On the estimation of epidemiological parameters from serological survey data using Bayesian mixture modelling

*Sulla stima di parametri epidemiologici da dati da indagini sierologiche tramite l’uso di modelli di mistura bayesiani*

Emanuele Del Fava, Piero Manfredi, and Ziv Shkedy[[1]](#footnote-1)

**Abstract** In the context of serological surveys for the estimation of important epidemiological functions, mixture models offer an alternative and more accurate approach than the conventional one based on the diagnostic assay’s cut-off points. In this work, we propose an innovative Bayesian mixture model for the estimation of flexible models for the age-specific seroprevalence and force of infection (or incidence rate). In order to account for the possible waning of immunity by age, we propose a Bayesian variable selection approach to determine the best age-specific model for the mean and the variance of the seropositive subpopulation. Our methodology is applied to a sample of antibody titres to varicella from Italy. Our results confirm that mixture models are a flexible and highly customisable tool, adapt to be systematically used in serological surveys.

**Abstract** *Nel contesto delle indagini sierologiche per la stima di importanti parametri epidemiologici, i modelli di mistura rappresentano un approccio alternativo e più accurato di quello basato sull’uso dei valori critici associati ai test diagnostici. In questo lavoro, proponiamo un innovativo modello di mistura bayesiano, finalizzato alla stima di modelli flessibili per la sieroprevalenza e la forza dell’infezione (o tasso d’incidenza) dipendenti dall’età. Allo scopo di modellare il possibile declino degli anticorpi con l’età, proponiamo un metodo di selezione bayesiana di variabili per determinare il miglior modello, dipendente dall’età, per la media e la varianza della sottopopolazione dei sieropositivi. La nostra metodologia è applicata ad un campione di titoli anticorpali contro la varicella dall’Italia. I nostri risultati confermano che i modelli di mistura sono uno strumento flessibile e facilmente configurabile, da utilizzarsi in maniera sistematica nelle indagini sierologiche.*

**Key words:** mixture models, serological data, Bayesian variable selection, MCMC

Introduction

Serological surveys, which are usually employed to quantify the antibody titre against a specific antigen, are among the most direct and informative techniques that are available to investigate the dynamics of the level of immunity protection in a certain population [12]. Despite that, the data resulting from these surveys remain somehow unexploited, for various reasons. One of them is that the assessment of the immunity profile in a population, the so-called seroprevalence of prevalence of immune individuals, is still often performed by dichotomisation of individual antibody titres measured by the diagnostic serological assay. This means that all the information contained in the individual antibody titre, such as the strength and the individual heterogeneity of the immunological response, is lost, as it is replaced by a binary variable giving the infection status, namely, whether the subject is seronegative or seropositive (showing evidence of past infection or vaccination). Moreover, the use of fixed cut-offs in serological surveys may be sub-optimal as the method is prone to misclassification or inconclusive classification [2, 10]. Conversely, mixture models are showing to be a more appropriate tool for the classification of serological antibody titres and for the estimation of age-specific epidemiological parameters, both from the statistical and the epidemiological point of view, in the case of infections in the pre-vaccination and post-vaccination state [3, 4, 14, 15].

The objective of this work is to provide an example of how to accurately estimating flexible models for the age-specific seroprevalence and force of infection (FOI) for varicella, using Bayesian mixture models for the classification of antibody titres and Bayesian variable selection for estimating the parameters associated with the seropositive subpopulation while accounting for possible immunity waning by age.

Data and Methods

Data on antibody titres for varicella zoster virus (VZV) in Italy were collected between 2003 and 2004 at national level [8]. Serological tests for VZV-specific IgG were performed using a commercial enzyme linked immunosorbent assay (ELISA), according to the manufacturer’s guidelines. A sample of 2517 subjects, from 1 to 79 years of age, was collected. For each subject, the antibody titre (evaluated quantitatively as an antibody concentration and expressed as an optical density (OD) measured in mUI/mL) and the age were collected for the analysis. Children under 10 years old were oversampled.

We use Bayesian mixture models [5] in order to estimate the age-specific seroprevalence and the age-specific FOI directly from VZV antibody titres. We consider the population to be at demographic and epidemiological equilibrium. Since data were collected in a pre-vaccination period, we can safely assume that each serological sample is drawn from a population consisting of just two subpopulations, one for the seronegative and one for the seropositive individuals. We then assume that the individual antibody titre, after a logarithmic transformation, i.e. $Y\_{i}=log\_{10}(OD\_{i}+1)$, is distributed as a mixture of two skew-normal distributions, with mixture weights depending on the age $a$ of the individuals,

$$Y\_{i}\left(a\right)=\left(1-π\_{i}\left(a\right)\right)SN\left(μ\_{1},σ\_{1}^{2},γ\_{1}\right)+π\_{i}\left(a\right)SN\left(μ\_{2i}(a),σ\_{2i}^{2}(a),γ\_{2}\right)$$

where $μ\_{k}$, $σ\_{k}^{2}$, $γ\_{k}$, $k=1,2$ denote the mean, the variance, and the skewness parameters of the two mixture components, respectively. The mean and the variance of the seronegative component ($μ\_{1}$, $σ\_{1}^{2}$) are assumed to be age-independent, while those for the seropositive component ($μ\_{2}$, $σ\_{2}^{2}$) are allowed to vary by age, in order to account for possible waning of the antibody titre. The skew-normal distribution generalises the normal distribution by allowing for skewness through a specific parameter, $γ\_{k}$, which we assume to be independent of age [1, 7]. A positive (negative) value of $γ\_{k}$ implies a distribution skewed to the right (left), thus a distribution with an excess of extremely high (low) antibody titres. The mixture weight of the immune component, $π\left(a\right)$, represents the age-specific seroprevalence, which is the expected proportion of immune individuals at exact age $a$ in the given population [6]. Model estimation and inference are carried out by using Bayesian Markov Chain Monte Carlo (MCMC) methods.

For the parameters $θ^{SP}\left(a\right)=(μ^{SP}\left(a\right),τ^{SP}\left(a\right))$, where $τ$ is the precision, i.e., the reciprocal of the variance $σ^{2}$, we specify an age-specific piecewise-constant model, where the parameter $θ^{SP}$ is constant within each of the $T$ considered age groups $(a\_{[t-1]},a\_{[t]})$:

$$θ^{SP}\left(a\right)=θ\_{1}^{SP}+\sum\_{t=2}^{T}θ\_{[t]}^{SP}, a\in \left(a\_{\left[t-1\right]},a\_{\left[t\right]}\right).$$

Since we do not have any a priori knowledge about the direction of the changes across age of $θ^{SP}$, we use an unrestricted model, i.e. the age-dependent parameters $θ\_{[t]}^{SP}$ may either increase or decrease with respect to the preceding value $θ\_{[t-1]}^{SP}$. Since it would be too computer-intensive to fit all possible models and then select the best one using a selection criterion, we propose a Bayesian Variable Selection (BVS) approach to estimate the posterior probability of each possible model, and, in particular, the one for the model with the constant parameter [9, 11, 13].

As regards the seroprevalence $π\left(a\right)$, we specify a Beta prior distribution for each age group $j$, namely, $π\_{[j]}\~Beta(α\_{[j]},β\_{[j]})$, under the monotonically non-decreasing constraint $π\_{[j-1]}\leq π\_{[j]}\leq π\_{[j+1]}$, with the hyperparameters $α\_{[j]}$ and $β\_{[j]}$ being given weakly-informative non-negative prior distributions. The order constraint is necessary to obtain a nonnegative estimate of the FOI. The ensuing posterior distribution of the age-specific seroprevalence is still distributed as a Beta, namely, $Beta(Y\_{[j]}+α\_{[j]},n\_{[j]}-Y\_{[j]}+β\_{[j]})$, under the same order constraint. Given the estimate of the seroprevalence, the FOI is successively estimated by ${π\_{[j]}^{'}}/{(1-}π\_{[j]})$, where the first derivative of the prevalence at the numerator is approximated by ${\left(4π\_{[j+1]}-3π\_{[j]}-π\_{[j+2]}\right)}/{2}$ and then smoothed.

Results

The VZV antibody titres data show a clear polarised distribution between the seronegative and the seropositive individuals (Figure 1A), with the seronegative cases being concentrated among children and, to a lesser extent, young adults. There is evidence for waning of the antibody protection after 40 years, as implied by the decrease in the mean of the seropositive component, with a posterior probability of 0.47 for this model (Figure 1A). There is also some evidence for a two-step waning model, showing a decrease in the 10-20 age group and again after 40 years of age, with a posterior probability of 0.38. Conversely, for what concerns the variance of the component, we do not reject the null hypothesis of the null model, as its posterior probability is 0.25. The negative estimates of the skewness for both components ($γ^{SN}=-1.81, γ^{SP}=-2.59)$ imply an excess of high-value antibody titres.

The estimate of the seroprevalence (Figure 1B) shows a steep linear increase in the first ten years of age, up to around 80 % (much lower than in other European countries), followed by a slower increase in the following years. This pattern is reflected by the estimated age-specific FOI (Figure 1C), which peaks between 5 and 10 years (during primary school) and then declines to a plateau.

**Table 1:** Mean and standard deviation of the seronegative (SN) and the seropositive (SP) components of the skew-normal mixture model fitted to VZV data.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***Age group*** | $$μ^{SN}$$ | $$σ^{SN}$$ | $$μ^{SP}\left(a\right)$$ | $$σ^{SP}(a)$$ |
| 1-10 | 1.40(1.37, 1.43) | 0.34(0.31, 0.37) | 3.52(3.44, 3.58) | 0.62(0.56, 0.70) |
| 10-20 | - | - | 3.47(3.42, 3.54) | 0.63(0.58, 0.71) |
| 20-40 | - | - | 3.47(3.40, 3.54) | 0.63(0.58, 0.71) |
| 40+ | - | - | 3.33(3.27, 3.39) | 0.63(0.58, 0.71) |

****

**Figure 1:** Fit of the mixture model to VZV data for Italy: A) scatter plot by age of antibody titres with over imposed the constant mean of the seronegative component and the age-dependent mean of the seropositive component (with 95% credible interval); B) age-specific seroprevalence with 95% credible interval; C) age-specific FOI with 95% credible interval.

Conclusions

In this work, we employed Bayesian mixture models to estimate key epidemiological parameters, such as the seroprevalence and the force of infection directly from antibody levels. Contrarily to the fixed cut-off approach, which leads to the estimation of these parameters based on binary infection status data, the mixture model adapts directly to the antibody data, without losing the information contained therein. Rather, the method allows for both the classification of individuals among different serological groups (by giving each subject an age-dependent probability of belonging to a specific group), also of those cases that would be classified as inconclusive under the fixed cut-off approach, and the estimation of the epidemiological parameters of interest. Moreover, the employment of a Bayesian MCMC approach allows to derive credible intervals around all the model parameters.

The model is also customisable enough to include a model conditional on age for the mixture parameters. The use of the BVS approach is computationally feasible and allows to show the support for all possible models in terms of posterior probability, even though one must pay attention to the variables under consideration, as the computational time dramatically increases with their number.

For all these reasons, we believe that the mixture modelling approach represents a flexible and highly customisable approach that should be seriously considered as the optimal way of analysing serological survey data. As a consequence, we claim that more attention should be devoted to the design of the serological surveys, both for what concerns the determination of the sample size by age and the measurement of the antibody titres.

References

1. Azzalini, A.: A Class of Distributions Which Includes the Normal Ones. Scand. J. Statist. **12**, 171—178 (1985) doi:10.2307/4615982.
2. Bollaerts, K., Aerts, M., Shkedy, Z., Faes, C., Van der Stede, Y., Beutels, Ph., Hens, N.: Estimating the Population Prevalence and Force of Infection Directly From Antibody Titres. Statist. Mod. **12**, 441—462 (2012) doi:10.1177/1471082X12457495.
3. Del Fava, E., Mirinaviciute, G., Flem, E., Freiesleben de Blasio, B., Scalia Tomba, G., Manfredi, P.: Estimating Age-Specific Immunity and Force of Infection of Varicella Zoster Virus in Norway Using Mixture Models. PLoS ONE **11**, e0163636—12 (2016) doi:10.1371/journal.pone.0163636.
4. Del Fava, E., Shkedy, Z., Bechini, A., Bonanni, P., Manfredi, P.: Towards Measles Elimination in Italy: Monitoring Herd Immunity by Bayesian Mixture Modelling of Serological Data. Epidemics **4**, 124—31 (2012) doi:10.1016/j.epidem.2012.05.001.
5. Diebolt, J., Robert, C.P.: Estimation of Finite Mixture Distributions Through Bayesian Sampling. J. R. Statist. Soc. B **56**, 363—75 (1994)
6. Evans, R.B., Erlandson, K.: Robust Bayesian Prediction of Subject Disease Status and Population Prevalence Using Several Similar Diagnostic Tests. Statist. Med. **23**, 2227—2236 (2004) doi:10.1002/sim.1792.
7. Frühwirth-Schnatter, S., Pyne, S.: Bayesian Inference for Finite Mixtures of Univariate and Multivariate Skew-Normal and Skew-T Distributions. Biostatistics **11**, 317—336 (2010) doi:10.1093/biostatistics/kxp062.
8. Gabutti, G., Rota, M.C., Guido, M., De Donno, A., Bella, A., Ciofi degli Atti, M.L., Crovari, P.: The Epidemiology of Varicella Zoster Virus Infection in Italy. BMC Pub. Health **8**, 372 (2008) doi:10.1186/1471-2458-8-372.
9. George, E., McCulloch, R.E.: Variable Selection via Gibbs Sampling. J. Am. Statist. Assoc. **88**, 881—889 (1993)
10. Hardelid, P., Williams, D., Dezateux, C., Tookey, P.A., Peckham, C.S., Cubitt, W.D., Cortina-Borja, M.: Analysis of Rubella Antibody Distribution From Newborn Dried Blood Spots Using Finite Mixture Models. Epidemiol. Infect. **136**, 1698 (2008) doi:10.1017/S0950268808000393.
11. Kasim, A., Shkedy, Z., Kato, B.S.: Estimation and Inference Under Simple Order Restrictions: Hierarchical Bayesian Approach. In: Dan L., Shkedy, Z., Yekutieli, D., Amaratunga, D., Bijnens (eds.) Modeling Dose-Response Microarray Data in Early Drug Development Experiments Using R. Springer, Heidelberg.
12. Metcalf, C.J.E., Farrar, J., Cutts, F.T., Basta, N.E., Graham, A.L, Lessler, J.T., Ferguson, N.M., Burke, D.S., Grenfell, B.T.: Use of Serological Surveys to Generate Key Insights Into the Changing Global Landscape of Infectious Disease. Lancet **388**, 728–730 (2016) doi:10.1016/S0140-6736(16)30164—7.
13. O'Hara, R.B., Sillanpää, M.J.: A Review of Bayesian Variable Selection Methods: What, How and Which. Bayes. Anal. **4**, 85—117 (2009) doi:10.1214/09-BA403.
14. Rota, M.C., Massari, M., Gabutti, G., Guido, M., De Donno, A., Ciofi degli Atti, M.L.: Measles Serological Survey in the Italian Population: Interpretation of Results Using Mixture Model. Vaccine **26**, 4403—4409 (2008) doi:10.1016/j.vaccine.2008.05.094.
15. Vyse, A.J., Gay, N.J., Hesketh, L.M., Morgan-Capner, P., Miller, E.: Seroprevalence of antibody to varicella zoster virus in England and Wales in children and young adults. Epidemiol. Infect. **132**, 1129-1134 (2004)
1. Emanuele Del Fava, Bocconi University, Milan; email: emanuele.delfava@unibocconi.it

 Piero Manfredi, University of Pisa, Pisa, Italy; email: piero.manfredi@unipi.it

 Ziv Shkedy, Hasselt University, Diepenbeek, Belgium; email: ziv.shkedy@uhasselt.be

 [↑](#footnote-ref-1)