Analysis of rubella antibody distribution from newborn dried blood spots using finite mixture models

P. HARDELID[1]*, D. WILLIAMS[1], C. DEZATEUX[1], P. A. TOOKEY[1],
C. S. PECKHAM[1], W. D. CUBITT[2] AND M. CORTINA-BORJA[1]

[1] *MRC Centre of Epidemiology for Child Health, UCL Institute of Child Health, London, UK*
[2] *Department of Virology, Great Ormond Street Hospital NHS Trust, London, UK*

SUMMARY

Eluted dried blood spot specimens from newborn screening, collected in 2004 in North Thames and anonymously linked to birth registration data, were tested for maternally acquired rubella IgG antibody as a proxy for maternal antibody concentration using an enzyme-linked immunosorbent assay. Finite mixture regression models were fitted to the antibody concentrations from 1964 specimens. The Bayesian Information Criterion (BIC) was used as a model selection criterion to avoid over-fitting the number of mixture model components. This allowed investigation of the independent effect of maternal age and maternal country of birth on rubella antibody concentration without dichotomizing the outcome variable using cut-off values set *a priori*. Mixture models are a highly useful method of analysis in seroprevalence studies of vaccine-preventable infections in which preset cut-off values may overestimate the size of the seronegative population.

INTRODUCTION

Results of antibody assays may be difficult to interpret when used in population-based seroepidemiological studies where quantification of previous exposure to infectious disease, either through natural infection or vaccination, is of interest. This is partly a consequence of the difficulty in applying a meaningful cut-off value to continuous distributions of antibody concentrations in order to define seronegativity [1]. If a sample falls below a cut-off value set by the assay manufacturer, this does not necessarily indicate non-immunity, as IgG antibody assays for vaccine-preventable diseases are most often employed in screening to indicate a need for immunization. In this

* Author for correspondence: Ms P. Hardelid, MRC Centre of Epidemiology for Child Health, UCL Institute of Child Health, 30 Guilford Street, London WC1N 1EH, UK.
(Email: p.hardelid@ich.ucl.ac.uk)

context, cut-off values are likely to be set high to avoid false negatives [2]. It follows that using fixed cut-off values for seroepidemiological studies implies an at best artificial, but in some cases completely arbitrary, measure of the population at risk of infection. In addition, since the majority of antibody assays are optimized for measuring antibody concentrations in serum, cut-off values provided by the assay manufacturer may not apply if the assays have been modified for use on other test media [3–5].

An alternative approach is to fit latent-class finite mixture models [6] to continuous distributions of antibody concentrations in order to characterize different exposure categories in the population, e.g. previously infected or vaccinated, and unexposed. The basic assumption in a mixture model is that the study population consists of individuals in different exposure subpopulations. The number of subpopulations is either assumed to be known *a priori* or,

more often, is a quantity of interest to be determined from the data. In either case, mixture models provide a convenient framework to test hypotheses relating to the number of exposure groups in the population and their characteristics. Mixture models have been reported in the literature for analyses of antibody concentrations for various infections [1, 3, 7, 8].

In this study, we fit latent-class finite mixture regression models to maternally acquired rubella IgG antibody concentrations obtained from testing residual dried blood spots from newborn screening in North London and surrounding counties in 2004. Since newborn antibody concentration is used as a proxy for maternal antibody concentration [9, 10], we hypothesize that there would be three subpopulations, one seronegative group, a second group with intermediate antibody concentrations that has been vaccinated or infected but mounted a weaker antibody response, and a third group with higher antibody concentrations following infection or vaccination, which is possibly more recent. This model can be expressed as:

$$f(y) = \pi_l f_l(y) + \pi_m f_m(y) + \pi_h f_h(y), \qquad (1)$$

where y is antibody concentration, the parameters (π_l, π_m, π_h) are the component weights, and (f_l, f_m, f_h) are the density functions of the three components.

Maternal age and country of birth have been found to be important predictors of maternal rubella antibody concentration [11]. In order to examine the independent effects of these two variables on maternal antibody concentration, we fit latent-class finite mixture regression models of the form

$$f(y) = \sum_{k=1}^{K} \pi_k(X_1, \beta_{1k}) f_k(y; X_2 \beta_{2k}; X_3 \beta_{3k}), \qquad (2)$$

where y is the antibody concentration, K is the number of components, X_1, X_2 and X_3 are design matrices for the component weights (π), the location and scale parameters of the kth component respectively. $\{\beta_{1k}\}$ is the vector of parameters for the linear predictor in each component weight, and $\{\beta_{2k}\}$ and $\{\beta_{3k}\}$ are the location and scale parameters. We used the identity, logarithmic and multinomial link functions to fit the regression parameters for location, scale and component weights respectively. We are not aware of any papers reporting the use of this class of mixture regression models to analyse antibody concentrations. We demonstrate their usefulness in quantifying the independent effect of two or more predictors on antibody concentration without dichotomizing this variable.

METHODS

Samples

We measured rubella IgG antibody concentration in dried blood spots, left over from the North Thames newborn screening programme. The newborn screening programme in North Thames covers inner and outer North London, Bedfordshire, Hertfordshire and Essex, an area with an ethnically diverse population. All newborns are offered screening at around 5–8 days of age and coverage is estimated to be close to 100 % [12]. Four spots of blood are collected on a Schleicher & Schuell 903® filter paper card (Schleicher & Schuell GmbH, Dassel, Germany) from the baby's heel by a midwife or health visitor who then sends the specimens to the screening laboratory. Blood spot samples are punched from the cards, and tested for biomarkers for inherited or congenital conditions including phenylketonuria and congenital hypothyroidism.

When metabolic screening has been completed, an extra sample is punched from the screening cards and tested for anti-HIV antibody as part of the newborn unlinked anonymous HIV surveillance programme. Anonymized data linkage between screening card data and birth registration records held by the Office for National Statistics was established for this programme. The data linkage and HIV testing algorithm have been described in detail elsewhere [13, 14].

Laboratory methods

Two extra samples were punched from all blood spot cards sent to the screening laboratory between 1999 and 2006 as part of a study to serotype samples found to be HIV positive [15]. As part of the HIV testing protocol, dried blood spot punches are placed in 96-well plates, and eluted in PBS-Tween 20 solution. HIV-positive samples had to be excluded from the analyses, since the remaining eluate is used for HIV confirmatory testing. Following HIV tests, eluates were kept at −20 °C for up to 20 months before rubella antibody testing. Plates containing samples received in the screening laboratory in the third quarter of 2004 in North Thames were randomly sampled for rubella IgG testing. A commercial enzyme-linked immunosorbent assay (ELISA; Diesse, Siena, Italy) was used to test for rubella IgG antibody.

This ELISA has been found to be valid for use on eluted dried blood spots [16]. Dried blood spot eluate samples (50 μl each) were tested for rubella IgG antibody according to the manufacturer's instructions. The ELISA results are quantitative, given as a ratio between the optical density reading of the sample and a plate-specific reference sample provided by the manufacturer. The optical densities can be converted into antibody concentrations in international units per millilitre (IU/ml) using a calibration curve, which is constructed from four plate-specific reference sample values with pre-determined antibody concentrations provided by the manufacturer.

Demographic data

Data on maternal age were available from screening records, and on maternal country of birth through data linkage to birth registration data. Maternal country of birth was coded into a binary variable ('UK-born' and 'born abroad').

Statistical methods

In order to amplify the values of the optical density ratios between 0 and 1, which are of largest epidemiological and clinical significance, we modelled the distribution of the natural logarithm of the optical density ratio [ln(ODR)]. Low optical density values correspond to seronegative status.

We used the FLEXMIX [17] and GAMLSS.MX [18] packages in the R statistical software version 2.5.1 [19] to fit continuous finite mixture models based on a latent-class regression approach. Both packages employ an expectation-maximization (EM) algorithm that uses maximum likelihood to estimate the model parameters and calculates the probability of belonging to each component for every observation. Hence the size of each component can be estimated as follows: the probability that observation i belongs to class j is given by

$$\Pr[j|X_{i1}, X_{i2}, X_{i3}; y_i; \beta_1, \beta_2, \beta_3] = \frac{\pi_j(X_{i1}, \beta_{1j})\, f(y_i)}{\sum_{k=1}^{K} \pi_k(X_{i1}, \beta_{1k})\, f(y_i)}, \tag{3}$$

where $j = 1, 2, \ldots, K$, and $i = 1, 2, \ldots, n$. We used this expression (known as Bayes' rule) to assign each observation to the class with the maximum probability.

We used the diagnostic tools provided by these packages to check the models' goodness-of-fit properties. We based our model selection strategy on the Bayesian Information Criterion (BIC) [17, 20] defined as

$$\mathrm{BIC} = -2 \ln L(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3) + p \ln(n), \tag{4}$$

where L is the value of the likelihood evaluated at the estimates $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$, p is the number of model parameters and n is the number of independent observations. The BIC is more conservative against overparameterization than other model selection criteria, which is particularly important when fitting mixture models as the optimal number of components may otherwise be overestimated [17].

We fitted mixture models to the full distribution of the natural logarithms of the ODRs available, assuming normal and gamma distributions for the components allowing for up to five components and compared the goodness of fit of these models using BIC. In order to fit gamma-distributed components we added a constant to the ln(ODR) values to ensure they were positive. Up to five sets of maximum-likelihood estimates were obtained for each regression model, the optimal model for each set of covariates being that which minimized BIC. Increasing the number of components beyond five resulted in empty components. The finite mixture regression models were then compared using a stepwise procedure to account for one covariate for the mean at a time, then including both variables, and finally including an interaction term. Maternal age was modelled both as a linear and as a quadratic term in the linear predictors defining the components' means. Again, the optimal model was that for which BIC was minimized. The optimal regression model was also refitted with covariates for the component weights and variances, using a stepwise approach, adding more covariates, to assess whether the models' fit improved.

Ethics approval

Ethics approval was obtained from the London Multicentre Research Ethics Committee (MREC).

RESULTS

We analysed ln(ODR)s from 1964 dried blood spot samples out of 1979 samples tested. Eleven samples were excluded from analyses as they were HIV positive, and a further four because parents had opted out of HIV testing or newborn screening. Linkage to birth registration data was achieved for 1943 of 1964
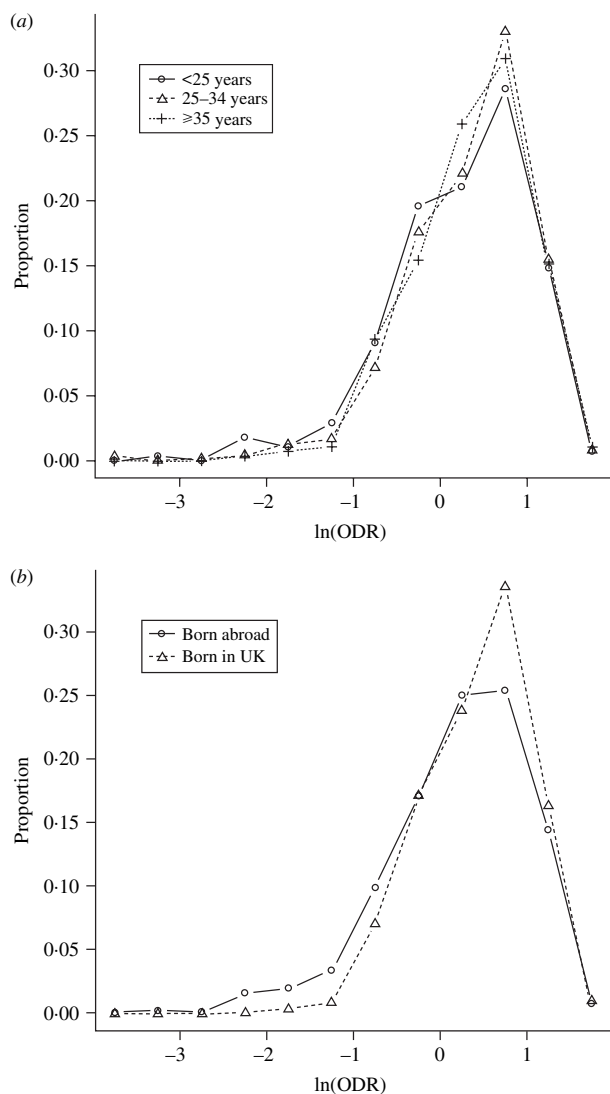
Fig. 1. Distribution of the natural logarithm of the optical density ratio [ln(ODR)]: (*a*) by maternal age group (*n* = 1265); (*b*) by maternal country of birth (*n* = 1811).

Table 1. *Model selection procedure for univariable finite mixture models* (*n* = *1964*)

No. of components in model	D.F.	BIC
Normal components		
1	2	4355·2
2	5	4130·2
3	8	4055·1
4	11	4067·8
5	14	4090·3
Gamma components		
1	2	4875·7
2	5	4248·0
3	8	4086·4
4	11	4075·8
5	14	4083·3

D.F., Degrees of freedom; BIC, Bayesian Information Criterion.

with three normally distributed components had the lowest BIC; the three normal density functions correspond to the three hypothesized groups of exposure to rubella infection (Table 2). Note that the component weights are not equal to the proportions of samples allocated to each component since the former are the estimates of the parameters π_l, π_m and π_h and the latter were calculated using equation (3).

Figure 2 shows the densities of each component of this model, as well as the density of the total sample of ln(ODR).

To account for the variation in antibody concentration by maternal age and country of birth, we fitted mixture regression models adjusting for these covariates. In order to compare models with different covariates using the BIC, all samples for which either maternal country of birth or maternal age was missing had to be excluded, leaving 1236 samples for analysis. A summary of these regression models can be found in Table 3.

The BIC was minimized for the model where only maternal country of birth is included as a covariate (model 3). This model was therefore refitted with covariates for the component weights (π) and variance parameters using the same combinations of covariates as described in Table 3.

A model in which maternal age was included as a covariate for the component weights (with no covariates for the variances) improved very marginally on the optimal model described in Table 3 and resulting in a BIC of 2503·6. The parameter estimates for the linear predictors of this model are shown in Table 4*a*;

samples (98·9%). Data on maternal age was available for 1265 of these linked samples (65·1%). Mean maternal age was 29·5 years (s.d. = 5·9 years). Maternal country of birth was available for 1811 samples (93·2%): 792 (43·7%) mothers were born abroad. The distribution of ln(ODR) by maternal age group (< 25, 25–34 and ≥ 35 years) and country of birth are shown in Figure 1(*a*, *b*).

Table 1 shows the model selection process for the finite mixture models using the full population of 1964 samples, varying the number of mixture model components, assuming normal or gamma distributions of the components. The former appeared to provide a better fit, and therefore all subsequent models were based on this assumption. The model

Table 2. *Summary of the three components in the univariable model with lowest Bayesian Information Criterion (means and confidence intervals have been anti-logged)*

Component	No. of observations in component	Samples (%)	Component weights (π)	Mean ODR	S.D.	95% CI for mean
Low	71	3·6%	6·7%	0·33	0·94	(0·18–0·62)
Medium	897	45·7%	47·4%	1·03	0·51	(0·98–1·08)
High	996	50·7%	45·9%	2·31	0·34	(2·24–2·40)

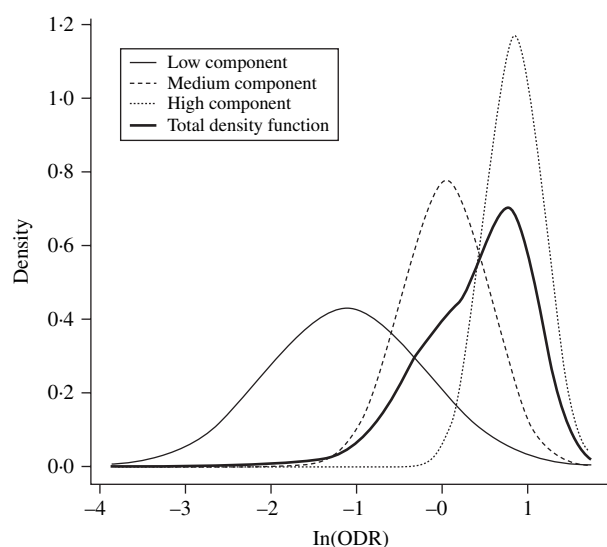ODR, Optical density ratio, S.D., standard deviation; CI, confidence interval.

Fig. 2. The density functions of the three components of the finite mixture model with lowest Bayesian Information Criterion and the density function of the total distribution of the natural logarithm of the optical density ratio [ln(ODR)] ($n = 1964$).

note that the functions for the component weights are given on the multinomial scale, considering the low component as baseline. Table 4b gives the mean and confidence interval (CI) for the components by maternal country of birth.

The mean ODR for the low and medium mixture model components is significantly lower for women born abroad than for UK-born women. However, for UK-born women, the mean of the medium component is not significantly different to that of the low component. Figure 3 shows how the component weights change as a function of maternal age. These curves were obtained by applying the inverse multinomial function to the linear predictors in Table 4a. Adding a quadratic effect for maternal age to these probability functions did not improve the goodness of fit.

As maternal age increases, women are less likely to be assigned to the low component. The probability of being assigned to the medium component increases with maternal age, however, the proportion in the higher component changes little across maternal ages. For a mother of the mean age in our sample (29·5 years), the component weights for the low, medium and high components are therefore 5·7%, 54·1% and 40·2% respectively.

DISCUSSION

Left-over dried blood spots from newborn screening provide a large, relatively inexpensive, and near-universal sampling frame within which rubella antibody concentrations among newly delivered women can be investigated. However, because this implies using newborn antibody concentration as a proxy for maternal antibody concentration, using cut-off values to define the seronegative population would not be appropriate for our data. We therefore used finite mixture models to analyse the natural logarithms of the ODRs, as this avoids dichotomizing the outcome variable of interest.

Several authors [1, 3] have commented on the usefulness of finite mixture models for analysing results of seroprevalence studies in which the specificity of the antibody assay used is low. This is supported by our findings; using this approach and assuming that the samples allocated to the component with the lowest mean antibody concentration are seronegative (Table 2), 3·6% of samples would be considered antibody negative (95% CI 2·8–4·4). This is similar to the proportion of women who were described as susceptible through antenatal screening in London in 2004 [21]. By applying the cut-off recommended by the assay manufacturers (of samples with an antibody level <7 IU/ml classed as sero-negative), the proportion of samples that would be

Table 3. *Model selection strategy for mixture regression models (n = 1236)*

Model	Model covariates	No. of components	D.F.	Log-likelihood	BIC
1	No covariates	3	8	−1227·1	2511·2
2	Maternal age only	3	11	−1221·1	2520·6
3	Maternal country of birth only	3	11	−1212·7	2503·7
4	Maternal country of birth + maternal age	3	14	−1206·3	2512·3
5	Maternal country of birth + maternal age + interaction term	3	17	−1204·4	2529·9
6	Maternal age² + maternal age only	3	14	−1219·0	2523·7
7	Maternal age² + maternal age + maternal country of birth	3	17	−1202·3	2525·6
8	Maternal age² + maternal age + maternal country of birth + interaction term	3	23	−1199·1	2561·9

D.F., Degrees of freedom; BIC, Bayesian Information Criterion.

Table 4a. *Linear predictors for the optimal model*

Component	π	μ	σ
Low	0	−1·53 + 1·15 * UK-born	0·74
Medium	−1·88 + 0·14 * maternal age	−0·04 + 0·19 * UK-born	0·52
High	−1·88 + 0·13 * maternal age	0·84 + 0·02 * UK-born	0·32

Table 4b. *Summary of the optimal mixture regression model with maternal age as a covariate for the component weights (π) and maternal country of birth as a covariate for the mean of the components (means and confidence intervals have been anti-logged)*

Component	No. in each component (%)*	Mean ODR for women born in UK	95% CI	Mean ODR for women born abroad	95% CI
Low	48 (3·9%)	0·68	(0·41–1·15)	0·22	(0·17–0·27)
Medium	608 (49·2%)	1·17	(1·02–1·34)	0·96	(0·90–1·02)
High	580 (46·9%)	2·38	(2·16–2·62)	2·32	(2·22–2·43)

ODR, Optical density ratio; CI, confidence interval.
* The number of samples in each component was estimated using the posterior probabilities derived using equation (3).

considered seronegative is 15·3% (95% CI 13·8–17·0), which is a gross overestimate. This emphasizes the need for caution in applying cut-off values determined *a priori*, which are intended for use in a clinical setting, without consideration for the particular context in which they are applied, or of the specimens used in the serosurvey.

As Figure 2 shows, there is some overlap among the components of the mixture model. This implies that some samples (e.g. with log optical densities between −1 and 0), which may have been assigned to the medium antibody component also have a non-negligible probability of being assigned to the low antibody component. This raises some questions about the validity of determining a cut-off value based on this particular mixture model.

Since data on maternal country of birth were available from data linkage between the newborn
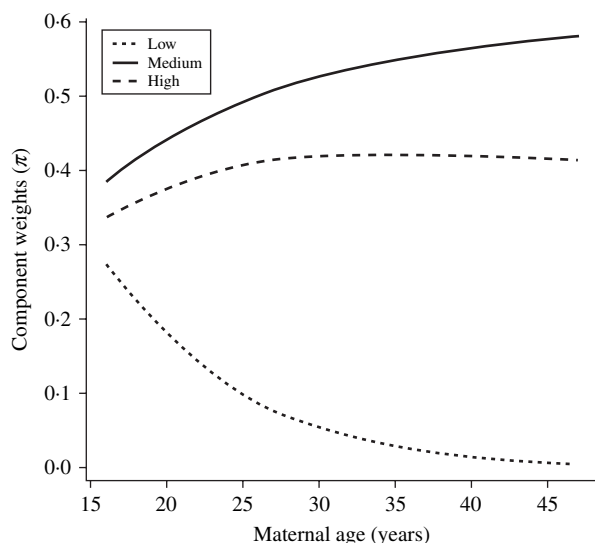
Fig. 3. Component weights (π) by maternal age for the optimal model described in Tables 4a and 4b.

screening records and birth registration data, and maternal age could be obtained from screening datasets, finite mixture regression models could be fitted, allowing the means, component weights and variances in the model to vary by risk factors. In previous studies where finite mixture models have been used, only age group has been of interest as a predictor of antibody level [1, 3, 8], and mixture models are therefore often stratified in order to determine the proportion of seronegative samples in each age group. This is important in order to establish whether further vaccination is required due to higher susceptibility in some cohorts. However, maternal country of birth or ethnic group has been reported to be a highly significant predictor of rubella susceptibility among pregnant women in the United Kingdom [11, 22]. Our method allows hypotheses about the independent, as well as the combined effects, of age and country of birth on antibody concentration to be tested.

In our optimal model, the means of the low and medium components for women born abroad were significantly lower than those for women born in the United Kingdom, yet the component weights were not found to vary by country of birth. One possible reason for this is that women born in the United Kingdom with low antibody concentrations may have been exposed to the virus through vaccination several years previously and therefore their antibody concentrations may have waned without 'boosting' through contact with wild virus. This is

in contrast to women born abroad with low antibody levels who may not have had previous virus contact. For UK-born women, the difference in means between the low and the medium component was not statistically significant. A larger sample size will allow the component means to be more precisely estimated, as well as the inclusion of a more detailed maternal country-of-birth variable to identify specific groups at high risk of infection during pregnancy.

We also found that the weight for the low component decreased as maternal age increased, but increased with maternal age for the medium component, implying that younger pregnant women are more likely to be seronegative, as defined by the mixture models. Vyse *et al.* [1] also found a similar age pattern of rubella seronegativity among women in England and Wales. This could be a consequence of the antenatal screening programme, where women who screened negative in a previous pregnancy were vaccinated post partum.

With the possibility of adjusting for several covariates, it is important that the model selection process avoids models which overfit the number of components as well as the number of covariates. The FLEXMIX and GAMLSS.MX R-packages allow for the calculation of BIC as a model selection statistic. In comparison to using log-likelihood only, the BIC penalizes both for sample size and the number of parameters, and is therefore more conservative in terms of the optimal number of components and variables included (Table 3). The use of BIC for model selection in finite mixture regression models depends, as all likelihood-based procedures, on all competing models being fitted to the same dataset. Because maternal age was missing for a large proportion of samples, this resulted in the sample sizes for the regression models including this covariate being reduced by about 40%.

Several previous papers [1, 3, 8] in which mixture models were applied to results of seroprevalence studies have assumed normal distributions for the mixture model components. We tested the assumption that gamma-distributed mixture model components provided a better fit, however, this was found not to be the case.

In this paper we have distinguished between the estimates of the component weights ($\pi_1, \pi_2, ..., \pi_K$) and the estimated proportion of observations in each component, which is calculated based on the posterior probabilities given in equation (3). This is the only

feasible approach for finite mixture regression models fitted with covariates for the component weights. Indeed, this procedure should be used even in finite mixture regression models without covariates for the component weights since the posterior probabilities allow the assessment of how well defined the components are. The posterior probabilities also allow the analysis of the characteristics of the individuals whose samples have been assigned to a particular component. Using our approach, it can be shown that 60·4% of women whose samples were assigned to the low component described in Table 4*b* were born abroad. This would not have been possible if component weights had been used to estimate the size of the components.

The use of residual dried blood spots from newborn screening anonymously linked to birth registration data provides a near-universal sampling frame from a large and ethnically diverse area such as North Thames. This allows differences in maternal rubella antibody concentrations by maternal age and country of birth to be investigated. We have demonstrated that latent-class finite mixture regression models are a powerful statistical method for analysing antibody concentrations from serosurveys in which assay cut-off values are not applicable. Fitting these models allows inclusion of linear predictors for antibody concentration and the identification of groups of pregnant women who may require vaccination. Newborn dried blood spots can therefore be used to obtain more detailed estimates of the proportion of women who are seronegative in different age and country-of-birth categories.

ACKNOWLEDGEMENTS

DECLARATION OF INTEREST

None.

REFERENCES

1. **Vyse AJ, *et al.*** Interpreting serological surveys using mixture models: the seroepidemiology of measles, mumps and rubella in England and Wales at the beginning of the 21st century. *Epidemiology and Infection* 2006; **134**: 1303–1312.

2. **Ades AE.** Evaluating screening-tests and screening programs. *Archives of Disease in Childhood* 1990; **65**: 792–795.

3. **Gay NJ, *et al.*** Improving sensitivity of oral fluid testing in IgG prevalence studies: application of mixture models to a rubella antibody survey. *Epidemiology and Infection* 2003; **130**: 285–291.

4. **Sheppard C, *et al.*** Development and evaluation of an antibody capture ELISA for detection of IgG to Epstein-Barr virus in oral fluid samples. *Journal of Virological Methods* 2001; **93**: 157–166.

5. **Talukder Y, *et al.*** Development and evaluation of Varicella zoster virus ELISA for oral fluid suitable for epidemiological studies. *Journal of Virological Methods* 2005; **128**: 162–167.

6. **McLachlan G, Peel D**. *Finite Mixture Models*. New York: John Wiley and Sons, 2000.

7. **Baughman AL, *et al.*** Mixture model analysis for establishing a diagnostic cut-off point for pertussis antibody levels. *Statistics in Medicine* 2005; **25**: 2994–3010.

8. **Gay NJ.** Analysis of serological surveys using mixture models: application to a survey of parvovirus B19. *Statistics in Medicine* 1996; **15**: 1567–1573.

9. **Leineweber B, *et al.*** Transplacentally acquired immunoglobulin G antibodies against measles, mumps, rubella and varicella-zoster virus in preterm and full term newborns. *Pediatric Infectious Diseases Journal* 2004; **23**: 361–363.

10. **Goncalves G, *et al.*** Levels of rubella antibody among vaccinated and unvaccinated Portuguese mothers and their newborns. *Vaccine* 2006; **24**: 7142–7147.

11. **Tookey PA, Cortina-Borja M, Peckham CS.** Rubella susceptibility among pregnant women in North London, 1996–1999. *Journal of Public Health Medicine* 2002; **24**: 211–216.

12. **Ades AE, *et al.*** Coverage of neonatal screening: failure of coverage or failure of information system. *Archives of Disease in Childhood* 2001; **84**: 476–479.

13. **Ades AE, *et al.*** Effect of the worldwide epidemic on HIV prevalence in the United Kingdom: record linkage in anonymous neonatal seroprevalence surveys. *Aids* 1999; **13**: 2437–2443.

14. **Cortina-Borja M, *et al.*** HIV prevalence in pregnant women in an ethnically diverse population in the UK: 1998–2002. *AIDS* 2004; **18**: 535–540.

15. **Cortina-Borja M, *et al.*** HIV-1 subtypes in pregnant women in the UK. *International Journal of STD AIDS* 2007; **18**: 160–162.

16. **Hardelid P, *et al.*** Agreement of rubella IgG antibody measured in serum and dried blood spots using two

commercial enzyme-linked immunosorbent assays. *Journal of Medical Virology* (in press).

17. **Leisch F.** FlexMix: A general framework for finite mixture models and latent class reduction in R. *Journal of Statistical Software* 2004; **11**: 1–18.

18. **Rigby RA, Stasinopoulos DM.** Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society Series C: Applied Statistics* 2005; **54**: 507–544.

19. **R Development Core Team.** R: A language and environment for statistical computing. 2005. Vienna, Austria, R Foundation for Statistical Computing.

20. **Kuha J.** AIC and BIC – comparisons of assumptions and performance. *Sociological Methods & Research* 2004; **33**: 188–229.

21. **Giraudon I,** *et al.* Antenatal infection screening surveillance in London 2000 to 2004. Health Protection Agency, London Regional Epidemiology Unit, 2005.

22. **Miller E.** Rubella in the United Kingdom. *Epidemiology and Infection* 1991; **107**: 31–42.