South African Statistical Journal
Vol. 59, No. 1, pp. 37–58
https://doi.org/10.37920/sasj.2025.59.1.3

A Bayesian mixture model accounting for individual heterogeneity in response to pathogenic infection

Adelino Martins[1], *Niel Hens*[2,3] *and Steven Abrams*[2,4]

[1]Department of Mathematics and Informatics, Eduardo Mondlane University, Maputo, Mozambique
[2]Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Data Science Institute, UHasselt, Diepenbeek, Belgium
[3]Centre for Health Economics Research and Modelling Infectious Diseases, Vaccine and Infectious Disease Institute, Antwerp, Belgium
[4]Global Health Institute, Department of Family Medicine and Population Health, University of Antwerp, Antwerp, Belgium

The analysis of multivariate serological data derived from blood serum samples and tested for the presence of antibodies against multiple pathogens gained attention in recent years. Despite the common use of a so-called threshold approach to classify individuals as seronegative or -positive, limitations of such an approach have been reported in the literature, with the subjective choice of the threshold being the most important. Here, we consider a Bayesian mixture approach to model continuous IgG antibody concentrations directly while accounting for the presence of individual heterogeneity and implied association between antibody titer levels for two infections. We fitted the proposed model to Belgian bivariate serological data on the varicella-zoster virus (VZV) and parvovirus B19 (PVB19). Given the existing body of evidence with respect to possible reinfections with PVB19, we investigated whether models explicitly accounting for waning of humoral immunity improved model fit. Our results showed that although after a steep rise with age, the observed seroprevalence for PVB19 decreases between the ages of 20 and 40, the mean IgG antibody concentration remains constant with age among individuals in the seropositive component. This could provide evidence of a direct impact of reinfections with PVB19 on the observed IgG antibody levels, while individuals with loss of humoral immunity after natural infection imply an increase in susceptibility. For VZV, the mean IgG antibody levels slightly decrease with increasing age among seropositive individuals, indicating only very limited waning of humoral immunity as age-dependent seroprevalence estimates are monotonically increasing with increasing age. In general, based on our analyses, we showed that mixture models provide additional insights concerning the waning of humoral immunity as compared to more traditional frailty approaches, which focus on estimating the seroprevalence solely while the model is sufficiently flexible to capture observed dynamics in IgG antibody decay.

Keywords: Antibody waning, Frailty, Markov chain Monte Carlo, Mixture models, Serological survey data.

Corresponding author: Adelino Martins (hawitamartins53@gmail.com)
MSC2020 subject classifications: 62F15, 62H12, 62P10, 62N99

1. Introduction

In modelling infectious diseases, serological survey data constitute an important data source for the estimation of key epidemiological parameters describing transmission dynamics and past infection for the pathogen(s) under study. In particular, residual serum samples are collected and tested for the presence of, for example, immunoglobulin G (IgG) antibodies which are formed in response to an infecting organism (Hens et al., 2012). The analysis of such data can be performed either by modelling continuous IgG antibody titer concentrations directly or more commonly by using a dichotomised version of the serological data (i.e., using the so-called threshold approach) thereby classifying individuals as seropositive or -negative based on a threshold value often supplied by the test manufacturer (Bollaerts et al., 2012; Hens et al., 2012).

Limitations of the threshold approach have been reported in the literature with the subjective choice of the threshold being the most important one (Hardelid et al., 2008; Hens et al., 2012; Bollaerts et al., 2012). Therefore, direct modelling of the continuous IgG antibody titer concentrations based on a finite mixture model avoids the specification of one (or more) threshold value(s) (Bollaerts et al., 2012). Several authors have considered mixture models for the analysis of continuous antibody titer concentrations for various infections (see, e.g., Gay, 1996; Gay et al., 2003; Baughman et al., 2006; Vyse et al., 2006; Bollaerts et al., 2012; Vink et al., 2016). However, none of these authors accounted for individual heterogeneity in the acquisition of infections when relying on mixture models (Coutinho et al., 1999; Hens et al., 2009).

In recent years, it has been demonstrated that ignoring such individual heterogeneity leads to bias when estimating epidemiological parameters such as the basic reproduction number and critical vaccination coverage (see, e.g., Martins et al., 2019). An additional complexity when modelling serological data is the fact that after an initial humoral immune response following infection, IgG antibody levels may wane with time since infection (Held et al., 2019). In the absence of longitudinal persistence data, which would enable a direct investigation of the evolution of IgG antibody levels and inference related to waning rates, presumed infection dynamics need to be contrasted with observed cross-sectional seroprevalence data under the assumption of time homogeneity (Goeyvaerts et al., 2011) or with serial serological data (i.e., comprising several cross-sectional serological surveys) in case of time heterogeneity. Here we will focus on a single cross-sectional serological survey studying two pathogens under the assumption of endemicity. In serological survey data, one typically measures IgG antibody levels for multiple pathogens based on a serum sample of an individual. Given the possible association between antibody titer measurements, a bivariate or paired assessment (i.e., relying on data for two pathogens) requires the use of a model accommodating association in occurrence of past infection and the imposed antibody levels upon infection.

In this manuscript, we propose novel methodology to model IgG antibody titer data on two pathogens by means of a finite mixture model, while explicitly accommodating individual heterogeneity in infection risks, thereby building upon the work of Abrams (2015) and Bollaerts et al. (2012), respectively. Furthermore, the model allows for capturing any decay in mean antibody titer concentration with age, at least for those individuals who experienced infection in the past. Parameter estimation is performed in a Bayesian framework using Markov chain Monte Carlo (MCMC) sampling.

Our manuscript is organised as follows. We start by introducing a motivating example of bivariate serological data on varicella-zoster virus (VZV) and parvovirus B19 (PVB19) from Belgium in

Section 2. In Section 3, we present the basic definition of bivariate finite mixture models, extensions towards incorporating individual heterogeneity in infection risk and modelling association in the acquisition of the two pathogens under study, and the inclusion of waning immunity dynamics in the model. Furthermore, details concerning the likelihood function and the Bayesian MCMC approach are provided therein. The results of fitting the proposed models to VZV and PVB19 serology are shown in Section 4. Finally, a discussion with regard to the underlying assumptions and potential avenues of further research are included in Section 5.

2. Motivating example

The methodology presented in this manuscript is applied to data collected during a single cross-sectional serological survey on VZV and PVB19 in Belgium anno 2001–2003. Parvovirus B19 is the infectious agent of erythema infectiosum, also known as slapped cheek syndrome or fifth disease (Broliden et al., 2006). The disease is usually mild in children and teenagers, but infection during pregnancy has been associated with miscarriage, intrauterine fetal death, fetal anemia, and non-immune hydrops (Tolfvenstam et al., 2001). Individuals with PVB19 IgG antibodies are generally considered immune to recurrent infection, although reinfection is possible in a minority of cases (Lehmann et al., 2003). VZV is one of eight herpes viruses known to affect humans and vertebrates (Ray and Ryan, 2004). Infection with VZV causes two clinically distinct diseases, namely varicella and herpes zoster disease upon reactivation of the virus after it becoming dormant in the human body (Gnann Jr, 2002). From a humoral immunity perspective, VZV is presumed to provide lifelong detectable and high IgG antibodies (Ray and Ryan, 2004). Blood samples were collected for 3379 different individuals and tested for the presence of IgG antibodies against PVB19 and VZV (Hens et al., 2012). For each individual, the antibody level (expressed as log10 mUI/mL) and age are available. Moroever, in our data application, analyses are confined to the data for which pairs of IgG concentrations are available for both VZV and PVB19 infections. Hence, in total, 2382 complete profiles were identified for individuals with known immunological status with respect to both VZV and PVB19.

Figure 1 displays (1) log-transformed IgG antibody levels for VZV and PVB19 plotted against age (in years), and corresponding histograms collapsed over the age dimension with equivocal ranges depicted using vertical dashed lines (i.e., determined by two thresholds), and (2) the age-specific seroprevalence, determined based on a prespecified cut-off value by the manufacturer of the ELISA test, for VZV and PVB19 with the sise of the dots proportional to the corresponding number of observations. The observed seroprevalence for PVB19 decreases for individuals around the age of 20 years to approximately 40 years and starts to increase again thereafter (see, e.g., lower right panel of Figure 1). As mentioned previously, this decrease in seroprevalence could be the result of waning of humoral immunity leading to potential PVB19 reinfections or re-exposures at higher ages.

3. Methodology

In this section, we briefly review the threshold approach as described by Bollaerts et al. (2012) and Hens et al. (2012). Next, bivariate finite mixture models are introduced and discussed focusing on directly modelling antibody titer concentrations while accommodating individual heterogeneity in the acquisition of infection. Antibody waning is further discussed in Section 3.6.

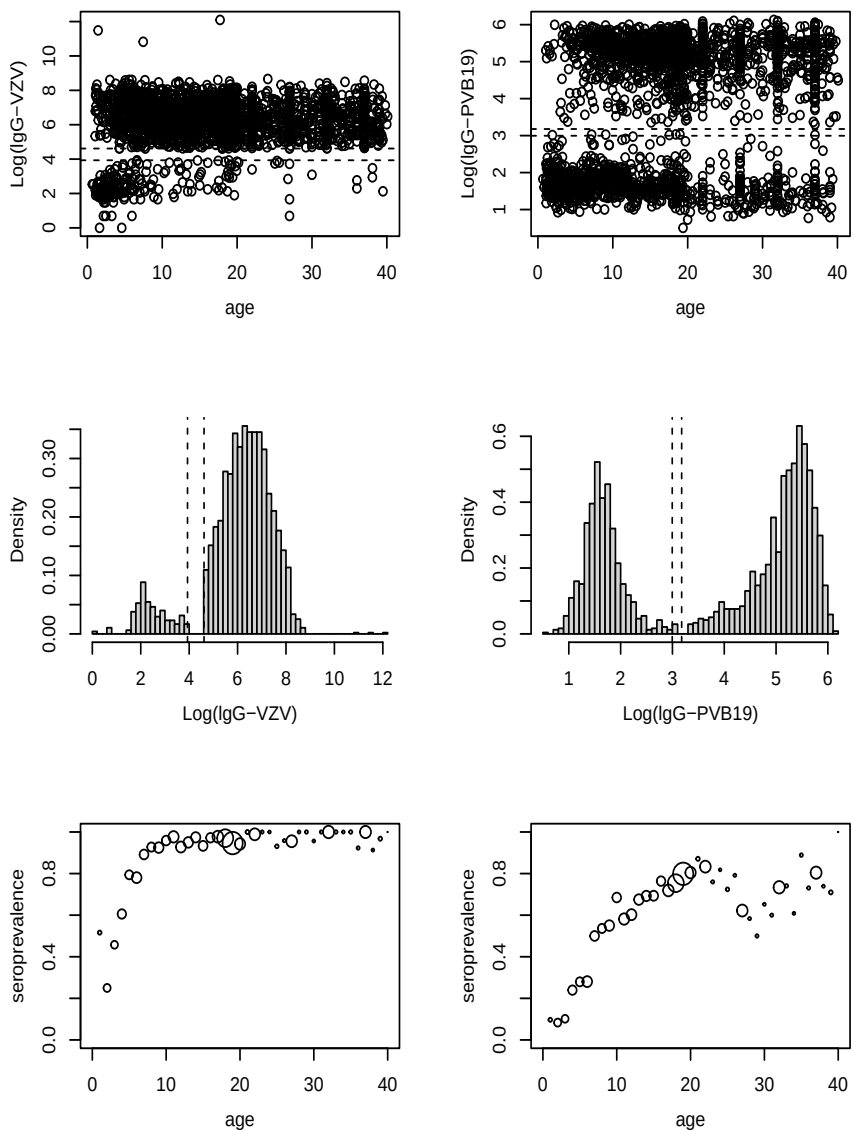

Figure 1. Log-transformed IgG antibody concentrations for VZV (upper left panel) and PVB19 (upper right panel) plotted against age (in years), and corresponding histograms collapsed over the age dimension (middle panels) with equivocal ranges depicted using vertical dashed lines. Age specific seroprevalence for VZV (lower left panel) and PVB19 (lower right panel) with the size of the dots proportional to the corresponding number of observations.

3.1 Threshold approach

Cross-sectional serological survey data are often used to quantify the (sero)prevalence of a certain disease in a population. More specifically, serological data is derived from blood serum samples which are tested for the presence of (IgG) antibodies against one or more pathogens. Such a determination of the presence of antibodies relies on a direct quantification of the antibody titer concentration and individuals having an antibody level exceeding a specific threshold value (often supplied by the manufacturer of the test) are then classified as seropositive (Bollaerts et al., 2012; Hens et al., 2012). Alternatively, persons without humoral immunity against the pathogen under study are considered seronegative. However, one of the disadvantages of this approach is the ad hoc specification of threshold values which is prone to misclassification errors. To overcome the problem of misclassification, prevalence estimates can be corrected based on sensitivity and specificity estimates for the diagnostic test under consideration as proposed by Rogan and Gladen (1978). In case of two threshold values, equivocal (inconclusive) cases are either excluded or all viewed as either seronegative or -positive (Bollaerts et al., 2012; Hens et al., 2012).

3.2 Finite mixture models

We consider a continuous perspective focusing on modelling antibody titer concentrations directly using a bivariate finite mixture model as an alternative to the aforementioned threshold approach, including individual-specific frailty terms to describe heterogeneity in the acquisition of infections and to model association in the acquisition of the two pathogens under study. Mixture models have the advantage that there is no need to predetermine a cut-off value to determine whether an individual has been previously infected or not and no inconclusive (equivocal) observations are discarded (Bollaerts et al., 2012; Hens et al., 2012).

In particular, we focus on bivariate serological data. Let $\boldsymbol{Y} = (Y_1, Y_2)$ represent a bivariate random vector with Y_1 and Y_2 being log-transformed IgG antibody concentrations for pathogen 1 and 2, respectively. Let $\{(\boldsymbol{y}_i, \boldsymbol{x}_i);\ i = 1, 2, ..., n\}$ denote the observed data, where $\boldsymbol{y}_i = (y_{i1}, y_{i2})$ is a realisation of $\boldsymbol{Y}$ for individual i, and $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2})$ represents a vector of individual-specific covariate values. More specifically, $\boldsymbol{x}_{ij}$ is a $(1 \times p_j)$ row vector, $j = 1, 2$, containing individual- and pathogen-specific covariate information. We further assume that the conditional probability density function of $\boldsymbol{Y}_i$, given $\boldsymbol{x}_i$, is characterised by a finite mixture model of the form

$$\begin{aligned} f(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\Psi}, \boldsymbol{\theta}) = \ & \pi_{00}(\boldsymbol{x}_i|\boldsymbol{\theta}) f_{00}(\boldsymbol{y}_i|\psi_{00}) + \pi_{01}(\boldsymbol{x}_i|\boldsymbol{\theta}) f_{01}(\boldsymbol{y}_i|\psi_{01}) \\ & + \pi_{10}(\boldsymbol{x}_i|\boldsymbol{\theta}) f_{10}(\boldsymbol{y}_i|\psi_{10}) + \pi_{11}(\boldsymbol{x}_i|\boldsymbol{\theta}) f_{11}(\boldsymbol{y}_i|\psi_{11}), \end{aligned} \tag{1}$$

where $f_{kq}(\boldsymbol{y}_i|\psi_{kq}) = f_{kq}(y_{i1}, y_{i2}|\psi_{kq})$ are referred to as component-specific bivariate continuous density functions, $\pi_{kq}(\boldsymbol{x}_i|\boldsymbol{\theta})$, $k, q = 0, 1$, are mixture weights (or mixing proportions) with $\sum_{k=0}^{1} \sum_{q=0}^{1} \pi_{kq}(\boldsymbol{x}_i|\boldsymbol{\theta}) = 1$, and $\boldsymbol{\Psi} = (\psi_{kq})_{k,q}$ and $\boldsymbol{\theta}$ represent the vectors of all distinct parameters occurring in the component densities as well as mixing proportions, respectively, and that require estimation (McLachlan and Peel, 2000). In the next section, the mixture model in (1) is refined to account for individual heterogeneity in infection risk.

3.3 Individual heterogeneity in response to infection

Individual heterogeneity in the acquisition of infections has been studied extensively in the past (Coutinho et al., 1999; Farrington et al., 2001; Kanaan and Farrington, 2005), albeit that one often

relies on dichotomised cross-sectional serological survey data. Heterogeneity can be accounted for statistically by introducing individual-level random effects, also referred to as frailty terms in a survival context, in the model structure. A direct generalisation of the frailty approach for the aforementioned binary setting (based on classification of individuals in seropositive and -negative individuals; see Section 3.1) to a continuous IgG antibodies model (1) yields the following conditional probability density function for $\boldsymbol{Y}_i|\boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{z}_i^*$:

$$f(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{z}_i^*, \boldsymbol{\Psi}, \boldsymbol{\theta}) = \pi_{00}(\boldsymbol{x}_i|\boldsymbol{z}_i, \boldsymbol{\theta}) f_{00}(\boldsymbol{y}_i|\boldsymbol{z}_i^*, \psi_{00}) + \pi_{01}(\boldsymbol{x}_i|\boldsymbol{z}_i, \boldsymbol{\theta}) f_{01}(\boldsymbol{y}_i|\boldsymbol{z}_i^*, \psi_{01}) + \pi_{10}(\boldsymbol{x}_i|\boldsymbol{z}_i, \boldsymbol{\theta}) f_{10}(\boldsymbol{y}_i|\boldsymbol{z}_i^*, \psi_{10}) + \pi_{11}(\boldsymbol{x}_i|\boldsymbol{z}_i, \boldsymbol{\theta}) f_{11}(\boldsymbol{y}_i|\boldsymbol{z}_i^*, \psi_{11}), \quad (2)$$

where $\boldsymbol{z}_i = (z_{i1}, z_{i2})$ and $\boldsymbol{z}_i^* = (z_{i1}^*, z_{i2}^*)$ are vectors of individual- and pathogen-specific frailty terms introduced at the level of the mixing proportions and mixture densities, respectively. In this manuscript, the random vectors $\boldsymbol{z}$ and $\boldsymbol{z}^*$ are assumed to be independent.

3.4 Specification of mixing probabilities

The mixture probabilities $\pi_{kq}(\boldsymbol{x}_i|\boldsymbol{z}_i, \boldsymbol{\theta})$, $k, q = 0, 1$, in (2) can be interpreted as (i) the proportion of individuals susceptible to both infections, $\pi_{00}(\boldsymbol{x}_i|\boldsymbol{z}_i, \boldsymbol{\theta})$, (ii) the proportion of individuals infected (in the past) with infection 1, but susceptible to infection 2, $\pi_{10}(\boldsymbol{x}_i|\boldsymbol{z}_i, \boldsymbol{\theta})$, (iii) or vice versa in case of $\pi_{01}(\boldsymbol{x}_i|\boldsymbol{z}_i, \boldsymbol{\theta})$, and (iv) the proportion of individuals with past infection for both pathogens, i.e., $\pi_{11}(\boldsymbol{x}_i|\boldsymbol{z}_i, \boldsymbol{\theta})$. Without loss of generality, we confine attention here to $\boldsymbol{x}_i = (a_{i1}, a_{i2}) \equiv (a_i, a_i)$ given the presence of univariate monitoring times $a_{i1} = a_{i2} = a_i$, i.e., the age of individuals at the cross-sectional sampling time, in the motivating example. In what follows we briefly describe two different classes of models, a mechanistic frailty model and a bivariate random effects model to estimate the mixing probabilities.

3.4.1 Mechanistic approach using frailty models

We first consider a frailty approach to model the mixture probabilities to explicitly link our approach to survival models that have been used in the past. In this model, individual- and pathogen-specific frailty terms $\boldsymbol{z}_i$ are assumed to act multiplicatively on an age-specific baseline force of infection $\lambda_{0j}(\cdot)$ as follows:

$$\lambda_j(a_i|z_{ij}, \boldsymbol{\theta}_j) = z_{ij}\lambda_{0j}(a_i|\boldsymbol{\theta}_j),$$

thereby implying the conditional univariate survival functions (assuming lifelong immunity after infection)

$$S_j(a_i|z_{ij}, \boldsymbol{\theta}_j) = \exp\left(-\int_0^{a_i} \lambda_j(u|z_{ij}, \boldsymbol{\theta}_j)\mathrm{d}u\right) = \exp\left[-z_{ij}\Lambda_{0j}(u|\boldsymbol{\theta}_j)\right],$$

where $\Lambda_{0j}(u|\boldsymbol{\theta}_j)$ refers to the cumulative or integrated baseline hazard function defined as

$$\Lambda_{0j}(u|\boldsymbol{\theta}_j) = \int_0^{a_i} \lambda_{0j}(u|\boldsymbol{\theta}_j)\mathrm{d}u.$$

Under the assumption of conditional independence, the joint conditional survival function is given by

$$S_{12}(a_i|\boldsymbol{z}_i, \boldsymbol{\theta}) = S_1(a_i|z_{i1}, \boldsymbol{\theta}_1)S_2(a_i|z_{i2}, \boldsymbol{\theta}_2),$$

Table 1. Specification of the joint unconditional survival function using different bivariate gamma frailty distributions. The parameters $\alpha_1 > 1, \alpha_2 > 1$ and $k_0, k_1, k_2, n_1, n_2, n_3, \gamma$ are real-valued positive parameters; under the constraint of unit frailty means, we have frailty variances $\sigma^2_{1f} = (k_0 + k_1)^{-1}$ and $\sigma^2_{2f} = (k_0 + k_2)^{-1}$ with correlation between the frailties $\rho = k_0[(k_0 + k_1)(k_0 + k_2)]^{-1/2}$; for the Loáiciga-Leipnik distribution, we set $k_0 = 0$, $k_1 = \alpha_1\gamma$ and $k_2 = \alpha_2\gamma$, and for the Van den Berg distribution $k_1 = n_1$ and $k_2 = n_2$.

Gamma frailty distribution	Joint unconditional survival function $S_{12}(a_i\|\boldsymbol{\Theta})$
Shared	$\left[1 + \sigma_f^2 \{\Lambda_{01}(a_i\|\boldsymbol{\theta}_1) + \Lambda_{02}(a_i\|\boldsymbol{\theta}_2)\}\right]^{-1/\sigma_f^2}$
Yashin	$\left[1 + \sigma^2_{1f}\Lambda_{10}(a_i\|\boldsymbol{\theta}_1) + \sigma^2_{2f}\Lambda_{20}(a_i\|\boldsymbol{\theta}_2)\right]^{-k_0} \left[(1 + \sigma^2_{1f}\Lambda_{10}(a_i\|\boldsymbol{\theta}_1)\right]^{-k_1} \left[1 + \sigma^2_{2f}\Lambda_{20}(a_i\|\boldsymbol{\theta}_2)\right]^{-k_2}$
Kibble-Wicksell	$\left\{\left[(1 + \sigma_f^2\Lambda_{10}(a_i\|\boldsymbol{\theta}_1)\right]\left[1 + \sigma_f^2\Lambda_{20}(a_i\|\boldsymbol{\theta}_2)\right] - \rho\sigma_f^4\Lambda_{10}(a_i\|\boldsymbol{\theta}_1)\Lambda_{20}(a_i\|\boldsymbol{\theta}_2)\right\}^{-1/\sigma_f^2}$
Loáiciga-Leipnik	$\left\{\left[1 + \sigma^2_{1f}\Lambda_{10}(a_i\|\boldsymbol{\theta}_1)\right]^{\alpha_1}\left[1 + \sigma^2_{2f}\Lambda_{20}(a_i\|\boldsymbol{\theta}_2)\right]^{\alpha_2} - \rho\sqrt{\alpha_1\alpha_2}\sigma^2_{1f}\Lambda_{10}(a_i\|\boldsymbol{\theta}_1)\sigma^2_{2f}\Lambda_{20}(a_i\|\boldsymbol{\theta}_2)\right\}^{-\gamma}$
Van den Berg	$\left[1 - \frac{\rho_c n_3^{-1}\sqrt{n_1 n_2}\sigma^2_{1f}\Lambda_{10}(a_i\|\boldsymbol{\theta}_1)\sigma^2_{2f}\Lambda_{20}(a_i\|\boldsymbol{\theta}_2)}{(1 + \sigma^2_{1f}\Lambda_{10}(a_i\|\boldsymbol{\theta}_1))(1 + \sigma^2_{2f}\Lambda_{20}(a_i\|\boldsymbol{\theta}_2))}\right]^{-n_3} \left[1 + \sigma^2_{1f}\Lambda_{10}(a_i\|\boldsymbol{\theta}_1)\right]^{-n_1} \left[1 + \sigma^2_{2f}\Lambda_{20}(a_i\|\boldsymbol{\theta}_2)\right]^{-n_2}$

with $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. Inference is based on a marginalisation of the likelihood function given in Section 3.7, thereby integrating out the unobserved individual-specific frailty terms. Marginalised mixing probabilities are defined as follows (Farrington et al., 2001):

$$\begin{cases} \pi_{00}(a_i|\boldsymbol{\Theta}) &= S_{12}(a_i|\boldsymbol{\Theta}) \\ \pi_{01}(a_i|\boldsymbol{\Theta}) &= S_1(a_i|\boldsymbol{\Theta}_1) - S_{12}(a_i|\boldsymbol{\Theta}) \\ \pi_{10}(a_i|\boldsymbol{\Theta}) &= S_2(a_i|\boldsymbol{\Theta}_2) - S_{12}(a_i|\boldsymbol{\Theta}) \\ \pi_{11}(a_i|\boldsymbol{\Theta}) &= 1 - S_1(a_i|\boldsymbol{\Theta}_1) - S_2(a_i|\boldsymbol{\Theta}_2) + S_{12}(a_i|\boldsymbol{\Theta}) \end{cases}$$

where $S_j(a_i|\boldsymbol{\Theta}_j)$, $j = 1, 2$, and $S_{12}(a_i|\boldsymbol{\Theta})$ are univariate and joint unconditional survival functions, respectively. Note that, $\boldsymbol{\Theta}_j = (\boldsymbol{\theta}_j, \boldsymbol{\xi}_j)$ and $\boldsymbol{\Theta} = (\boldsymbol{\theta}, \boldsymbol{\xi})$ include the vectors $\boldsymbol{\xi}_j$, $j = 1, 2$, with $\boldsymbol{\xi} = (\boldsymbol{\xi}_1, \boldsymbol{\xi}_2)$ associated with the frailty distribution of $\boldsymbol{z}$.

The functional form of the univariate and joint unconditional survival functions is determined by the selected bivariate frailty distribution. For identifiability reasons, the mean of the frailty distribution is typically constrained to be **1** and variance parameters σ^2_{jf} are estimated from the data at hand. Depending on the chosen frailty model/distribution, a correlation coefficient ρ expresses the strength of association among pathogen-specific frailty terms. For example, Martins et al. (2019) proposed a generalisation of the additive correlated gamma model introduced by Yashin et al. (1995), implying a specific bivariate gamma frailty distribution. The proposed bivariate model includes several well-known gamma frailty distributions as special cases. A summary of the different models (or distributions) and the implied joint unconditional survival function is provided in Table 1.

3.4.2 Phenomenological approach using bivariate generalised linear mixed models

As an alternative to the mechanistic approach, a bivariate generalised linear mixed model (BGLMM) can be used to model the mixture probabilities and to account for potential association in acquisition

of both infections. In general, specification of the model is as follows:

$$g\left[\pi_{ij}(a_i|z_{ij},\boldsymbol{\Theta})\right] = m(a_i|\boldsymbol{\Theta}) + z_{ij}, \text{ and } \boldsymbol{z}_i = (z_{i1}, z_{i2})' \sim N\left(\boldsymbol{\mu}_{\boldsymbol{z}}, \boldsymbol{\Sigma}_{\boldsymbol{z}}\right), \; i = 1, \ldots, n, \; j = 1, 2, \tag{3}$$

where $m(\cdot|\cdot)$ refers to the linear predictor including covariate effects, here a_i, $g(\cdot)$ denotes the link function (such as logit, probit or complementary log-log link functions), and with the mean vector $\boldsymbol{\mu}_{\boldsymbol{z}} = (0, 0)'$ and variance-covariance matrix

$$\boldsymbol{\Sigma}_{\boldsymbol{z}} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1^2\sigma_2^2 \\ \rho\sigma_1^2\sigma_2^2 & \sigma_2^2 \end{bmatrix}.$$

To allow for sufficient flexibility in model (3), we consider a generalised additive mixed model approach to describe a potential non-linear though smooth effect of age at the linear predictor scale (i.e., with $m(a_i|\boldsymbol{\Theta})$ a smooth function of age (Ruppert et al., 2003; Wood, 2017). There are several basis functions that can be used, such as B-splines, truncated polynomial splines, natural cubic splines etc. In this paper, we consider penalised splines with truncated power basis functions of degree p to estimate $m(a_i|\boldsymbol{\Theta})$ (Wood, 2017; Hens et al., 2012):

$$m(a_i|\boldsymbol{\Theta}) = \beta_0 + \beta_1 a_i + \cdots + \beta_p a_i^p + \sum_{k=1}^{K} u_k |a_i - \kappa_k|_+^p, \tag{4}$$

with $|a_i - \kappa_k|_+^p = (a_i - \kappa_k)^p$ if $a_i > \kappa_k$ and 0 otherwise, $\boldsymbol{\Theta} = \left(\beta_0, \beta_1, \ldots, \beta_p, u_1, \ldots, u_K\right)$ represents the vector of parameters to be estimated, and $\kappa_1, < \kappa_2 < \cdots < \kappa_k$ are fixed knots. Due to the penalisation in the penalised splines approach, the number of knots and their placement are not of importance, with the number of knots sufficiently large to govern a sufficient degree of smoothness (Eilers and Marx, 2010). In our data application, we considered $p = 3$ and $K = 20$ knots. We refer to Wood (2017) for a more in-depth discussion on generalised additive models and smoothing.

Model (3) with smoothing function (4) is in fact a logistic random effects model in which the spline coefficients can be estimated within the framework of a mixed effects model (Wood, 2017; Hens et al., 2012). Here, the coefficients $\boldsymbol{u} = (u_1, \ldots, u_K)'$ are random effects with $\boldsymbol{u} \sim N\left(\boldsymbol{0}, \sigma_u^2 \boldsymbol{I}\right)$, and we further assume that u and z are independent. The advantage of such model representation is the automatic selection of the smoothing parameter (expressed as $\frac{1}{\sigma_u}$) through the estimation procedure (Wood, 2017; Hens et al., 2012). Note that, the specification of model in (3) implies a different frailty approach in which the frailties do not act multiplicatively on a baseline force of infection, and the underlying frailty distribution is lognormal. Therefore, under the phenomenological approach, the conditional hazard function for immunising infections is given by

$$\lambda_j\left(a_i|z_{ij},\boldsymbol{\Theta}\right) = \exp\left[\gamma_j\left(a_i|z_{ij},\boldsymbol{\Theta}\right)\right]\gamma_j'\left(a_i|z_{ij},\boldsymbol{\Theta}\right), \tag{5}$$

where $\gamma_j'(\cdot|\cdot)$ denotes the derivative of the function $\gamma_j(\cdot|\cdot)$ with respect to a_i, and $\gamma_j\left(a_i|z_{ij},\boldsymbol{\Theta}\right) = \log\left[\Lambda_j\left(a_i|z_{ij},\boldsymbol{\Theta}\right)\right]$ in terms of the cumulative hazard function $\Lambda_j(.|.)$. The choice of the link function $g(.)$ determines the form of $\Lambda_j\left(a_i|z_{ij},\boldsymbol{\Theta}\right)$, for example, let $g(.)$ be a complementary log-log link function, then (see (5))

$$\Lambda_j\left(a_i|z_{ij},\boldsymbol{\Theta}\right) = \exp\left\{g\left[\pi_{ij}(a_i|z_{ij},\boldsymbol{\Theta})\right]\right\} \text{ and } \lambda_j\left(a_i|z_{ij},\boldsymbol{\Theta}\right) = \exp\left\{g\left[\pi_{ij}(a_i|z_{ij},\boldsymbol{\Theta})\right]\right\} m'(a_i|\boldsymbol{\Theta}).$$

3.5 Mixture distributions

In the previous subsections, we proposed either a generalisation of the frailty approach applied in the context of bivariate binary serological survey data to be applicable in the context of continuous IgG antibody titer concentrations for two pathogens measured for the same individuals, or a bivariate generalised additive mixed model formulation to model mixture proportions. Next to individual heterogeneity in the infection risk, individuals are likely to experience a different (initial) humoral immune response following infection with one of the pathogens under study. Although this intrinsic variability in immune response is captured by the shape of the mixture densities, one could opt to include a shared random effect $\boldsymbol{z}_i^* \equiv z_i^*$ at the level of the mean antibody titer concentrations for individual i to disentangle random noise in the measurement process from variability in immune response. Indeed, assuming that $\left(\boldsymbol{Y}_i | z_i^*, K = k, Q = q\right) \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with

$$\begin{aligned} Y_{i1} &= \mu_{kq1} + z_i^* + \epsilon_{ikq1}, \\ Y_{i2} &= \mu_{kq2} + z_i^* + \epsilon_{ikq2}, \end{aligned}$$

and $\boldsymbol{\mu} = (\mu_{kq1}, \mu_{kq2})$, $z_i^* \sim N(0, (\sigma^*)^2)$, $\epsilon_{ikq1} \sim N(0, s_1^2)$, $\epsilon_{ikq2} \sim N(0, s_2^2)$ and assuming independence between z_i^* and these error terms, implies the following variance-covariance structure:

$$\boldsymbol{\Sigma} = \begin{bmatrix} (\sigma^*)^2 + s_1^2 & s_{12} \\ s_{12} & (\sigma^*)^2 + s_2^2 \end{bmatrix},$$

where s_{12} is the covariance between the pathogen-specific measurement errors. Needless to say, allowing the random effect z_i^* to be infection-specific would not be identifiable. In the remainder of the manuscript, we will model $\epsilon_{ikqj}^* = z_i^* + \epsilon_{ikqj}$ altogether by estimating component-specific variance-covariance parameters.

In our data application, we will focus on bivariate normal or skew-normal mixture distributions in the mixture model for the log-transformed antibody titer concentrations. Furthermore, for identifiability reasons, we assume that the component mean vectors $\boldsymbol{\mu}_{kq} = \left(\mu_{kq1}, \mu_{kq2}\right)$ satisfy the following conditions:

$$\mu_{001} \leq \mu_{101} = \mu_{111} \quad \text{and} \quad \mu_{002} \leq \mu_{012} = \mu_{112}. \tag{6}$$

This order restriction avoids label-switching during parameter estimation and is inspired by the fact that mean antibody titer concentrations are larger for seropositive individuals (Evans and Erlandson, 2004; McLachlan and Peel, 2000). To complete the specification of the component densities, two different assumptions about the structure of the component-specific variance-covariance matrices

$$\Sigma_{kq} = \begin{bmatrix} \sigma_{kq1}^2 & \rho_{kq}\sigma_{kq1}^2\sigma_{kq2}^2 \\ \rho_{kq}\sigma_{kq1}^2\sigma_{kq2}^2 & \sigma_{kq2}^2 \end{bmatrix},$$

are considered: (i) constant variance-covariance matrices, i.e., $\Sigma_{00} = \Sigma_{01} = \Sigma_{10} = \Sigma_{11} = \Sigma$, or (ii) Σ_{kq} with $\sigma_{111}^2 = \sigma_{101}^2 \equiv \sigma_{1.1}^2$, $\sigma_{112}^2 = \sigma_{012}^2 \equiv \sigma_{.12}^2$ and correlation coefficients ρ_{00}, ρ_{01}, ρ_{10}, and ρ_{11} estimated from the data. The subscripts of the means μ_{kqj}, variances σ_{kqj}^2, and correlation coefficients ρ_{kq} indicate to which mixture component the model parameter belongs.

3.6 Waning of humoral immunity

After an individual is infected, the body generally depends on its immune system to strive against infection and to give resistance to the disease (Ray and Ryan, 2004). Although next to humoral immunity, a cellular immune response is also part of the overall immune response to pathogenic invasion of the human body, in this paper we primarily focus on humoral immunity. Consequently, here and elsewhere we refer to humoral immunity when talking about an immune response. Knowledge about humoral immunity dynamics and consequently susceptibility is indispensable because the infection can only spread in the population if an infected individual makes effective contact with susceptible individuals (Held et al., 2019). Some infectious diseases confer lifelong immunity (examples include several childhood infections such as VZV and measles) while others only give rise to temporary humoral immunity. In the latter case, individuals lose their humoral immunity over time (though not necessarily their immunity). One of the potential factors which may give rise to repeated outbreaks is waning of acquired humoral immunity (Barbarossa and Röst, 2015).

In the previous sections, we introduced a mixture model for IgG antibody concentrations thereby focusing on the underlying mixture distribution and its parameters with age as a proxy for the time since infection and adding more complexity into the model to describe the antibody waning process. Based on a single cross-sectional serological survey, waning dynamics of IgG antibody titers cannot be inferred directly given the lack of a temporal perspective on antibody kinetics. More specifically, a moderately high IgG antibody can be the result of a higher level that decayed with time since infection or the result of a recent infection inducing an average amount of generated antibodies. Consequently, the assessment of antibody kinetics is a population-averaged one in which the evolution of the overall antibody levels among positive individuals can be studied.

In this serological data application, while VZV infection is assumed to confer lifelong immunity (Abrams et al., 2018), hypotheses of waning of IgG antibodies and reinfections with PVB19 have been advocated in the literature (Schoub et al., 1993; Gay, 1996; Vyse et al., 2006). This has been even further exemplified by the observed decrease in seroprevalence for PVB19 between the age of 20 years to 40 years after which the seroprevalence increases again (see Figure 2). It is noteworthy that such a decrease is also observed in PVB19 seroprevalence in other countries (Goeyvaerts et al., 2011). Here, antibody waning is accounted for by allowing the mean IgG antibody concentration in the seropositive component μ_{11} and μ_{12} to be age-dependent through stratification, whereas model parameters associated with other mixture components are considered constant with age.

Here, we provide a graphical exploration of antibody titer dynamics at the individual level, including individual variation in humoral immune response, and the corresponding effects at the population level. We focus on a non-vaccinated population where humoral immunity is solely a consequence of natural infection. We further ignore maternal passive immunity, i.e., the humoral immune response that results from antibodies passed on from the mother to the child during pregnancy. We investigated three different scenarios to obtain a detailed description of the effects of waning of humoral immunity on both the mixture component associated with seropositive individuals and the corresponding seroprevalence defined using a predefined threshold value τ. In the first scenario (Scenario 1), we consider an infection that confers lifelong humoral immunity at a constant level after humoral immunity build-up, i.e., infection without waning of the humoral immune response. In Scenario 2, we examine a situation in which an infection induces humoral immunity that wanes over time, but without reinfection after loss of humoral immunity. In this scenario, after an individual is infected

and humoral immunity is built up, the IgG levels decrease over time and remain at a level certainly above a predefined threshold value τ. In the third scenario (Scenario 3), we illustrate an infection process with waning of antibody levels after which reinfections are possible to occur.

Figure 2 depicts the impact of all scenarios described above, and the relation with the model approach considered here. For each scenario, the figure shows seropositivity plotted across age (lower panels) and the distribution of log-transformed IgG antibodies (upper panels) measured across the population. Next to that, the middle panels of the figure display the longitudinal evolution of IgG antibody concentrations of four randomly selected individuals with dots indicating the log-antibody titer concentrations observed at data collection (i.e., the monitoring time, here represented by vertical dashed lines, which are typically equal for different pathogens in a cross-sectional serological survey). Serological data consisting of individual IgG antibody concentrations are then translated into a binary immunological status with seropositive (i.e., with IgG levels on the right-hand side of the predefined threshold value – vertical dashed grey line, here denoted protective) or seronegative individuals (i.e., IgG values on the left-hand side of the protective line).

In Scenario 2 (waning without reinfection) no impact on the seroprevalence is observed in the sense that the proportion of seropositive individuals increases monotonically with age. However, in Scenario 3 (waning and potential reinfection) the seroprevalence shows a decrease between the ages of 30 and 50, after a steep monotone rise with age, and starts to increase thereafter again. In general, Scenarios 2 and 3 show similar seroprevalence curves compared to the observed seroprevalence with respect to VZV and PVB19, respectively. In terms of analysis strategies based on the mixture modelling approach considered here, the waning process can be accounted for depending on the scenarios described above. For instance, if humoral immunity induced by an infection is characterised by Scenario 1, the model parameters associated with the mixture components are age-invariant. In contrast, in Scenarios 2 and 3 the waning of IgG antibodies can be accounted for by allowing the mean of the seropositive mixture component to vary with age.

3.7 Bayesian inference

In order to estimate the model parameters of the proposed finite mixture model, we rely on Bayesian inference using MCMC sampling. The marginalised likelihood function for the observed vector of log-transformed IgG antibody levels $\boldsymbol{y}$ can be rendered as

$$L(\boldsymbol{\Psi}, \boldsymbol{\theta}|\mathbf{y}_1, \mathbf{y}_2, a_i) = \prod_{i=1}^{n} \left\{ \prod_{k=0}^{1} \prod_{q=0}^{1} \left[\pi_{kq}(a_i|\theta) f_{kq}(y_i|\psi_{kq}) \right] \right\}.$$

In a practical Bayesian inference, parameters $\boldsymbol{\Psi}$ and $\boldsymbol{\theta}$ are assumed to be random, and specification of the prior distribution $\mathrm{P}(\boldsymbol{\Psi}_0, \boldsymbol{\theta}_0)$ is required (Ntzoufras, 2011). Moreover, all information contained in the data $\mathbf{y}$ about $\boldsymbol{\Psi}$ and $\boldsymbol{\theta}$ is summarised in terms of the posterior density $\mathrm{P}(\boldsymbol{\Psi}, \boldsymbol{\theta}|\mathbf{y}, a_i)$, which is derived using Bayes's theorem (Frühwirth-Schnatter, 2006):

$$\mathrm{P}(\boldsymbol{\Psi}, \boldsymbol{\theta}|\mathbf{y}_i, a_i) \propto L(\boldsymbol{\Psi}|\mathbf{y}_1, \mathbf{y}_2, a_i)\mathrm{P}(\boldsymbol{\Psi}_0)\mathrm{P}(\boldsymbol{\theta}_0). \tag{7}$$

The prior distributions for the model parameters in (7) are chosen to make the distribution proper but diffuse with large variances. We also assume that the priors for all parameters are independent

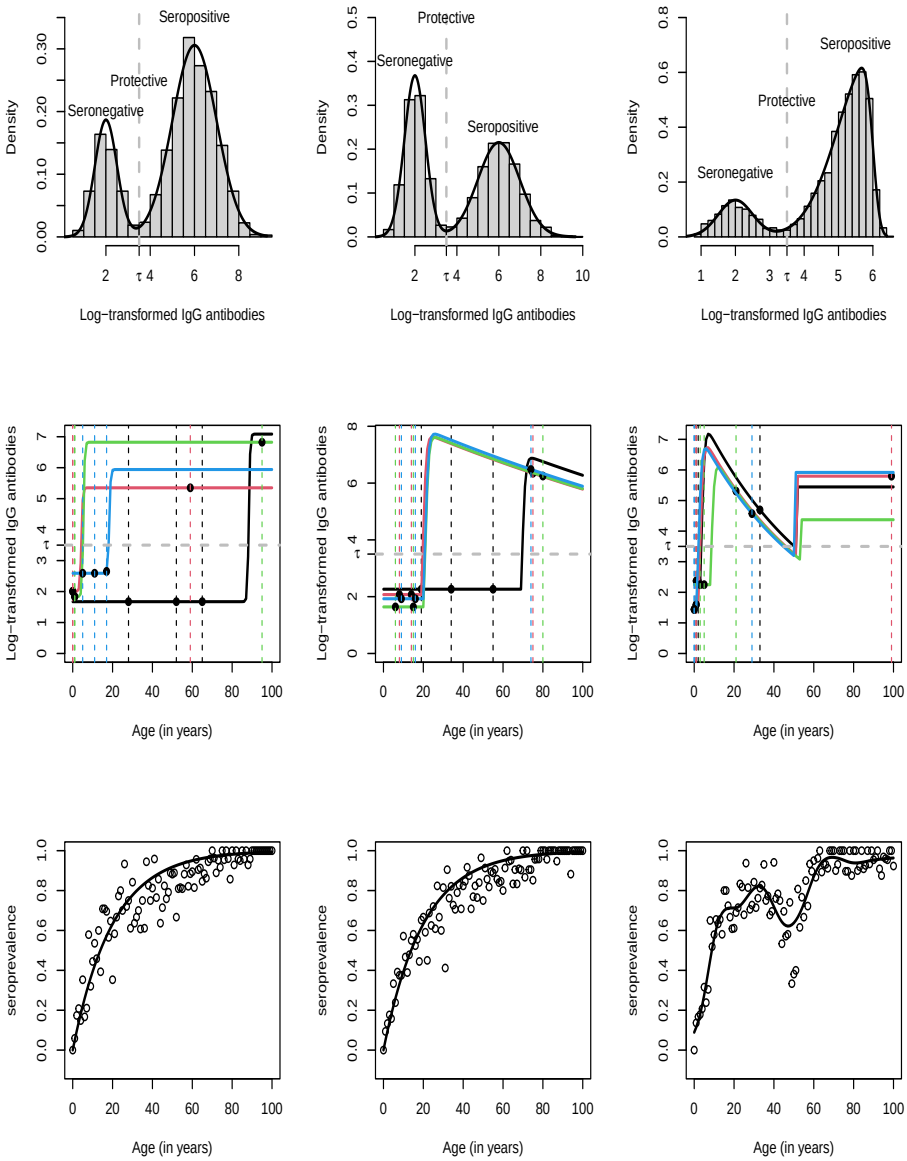

Figure 2. Impact of waning of humoral immunity and relation with our modelling approach. Distribution of log-transformed IgG antibody concentrations collapsed over the age dimension (upper panels). Longitudinal evolution of IgG antibody levels of four randomly selected individuals with vertical dashed lines indicating individual age together with IgG antibody titer concentrations measured at data collection (middle panels). Age-specific seroprevalence (lower panels). Left to right: Scenario 1, Scenario 2 and Scenario 3.

such that the joint prior density in (7) equal the product of the marginal prior distributions, $P(\boldsymbol{\Psi}_0)$ and $P(\boldsymbol{\theta}_0)$.

A vague normal prior with mean zero and variance equal to 1000 is considered for all model parameters with support $]-\infty, +\infty[$. A gamma prior distribution with mean one and variance 100 was assumed for all unknown parameters with support $]0, +\infty[$ and a uniform prior on the unit interval was chosen for non-negative correlation coefficients (by construction of the frailty model) or on the interval $[-1, 1]$ otherwise. Prior distributions for the mean vectors of the bivariate component densities are specified in order to satisfy the restriction imposed in (6), i.e., $\mu_{001} \sim N(0, 1.0 \times 10^3)$, $\mu_{002} \sim N(0, 1.0 \times 10^5)$, $\mu_{012} \sim N(0, 1.0 \times 10^3)\mathrm{T}(\mu_{002},)$, and $\mu_{101} \sim N(0, 1.0 \times 10^3)\mathrm{T}(\mu_{001},)$. Finally, instead of selecting the commonly used Wishart prior distribution for Σ_{kq}, we follow the idea of Turner et al. (2019) thereby specifying a prior distribution for each of the individual elements of the variance-covariance matrix Σ_{kq}.

3.8 Model selection

For model comparison, selection, or averaging, one can measure the predictive accuracy of the fitted Bayesian model (Geisser and Eddy, 1979) using leave-one-out cross-validation (LOO) or the Watanabe-Akaike Information Criterion (WAIC, Watanabe and Opper, 2010), which are methods to estimate out-of-sample predictive accuracy (Vehtari et al., 2017). Disadvantages of the use of LOO, however, have been reported in the literature (Peruggia, 1997; Epifani et al., 2008). More recently, Vehtari et al. (2017) proposed a more efficient approximation to LOO using Pareto-smoothed importance sampling (PSIS). Both the PSIS-LOO and WAIC are easily computed using the loglikelihood evaluated at the posterior MCMC draws of the model parameters from a converged chain (Vehtari et al., 2017) and are implemented in the R package loo. For model comparison, we use this package and we consider both PSIS-LOO and WAIC to select the best model. The model deviance is reported as well.

3.9 Software

MCMC samples of the joint posterior distribution of the model parameters are obtained using Gibbs sampling via the JAGS (Just Another Gibbs Sampler) function and the R2jags package (Plummer, 2003). The JAGS function in R2jags package was used specifically to obtain the results of the final models. As NIMBLE's (Numerical Inference for statistical Models for Bayesian and Likelihood Estimation) package (de Valpine et al., 2023) proved to be fast software in mixture model and it provides option to return a WAIC value (Beraha et al., 2021), we used for model selection. Both packages the R2jags and NIMBLE are implemented in the statistical software R, version 4.3.0 (R Development Core Team 2023). The R program to fit the proposed model can be found in the Supplementary Material (see Appendix D).

4. Data application

We apply the finite mixture model to the bivariate log-transformed antibody titer concentrations for VZV ($i = 1$) and PVB19 ($i = 2$) introduced in Section 2. In particular, we fitted both the bivariate normal and the skew-normal mixture models while exploring various modelling strategies with respect to the seropositive mixture component densities and mixing probabilities. In our data

Table 2. Posterior means, posterior standard deviation (SD) and 95% credible interval for the model parameters obtained by fitting the bivariate mixture models with mixing proportions based on the BGLMM model with logit link functions. *Since the posterior distribution of the variance component is skewed, the posterior median are used as summary measure.

			Normal mixture model		Skew-normal mixture model	
	Notation	Parameter	Estimate (SD)	95% CI	Estimate (SD)	95% CI
Component densities	Ψ	μ_{001}	2.355 (0.049)	[2.260, 2.453]	2.361 (0.049)	[2.267, 2.458]
		μ_{002}	1.661 (0.013)	[1.635, 1.686]	1.641 (0.012)	[1.617, 1.665]
		μ_{111}				
		0.6 – 20	6.650 (0.044)	[6.562, 6.738]	6.648 (0.045)	[6.561, 6.735]
		10 – 20	6.378 (0.031)	[6.321,6.439]	6.379 (0.031)	[6.317, 6.440]
		20 – 35	6.333 (0.040)	[6.255, 6.410]	6.339 (0.040)	[6.261, 6.417]
		35 – 40	6.263 (0.069)	[6.132, 6.398]	6.264 (0.068)	[6.131, 6.398]
		μ_{112}				
		0.6 – 10	5.273 (0.039)	[5.197, 5.349]	5.884 (0.017)	[5.835, 5.936]
		10 – 20	5.163 (0.022)	[5.120, 5.206]	5.838 (0.024)	[5.805, 5.871]
		20 – 35	5.157 (0.028)	[5.101, 5.212]	5.883 (0.024)	[5.837, 5.930]
		35 – 40	5.043 (0.047)	[4.951, 5.130]	5.880 (0.035)	[5.812, 5.948]
		σ^2_{001}	0.410 (0.045)	[0.337, 0.507]	0.412 (0.045)	[0.333, 0.509]
		σ^2_{002}	0.143 (0.008)	[0.129, 0.159]	0.123 (0.007)	[0.100, 0.126]
		σ^2_{111}	0.889 (0.030)	[0.834, 0.949]	0.882 (0.029)	[0.828, 0.940]
		σ^2_{112}	0.335 (0.014)	[0.310, 0.363]	0.024 (0.004)	[0.017, 0.032]
		ρ_{00}	0.039 (0.080)	[-0.131, 0.190]	0.019 (0.110)	[-0.196, 0.238]
		ρ_{01}	-0.118 (0.131)	[-0.364, 0.145]	-0.236 (0.239)	[-0.678, 0.247]
		ρ_{10}	0.159 (0.038)	[0.083, 0.232]	0.202 (0.035)	[0.133, 0.269]
		ρ_{11}	0.247 (0.026)	[0.197,0.297]	0.589 (0.054)	[0.485, 0.690]
Skewness parameters		α	-		-6.073 (0.587)	[-7.329, -5.010]
		ω	-		0.895 (0.043)	[0.817, 0.987]
Mixing proportions	ξ	σ^2	8.716 (5.616, 9.999)*		8.468 (4.861, 9.999)*	
		ρ	1.000 (–)		1.000 (–)	
Model selection		LOO	9760.64		**6046.22**	
		WAIC	9768.81		**6097.31**	
		Deviance	9571.41		**4195.21**	

analysis, we considered two different assumptions for the waning of humoral immunity. To be more specific, first we fitted several models while assuming that after an individual is infected and humoral immunity is built up, antibody levels remain constant over time. Hence, model parameters associated with the mixture components are assumed independent of age. In the second situation, we fitted various models in which antibody waning is accounted for by allowing the mean IgG antibody concentrations in the seropositive component to be age-dependent according to the specification of different age groups. For each model strategy, we assumed (i) a common variance-covariance matrix for the mixture component-specific variance-covariance matrix and (ii) different variance-covariance matrices. For model comparison, we based model selection on the PSIS-LOO (Vehtari et al., 2017), WAIC (Watanabe and Opper, 2010) and deviance. The skew-normal distribution was only assumed for component densities related to the PVB19 positive mixture component, as the distribution of log-transformed PVB19 antibody levels showed left-skewed behavior.

The mean vectors of the component densities corresponding to seropositive individuals for

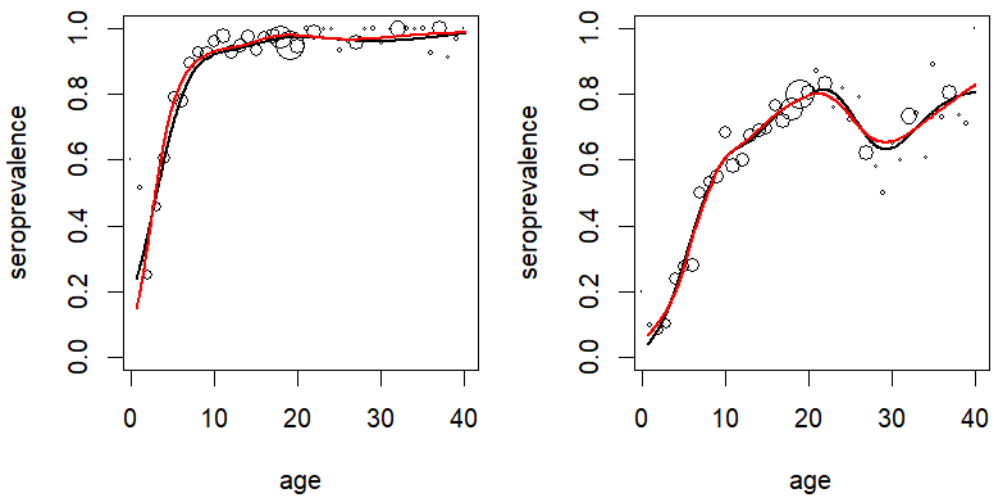

Figure 3. Observed seroprevalence (dots with size proportional to the number of observations), estimated seroprevalence for VZV infection (left panel) and PVB19 infection (right panel) in Belgium based on mixture skew-normal model with bivariate generalised linear mixed model (black line) and threshold approach (red line).

at least one pathogen are modelled using a piecewise constant function of age with age groups: $0.6-10$, $10-20$, $20-35$ and $35-40$. The age groups considered here are similar to those used by Goeyvaerts et al. (2011), who considered a Maternally derived immunity-Susceptible-Infectious-Recovered-Waned-boosting (MSIRWb) compartmental model with age-dependent antibody waning immunity with the optimal cut-off value $H = 35$ years. More specifically, these authors found that after an individual recovers from PVB19 infection, the decay in antibody levels may depend on age, more likely after age 35 years. We also explored the use of other age categories, however, conclusions did not change (results not shown here - see Table B.6 in the Supplementary Material).

The age-dependent mixture probabilities are modelled using two different model strategies. First we consider a mechanistic approach that mimics the underlying infection process through the specification of a hazard model (cfr. frailty model), though interpretation relies on the restrictive assumption of lifelong immunity after infection. More specifically, we consider correlated gamma frailty models (see Table 1) with different parametric baseline hazard functions. Secondly, we adopted a purely statistical or phenomenological approach in which the mixture probabilities are modelled using a BGLMM. These mixture probabilities as a function of age were modelled using penalised splines with truncated polynomial basis functions and $K = 20$ knots, with the $k-th$ knot placed at the $k/(K+1)$ sample quantile of the age distribution (Ruppert, 2002). As advocated previously, two different assumptions about the structure of the component-specific variance-covariance matrices were

considered. As MCMC sampling showed to be very time-consuming, a single chain with 100 000 iterations was run with a burn-in of 35 000 iterations. The remaining 65 000 samples were used for posterior inference related to the model parameters. Convergence of the chain was assessed using graphical diagnostic tools such as trace plots, histograms of posterior samples (i.e., representing the posterior density), and auto-correlation plots (see Figures C1–C6 in Appendix C of the Supplementary Material).

In general, the models with component-specific variance-covariance matrices fit the observed data better. In addition, the models ignoring potential antibody waning perform worse as compared to those with age-dependent means thereby providing evidence in favor of waning of humoral immunity (see model comparison in Table B.1-Table B.5 in Appendix B of the Supplementary Material). Among all fitted models considered here, results are presented for the bivariate mixture models with mixing proportions based on the BGLMM with logit-link function (which outperformed all fitted models based on the PSIS-LOO, WAIC and Deviance). Posterior means, standard deviation in parenthesis and the 95% credible intervals of the mixture probabilities and component density parameters with regard to normal component densities or skew-normal densities are shown in Table 2.

We observe that some of the estimated model parameters and their standard deviations are very similar in both models. However, the estimated means for seropositive individuals with regard to PVB19 are slightly higher in the skew-normal mixture model across all age groups, as the model accounts for negative skewness ($\alpha = -6.073$ with posterior standard deviation 0.587). The negative skewness is consistent with the pronounced left-skewed shape observed in the PVB19 seropositive population. These findings are in line with the results of earlier work of Del Fava (2012), albeit in a univariate setting ignoring individual heterogeneity and association in acquisition of PVB19 (and VZV). The model assuming a skew-normal distribution for the antibody levels of PVB19 seropositive individuals outperforms the bivariate normal mixture model. Figure 4 shows the distributions of the log-transformed IgG antibodies for VZV and PVB19 together with the estimated mixture density. The estimated density functions for VZV agree closely between normal and skew-normal models, but differ for PVB19. In addition, Figure 3 provides the estimated seroprevalence curves based on the mixture model (black line) and threshold (red line) approaches. Visual comparison of the figures shows that the estimated seroprevalence curves for VZV and PVB19 based on the mixture model agree closely with threshold approach.

5. Discussion

In this manuscript, we proposed the use of a finite mixture model to describe bivariate continuous antibody titer data in the presence of individual heterogeneity and implied association regarding the acquisition of two infections. More specifically, bivariate cross-sectional serological survey data is analysed using a bivariate generalised linear mixed models and a well-known frailty approach previously introduced in the context of dichotomised data and now refined to encompass relevant dynamics in humoral immunity responses following infection. In particular, the frailty terms are incorporated at the level of the mixing probabilities and/or at the level of the mixture densities, thereby leading to a generalisation of existing frailty models considered in the field of infectious disease modelling. At the level of mixing probabilities, the latent frailties are assumed to act multiplicatively on the age-specific baseline force of infection and the time of infection (i.e., the age of

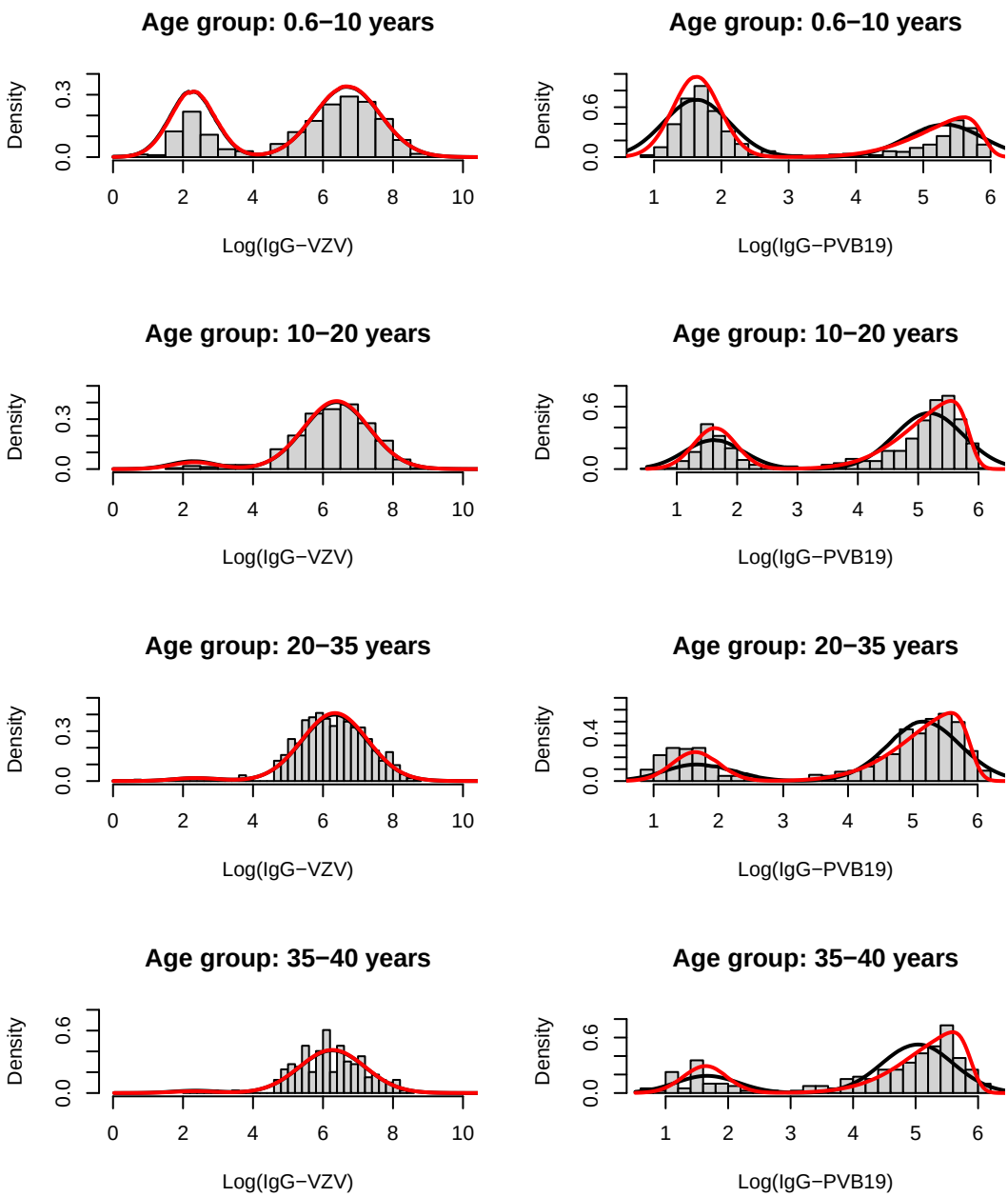

Figure 4. Histogram together with estimated density functions for VZV (left panel) and PVB19 (right panel) log-transformed IgG antibodies by age group. Bivariate normal (black line) and skew-normal (red line) mixture models with different variance-covariance matrices.

individuals at the cross-sectional sampling time, in our data application) are assumed conditionally independent given the frailty terms. Moreover, a shared random effect is introduced at the level of the mean IgG antibody concentration and is assumed to be independent of the frailty terms introduced at the level of the mixing proportions. The random effects introduced in the component densities, however, allow us to capture the differences in humoral immune response following infection, and also to disentangle random noise in the measurement process from variability in immune response. However, when doing so, this leads to identifiability problems and consequently the random effects are modelled together by estimating component-specific variance-covariance parameters.

Different submodels of the general model are applied to bivariate serological data on PVB19 and VZV from Belgium, in which we assumed no disease-related mortality. Given previous evidence of waning of humoral immunity for PVB19, a decay in mean antibody titer concentrations in its seropositive component is allowed for. Typically, without a longitudinal cohort study that would enable us to investigate antibody waning directly, the decay in IgG antibody levels was accounted for by allowing the mean IgG antibody concentration in the seropositive population to be age-dependent through stratification. Our analysis showed that the estimated seroprevalence for PVB19 is characterised by a steep increase with increasing age, as a result of infections among young children, followed by a decrease between the age of 20 to 40 years, after which the seroprevalence increases again. Moreover, the evolution of the mean antibody titer concentration is rather constant across age groups, indicating that despite a decay in humoral immunity at the individual-level, population-level mean antibody titer levels remain unchanged because of reinfections with PVB19 among 20-40 year olds. Given the risk of spontaneous abortion after PVB19 infection during pregnancy, waning of humoral immunity in 20-40 year olds could be responsible for an excess of miscarriage and fetal death. For VZV, the seroprevalence is monotonically increasing, indicating that varicella infection is responsible for high levels of humoral immunity persisting for life. The mean antibody levels show a slight decrease with increasing age among seropositive individuals, however, not to an extent that seroprotection is not ensured for life.

Although normal mixture components are often considered as a default option in many applications in which a mixture model formulation is considered, applying a model with skewed-normal mixture densities outperformed the normal densities for the data at hand. Needless to say, the left-skewness in the positive components for PVB19, irrespective of age, could be induced by a limit of detection regarding the quantification of the respective antibody titer concentrations. From a statistical perspective, one could easily adjust the likelihood function to encompass the right-censored nature of certain observations. However, information on whether observations are censored or not is lacking in the dataset. An additional limitation is the fact that a high IgG antibody titer concentration could represent a relatively recent infection taking into account the gradual build-up of humoral immunity levels. Similarly, lower antibody titer concentrations could indicate historical past infection, a low antibody response to a recent infection, or a recent but mild infection. However, with a single cross-sectional serological survey, these scenarios cannot be distinguished (Nhat et al., 2017). Typically such data relies on assumptions of time homogeneity to estimate key epidemiological parameters, at least in pre-vaccination settings(Held et al., 2019).

We adopted mechanistic and phenomological modelling approaches to estimate the age-dependent mixture probabilities. These mixture probabilities relate to the true prevalence of individuals with past infection experience and avoids issues related to the estimation of the prevalence based on

seroprevalence derived from dichotomisation of the continuous antibody concentrations. The mechanistic approach, although attractive from an interpretation point of view, is not preferred here due to its implicit assumption of lifelong humoral immunity after infection, which has been questioned before for PVB19 (Abrams et al., 2018). An in-depth investigation of infection dynamics could serve as a further extension of the approach outlined in this paper, thereby combining continuous antibody dynamics with individual heterogeneity in humoral immune response following infection. Whereas the phenomenological approach provides a smooth estimate of the age-dependent (sero)prevalence, these estimates do not translate directly to an estimate of the underlying force of infection. Needless to say, due to the non-monotonicity of the implied seroprevalence estimate, a naive estimate of the force of infection obtained by taking the derivative of the seroprevalence would be negative (in certain age intervals) (Hens et al., 2012). Note, however, that such an estimation of the force of infection relies on lifelong persistence of humoral immunity as well. This is perceived as a disadvantage of a phenomological approach as the statistical model does not explicitly account for potential reinfections and their repercussions on antibody dynamics.

The refined mixture model was implemented in the Bayesian paradigm and inference regarding the model parameters is based on MCMC sampling. We caution readers regarding the interpretation of the model selection criterion considered here (PSIS-LOO and WAIC), as assessment of the goodness-of-fit of the models and model selection in the application of mixture models is a challenging aspect (Grimm et al., 2021). The PSIS-LOO is estimated using Pareto smoothed importance sampling (PSIS). Essentially, the approximation is more accurate by fitting a Pareto distribution to the upper tail of the distribution of the importance weights (Vehtari et al., 2017). To conclude, we wish to mention that in general, based on our analyses, we showed that the mixture model provides additional insights concerning waning of IgG antibody concentrations as compared to more traditional frailty approaches while the model is sufficiently flexible to capture observed dynamics in IgG antibodies. Furthermore, the model accounts for association in the acquisition of the pathogens under study through the specification of random effects. This is an advantage of using our proposed Bayesian bivariate finite mixture models rather than a binary classification of IgG serological status, distinguishing seropositive individuals from seronegative individuals based on subjective thresholds.

Acknowledgements

The authors thank the Flemish Interuniversity Council (VLIR-UOS) and the Special Research Fund (Bijzonder Onderzoeksfonds, BOF), for the PhD financial support that enabled completion of this work. This project has received funding from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (grant agreement 682540 – TransMID).

Supplemental material

Supplemental material is available for this article in a Github repository available at
https://github.com/StevenAbramsLUCP2479/Bayesian_mixture_model.git

References

Abrams, S. (2015). *Statistical models for estimating individual heterogeneity in acquisition of infectious diseases and outbreak risk in highly vaccinated populations*. Ph.D. thesis, Hasselt University.

Abrams, S., Wienke, A., and Hens, N. (2018). Modelling time varying heterogeneity in recurrent infection processes: An application to serological data. *Journal of the Royal Statistical Society: Series C*, **67**, 687–704.

Barbarossa, M. V. and Röst, G. (2015). Mathematical models for vaccination, waning immunity and immune system boosting: A general framework. *In BIOMAT 2014: International Symposium on Mathematical and Computational Biology*. World Scientific, 185–205.

Baughman, A. L., Bisgard, K. M., Lynn, F., and Meade, B. D. (2006). Mixture model analysis for establishing a diagnostic cut-off point for pertussis antibody levels. *Statistics in Medicine*, **25**, 2994–3010.

Beraha, M., Falco, D., and Guglielmi, A. (2021). JAGS, NIMBLE, Stan: a detailed comparison among Bayesian MCMC software. *arXiv preprint arXiv:2107.09357*.

Bollaerts, K., Aerts, M., Shkedy, Z., Faes, C., Van der Stede, Y., Beutels, P., and Hens, N. (2012). Estimating the population prevalence and force of infection directly from antibody titres. *Statistical Modelling*, **12**, 441–462.

Broliden, K., Tolfvenstam, T., and Norbeck, O. (2006). Clinical aspects of parvovirus B19 infection. *Journal of Internal Medicine*, **260**, 285–304.

Coutinho, F., Massad, E., Lopez, L., Burattini, M., Struchiner, C., and Azevedo-Neto, R. (1999). Modelling heterogeneities in individual frailties in epidemic models. *Mathematical and Computer Modelling*, **30**, 97–115.

de Valpine, P., Paciorek, C., Turek, D., Michaud, N., Anderson-Bergman, C., Obermeyer, F., Wehrhahn Cortes, C., Rodrìguez, A., Temple Lang, D., and Paganin, S. (2023). *NIMBLE User Manual*. doi:10.5281/zenodo.1211190. R package manual version 1.0.1.

Del Fava, E. (2012). *Statistical methods for modeling of drug-related and close-contact infections*. Ph.D. thesis, Universiteit hasselt, School voor Informatietechnologie, Maastricht University.

Eilers, P. H. and Marx, B. D. (2010). Splines, knots, and penalties. *Wiley Interdisciplinary Reviews: Computational Statistics*, **2**, 637–653.

Epifani, I., MacEachern, S. N., and Peruggia, M. (2008). Case-deletion importance sampling estimators: Central limit theorems and related results. *Electronic Journal of Statistics*, **2**, 774–806.

Evans, R. B. and Erlandson, K. (2004). Robust Bayesian prediction of subject disease status and population prevalence using several similar diagnostic tests. *Statistics in Medicine*, **23**, 2227–2236.

Farrington, C. P., Kanaan, M. N., and Gay, N. J. (2001). Estimation of the basic reproduction number for infectious diseases from age-stratified serological survey data. *Journal of the Royal Statistical Society: Series C*, **50**, 251–292.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer, New York, NY.

Gay, N., Vyse, A., Enquselassie, F., Nigatu, W., and Nokes, D. (2003). Improving sensitivity of

oral fluid testing in IgG prevalence studies: Application of mixture models to a rubella antibody survey. *Epidemiology & Infection*, **130**, 285–291.

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

Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, **74**, 153–160.

Gnann Jr, J. W. (2002). Varicella-zoster virus: Atypical presentations and unusual complications. *The Journal of Infectious Diseases*, **186**, S91–S98.

Goeyvaerts, N., Hens, N., Aerts, M., and Beutels, P. (2011). Model structure analysis to estimate basic immunological processes and maternal risk for parvovirus B19. *Biostatistics*, **12**, 283–302.

Grimm, K. J., Houpt, R., and Rodgers, D. (2021). Model fit and comparison in finite mixture models: A review and a novel approach. *Frontiers in Education*, **6**, 613645.

Hardelid, P., Williams, D., Dezateux, C., Tookey, P., Peckham, C., Cubitt, W., and Cortina-Borja, M. (2008). Analysis of rubella antibody distribution from newborn dried blood spots using finite mixture models. *Epidemiology & Infection*, **136**, 1698–1706.

Held, L., Hens, N., D O'Neill, P., and Wallinga, J. (2019). *Handbook of Infectious Disease Data Analysis*. Chapman & Hall, Boston, MA.

Hens, N., Shkedy, Z., Aerts, M., Faes, C., Van Damme, P., and Beutels, P. (2012). *Modeling Infectious Disease Parameters based on Serological and Social Contact Data: A Modern Statistical Perspective*. Springer, New York, NY.

Hens, N., Wienke, A., Aerts, M., and Molenberghs, G. (2009). The correlated and shared gamma frailty model for bivariate current status data: An illustration for cross-sectional serological data. *Statistics in Medicine*, **28**, 2785–2800.

Kanaan, M. and Farrington, C. (2005). Matrix models for childhood infections: A Bayesian approach with applications to rubella and mumps. *Epidemiology & Infection*, **133**, 1009–1021.

Lehmann, H. W., von Landenberg, P., and Modrow, S. (2003). Parvovirus B19 infection and autoimmune disease. *Autoimmunity Reviews*, **2**, 218–223.

Martins, A., Aerts, M., Hens, N., Wienke, A., and Abrams, S. (2019). Correlated gamma frailty models for bivariate survival time data. *Statistical Methods in Medical Research*, **28**, 3437–3450.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York, NY.

Nhat, N. T. D., Todd, S., De Bruin, E., Thao, T. T. N., Vy, N. H. T., Quan, T. M., Vinh, D. N., Van Beek, J., Anh, P. H., and Lam, H. M. (2017). Structure of general-population antibody titer distributions to influenza A virus. *Scientific Reports*, **7**, 1–9.

Ntzoufras, I. (2011). *Bayesian Modeling Using WinBUGS*, volume 698. Wiley, Hoboken, NJ.

Peruggia, M. (1997). On the variability of case-deletion importance sampling weights in the Bayesian linear model. *Journal of the American Statistical Association*, **92**, 199–207.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *In Proceedings of the 3rd international workshop on distributed statistical computing*. Vienna, Austria, 1–10.

Ray, C. G. and Ryan, K. J. (2004). *Sherris Medical Microbiology: An Introduction to Infectious*

Diseases. McGraw-Hill, New York, NY.

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

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, **11**, 735–757.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric regression*. Cambridge University Press, Cambridge.

Schoub, B., Blackburn, N., Johnson, S., and McAnerney, J. (1993). Primary and secondary infection with human parvovirus B19 in pregnant women in South Africa. *South African Medical Journal*, **83**, 505–506.

Tolfvenstam, T., Papadogiannakis, N., Norbeck, O., Petersson, K., and Broliden, K. (2001). Frequency of human parvovirus B19 infection in intrauterine fetal death. *The Lancet*, **357**, 1494–1497.

Turner, B. M., Forstmann, B. U., and Steyvers, M. (2019). *Joint Models of Neural and Behavioral Data. Computational Approaches to Cognition and Perception*. Springer, Cham.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, **27**, 1413–1432.

Vink, M. A., Berkhof, J., van de Kassteele, J., van Boven, M., and Bogaards, J. A. (2016). A bivariate mixture model for natural antibody levels to human papillomavirus types 16 and 18: Baseline estimates for monitoring the herd effects of immunization. *PloS One*, **11**, e0161109.

Vyse, A. J., Gay, N., Hesketh, L., 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 & Infection*, **134**, 1303–1312.

Watanabe, S. and Opper, M. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, **11**.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R*. Chapman and Hall, Boca Raton, FL.

Yashin, A. I., Vaupel, J. W., and Iachine, I. A. (1995). Correlated individual frailty: An advantageous approach to survival analysis of bivariate data. *Mathematical Population Studies*, **5**, 145–159.

Manuscript received 2024-05-08, revised 2025-02-11, accepted 2025-02-11.