STATISTICS IN MEDICINE
Statist. Med. 2006; **25**:2994–3010
Published online 12 December 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2442

Mixture model analysis for establishing a diagnostic cut-off point for pertussis antibody levels[§]

Andrew L. Baughman[1,*,†], Kristine M. Bisgard[1,‡], Freyja Lynn[2] and Bruce D. Meade[2]

[1]*National Immunization Program, Centers for Disease Control and Prevention, Atlanta, GA, U.S.A.*
[2]*Center for Biologics Evaluation and Research, Food and Drug Administration, Bethesda, MD, U.S.A.*

SUMMARY

Previous studies of pertussis (whooping cough) that have derived diagnostic cut-off points for pertussis antibody levels have assumed a single distribution for antibody levels and have used small sample sizes. In a recent study of 5409 serum samples from the Third National Health and Nutrition Examination Survey (NHANES III), a finite mixture model was developed to examine the distribution of immunoglobulin G (IgG) antibody levels against pertussis toxin (PT), an antigen specific to the *Bordetella pertussis* bacterium. The mixture model identified three component populations with antibody levels greater than the quantitative assay's lower limit of quantitation (LLQ) and included a point distribution located at or below the LLQ to account for the excess number of antibody values that fell below the LLQ. The mixture model analysis accounted for the NHANES III design. A cut-off point for anti-PT IgG levels was chosen to have a 99 per cent model specificity based on the two overlapping normal distributions assumed for the two component populations with the highest antibody levels. This cut-off point may have a higher diagnostic sensitivity for acute *B. pertussis* infection than other cut-off points derived by assuming a single distribution for antibody levels. Published in 2005 by John Wiley & Sons, Ltd.

KEY WORDS: mixture model; diagnostic cut-off point; Third National Health and Nutrition Examination Survey; antibodies; pertussis

1. INTRODUCTION

Pertussis (whooping cough) is endemic in the U.S. and globally, and yet no routine, reliable, and accurate diagnostic test is currently available. Previous studies of pertussis that have derived diagnostic cut-off points for a positive antibody test have assumed a single distribution

*Correspondence to: Andrew L. Baughman, Centers for Disease Control and Prevention, 1600 Clifton Road, N.E., Mailstop E-61, Atlanta, GA 30333, U.S.A.
†E-mail: dbaughman@cdc.gov
‡Present address: Division of Health Studies, National Center for Environmental Health/Agency for Toxic Substances and Disease Registry, Atlanta, GA, U.S.A.

Received 29 March 2005
Published in 2005 by John Wiley & Sons, Ltd. *Accepted 23 September 2005*

for antibody levels and have used small sample sizes [1–3]. The cut-off points used in these investigations have been derived from an external control group that was assumed to have not been recently infected with *Bordetella pertussis*, the bacterium that causes pertussis, and to have not recently received pertussis vaccine. The data were commonly transformed by a logarithm function to make them normally distributed, and cut-off points for a positive test were derived as the 99 per cent upper tolerance limit (UTL, the upper 95 per cent confidence bound for the 99th percentile of the sampled population), or the mean $+2$ (or $+3$) standard deviations (SDs). In addition, the data were often left-censored with a known lower limit of detection (LLD). Censored observations were usually replaced by a level of one half of the LLD [3, 4]; this procedure can produce biased estimates of the geometric mean and geometric standard deviation [5].

In a recent large study of *B. pertussis* infection, a finite mixture model was developed to identify several groups with different hypothesized exposures to *B. pertussis*. Finite mixture models have been used to determine diagnostic cut-off points [6, 7] and to establish reference ranges [8]. Moulton and Halsey [9] developed a two-component mixture model to investigate predictors of antibody response following measles vaccination. The model included a component representing non- or low responders located entirely at or below the lower detection limit of the antibody assay, and a component representing responders with detectable antibody.

In this paper, we develop a similar finite mixture model to establish a diagnostic cut-off point for pertussis antibody levels. However, we use a different censoring level, allow the distribution of antibody levels above the censoring level to be a mixture of several distinct populations, and incorporate the sampling design of the study. Section 2 provides background information on the recent pertussis study. In Section 3, the finite mixture model is developed. Methods for calculating diagnostic cut-off points based on the mixture model are outlined in Section 4, and data on antibody levels from the recent pertussis study are analysed in Section 5. The methods and results are discussed in Section 6.

2. STUDY OF *B. PERTUSSIS* INFECTION

A recent study of *B. pertussis* infection used a standardized enzyme-linked immunosorbent assay (ELISA) to assess the distribution of antibody levels to *B. pertussis* antigens in a large sample representative of the U.S. population [10]. Based on these data, the investigators wanted to determine if a diagnostic cut-off point could be estimated for use in a routine diagnostic test to help detect *B. pertussis* infection in U.S. adolescents and adults with clinical symptoms of pertussis. Surplus sera were obtained for persons aged 10–49 years who participated in phase 2 (1991–1994) of the Third National Health and Nutrition Examination Survey (NHANES III), a cross-sectional survey that provided a nationally representative sample of the U.S. civilian, non-institutionalized population [11–14]. Serum immunoglobulin (Ig) G antibody levels against pertussis toxin (PT) were measured by ELISA methods [15] and quantitated with respect to a reference serum that was calibrated against the U.S. Reference Pertussis Serum, Human, Lot 3. The antibody levels were expressed in terms of ELISA units/ml (EU). The LLD, defined as the lowest amount of antibody detectable in the assay, was determined to be 1 EU. The lower limit of quantitation (LLQ) was conservatively estimated to be 20 EU, and was defined as the ELISA value above which the precision of estimated values stabilized at an acceptable level.

3. MIXTURE MODEL ANALYSIS

3.1. Rationale

Because pertussis remains an endemic disease [16], we expected some proportion of the NHANES III sample to have been recently infected with *B. pertussis*. During 1991–1994, pertussis outbreaks involving several hundred culture- or PCR-confirmed cases were documented throughout the U.S. [17–21]. However, in the U.S., pertussis vaccinations currently are routinely administered only to persons aged <7 years. Thus, to avoid confounding exposure to the bacteria with vaccination, we restricted the analysis to persons aged >9 years. Some persons do not have a detectable antibody response to either pertussis vaccination or *B. pertussis* infection [2], and among those persons who mount a detectable antibody response, antibody levels decline within a few years after vaccination [22, 23] or natural disease [24–26]. In addition, anti-PT IgG has been found to be a specific indicator for recent *B. pertussis* infection [27–29]. Given these considerations, we hypothesized that persons who provided serum represented one of several possible groups defined by previous exposure to *B. pertussis* infection(s): (a) no recent exposure to *B. pertussis* or vaccination; (b) past vaccination with waning immunity; (c) prior infection(s) with *B. pertussis*; or (d) acute or recent infection with *B. pertussis*. Distinct groups of persons with midrange antibody levels could result from individual variation in duration of time since last exposure to *B. pertussis*, in occurrence of asymptomatic *B. pertussis* infection and 'natural boosting,' or in producing IgG in response to *B. pertussis* infection (i.e. some persons with confirmed pertussis have little or no antibody response). Assuming that study subjects fell into one of several exposure groups, the distribution of quantitative results of the assay can be modelled as a mixture of several underlying populations.

3.2. Finite mixture model specification

We first considered the model that assumed a single distribution for antibody level. Then we developed the model that assumed a mixture distribution for antibody level.

3.2.1. Single left-censored normal distribution. Data on the concentration of antibody are often log-transformed to make them normally distributed (i.e. data on the original scale are lognormal). The logarithm to the base 10 transformation was used for convenience. Let X be a random variable denoting the antibody concentration level (EU), and let LLQ denote the lower limit of quantitation, a known constant. Let $Y=\log_{10}(X)$ be the transformed response and $T=\log_{10}(\text{LLQ})$ be the transformed LLQ, where $\log_{10}$ denotes logarithm to the base 10. Let n be the number of serum samples with results y_i, $i=1,\ldots,n$. Uncensored values, $y_i>T$, are assumed to come from a normal distribution with unknown mean μ and standard deviation σ:

$$P(Y=y_i)=\frac{1}{\sqrt{2\pi}\sigma}\exp\{-(y_i-\mu)^2/2\sigma^2\}$$

Values censored at T are assumed to come from the same normal distribution, with probability:

$$P(Y\leqslant T)=\Phi\{(T-\mu)/\sigma\}$$

where Φ is the cumulative distribution function of the standard normal distribution. To develop the censored data likelihood, we introduce an indicator function to denote whether the observed

value for subject i is censored or uncensored:

$$\delta_i = \begin{cases} 1, & y_i \leqslant T \\ 0, & y_i > T \end{cases}$$

The contribution of one observed value y_i to the likelihood function is then:

$$\{P(Y \leqslant T)\}^{\delta_i}\{P(Y=y_i)\}^{1-\delta_i} = [\Phi\{(T-\mu)/\sigma\}]^{\delta_i}[\exp\{-(y_i-\mu)^2/2\sigma^2\}/\sqrt{2\pi}\sigma]^{1-\delta_i} \quad (1)$$

The likelihood function for the censored data is

$$L(\mu,\sigma) = \prod_{i=1}^{n}[\Phi\{(T-\mu)/\sigma\}]^{\delta_i}[\exp\{-(y_i-\mu)^2/2\sigma^2\}/\sqrt{2\pi}\sigma]^{1-\delta_i} \quad (2)$$

3.2.2. Mixture of left-censored normal distributions. Let $j = 1,2,\ldots,J$ denote the index for the component population in the mixture distribution. The mixture distribution is modelled as the outcome of a multinomial random variable D with parameters π_j, where $P(D=j)=\pi_j$, $0<\pi_j<1$, and $\sum_{j=1}^{J}\pi_j = 1$. Because antibody levels $\leqslant$LLQ cannot be precisely measured and are not predictive of recent *B. pertussis* infection, values $\leqslant$LLQ were considered to come from a separate population of persons with little or no measurable antibody. Thus, the first component population ($D=1$) represents persons with anti-PT IgG levels at or below the LLQ. ELISA values at or below the LLQ are modelled as a point mass distribution with a single parameter, π_1. If $D=1$, then $Y \leqslant T$ with probability one, or $P(Y \leqslant T|D=1)=1$, and $P(Y=y_i|D=1)=0,\ i=1,\ldots,n$.

Log-transformed ELISA values above the LLQ are modelled as a mixture of left-censored normal distributions for the second and higher component populations ($D=2,\ldots,J$). We interpret the component population with the highest antibody levels ($D=J$) as the population of recently or acutely infected individuals. Because some culture-positive patients can have low levels of *B. pertussis* antibodies [2], we allow for the possibility that some portion of the values below the LLQ may have arisen from the component populations above the LLQ, and that the variances of the component populations above the LLQ may be unequal. Thus, conditional on $D=j,\ j=2,\ldots,J,\ Y$ is assumed to follow a normal distribution censored at the LLQ with a mean of μ_j and a variance of σ_j.

To develop the likelihood, we condition the probabilities in equation (1) on D and substitute values for the conditional probabilities:

$$\begin{aligned}
&\{P(Y \leqslant T)\}^{\delta_i}\{P(Y=y_i)\}^{1-\delta_i} \\
&= \left\{\sum_{j=1}^{J} P(Y \leqslant T|D=j)P(D=j)\right\}^{\delta_i}\left\{\sum_{j=1}^{J} P(Y=y_i|D=j)P(D=j)\right\}^{1-\delta_i} \\
&= \left\{P(D=1)+\sum_{j=2}^{J} P(Y \leqslant T|D=j)P(D=j)\right\}^{\delta_i}\left\{\sum_{j=2}^{J} P(Y=y_i|D=j)P(D=j)\right\}^{1-\delta_i} \\
&= \left[\pi_1+\sum_{j=2}^{J}\pi_j\Phi\{(T-\mu_j)/\sigma_j\}\right]^{\delta_i}\left[\sum_{j=2}^{J}\pi_j\exp\{-(y_i-\mu_j)^2/2\sigma_j^2\}/\sqrt{2\pi}\sigma_j\right]^{1-\delta_i}
\end{aligned}$$

Note that any observed value above the LLQ ($y_i > T$) may have arisen only from one of the component distributions beyond the first one, whereas a censored value ($y_i \leqslant T$) may have come from any of the component distributions. The likelihood function for the mixture model with J component populations is

$$L(\boldsymbol{\Psi}) = \prod_{i=1}^{n} \left[\left(1 - \sum_{j=2}^{J} \pi_j \right) + \sum_{j=2}^{J} \pi_j \Phi\{(T - \mu_j)/\sigma_j\} \right]^{\delta_i} \times \left[\sum_{j=2}^{J} \pi_j \exp\{-(y_i - \mu_j)^2/2\sigma_j^2\}/\sqrt{2\pi}\sigma_j \right]^{1-\delta_i} \tag{3}$$

where $\boldsymbol{\Psi} = (\pi_2, \pi_3, \ldots, \pi_J, \mu_2, \mu_3, \ldots, \mu_J, \sigma_2, \sigma_3, \ldots, \sigma_J)'$ is the vector containing all of the parameters. Note that when $\pi_2 = 1$, or when all of the ELISA values are assumed to follow a single distribution, the likelihood for the mixture model (3) reduces to that for the model that assumes a single distribution (2).

This mixture model for left-censored antibody level data is parallel to long-term survival or cure mixture models for right-censored failure-time data in which a proportion of patients are cured and never have the failure or event of interest [30, 31]. In a cure mixture model, failure times for cured patients are considered right-censored observations, and those for uncured patients are assumed to follow a survival distribution [32, 33].

3.3. Accounting for the NHANES III design

The NHANES III design was accounted for by incorporating sampling weights into the mixture model likelihood, and by implementing a jackknife variance estimator. NHANES III used a stratified, multistage, probability design [11, 12]. To produce national estimates, each sampled person in NHANES III was assigned a sampling weight that incorporated the probability of selection and included adjustments for non-coverage and non-response. We used the phase 2 sample final weights in the original data sources [13, 14]. Let w_i be the NHANES III sample weight for subject i, $i = 1, \ldots, n$. We rescaled the sample weights by dividing each weight by the average weight, so that the sum of the rescaled weights was equal to the sample size: $w_i^* = w_i/(\sum_{i=1}^{n} w_i/n)$. The rescaled sample weights were incorporated into the mixture model likelihood (3) to form a weighted likelihood [34, 35]:

$$WL(\boldsymbol{\Psi}) = \prod_{i=1}^{n} \left(\left[\left(1 - \sum_{j=2}^{J} \pi_j \right) + \sum_{j=2}^{J} \pi_j \Phi\{(T - \mu_j)/\sigma_j\} \right]^{\delta_i} \times \left[\sum_{j=2}^{J} \pi_j \exp\{-(y_i - \mu_j)^2/2\sigma_j^2\}/\sqrt{2\pi}\sigma_j \right]^{1-\delta_i} \right)^{w_i^*} \tag{4}$$

Because the sample design for each phase of NHANES III was a one-primary sampling unit (PSU)-per-stratum selection, variance estimation was carried out by pairing strata that were similar [12]. For phase 2, 23 pseudo-strata and a pair of PSU codes per stratum were

designed [13, 14]. The jackknife variance estimator [34] was used to estimate variances of parameter estimates in the mixture models.

3.4. Model fitting and selection

3.4.1. Maximum likelihood estimation. Parameters in each model were estimated by maximum likelihood. We used the Newton–Raphson (NR) algorithm [36] to calculate the maximum likelihood estimates (MLEs). The NR algorithm was carried out using SAS PROC NLP [37], which used analytic formulas for the gradient vector and Hessian matrix. The maximization was subject to the constraints that each multinomial proportion in the mixture model must lay between 0 and 1 ($0<\pi_j<1, j=2,\ldots,J$). In addition, each variance parameter was constrained to be greater than 0 ($\sigma_j>0, j=2,\ldots,J$). The MLEs for the mixture model can also be obtained by using the expectation-maximization (EM) algorithm (Appendix A).

3.4.2. Model selection. For mixture models, the likelihood ratio test that compares the model with j components to the model with $j+1$ components is not valid because regularity conditions do not hold for the likelihood ratio test statistic to have its usual asymptotic chi-squared distribution [30]. Models with increasing numbers of components were fit to determine the best model to cluster the antibody levels. The Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC) [30] were used as guidelines to decide how many components to include in the model. Smaller values of these statistics indicate better fits.

3.5. Appropriateness of the log transformation

After selecting a model, we evaluated whether the log-transformed antibody level data followed the normal mixture model (4) by applying a power transformation to the untransformed antibody level [38]:

$$x_i^{(\lambda)}=\begin{cases}(x_i^{\lambda}-1)/\lambda, & \lambda\neq 0\\ \ln(x_i), & \lambda=0\end{cases}$$

and to the censoring level:

$$\mathrm{LLQ}^{(\lambda)}=\begin{cases}(\mathrm{LLQ}^{(\lambda)}-1)/\lambda, & \lambda\neq 0\\ \ln(\mathrm{LLQ}), & \lambda=0\end{cases}$$

The weighted likelihood for the selected model of the power-transformed data is

$$WL(\boldsymbol{\Psi},\lambda)=\prod_{i=1}^{n}\left(\begin{array}{l}\left[\left(1-\sum_{j=2}^{k}\pi_j\right)+\sum_{j=2}^{k}\pi_j\Phi\{(\mathrm{LLQ}^{(\lambda)}-\mu_j)/\sigma_j\}\right]^{\delta_i}\\ \times\left[\left(\sum_{j=2}^{k}\pi_j\exp\{-(x_i^{(\lambda)}-\mu_j)^2/2\sigma_j^2\}/\sqrt{2\pi}\sigma_j\right)x_i^{\lambda-1}\right]^{1-\delta_i}\end{array}\right)^{w_i^*} \tag{5}$$

where k denotes the number of components in the selected model, and the term $x_i^{\lambda-1}$ in the likelihood is the Jacobian of the transformation from x_i to $x_i^{(\lambda)}$. The log likelihood was

maximized for a series of fixed values of λ to generate a profile log likelihood. The 95 per cent confidence interval for the maximum value of λ was used to assess whether the log transformation ($\lambda = 0$) provided the scale for our data that satisfied the normal distribution assumption [7, 39].

4. DIAGNOSTIC CUT-OFF POINTS

A cut-off point for diagnosing acute *B. pertussis* infection was calculated based on the selected mixture model. We assumed that the selected model correctly identified distinct populations and the component population with the highest antibody levels represented persons recently infected with *B. pertussis*. The specificity and sensitivity corresponding to each cut-off point were based on analyses of the two overlapping normal distributions estimated for the two component populations with the highest antibody levels to distinguish recently infected persons from other persons. Sensitivity is the proportion of truly infected patients classified as positive, and specificity is the proportion of truly non-infected patients classified as negative [40]. A cut-off point was chosen to attain a high specificity (99 per cent) under the model. For epidemiologic investigations, it is important to have a highly specific test to limit the number of false-positive results (many respiratory infections have symptomatology similar to *B. pertussis* infection). False-positive results for *B. pertussis* infections might cause unnecessary case or outbreak investigations, including administration of antibiotic treatments and prophylaxis among suspected cases.

We also evaluated alternative cut-off points that have been commonly used in previously published research. These alternative cut-off points assume a single healthy population; the log-transformed antibody levels of an external control group are assumed to arise from a single normal distribution: the mean $+2$ or $+3$ SDs, the 99th percentile, and the distribution-free 99 per cent UTL [41]. For each of these statistics, EU values $<$LLD were assigned a value of $\frac{1}{2}$LLD.

5. RESULTS

5.1. Study population

All analyses were restricted to persons aged 10–49 years in phase 2 of NHANES III who had sufficient quantity of serum. Of the 7747 persons aged 10–49 years who were interviewed, 94 per cent ($n = 7260$) were examined. Of those examined, 75 per cent ($n = 5419$) had sufficient quantity of serum. A total of 5409 serum samples had an anti-PT IgG value that met the laboratory acceptance criteria for a valid result. Of these, 881 (16 per cent) results were $<$LLD (1 EU), and 4577 (85 per cent) results were $\leqslant$LLQ (20 EU). Of the 5409 persons with valid samples, the median NHANES III sample weight was 7592 (range: 428–262 189). The sum of the sample weights was 125.3 million, which represents the number of persons in the target population represented by the sampled persons. Figure 1 shows the histogram of weighted anti-PT IgG levels for results $>$LLQ.

5.2. Model fitting

We fit a series of models that assumed increasing numbers of components using the weighted maximum likelihood approach (4), with starting values derived from clustering observations

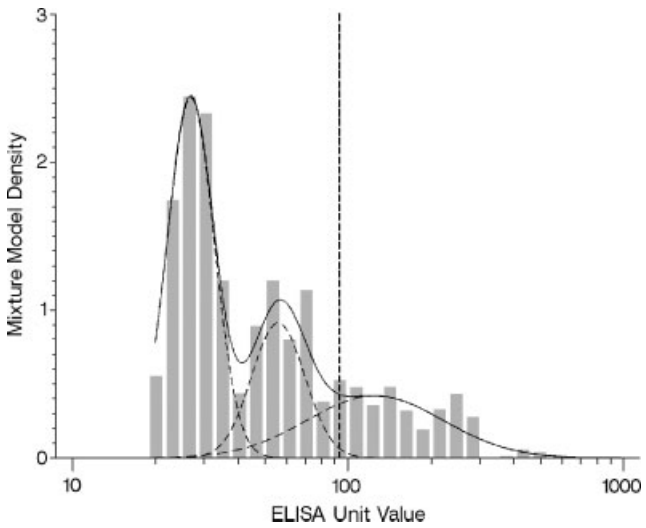

Figure 1. Four-component mixture model fit for weighted anti-PT IgG levels among persons 10–49 years of age ($n = 5409$). The histogram graphs observed data $>$LLQ. Observed values $\leqslant$LLQ (85 per cent of sample size) and one observed value >1000 EU are not shown. The solid line plots the mixture model fit for components 2, 3, and 4; component 1 is $\leqslant$LLQ and is not shown. The dashed lines indicate the underlying component populations in the mixture model. A reference dashed line is drawn at the selected cut-off point of 94 EU.

(Table I). For each model, we checked whether the parameter estimates were global optima by refitting the model using different sets of random starting values. For the three-, four-, and five-component models, some sets of random starting values resulted in spurious local maximizers [30], where at least one of the components located above the LLQ had a very small value for the mixing proportion or a much smaller standard deviation than those for the other components. Of the solutions to the estimating equations that were not spurious, some were on the boundary of the parameter space. For each model, after eliminating the spurious and boundary solutions, none of the solutions had a higher log likelihood than that found by using the K-means clustering starting values.

For the five-component model, both the NR and EM algorithms using the K-means clustering starting values resulted in the same spurious local maximizer (Table I). In the fourth component of this model, the weighted frequencies of a few observations (126 and 129 EU) were relatively high compared with those of nearby observations. Use of random starting values resulted in either boundary solutions or other spurious local maximizers with lower log likelihoods.

5.3. Number of components in mixture model

The lowest value of AIC occurred for the four-component model, and the lowest value of BIC occurred for the three-component model (Table I). However, the value of BIC for the four-component model was relatively close to that for the three-component model (4731.5 *versus* 4729.1). In addition, the four-component model more clearly identified a component

Table I. Model selection criteria for mixture models fit to weighted anti-PT IgG data.

Number of populations*	AIC	BIC	Component population number	Proportion in component population	Mean EU (mean EU ± 2 SDs)
1	4749.1	4762.2	1	1.000	4 (0.1, 106)
2	4747.7	4767.5	1	0.481	⩽LLQ
			2	0.519	10 (0.6, 160)
3	4689.5	4729.1	1	0.823	⩽LLQ
			2	0.053	27 (19, 37)
			3	0.124	51 (9, 303)
4	**4672.2**	**4731.5**	**1**	**0.838**	**⩽LLQ**
			2	**0.084**	**27 (18, 40)**
			3	**0.036**	**56 (36, 87)**
			4	**0.042**	**126 (40, 392)**
5	4672.3	4751.5	1	0.837	⩽LLQ
			2	0.083	27 (18, 39)
			3	0.044	57 (34, 95)
			4	0.003	128 (119, 137)
			5	0.033	143 (49, 416)

AIC, Akaike's Information Criterion; BIC, Bayesian Information Criterion; EU, ELISA units/ml; SD, standard deviation; LLQ, lower limit of quantitation (20 EU).
*Selected model is in boldface.

with the highest antibody levels that we speculated represented persons with recent *B. pertussis* infection. The fourth component had a much higher mean than the third component. Its relatively large variance was plausible because persons infected with *B. pertussis* have heterogeneous anti-PT IgG responses [26]. Thus, the four-component model was chosen to represent the distribution of anti-PT IgG levels. The mixture density estimated by the four-component model indicated that the model reflected the frequency distribution of ELISA values $>$LLQ (Figure 1).

The first population in the four-component model, with antibody levels $\leqslant$LLQ, included persons with a non-detectable or low level of anti-PT IgG. The second and third component populations included persons with moderate levels of antibody. The fourth component population included persons who had relatively high levels of antibody. Results from the mixture model suggested that the group with the highest antibody levels comprised 4.23 per cent (95 per cent CI, 2.27–6.19 per cent) of the U.S. population aged 10–49 years.

We evaluated the statistical significance of each parameter estimate in the four-component model. For each parameter estimate, we applied the jackknife variance estimation procedure to estimate the variance of the parameter estimate, and then performed a t-test. All of the parameter estimates in the model were statistically significant ($P<0.05$) (Table II).

5.4. Evaluation of log transformation

To evaluate whether the log transformation provided a suitable scale for the pertussis antibody level data, we examined the log likelihood for the four-component model as a function

Table II. Parameter estimates and standard errors for the four-component mixture model fit to weighted anti-PT IgG data (logarithm to the base 10 scale).

Distribution	Parameter	Estimate (SE)	P-value*
Mixing†	π_2	0.084 (0.011)	<0.001
	π_3	0.036 (0.013)	0.005
	π_4	0.042 (0.009)	<0.001
Second component	μ_2	1.429 (0.011)	<0.001
	σ_2	0.085 (0.010)	<0.001
Third component	μ_3	1.747 (0.030)	<0.001
	σ_3	0.096 (0.023)	<0.001
Fourth component	μ_4	2.099 (0.102)	<0.001
	σ_4	0.247 (0.043)	<0.001

SE, standard error.
*t-test (one-tailed) of the null hypothesis that the parameter is equal to zero. The t-test is based on a denominator degrees of freedom of 23.
†The first component population is modelled as the proportion of the target population that is at or below the lower limit of quantitation (20 ELISA units/ml) ($\hat{\pi}_1 = 1 - \sum_{j=2}^{4} \hat{\pi}_j = 0.838$).

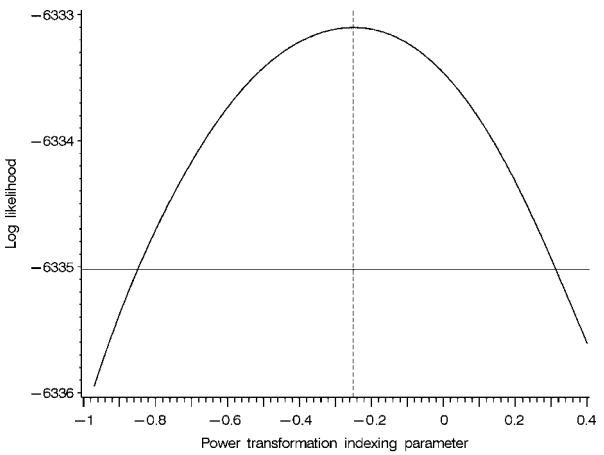

Figure 2. Profile log likelihood *versus* power transformation indexing parameter. A vertical reference dashed line is drawn at the maximum. A horizontal reference line is drawn to denote the 95 per cent confidence interval for the maximum.

of the indexing parameter of the power transformation (5). The maximum value occurred at $\lambda = -0.250$ with a 95 per cent confidence interval that contained zero (log transformation) (Figure 2). Therefore, the log transformation was a reasonable transformation for our chosen mixture model.

Table III. Diagnostic cut-off points for acute pertussis infection based on the four-component mixture model or previously published methods applied to weighted anti-PT IgG data.

Basis of cut-off point and statistic		Cut-off point (EU)*
Mixture model with specificity, sensitivity†		
0.950	0.783	80
0.990	**0.697**	**94**
0.999	0.586	111
0.9999	0.489	128
Previously published methods		
Mean + 2 SDs		94
Mean + 3 SDs		436
99th percentile		198
99% UTL		232

EU, ELISA units/ml; SD, standard deviation; UTL, upper tolerance limit (distribution-free).
*Selected cut-off point is in boldface.
†Specificity and sensitivity were based on the two overlapping normal distributions estimated for the two component populations with the highest antibody levels.

5.5. *Cut-off points for diagnosing B. pertussis infection*

Based on a 99 per cent specificity with respect to the third and fourth groups in the four-component mixture model, the diagnostic cut-off point for recent infection was chosen to be 94 EU and resulted in a sensitivity of 70 per cent (Table III). In contrast, the cut-off point for acute infection defined by the mean +3 SDs was 436 EU, and the cut-off defined by the 99 per cent UTL was 232 EU, both higher than the cut-off established using the mixture model.

The model-based cut-off point of 94 EU is well above the low end of the distribution for the fourth group in the mixture model (Figure 1). For this reason, we calculated a second cut-off point of 49 EU (the fifth percentile of the fourth group) to define a range of ELISA values that might possibly be associated with acute or recent *B. pertussis* infection. The interval 49–93 EU could be used to define an indeterminate group for acute *B. pertussis* infection.

6. DISCUSSION

A finite mixture model was developed to establish a diagnostic cut-off point for anti-PT IgG levels so that a standardized ELISA could be used to diagnose *B. pertussis* infections in U.S. adolescents and adults with an acute cough illness. Traditional approaches to developing a diagnostic cut-off value for anti-PT IgG levels have assumed one healthy homogenous population that is lognormally distributed and used the 99 per cent UTL or mean +2 (or +3) SDs to define the diagnostic cut-off value [1–3]. Such criteria have been applied to a control group (e.g. unexposed persons, healthy volunteers, blood donors) to help diagnose pertussis in numerous pertussis studies [42]. Another method defined the cut-off point as the maximum

observed value in the control group to ensure maximum specificity [43]. These traditional approaches to deriving diagnostic cut-off values are unsatisfactory for several reasons. Antibody levels in the control group may not be lognormally distributed, even if none of the group had been recently exposed to disease or vaccination. Additionally, a positive value indicates only that the study subject's antibody level is extremely unlikely to have arisen if the subject had been a member of the control (unexposed) group, irrespective of the range of values among diseased persons. High antibody levels may be due to disease or vaccination, and levels decline gradually over several months to years, which may account for the different midrange populations seen in the distribution of anti-PT IgG levels. The possibility of more than one underlying population is not taken into account in the traditional approaches.

A more comprehensive approach to discriminating between persons who have and have not been recently infected with *B. pertussis* requires knowledge of the antibody distributions of persons in these groups [44, 45]. Our mixture model analysis assumed the existence of more than one exposure group in the NHANES III population. In the four-component mixture model for anti-PT IgG levels, persons in the fourth group were presumed to have been recently infected with *B. pertussis* (including persons who had mild or asymptomatic infection), whereas those in the third group likely had less recent infection. The size of the NHANES III sample tested and the presence of endemic disease in the U.S. would make it likely that we would detect recently exposed adolescents and adults in our study. The cut-off point chosen to distinguish between the third and fourth groups with high model specificity (99 per cent) resulted in moderate sensitivity (70 per cent); a moderate sensitivity was expected because some persons with recent or acute *B. pertussis* infection produce little or no measurable anti-PT IgG [46, 47]. This approach yielded a diagnostic cut-off point of 94 EU of anti-PT IgG.

In addition to identifying multiple exposure groups, our mixture model accounted for the preponderance of observed antibody levels (85 per cent) that fell below the lower limit of quantitation. This large proportion of observations lying below the censoring level was much higher than would be expected under the assumption of a single left-censored lognormal distribution. In general, any limit could be used as the censoring level, including a limit that is known to represent protection from disease or response to vaccination. Deriving a diagnostic cut-off point from a mixture model with a point mass distribution located entirely at or below the censoring level may be more appropriate than arbitrarily assigning some value for results that fall below the censoring level, and then calculating a diagnostic cut-off point based on the resulting single distribution. Power transformations may also be used to explore the scale for antibody level data.

Another feature of our mixture model analysis was accounting for the probability sampling design of NHANES III. Because the objective of the analysis was to obtain an unbiased estimate of the diagnostic cut-off point, we used weighted maximum likelihood estimation, which has been used previously for analysing survey data with complex probability sampling designs [34, 35]. We also used a relatively simple jackknife procedure for obtaining variance estimates. Jackknife variance estimation is a commonly used method for analysing health surveys [34] and is easier to implement than a linearization method for complex surveys [35].

The mixture model analysis of NHANES III data made several assumptions. The assumption that the fourth population in the mixture model represented individuals with recent or acute *B. pertussis* infection could not be evaluated because no data were available to indicate that persons classified as positive to anti-PT IgG had been recently infected with *B.*

pertussis. The NHANES III data set contained no diagnostic or clinical information to indicate the participants' actual exposure states. Another assumption was that the ELISA values were lognormally distributed for component populations in the mixture models; the results of applying the power transformation to our data suggested that this was a reasonable assumption. We also assumed that a single serologic value can be used to identify infected persons over the age range of the data (10–49 years). The age range was chosen to avoid possible confounding due to anti-PT IgG response to vaccination. Persons aged 10–49 years with an elevated anti-PT IgG value were likely to have been exposed to *B. pertussis* infection rather than pertussis vaccination because the last dose of pertussis vaccine was recommended to be given at 4–6 years of age during the time of the survey (1991–1994) [48]. We are currently planning a study to define the diagnostic accuracy of our cut-off point and determine whether a range of ELISA values below the cut-off can identify additional cases of pertussis.

Several aspects of fitting mixture models deserve comment. We were able to fit a mixture model to our data set because enough data points occurred in the right tail of the antibody level distribution and the component populations were well-separated. However, maximum likelihood estimation may not give accurate parameter estimates if the component distributions are not well separated or the sample size is small [49]. For the five-component model, a spurious local maximizer was the only possible solution with five components regardless of which fitting algorithm or set of starting values was used. The Newton–Raphson algorithm was relatively easy to carry out, including refitting the models with different starting values to determine whether global convergence of parameter estimates had been achieved. A potentially useful alternative approach to fitting is a hybrid EM/Newton–Raphson algorithm to first take advantage of the good global convergence properties of the EM algorithm and then exploit the rapid local convergence of the Newton–Raphson algorithm [30, 50]. Such an algorithm would entail employing the EM algorithm for several iterations to improve the starting values, then switching to the Newton–Raphson algorithm for the remaining iterations.

In summary, a mixture model can identify more than one exposure group and provide a useful framework for deriving diagnostic cut-off points for antibody level distributions. A mixture model can also account for an excess number of antibody values that fall below a censoring level, and incorporate complex sampling designs. We believe that the mixture model approach to establishing a diagnostic cut-off point for antibody levels is more appropriate than the traditional approach of assuming a single population. Mixture model analysis of antibody levels should be considered for establishing diagnostic criteria for pertussis and other diseases.

APPENDIX A: EXPECTATION-MAXIMIZATION ALGORITHM

This appendix develops the EM algorithm [36] for the mixture model with J components using the weighted likelihood approach (4). Let z_{ij} denote the indicator variable for the unobserved or missing data:

$$z_{ij} = \begin{cases} 1 & \text{if } y_i \text{ arose from component } j, \quad i = 1, \ldots, n; j = 1, \ldots, J \\ 0 & \text{otherwise} \end{cases}$$

The complete-data weighted likelihood is

$$WL_C(\boldsymbol{\Psi}) = \prod_{i=1}^{n} \left(\left[\left(1 - \sum_{j=2}^{J} \pi_j \right)^{z_{i1}} \prod_{j=2}^{J} [\pi_j \Phi\{(T-\mu_j)/\sigma_j\}]^{z_{ij}} \right]^{\delta_i} \times \left[\prod_{j=2}^{J} [\pi_j \exp\{-(y_i-\mu_j)^2/2\sigma_j^2\}/\sqrt{2\pi}\sigma_j]^{z_{ij}} \right]^{1-\delta_i} \right)^{w_i^*}$$

A.1. E-step

Estimates of the missing data are:

$$\hat{z}_{ij} = \begin{cases} \dfrac{\pi_j^{(k)} \Phi\{(T-\mu_j^{(k)})/\sigma_j^{(k)}\}}{\left(1 - \sum_{j=2}^{J} \pi_j^{(k)}\right) + \sum_{j=2}^{J} \pi_j^{(k)} \Phi\{(T-\mu_j^{(k)})/\sigma_j^{(k)}\}}, & \delta_i = 1 \\ \dfrac{\pi_j^{(k)} \exp\{-(y_i-\mu_j^{(k)})^2/2(\sigma_j^{(k)})^2\}/\sqrt{2\pi}\sigma_j^{(k)}}{\sum_{j=2}^{J} \pi_j^{(k)} \exp\{-(y_i-\mu_j^{(k)})^2/2(\sigma_j^{(k)})^2\}/\sqrt{2\pi}\sigma_j^{(k)}}, & \delta_i = 0 \end{cases} \qquad j = 2,\ldots,J$$

$$\hat{z}_{i1} = \begin{cases} 1 - \sum_{j=2}^{J} \hat{z}_{ij}, & \delta_i = 1 \\ 0, & \delta_i = 0 \end{cases}$$

This is the posterior probability that the observed value y_i belongs to the jth component of the mixture.

A.2. M-step

Compute the MLEs:

$$\mu_j^{(k+1)} = \frac{\sum_{i=1}^{n} w_i^*(1-\delta_i)\hat{z}_{ij} y_i - \sum_{i=1}^{n} w_i^* \delta_i \hat{z}_{ij} [\phi\{(T-\mu_j^{(k)})/\sigma_j^{(k)}\}/\Phi\{(T-\mu_j^{(k)})/\sigma_j^{(k)}\}]\sigma_j^{(k)}}{\sum_{i=1}^{n} w_i^*(1-\delta_i)\hat{z}_{ij}}$$

$j = 2,\ldots,J$, and

$$\sigma_j^{(k+1)} = \frac{-b + \sqrt{b^2 - 4ac}}{2a}, \quad j = 2,\ldots,J$$

where

$$a = \sum_{i=1}^{n} w_i^*(1-\delta_i)\hat{z}_{ij}$$

$$b = \sum_{i=1}^{n} w_i^* \delta_i \hat{z}_{ij} [\phi\{(T-\mu_j^{(k)})/\sigma_j^{(k)}\}/\Phi\{(T-\mu_j^{(k)})/\sigma_j^{(k)}\}](T-\mu_j^{(k)})$$

$$c = -\sum_{i=1}^{n} w_i^*(1-\delta_i)\hat{z}_{ij}(y_i-\mu_j^{(k)})^2$$

and ϕ denotes the probability density function of the standard normal distribution.

The E- and M-steps are alternated repeatedly until the difference $WL(\boldsymbol{\Psi}^{(k+1)}) - WL(\boldsymbol{\Psi}^{(k)})$ changes by less than 1×10^{-6}.

A.3. Starting values

Starting values for the mean and standard deviation in the model that assumed a single left-censored normal distribution can be obtained by calculating the weighted mean and standard deviation of all log-transformed observations, after assigning EU values $<$LLD a value of $\frac{1}{2}$LLD. Starting values for the parameters in the mixture model can be obtained by first clustering the uncensored values using the K-means clustering method [51] to obtain initial estimates of the weighted prevalence, mean, and standard deviation for each assumed component $>$LLQ. The K-means clustering method can be implemented using SAS PROC FASTCLUS [52]. These initial parameter estimates are then used in the first E-step to derive the initial values $z_{ij}^{(0)}$, as well as in the first M-step.

Another way of specifying an initial set of parameter values is to generate random starting values by randomly dividing the uncensored data into $J-1$ groups. For each uncensored value, we randomly generate an integer from the set $2,\ldots,J$, and then form the initial values:

$$z_{ij}^{(0)} = \begin{cases} 1 & \text{if random integer} = j, \quad j = 2,\ldots,J, \\ 0, & j = 1 \end{cases}$$

All censored values are assigned to the first component:

$$z_{ij}^{(0)} = \begin{cases} 1, & j = 1 \\ 0, & j = 2,\ldots,J \end{cases}$$

In the first M-step, initial values $\mu_j^{(0)}$ and $\sigma_j^{(0)}$ are obtained by calculating the weighted mean and standard deviation using those uncensored values for which $z_{ij}^{(0)} = 1$.

ACKNOWLEDGEMENTS

We are grateful to Larry Moulton (The John Hopkins University Bloomberg School of Public Health, Baltimore) for sharing his SAS PROC NLP code; Barry Graubard (National Cancer Institute, Bethesda) for advice regarding statistical analysis of NHANES III data; and John Moran (NIP, CDC), Trudy Murphy (NIP, CDC), Susan Chu (NIP, CDC), Phil Smith (NIP, CDC), Jingyee Kou (Center for Biologics Evaluation and Research, Food and Drug Administration), and George Reed (National Eye Institute, National Institutes of Health) for useful comments on earlier versions of the manuscript. We also thank two referees whose comments substantially improved the manuscript.

REFERENCES

1. Marchant CD, Loughlin AM, Lett SM, Todd CW, Wetterlow LH, Bicchieri R, Higham S, Etkind P, Silva E, Siber GR. Pertussis in Massachusetts, 1981–1991: incidence, serologic diagnosis, and vaccine effectiveness. *The Journal of Infectious Diseases* 1994; **169**:1297–1305.
2. Yih WK, Lett SM, des Vignes FN, Garrison KM, Sipe PL, Marchant CD. The increasing incidence of pertussis in Massachusetts adolescents and adults, 1989–1998. *The Journal of Infectious Diseases* 2000; **182**:1409–1416.
3. Strebel P, Nordin J, Edwards K, Hunt J, Besser J, Burns S, Amundson G, Baughman A, Wattigney W. Population-based incidence of pertussis among adolescents and adults, Minnesota, 1995–1996. *The Journal of Infectious Diseases* 2001; **183**:1353–1359.

4. Cattaneo LA, Reed GW, Haase DH, Wills MJ, Edwards KM. The seroepidemiology of *Bordetella pertussis* infections: a study of persons ages 1–65 years. *The Journal of Infectious Diseases* 1996; **173**:1256–1259.
5. Hornung RW, Reed LD. Estimation of average concentration in the presence of nondetectable values. *Applied Occupational and Environmental Hygiene* 1990; **5**:46–51.
6. Parker RA, Erdman DD, Anderson LJ. Use of mixture models in determining laboratory criterion for identification of seropositive individuals: application of parvovirus B19 serology. *Journal of Virological Methods* 1990; **27**:135–144.
7. Thompson TJ, Smith PJ, Boyle JP. Finite mixture models with concomitant information: assessing diagnostic criteria for diabetes. *Applied Statistics* 1998; **47**:393–404.
8. Lawrence CJ, Trewin VF. The construction of biochemical reference ranges and the identification of possible adverse drug reactions in the elderly. *Statistics in Medicine* 1991; **10**:831–837.
9. Moulton LH, Halsey NA. A mixture model with detection limits for regression analyses of antibody response to vaccine. *Biometrics* 1995; **51**:1570–1578.
10. Baughman AL, Bisgard KM, Edwards KM, Guris D, Decker MD, Holland K, Meade BD, Lynn F. Establishment of diagnostic cutoff points for levels of serum antibodies to pertussis toxin, filamentous hemagglutinin, and fimbriae in adolescents and adults in the United States. *Clinical and Diagnostic Laboratory Immunology* 2004; **11**:1045–1053.
11. National Center for Health Statistics. Plan and operation of the Third National Health and Nutrition Examination Survey, 1988–94. *Vital and Health Statistics* 1994; **1**(32).
12. Ezzati TM, Massey JT, Waksberg J, Chu A, Maurer KR. Sample design: third National Health and Nutrition Examination Survey. *Vital and Health Statistics* 1992; **2**(113).
13. U.S. Department of Health and Human Services (DHHS). National Center for Health Statistics. *Third National Health and Nutrition Examination Survey*, *1988–1994*, *NHANES III Household Youth Data File* (*CD-ROM*). Public Use Data File Documentation Number 77550. Centers for Disease Control and Prevention: Hyattsville, MD, 1996.
14. DHHS. *Third National Health and Nutrition Examination Survey*, *1988–1994*, *NHANES III Household Adult Data File* (*CD-ROM*). Public Use Data File Documentation Number 77560. Centers for Disease Control and Prevention: Hyattsville, MD, 1996.
15. Meade BD, Deforest A, Edwards KM, Romani TA, Lynn F, O'Brien CH, Swartz CB, Reed GF, Deloria MA. Description and evaluation of serologic assays used in a multicenter trial of acellular pertussis vaccines. *Pediatrics* 1995; **96**:570–575.
16. Centers for Disease Control and Prevention. Pertussis—United States, 1997–2000. *Morbidity and Mortality Weekly Report* 2002; **51**:73–76.
17. Christie CD, Marx ML, Marchant CD, Reising SF. The 1993 epidemic of pertussis in Cincinnati—resurgence of disease in a highly immunized population of children. *The New England Journal of Medicine* 1994; **331**:16–21.
18. Hardy IR, Strebel PM, Wharton M, Orenstein WA. The 1993 pertussis epidemic in Cincinnati [letter]. *The New England Journal of Medicine* 1994; **331**:1455–1456.
19. Rosenthal S, Strebel P, Cassiday P, Sanden G, Brusuelas K, Wharton M. Pertussis infection among adults during the 1993 outbreak in Chicago. *The Journal of Infectious Diseases* 1995; **171**:1650–1652.
20. Kenyon TA, Izurieta H, Shulman ST, Rosenfeld E, Miller M, Daum R, Strebel PM. Large outbreak of pertussis among young children in Chicago, 1993: investigation of potential contributing factors and estimation of vaccine effectiveness. *The Pediatric Infectious Disease Journal* 1996; **15**:655–661.
21. Guris D, Strebel PM, Bardenheier B, Brennan M, Tachdjian R, Finch E, Wharton M, Livengood JR. Changing epidemiology of pertussis in the United States: increasing reported incidence among adolescents and adults, 1990–1996. *Clinical Infectious Diseases* 1999; **28**:1230–1237.
22. Lambert HJ. Epidemiology of a small pertussis outbreak in Kent County, Michigan. *Public Health Reports* 1965; **80**:365–369.
23. Le T, Cherry JD, Chang S-J, Knoll MD, Lee ML, Barenkamp S, Bernstein D, Edelman R, Edwards KM, Greenberg D, Keitel W, Treanor J, Ward JI. Immune responses and antibody decay after immunization of adolescents and adults with an acellular pertussis vaccine: the APERT study. *The Journal of Infectious Diseases* 2004; **190**:535–544.
24. Mortimer EA. Pertussis and its prevention: a family affair. *The Journal of Infectious Diseases* 1990; **161**:473–479.
25. Versteegh FGA, Schellekens JFP, Nagelkerke AF, Roord JJ. Laboratory-confirmed reinfections with *Bordetella pertussis. Acta Paediatrica* 2002; **91**:95–99.
26. Teunis PFM, van der Heijden OG, de Melker HE, Schellekens JFP, Versteegh FGA, Kretzschmar MEE. Kinetics of the IgG antibody response to pertussis toxin after infection with *B. pertussis. Epidemiology and Infection* 2002; **129**:479–489.
27. Arico B, Rappuoli R. *Bordetella parapertussis* and *Bordetella bronchiseptica* contain transcriptionally silent pertussis toxin genes. *Journal of Bacteriology* 1987; **169**:2847–2853.

28. Schmitt-Grohe S, Cherry JD, Heininger U, Uberall MA, Pineda E, Stehr K. Pertussis in German adults. *Clinical Infectious Diseases* 1995; **21**:860–866.
29. Hodder SL, Cherry JD, Mortimer EA, Ford AB, Gornbein J, Papp K. Antibody responses to *Bordetella pertussis* antigens and clinical correlations in elderly community residents. *Clinical Infectious Diseases* 2000; **31**:7–14.
30. McLachlan G, Peel D. *Finite Mixture Models*. Wiley: New York, 2000.
31. McLachlan GJ, McGiffin DC. On the role of finite mixture models in survival analysis. *Statistical Methods in Medical Research* 1994; **3**:211–226.
32. Yamaguchi K. Accelerated failure-time regression models with a regression model of surviving fraction: an application to the analysis of 'permanent employment' in Japan. *Journal of the American Statistical Association* 1992; **87**:284–292.
33. Peng Y, Dear KBG, Denham JW. A generalized *F* mixture model for cure rate estimation. *Statistics in Medicine* 1998; **17**:813–830.
34. Korn EL, Graubard BI. *Analysis of Health Surveys*. Wiley: New York, 1999.
35. Wedel M, ter Hofstede F, Steenkamp J-BEM. Mixture model analysis of complex samples. *Journal of Classification* 1998; **15**:225–244.
36. McLachlan GJ, Krishnan T. *The EM Algorithm and Extensions*. Wiley: New York, 1997.
37. SAS Institute Inc. The NLP procedure. In *SAS/OR User's Guide*: *Mathematical Programming*, *Version 8*. SAS Institute Inc.: Cary, NC, 1999.
38. Emerson JD, Stoto MA. Transforming data. In *Understanding Robust and Exploratory Data Analysis*, Hoaglin DC, Mosteller F, Tukey JW (eds). Wiley: New York, 1983; 97–128.
39. Gutierrez RG, Carroll RJ, Wang N, Lee G-H, Taylor BH. Analysis of tomato root initiation using a normal mixture distribution. *Biometrics* 1995; **51**:1461–1468.
40. Zhou XH, Obuchowski NA, McClish DK. *Statistical Methods in Diagnostic Medicine*. Wiley: New York, 2002; 15–56.
41. Hahn GJ, Meeker WQ. *Statistical Intervals*: *A Guide for Practitioners*. Wiley: New York, 1991.
42. Meade BD, Mink CM, Manclark CR. Serodiagnosis of pertussis. In *Proceedings of the 6th International Symposium on Pertussis*, Manclark CR (ed.), DHHS no. (FDA) 90-1164, 1990; 322–329.
43. Steketee RW, Burstyn DG, Wassilak SGF, Adkins WN, Polyak MB, Davis JP, Manclark CR. A comparison of laboratory and clinical methods for diagnosing pertussis in an outbreak in a facility for the developmentally disabled. *The Journal of Infectious Diseases* 1988; **157**:441–449.
44. Kramer MS. *Clinical Epidemiology and Biostatistics*. Springer: New York, 1988; 201–219.
45. Murphy EA, Abbey H. The normal range—a common misuse. *Journal of Chronic Diseases* 1967; **20**:79–88.
46. van der Zee A, Agterberg C, Peeters M, Mooi F, Schellekens J. A clinical validation of *Bordetella pertussis* and *Bordetella parapertussis* polymerase chain reaction: comparison with culture and serology using samples from patients with suspected whooping cough from a highly immunized population. *The Journal of Infectious Diseases* 1996; **174**:89–96.
47. Lind-Brandberg L, Welinder-Olsson C, Lagergård T, Taranger J, Tollfors B, Zackrisson G. Evaluation of PCR for diagnosis of *Bordetella pertussis* and *Bordetella parapertussis* infections. *Journal of Clinical Microbiology* 1998; **36**:679–683.
48. Centers for Disease Control and Prevention. Pertussis vaccination: use of acellular pertussis vaccines among infants and young children–recommendations of the Advisory Committee on Immunization Practices (ACIP). *Morbidity and Mortality Weekly Report* 1997; **46**(RR-7):1–25.
49. Everitt BS, Hand DJ. *Finite Mixture Distributions*. Chapman & Hall: London, 1981; 44–45.
50. Aitkin M, Aitkin I. A hybrid EM/Gauss–Newton algorithm for maximum likelihood in mixture distributions. *Statistics and Computing* 1996; **6**:127–130.
51. Johnson RA, Wichern DW. *Applied Multivariate Statistical Analysis*. Prentice-Hall: Englewood Cliffs, NJ, 1992; 596–602.
52. SAS Institute Inc. The FASTCLUS procedure. In *SAS/STAT User's Guide*, *Version 6* (4th edn), vol. 1. SAS Institute Inc.: Cary, NC, 1989.