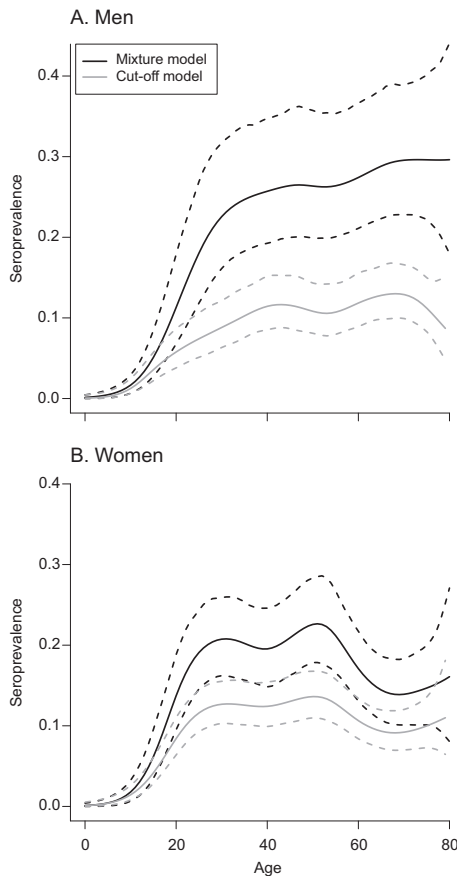


**Figure 1.** Results of the mixture model in which the seronegative component densities are the same but the seropositive component density is sex-specific. (A) Estimated seronegative and seropositive component densities for men and women. The dashed vertical line represents the cut-off value of 9LU/ml. (B) Seroprevalence of HPV16 for men and women with their 95% credible intervals (dashed lines).



steep increase in seroprevalence in both men and women can be seen starting from adolescence, coinciding with the age of sexual debut. Seroprevalence is stable or slightly increasing for men, whereas it decreases for women from age 50 years onwards.

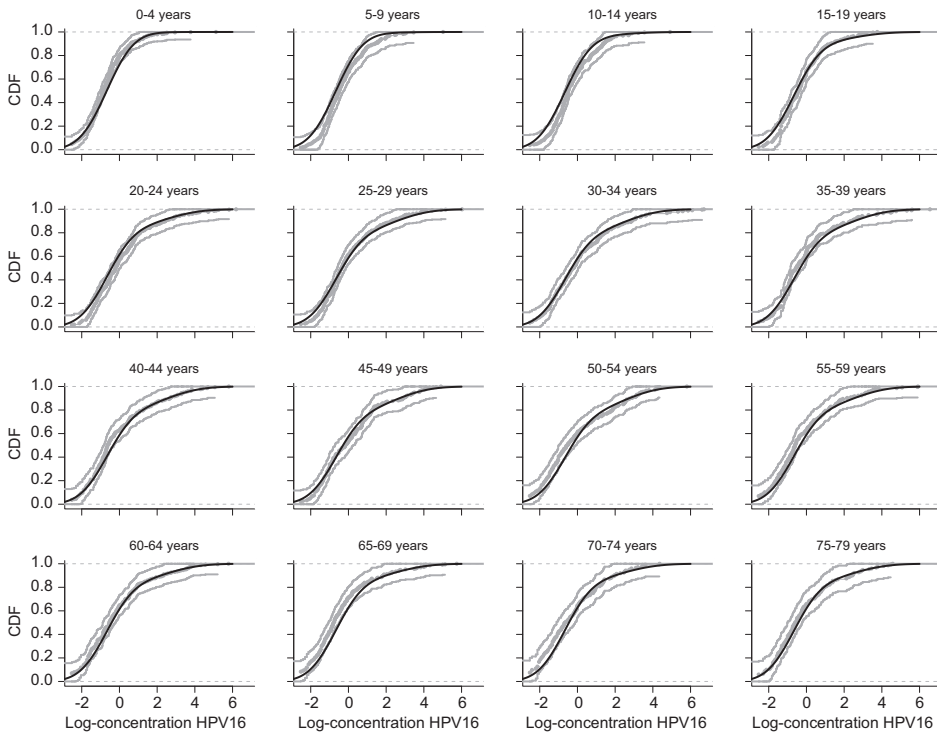
Defining seropositivity on the basis of a fixed cut-off value leads to a lower seroprevalence as compared with the mixture model (Figure 2). The mixture model is estimated at around 20% seropositivity in women between age 30 and 50 years, whereas only about 12% of women have antibody concentrations above the cut-off value in these age groups. Furthermore, the step-up in seroprevalence is more gradual in men when using the cut-off value, whereas the seroprevalence for women retains a similar shape in the two methods.



**Figure 2.** Comparison of the estimated seroprevalence using the mixture model and the cut-off value for men (A) and women (B). The dashed lines represent the 95% credible intervals.

## Model validation

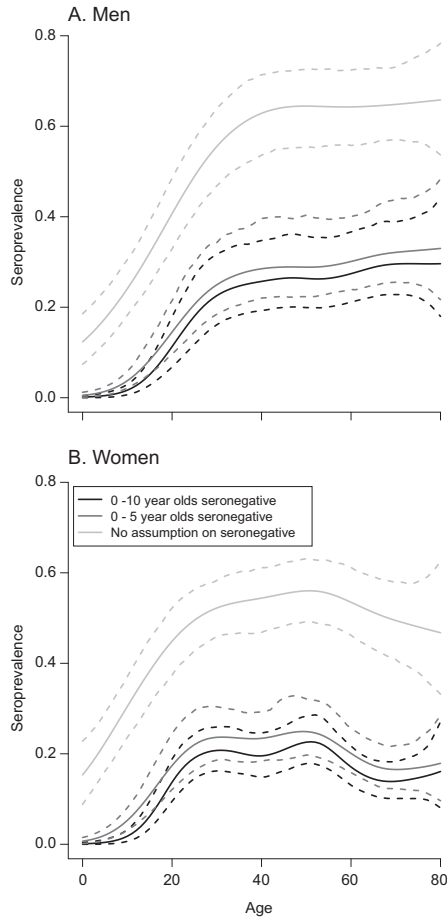
Figure 3 shows the empirical cumulative distribution function (CDF) of the HPV16 log-concentration together with the model-predicted CDF for women. The model-predicted CDF lies between the 95% confidence bounds of the empirical CDF. However, for 0-15 year old girls the model-predicted CDF lies at the boundary of this region, favoring higher concentrations than those observed for some age groups and lower concentrations for others. In particular, children aged 0-5 years have lower seronegative antibody concentrations than expected, while seronegative antibody concentrations among 5-10 year olds are higher than expected. In Appendix Figure 2 we provide similar figures and conclusions for men.



**Figure 3.** Model fit of the data for women. Gray lines denote the empirical cumulative distribution function and 95% confidence interval of the HPV16 log-concentration for women, per 5-year age group. The black line denotes the model-predicted cumulative distribution function (CDF).

## Sensitivity analysis

Figure 4 presents the results of a sensitivity analysis on the assumption that 0-10 year olds are strictly seronegative. The scenario in which 0-5 year olds are assumed to inform only the seronegative component is close to the base-case scenario. A marginal increase in seropositivity is seen, but the seroprevalence profile as well as the estimated mixture densities are similar to the base-case scenario (Table 2). If no prior assumptions on seronegativity are made, a large increase



**Figure 4.** Sensitivity analysis of the assumption that 0-10 year olds are seronegative for men (A) and women (B). Black line is the baseline scenario (0-10 year olds seronegative), dark gray lines represent the scenario that 0-5 year olds are assumed seronegative, light gray lines represent the scenario without assumptions on seronegativity. Dashed lines denote 95% credible intervals.

**Table 2.** Sensitivity analysis on the assumption of strict seronegativity in children.

	Seronegative component		Seropositive component	
	Men & Women		Men	Women
Assumption seronegativity				
None	N(-0.85,0.83 <sup>2</sup> )		N(0.42,1.86 <sup>2</sup> )	N(0.52,2.04 <sup>2</sup> )
Age 0-5 years	N(-0.72,1.10 <sup>2</sup> )		N(1.78,1.44 <sup>2</sup> )	N(2.26,1.52 <sup>2</sup> )
Age 0-10 years	N(-0.68,1.14 <sup>2</sup> )		N(1.90,1.43 <sup>2</sup> )	N(2.53,1.44 <sup>2</sup> )
Maternal antibodies				
Excluding age 0-0.5 years	N(-0.71,1.10 <sup>2</sup> )		N(1.76,1.47 <sup>2</sup> )	N(2.69,1.37 <sup>2</sup> )

in seroprevalence is seen, and the seropositive and seronegative mixture densities have more overlap (Table 2). Consequently, test sensitivity of the cut-off value becomes extremely poor (16% for men and 21% for women), and the estimated seroprevalence becomes implausibly high (with 18% seropositives among children).

Maternal antibodies against HPV16, if present, did not have an influence on the results. The model excluding the 0-0.5 year olds estimated almost identical component densities (Table 2), as well as seroprevalence figures, as the base-case scenario.

## DISCUSSION

Sero-epidemiologic studies offer a rich source for understanding infection dynamics, especially when mass vaccination programs against particular pathogens are in place.<sup>4,5</sup> This study presents a new approach for estimating age-specific seroprevalence profiles directly from antibody concentrations, which is particularly useful when studying infections with a weak serological response. Moreover, our method allows for investigation of covariate effects on the serological response, which makes it a useful tool to obtain new insights into the sero-epidemiology of the pathogen under study. Our application of this approach to a pre-vaccination serological survey of HPV16 in the Netherlands provides strong evidence that HPV16 seropositivity is associated with higher antibody concentrations in women compared with men. The estimated seroprevalence figures provide a benchmark for examining the effect of HPV vaccination in future serological surveys.

Sex-specific cut-off values to discriminate between seropositive and seronegative persons are not commonly applied in the analysis of serological data. We know of just one study that used different cut-off values for men and women in the case of HPV16 serology.<sup>24</sup> Nonetheless, given the anogenital nature of HPV16 infections, it makes sense that the humoral immune response would be sex-specific, as the infected epithelium differs between men and women.<sup>24,25</sup> For the virus, it is more difficult to infect the keratinized tissue of the penis than the soft tissue of the vagina or the anus. Recent publications suggest that the serological responses following an anal or vaginal HPV infection are quite similar, but that the serological response after a penile infection is much more subdued.<sup>26,27</sup> This may explain why, in a population-based study such as ours in which the large majority of men likely had only penile and no anal exposure to HPV16, lower antibody concentrations are observed in men.

Extrapolation of our model to include covariates other than sex is limited by data availability but straightforward in principle. Serosurveillance studies usually provide information on the joint occurrence of antibodies against a wide array of pathogens, and it would be interesting to investigate whether serological concentrations against HPV are modified by seropositivity against other sexually transmitted infections. For example, HIV infection is known to be a strong and independent determinant for HPV seropositivity,<sup>28</sup> but it is not known whether HPV infection leads to higher antibody concentrations in HIV-infected relative to HIV-negative persons.

Likewise, infection with *Chlamydia trachomatis* is hypothesized to promote HPV persistence by shifting the immune system from a cellular response, required to clear viral infection, to a humoral response, associated with higher antibody concentrations.<sup>29</sup> Unfortunately, antibody assays provide only partial information on the presence of *C. trachomatis*, and data on infection status are not routinely collected in population-based serosurveillance studies. Nevertheless, application of our model to other settings (e.g. sexually transmitted infections clinics) could help to clarify the possible interactions between sexually transmitted co-infections.

Serological data have been employed to calibrate HPV transmission models,<sup>30</sup> often with little knowledge about the underlying form of the seroprevalence profile. A major strength of our approach is the adoption of a flexible framework to model seropositivity as function of age. Seroprevalence profiles have been modeled in several ways, with different assumptions about the contact network for transmission, the seroconversion process, waning of antibodies, etc. (see Hens et al.<sup>31</sup> for a comprehensive overview). A variety of statistical models, ranging from fully parametric (e.g. the exponentially damped model<sup>32</sup> or fractional polynomials<sup>33</sup>) to semi- or non-parametric methods (e.g. (penalized) smoothing splines<sup>34</sup>) are available to capture age-specific patterns in serological data. These methods use only bivariate (seronegative/sero-positive) data and do not account for the misclassification bias that might occur. To our knowledge, we are the first to incorporate P-splines into a mixture model for serological data, which is attractive because it does not use an arbitrary cut-off, makes no global assumptions about the shape of the seroprevalence profile, and allows for rapid increases or decreases in age-dependent mixing proportions. A drawback of this procedure is the high dimensionality of the underlying B-spline basis, and finding optimal parameters is a computationally intensive process. Now that we have estimated the seroprevalence profile free of assumptions, it would be convenient to consider a parametric form that captures the characteristics of the smoothed seroprevalence as described by P-splines, but with only a few parameters. Alternatively, the mixture model could be extended to incorporate transmission model estimates of the proportion seropositive by age. Such an approach might be useful when changes in HPV16 seroprevalence over time are to be understood in terms of transmission dynamic models.

The HPV16 seroprevalence profile that we estimate for women agrees well with what has been observed in other Western countries, such as England and Australia.<sup>35, 25, 36</sup> In these studies, as well as in our analysis, female seroprevalence peaked between age 20 and 50 years at around 25%. However, seroprevalence profiles for men show mixed patterns. In England and the United States, HPV16 seroprevalence among men has previously been reported to increase with age but reaches only around 10% at the upper end – about a third of the level that we estimate. In Australia, male seroprevalence has been shown to have a pattern similar to that which we observed in women, with a peak seroprevalence at 15% and a declining trend after age 50 years. Comparing these studies with our results is difficult, because of population-specific differences in the epidemiology of HPV16 and differences in assay technology used across these studies. Nevertheless, all referenced studies used a fixed cut-off value to classify subjects as seropositive.

Therefore, misclassification of seropositive males might have played a role in the relatively low figures reported for men in almost all sero-epidemiological studies.

Seroprevalence estimates from the mixture model are higher than estimates obtained from prior classification of seropositive and seronegative individuals. This makes sense given the notion that test specificity is generally valued more than test sensitivity in the determination of a fixed cut-off value.<sup>6</sup> Using the cut-off value to estimate the seroprevalence, one does not account for persons who are incorrectly classified seronegative, and the seroprevalence will be underestimated. The modelled sensitivity of the cut-off value (59% for women) corresponds to the female seroconversion rates reported in the literature (50%-60%).<sup>9</sup> As these rates are based on data using a cut-off value to denote seropositivity, our results could be interpreted to mean that everyone seroconverts upon infection but that a fraction of the persons is misclassified. The next step is to incorporate infection dynamics to capture the relation between infection and seroconversion. For men, estimates on the basis of a cut-off value lead not only to a lower seroprevalence, but also to a delay in the step-up of seroprevalence as compared with women.<sup>11</sup> It seems that misclassification of seropositive men has an age-specific component, as the bias that arises from the use of a fixed cut-off value is most pronounced between 20 and 30 years of age. Apparently, HPV16-specific antibody concentrations are a bit lower in seropositive men at these ages. A possible explanation is that the humoral immune response of men is boosted upon repeated encounters with HPV16, leading to increased antibody concentrations over age.

There are additional age-dependent patterns in the data for which our model does not account. Results were shown to be highly dependent on the prior assumption that young children, aged 0-10 years, are strictly seronegative. Model validation showed that this assumption does not entirely capture the heterogeneity in antibody concentrations in this age group; children aged 0-5 years have lower seronegative antibody concentrations than 5-10 year olds. If we relaxed the assumption of strict seronegativity to hold only for 0-5 year olds, both the seronegative and the seropositive mixture component densities shifted to slightly lower values. Apparently, antibody concentrations among 5-10 year olds are somewhat higher than among 0-5 year olds, but not as high as among adults. Possibly, the antibody assay cross-reacts with other HPV-types. The multiplex assay includes in total 7 high-risk HPV-types, all sexually transmittable, so cross-reaction with these types is unlikely among children. More probable is cross-reaction with non-sexually transmitted HPV types – such as those that cause skin warts, which are very prevalent among children.<sup>37</sup> Note, however, that literature on this topic is equivocal.<sup>38</sup> We estimated a large shift toward lower values for the seropositive density if no prior knowledge on the seronegative mixture component was provided. It seems that an unsupervised model tends to explain the observed heterogeneity in lower antibody concentrations first, which makes sense given that most observed concentrations are likely seronegative and even slight deviations from log-normality are detected given the large sample size of our study. We could have accounted for more heterogeneity in the model, either by adding more mixture components or by allowing the mean and variance of the seropositive mixture density to depend on age. However, this might not improve comprehension of the results and could lead to selection of an over-fitted model.

The two-component mixture model, while not capable of capturing all data heterogeneity, gives a reasonable fit and allows for straightforward interpretation in terms of seroprevalence.

When monitoring changes in HPV infection dynamics in serological surveys conducted after introduction of the vaccination program, it would be interesting to extend the model to one in which multiple cross-sectional studies are analyzed to account for trends over time. Antibody concentrations among vaccinated persons are generally high compared with antibody levels induced by natural HPV infection.<sup>39</sup> Our model provides a natural framework for distinguishing between naturally infected and vaccinated girls, by including vaccination status as a covariate in the seropositive component density. This is possible only if information about vaccination status is available on the individual level in future serological surveys – which is recommended, especially because the Netherlands have decided to replace the 3-dose HPV16/18 vaccine scheme with a 2-dose scheme for vaccine-eligible cohorts from 2014 onwards.<sup>40</sup> Serological surveys provide a means to monitor antibody levels in vaccinated cohorts, complementary to longitudinal analysis of antibody levels in vaccinated persons – which is particularly relevant since data on the long-term immunogenicity of the 2-dose relative to the 3-dose HPV16/18 vaccine scheme is lacking.

To conclude, this study provides a new method to estimate seroprevalence as a function of age, without prior classification of seropositive individuals. It can be applied to infections with a weak serological response to infection or vaccination, as exemplified by an application to HPV16. We show that using a mixture model instead of a cut-off value to denote seropositivity may affect seroprevalence estimates and lead to a qualitatively different interpretation of infection dynamics.

## ACKNOWLEDGEMENTS

We thank Dr. S Waaijenborg for her contributions to the initial analysis of HPV16 serology, and Dr. H de Melker and Dr. F van der Klis for providing the antibody concentrations of HPV16 from the Pienter 2 survey.

## REFERENCES

1. Walboomers JM, Jacobs MV, Manos MM, et al. Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. *J Pathol.* 1999;189(1): 12-19.
2. Franceschi S, Denny L, Irwin KL, et al. Eurogin 2010 roadmap on cervical cancer prevention. *Int J Cancer.* 2011;128(12): 2765-2774.
3. Vink MA, Bogaards JA, van Kemenade FJ, et al. Clinical progression of high-grade cervical intraepithelial neoplasia: estimating the time to preclinical cervical cancer from doubly censored national registry data. *Am J Epidemiol.* 2013;178(7): 1161-1169.
4. Hens N, Shkedy Z and Aerts M. Modeling infectious disease parameters based on serological and social contact data: a modern statistical perspective. New York, Springer 2012.
5. Smits G, Mollema L, Hahne S, et al. Seroprevalence of mumps in The Netherlands: dynamics over a decade with high vaccination coverage and recent outbreaks. *PLoS One.* 2013;8(3): e58234.
6. Vyse AJ, Gay NJ, Hesketh LM, et al. Seroprevalence of antibody to varicella zoster virus in England and Wales in children and young adults. *Epidemiol Infect.* 2004;132(6): 1129-1134.
7. Mollers M, Vossen JM, Scherpenisse M, et al. Review: current knowledge on the role of HPV antibodies after natural infection and vaccination: implications for monitoring an HPV vaccination programme. *J Med Virol.* 2013;85(8): 1379-1385.
8. Stanley M. HPV - immune response to infection and vaccination. *Infect Agent Cancer.* 2010;5(19): 19.
9. Gravitt PE. The known unknowns of HPV natural history. *J Clin Invest.* 2011;121(12): 4593-4599.
10. af Geijersstam V, Kibur M, Wang Z, et al. Stability over time of serum antibody levels to human papillomavirus type 16. *J Infect Dis.* 1998;177(6): 1710-1714.
11. Scherpenisse M, Mollers M, Schepp RM, et al. Seroprevalence of seven high-risk HPV types in The Netherlands. *Vaccine.* 2012;30(47): 6686-6693.
12. Scherpenisse M, Mollers M, Schepp RM, et al. Changes in antibody seroprevalence of seven high-risk HPV types between nationwide surveillance studies from 1995-96 and 2006-07 in The Netherlands. *PLoS One.* 2012;7(11): e48807.
13. de Melker HE, Conyn-van Spaendonck MAE, Boot HJ, et al. Introduction of vaccination against cervical cancer [in Dutch]. *Ned Tijdschr van Geneesk.* 2009;153(A67): 1-5.
14. de Graaf H, Meijer S, Poelman J, et al. Sex Below the Age of 25: Sexual Health of Youth in the Netherlands in 2005 [in Dutch]. Delft, the Netherlands, Eburon Academic Publishers 2005.
15. Eilers PHC and Marx BD. Flexible smoothing with B-splines and penalties. *Stat. Sci.* 1996;11(2): 89-121.
16. Verbyla A, Cullis BR, Kenward MG, et al. The analysis of designed experiments and longitudinal data by using smoothing splines. *Appl Stat.* 1999;48(3): 269-311.
17. Plummer M. JAGS: A program for analysis of Bayesian graphical models usings Gibbs sampling. 2003.
18. Spiegelhalter DJ, Best NG, Carlin BP, et al. Bayesian measures of model complexity and fit. *J R Stat Soc Ser B (Statistical Methodol).* 2002;64(4): 583-639.
19. Plummer M. Penalized loss functions for Bayesian model comparison. *Biostatistics.* 2008;9(3): 523-539.
20. Lesaffre E (2012). Model building and assessment. Bayesian biostatistics. E. Lesaffre and A. B. Lawson. New York, John Wiley & Sons: 267-318.
21. Wood SN. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J R Stat Soc Ser B (Statistical Methodol).* 2011;73(1): 3-36.
22. R Core Team (2012). R: A language and environment for statistical computing. from [www.R-project.org](http://www.R-project.org).



23. Waaijenborg S, Hahne SJ, Mollema L, et al. Waning of maternal antibodies against measles, mumps, rubella, and varicella in communities with contrasting vaccination coverage. *J Infect Dis.* 2013;208(1): 10-16.
24. Carter JJ, Madeleine MM, Shera K, et al. Human papillomavirus 16 and 18 L1 serology compared across anogenital cancer sites. *Cancer Res.* 2001;61(5): 1934-1940.
25. Newall AT, Brotherton JM, Quinn HE, et al. Population seroprevalence of human papillomavirus types 6, 11, 16, and 18 in men, women, and children in Australia. *Clin Infect Dis.* 2008;46(11): 1647-1655.
26. Heiligenberg M, Alberts CJ, Waterboer T, et al. Route of sexual exposure is independently associated with seropositivity to HPV-16 and HPV-18 among clients of an STI clinic in the Netherlands. *J Infect Dis.* 2013;208(7): 1081-1085.
27. Vriend HJ, Bogaards JA, van der Klis FR, et al. Patterns of human papillomavirus DNA and antibody positivity in young males and females, suggesting a site-specific natural course of infection. *PLoS One.* 2013;8(4): e60696.
28. Mooij SH, van der Klis FR, van der Sande MA, et al. Seroepidemiology of high-risk HPV in HIV-negative and HIV-infected MSM: the H2M study. *Cancer Epidemiol Biomarkers Prev.* 2013;22(10): 1698-1708.
29. !!! INVALID CITATION !!!
30. Barnabas RV, Laukkanen P, Koskela P, et al. Epidemiology of HPV 16 and cervical cancer in Finland and the potential impact of vaccination: mathematical modelling analyses. *PLoS Med.* 2006;3(5).
31. Hens N, Aerts M, Faes C, et al. Seventy-five years of estimating the force of infection from current status data. *Epidemiol Infect.* 2010;138(6): 802-812.
32. Farrington CP. Modelling forces of infection for measles, mumps and rubella. *Stat Med.* 1990;9(8): 953-967.
33. Shkedy Z, Aerts M, Molenberghs G, et al. Modelling age-dependent force of infection from prevalence data using fractional polynomials. *Stat Med.* 2006;25(9): 1577-1591.
34. Keiding N, Begtrup K, Scheike TH, et al. Estimation from current-status data in continuous time. *Lifetime Data Anal.* 1996;2(2): 119-129.
35. Stone KM, Karem KL, Sternberg MR, et al. Seroprevalence of human papillomavirus type 16 infection in the United States. *J Infect Dis.* 2002;186(10): 1396-1402.
36. Desai S, Chapman R, Jit M, et al. Prevalence of human papillomavirus antibodies in males and females in England. *Sex Transm Dis.* 2011;38(7): 622-629.
37. Antonsson A, Green AC, Mallitt KA, et al. Prevalence and stability of antibodies to 37 human papillomavirus types—a population-based longitudinal study. *Virology.* 2010;407(1): 26-32.
38. Michael KM, Waterboer T, Sehr P, et al. Seroprevalence of 34 human papillomavirus types in the German general population. *PLoS Pathog.* 2008;4(6): e1000091.
39. Harper DM, Franco EL, Wheeler C, et al. Efficacy of a bivalent L1 virus-like particle vaccine in prevention of infection with human papillomavirus types 16 and 18 in young women: a randomised controlled trial. *The Lancet.* 2004;364(9447): 1757-1765.
40. Wijziging HPV-vaccinatieschema: 1 prik minder (2014). *Infectieziekten Bulletin.* 25(2): 53.

## APPENDIX

### R-code for estimating the mixture model parameters using JAGS

```
#
# Initialization
#

# Set working directory
setwd("<path>")

# Load packages
library(splines)
library(parallel)
library(rjags)

# Setup cluster
n.cores <- detectCores()
cl <- makeCluster(n.cores)

# JAGS function for parallel computing
jags.par <- function(X, data, inits, variable.names, n.iter, n.burnin,
thin) {
  library(rjags)
  setwd("<path>")
  jags.mod <- jags.model(file = "model.txt", data = data,
inits = inits, n.chains = 1, n.adapt = 500, quiet = TRUE)
  update(jags.mod, n.iter = n.burnin)
  coda.samples(model = jags.mod, variable.names = variable.names,
n.iter = n.iter, thin = thin)
}

# Mixed model P-spline setup function
PSplineSetup <- function(x, x.min = min(x), x.max = max(x),
k = 15, spline.deg = 3, diff.ord = 2) {

  # B-spline basis
  dx <- (x.max-x.min)/(k-spline.deg)
  knots <- seq(x.min-spline.deg*dx, x.max+spline.deg*dx, by = dx)
  B <- spline.des(knots = knots, x = x, ord = spline.deg+1, outer.
ok = TRUE)$design

  # Difference operator matrix
  D <- diff(diag(k), diff = diff.ord)

  # Re-parameterize B and D into X (fixed effects) and Z (random effects)
  X <- B%*%outer(knots[1:k], 0:(diff.ord-1), "^(^)")
  Z <- B%*%t(D)%*%solve(tcrossprod(D))
  return(list(X = X, Z = Z, n = nrow(X), q = ncol(X), m = ncol(Z)))
}

#
# Simulate log.hpv16 data
# (similar to the real HPV16 data)
#
```

```

# 8000 persons in total
n <- 8000
# About 50% is male (sex = 1)
p.M <- 0.5
# Component densities:
mu.M <- c(-1, 3); sigma.M <- c(1, 1.5)
mu.F <- c(-1, 5); sigma.F <- c(1, 1.5)
# Create dataframe
sim.data <- data.frame(
  age = sample(0:80, size = n, replace = TRUE),
  sex = sample(1:2, size = n, replace = TRUE, prob = c(p.M, 1-p.M)))
sim.data <- within(sim.data, {
  p.m <- SSlogis(age, Asym = 0.45, xmid = 20, scal = 3)
  p.f <- SSlogis(age, Asym = 0.25, xmid = 20, scal = 3)
  I.m <- rbinom(n, 1, p.m)
  I.f <- rbinom(n, 1, p.f)
  log.hpv16 <- ifelse(sex==1,
    (1-I.m)*rnorm(n, mu.M[1], sigma.M[1]) + I.m*rnorm(n, mu.M[2],
    sigma.M[2]),
    (1-I.f)*rnorm(n, mu.F[1], sigma.F[1]) + I.f*rnorm(n, mu.F[2],
    sigma.F[2]))
  rm(p.m, p.f, I.m, I.f)
})

# Detection limit of serological assay
lower.limit <- log(0.08)

# Data modifications
sim.data <- within(sim.data, {
  # Below detection limit: log.hpv16.cat = 0, above detection limit:
  # log.hpv16.cat = 1 (JAGS convention, see help dinterval)
  log.hpv16.cat <- ifelse(log.hpv16 <= lower.limit, 0, 1)

  # Below detection limit log.hpv16 = NA
  log.hpv16 <- ifelse(log.hpv16<=lower.limit, NA, log.hpv16)

  # Infection status: everyone under age 10 is 0 (=negative),
  # everyone older than 10 is NA (=to estimate)
  inf <- ifelse(age<10, 0, NA)
})

#
# P-spline setup
#

# Split datasets into males ("1") and females ("2")
sim.data.split <- with(sim.data, split(sim.data, f = sex))

# P-spline setup
psp.data.split <- lapply(
  x = sim.data.split,
  FUN = function(x) PSplineSetup(x = x$age-35, x.min = 0-35, x.max = 80-
35, k = 15))
# Predictions
psp.pred <- PSplineSetup(x = 0:80-35, x.min = 0-35, x.max = 80-35,
k = 15)

```

```

#
# Fit mixture model with JAGS
#

# Model
model.string <- "model {
# Likelihood data males
for (i in 1:n.P2.M) {
y.P2.M[i] ~ dnorm(mu.M[inf.P2.M[i]+1], tau.M[inf.P2.M[i]+1])
y.P2.M.cat[i] ~ dinterval(y.P2.M[i], lower.limit)
inf.P2.M[i] ~ dbern(p.P2.M[i])
logit(p.P2.M[i]) <- sum(X.P2.M[i, 1:q]*beta.P2.M[1:q]) +
sum(Z.P2.M[i, 1:m]*b.P2.M[1:m])
}
# Likelihood data females
for (i in 1:n.P2.F) {
y.P2.F[i] ~ dnorm(mu.F[inf.P2.F[i]+1], tau.F[inf.P2.F[i]+1])
y.P2.F.cat[i] ~ dinterval(y.P2.F[i], lower.limit)
inf.P2.F[i] ~ dbern(p.P2.F[i])
logit(p.P2.F[i]) <- sum(X.P2.F[i, 1:q]*beta.P2.F[1:q]) +
sum(Z.P2.F[i, 1:m]*b.P2.F[1:m])
}

# Predictions
for (i in 1:n.pred) {
logit(p.pred.P2.M[i]) <- sum(X.pred[i, 1:q]*beta.P2.M[1:q]) +
sum(Z.pred[i, 1:m]*b.P2.M[1:m])
logit(p.pred.P2.F[i]) <- sum(X.pred[i, 1:q]*beta.P2.F[1:q]) +
sum(Z.pred[i, 1:m]*b.P2.F[1:m])
}

# Prior random effects
for (j in 1:m) {
b.P2.M[j] ~ dnorm(0, tau.b.P2.M)
b.P2.F[j] ~ dnorm(0, tau.b.P2.F)
}

# Hyperprior random effects
tau.b.P2.M ~ dgamma(1, 0.01)
tau.b.P2.F ~ dgamma(1, 0.01)

# Prior fixed effects
for (j in 1:q) {
beta.P2.M[j] ~ dnorm(0, 0.01)
beta.P2.F[j] ~ dnorm(0, 0.01)
}

# Prior Normal components
# (assumed different between males and females for pos, equal for neg)
mu.M[1] ~ dnorm(0, 0.01)
mu.F[1] <- mu.M[1]
mu.M[2] <- mu.M[1]+d.mu.M
mu.F[2] <- mu.F[1]+d.mu.F
d.mu.M ~ dnorm(0, 0.01)I(0, )
d.mu.F ~ dnorm(0, 0.01)I(0, )
tau.M[1] ~ dgamma(1, 0.01)
tau.F[1] <- tau.M[1]

```

```

tau.M[2] ~ dgamma(1, 0.01)
tau.F[2] ~ dgamma(1, 0.01)
}"

```

```

# Data

```

```

data.list <- list(
  n.P2.M = psp.data.split[["1"]]$n,
  n.P2.F = psp.data.split[["2"]]$n,
  q = psp.data.split[["1"]]$q,
  m = psp.data.split[["1"]]$m,
  y.P2.M = sim.data.split[["1"]]$log.hpv16,
  y.P2.F = sim.data.split[["2"]]$log.hpv16,
  y.P2.M.cat = sim.data.split[["1"]]$log.hpv16.cat,
  y.P2.F.cat = sim.data.split[["2"]]$log.hpv16.cat,
  lower.limit = lower.limit,
  inf.P2.M = sim.data.split[["1"]]$inf,
  inf.P2.F = sim.data.split[["2"]]$inf,
  X.P2.M = psp.data.split[["1"]]$X,
  X.P2.F = psp.data.split[["2"]]$X,
  Z.P2.M = psp.data.split[["1"]]$Z,
  Z.P2.F = psp.data.split[["2"]]$Z,
  n.pred = psp.pred$n,
  X.pred = psp.pred$X,
  Z.pred = psp.pred$Z)

```

```

# Inits

```

```

inits.fun <- function() with(data.list, list(
  y.P2.M = ifelse(is.na(y.P2.M), lower.limit-0.1, NA),
  y.P2.F = ifelse(is.na(y.P2.F), lower.limit-0.1, NA),
  b.P2.M = rep(0, m),
  b.P2.F = rep(0, m),
  tau.b.P2.M = 1,
  tau.b.P2.F = 1,
  beta.P2.M = rep(0, q),
  beta.P2.F = rep(0, q),
  mu.M = c(0, NA),
  d.mu.M = 3,
  tau.M = rep(1, 2),
  mu.F = c(0, NA),
  d.mu.F = 3,
  tau.F = c(1, 1),
  .RNG.name = "base::wichmann-Hill", .RNG.seed = sample(1:10000, 1)
))

```

```

# Run JAGS

```

```

writeLines(model.string, "model.txt")
jags.model(file = "model.txt", data = data.list,
  inits = inits.fun, n.chains = 1, n.adapt = 500)
clusterExport(c1, varlist = "data.list")
post.mcmc <- as.mcmc.list(parSapply(c1, x = 1:4, FUN = jags.par,
  data = data.list,
  inits = inits.fun, n.iter = ceiling(50000/4), n.burnin = 500, thin = 10,
  variable.names = c(
    "tau.b.P2.M", "tau.b.P2.F",
    "beta.P2.M", "beta.P2.F",
    "mu.M", "mu.F",
    "tau.M", "tau.F",

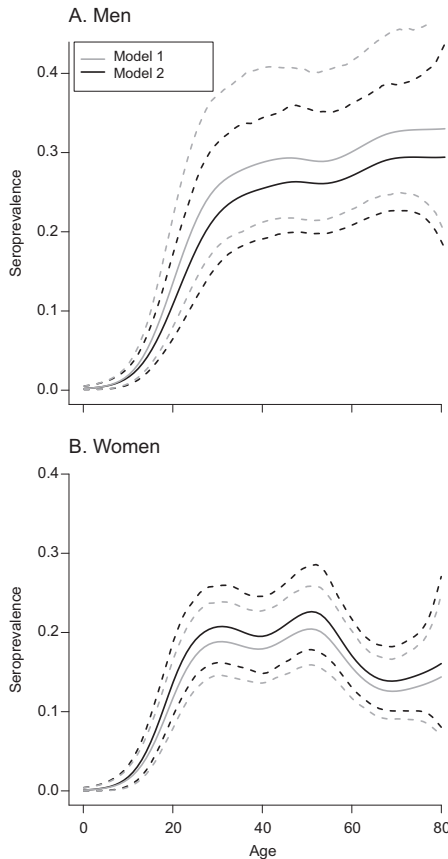
```

```
"p.pred.P2.M", "p.pred.P2.F",  
"deviance", "b.P1.F", "b.P2.F"  
)  
)
```

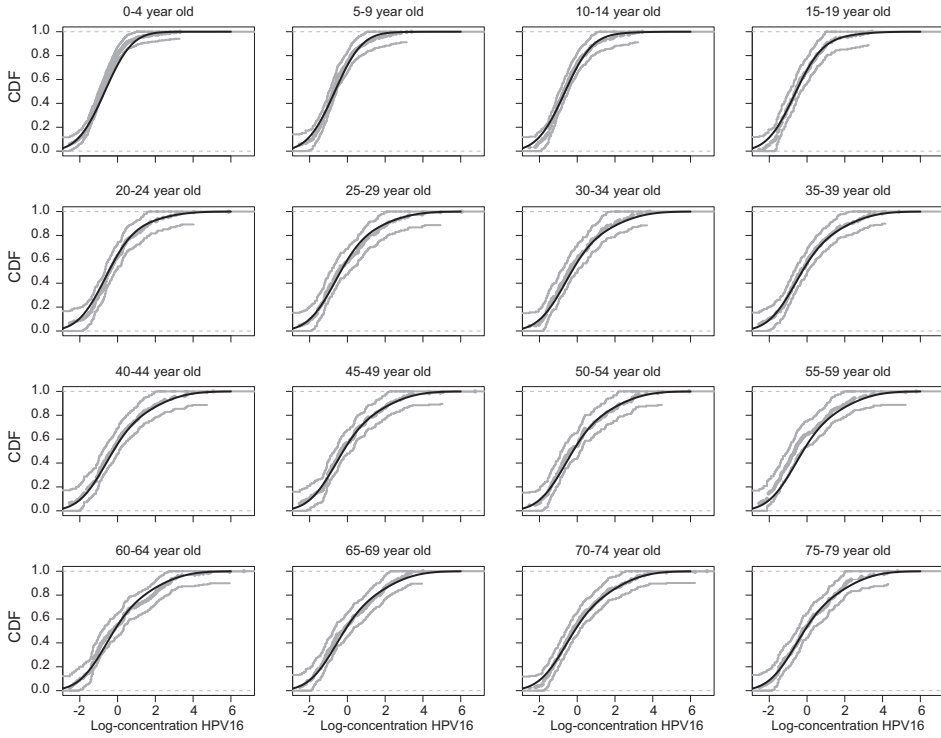
## SUPPLEMENTARY FIGURES AND TABLES

**Appendix Table 1.** Parameter estimates and 95% credible intervals of model 2 (seronegative component density the same for men and women, seropositive component density gender-specific). Both the seronegative and seropositive component density are normally distributed with given mean and standard deviation (sd).

Men	Seronegative component		Seropositive component	
	Mean	Sd	Mean	Sd
	-0.68 (-0.72, -0.64)	1.14 (1.11, 1.17)	1.90 (1.34, 2.24)	1.43 (1.24, 1.72)
Women	Seronegative component		Seropositive component	
	Mean	Sd	Mean	Sd
	-0.68 (-0.72, -0.64)	1.14 (1.11, 1.17)	2.53 (2.21, 2.81)	1.44 (1.28, 1.63)



**Appendix Figure 1.** Difference in seroprevalence when using Model 1 (both mixture component densities sex-specific) or Model 2 (only seropositive component density sex-specific). Left figure denotes male seroprevalence, right figure female seroprevalence.



**Appendix Figure 2.** Model fit of the data for men. Gray lines denote the empirical cumulative distribution function and 95% confidence interval of the HPV16 log-concentration for men, per 5-year age group. The black line denotes the model-predicted cumulative distribution function.