STATISTICS IN MEDICINE
Statist. Med. 2009; **28**:441–461
Published online 9 December 2008 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3470

Modeling conditional dependence between diagnostic tests: A multiple latent variable model[‡]

Nandini Dendukuri[1,2,*,†], Alula Hadgu[3] and Liangliang Wang[4]

[1]*Department of Epidemiology and Biostatistics, McGill University, Montreal, Que., Canada*
[2]*Technology Assessment Unit, McGill University Health Center, Montreal, Que., Canada*
[3]*Centers for Disease Control and Prevention, Atlanta, GA, U.S.A.*
[4]*Department of Statistics, University of British Columbia, Vancouver, BC, Canada*

SUMMARY

Applications of latent class analysis in diagnostic test studies have assumed that all tests are measuring a common binary latent variable, the true disease status. In this article we describe a new approach that recognizes that tests based on different biological phenomena measure different latent variables, which in turn measure the latent true disease status. This allows for adjustment of conditional dependence between tests within disease categories. The model further allows for the inclusion of measured covariates and unmeasured random effects affecting test performance within latent classes. We describe a Bayesian approach for model estimation and describe a new posterior predictive check for evaluating candidate models. The methods are motivated and illustrated by results from a study of diagnostic tests for *Chlamydia trachomatis*. Published in 2008 by John Wiley & Sons, Ltd.

KEY WORDS: Bayesian inference; sensitivity; specificity; conditional dependence

1. INTRODUCTION

Latent class models have been widely used to estimate disease prevalence and diagnostic test accuracy in the absence of results from a perfect diagnostic test [1]. The standard two-latent class model (TLCM), widely used in diagnostic testing applications, assumes that different diagnostic tests for the same disease are measuring the same binary latent variable, the true disease status, and that test results are independent conditional on the disease status. In this article, we describe

*Correspondence to: Nandini Dendukuri, Department of Epidemiology and Biostatistics, McGill University, Montreal, Que., Canada.
†E-mail: nandini.dendukuri@mcgill.ca

Contract/grant sponsor: Natural Sciences and Engineering Research Council of Canada
Contract/grant sponsor: CDC Research Participation Program
Contract/grant sponsor: Fonds de la Recherche en Santé du Québec

Received 14 November 2007
Accepted 10 September 2008

a multiple latent variable model (MLVM) that recognizes that tests based on different biological mechanisms may in fact be measuring different latent variables, which in turn are measures of the latent disease status. This approach is particularly useful when multiple imperfect diagnostic tests are analyzed together, and subsets of tests are thought to be measuring different latent variables, creating a dependence between tests within truly diseased and non-diseased classes.

We will illustrate our approach via an application on evaluation of diagnostic tests for *Chlamydia trachomatis* infection. *C. trachomatis* is the most common bacterial sexually transmitted disease (STD) in the United States [2]. Though typically asymptomatic, it is associated with serious sequelae such as infertility. Thus, it would be desirable for a diagnostic test for *C. trachomatis* to have high sensitivity (true-positive probability) in order to ensure that no cases are missed. On the other hand, high specificity (true-negative probability) is also desirable as a false-positive test could have serious social consequences. Evaluating new tests for this disease has proven problematic due to the absence of a gold-standard test that has 100 per cent sensitivity and specificity. The widely used reference standard is culture, a test that is believed to have poor sensitivity and near perfect specificity. The lack of a gold-standard test has led to inappropriate methods for diagnostic test evaluation, such as arbitrarily assuming that culture is a gold standard or using biased methods such as discrepant analysis [3].

We will illustrate our proposed method by application to a study where the following four widely used diagnostic tests for *C. trachomatis* were applied to asymptomatic women at STD clinics [2]:

1. The ligase chain reaction (LCR) and polymerase chain reaction (PCR) tests are nucleic-acid amplification tests (NAATs). NAATs are designed to measure the presence of *C. trachomatis* DNA. They are believed to be extremely sensitive, being able to detect DNA from as few as 1–10 organisms in a sample [4, 5]. However, published studies have raised important concerns regarding the NAAT evaluation process in general and their specificity in particular [6–8]. A disadvantage of these tests is that they cannot distinguish between viable and nonviable bacteria. They are also susceptible to cross-contamination in laboratory settings [9]. Thus, they could give a positive result because the patient had an earlier infection or because of the presence of residual DNA due to stochastic or systematic contamination in the laboratory. There is indirect evidence from studies of sex partners that the specificity of NAATs is less than 100 per cent [10, 11].
2. The culture test involves identifying viable *C. trachomatis* bacteria under a microscope. It is usually considered to have an almost perfect specificity, since the presence of a chlamydial inclusion is a clear demonstration of infection. However, culture is believed to have sub-optimal sensitivity, requiring between 3 and 100 organisms to be present in the sample in order to obtain a positive diagnosis [4].
3. The DNA probe test (DNAP) is also designed to measure *C. trachomatis* DNA. However, it is less sensitive than both NAATs and culture, requiring between 300 and 10 000 organisms in the sample in order to obtain a positive diagnosis [4]. Though in theory the DNAP test can also detect nonviable bacteria it is less likely to do so due to the much higher organism load required. Also, it is designed to be highly specific and does not have the same risk of cross-contamination as an NAAT [9].

The LCR and PCR tests used in our illustration were carried out on urine specimens while the DNAP and culture tests were carried out on cervical specimens. Obtaining a urine specimen is more convenient, but tests based on urine specimens have been reported to be less sensitive than those based on cervical specimens [12, 13].

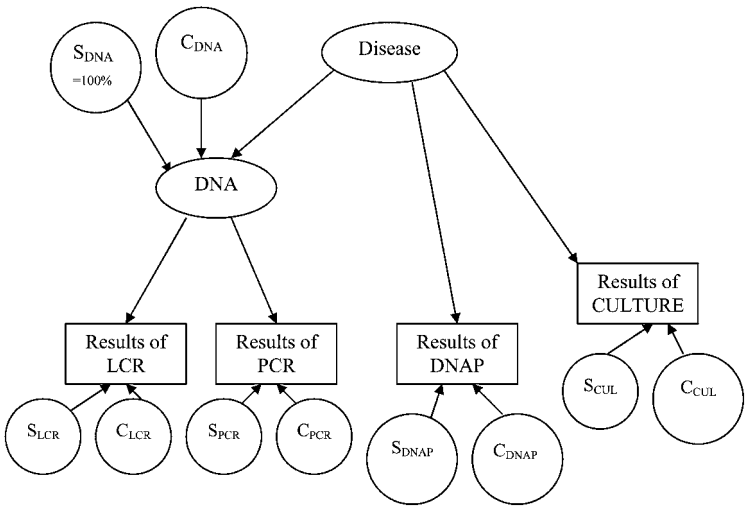

Figure 1. Diagrammatic representation of multiple latent variable model for *Chlamydia trachomatis* tests. Rectangles represent observed variables, ovals represent latent variables, and circles represent unknown parameters. S, sensitivity; C, specificity; LCR, ligase chain reaction; PCR, polymerase chain reaction; DNAP, DNA probe test; CUL, culture.

Our goal is to estimate the sensitivity and specificity of these four tests. We hypothesize that they can be classified into two types of tests measuring different latent variables, as illustrated in Figure 1. We hypothesize that the LCR and PCR tests measure the DNA latent variable, which is in turn a proxy for the true disease status. It is recommended that NAAT test results are reported in terms of whether DNA is detected or not rather than as positive or negative for *C. trachomatis* [9, 14]. We hypothesize that the DNAP and culture tests, on the other hand, measure the disease latent variable. Though the DNAP test measures DNA it does so at a much higher organism load. Thus, in practical terms, we considered DNAP closer to culture. The model in Figure 1 implies that LCR and PCR are conditionally dependent within truly diseased and truly non-diseased latent classes. It has been demonstrated that ignoring the conditional dependence between diagnostic tests could bias estimates of disease prevalence and test sensitivity and specificity [15, 16]. In order to model conditional dependence, the TLCM has been extended via the addition of fixed and random effects [17–21], addition of latent classes [22, 23], and the use of marginal models [24]. However, all extensions discussed so far assume that the common latent variable being measured by all tests is the binary disease status.

In Section 2 we describe the MLVM. In Section 3, we describe a posterior predictive check for the new model. Finally, in Section 4 we apply the new method to our motivating example on evaluating diagnostic tests for *C. trachomatis*. An R library to implement the methods described in this article is available from the corresponding author on request.

2. A MULTIPLE LATENT VARIABLE MODEL

We assume that results from P different diagnostic tests are available for each of the N subjects. Let T_{ip} denote the result of the ith subject on the pth test, and let $T_i = (T_{i1}, \ldots, T_{iP})$ denote the

vector of results for the ith subject. We assume that the P tests are of J types, i.e. they are measuring J latent variables $(l_1, \ldots, l_J)$ that are proxies for the disease of interest. Diagnostic tests, latent variables, and the true disease status are assumed to be dichotomous and to take values 1 and 0 to denote positive and negative, respectively. In the most basic version of the model, the different tests are assumed to be independent of each other conditional on the J latent variables and the latent true disease status D. Let $(L_1, \ldots, L_K)$ denote K latent classes defined by the possible combinations of the J latent variables and the true disease status (note that $K \leqslant 2^{J+1}$; see Section 4 for an illustration). The probability of observing T_i can be expressed as

$$P(T_i) = \sum_{D=0}^{1} \sum_{l_1=0}^{1} \cdots \sum_{l_J=0}^{1} P(T_{i1}, \ldots, T_{iP} | l_1, \ldots, l_J, D) \times P(l_1, \ldots, l_J, D)$$

$$= \sum_{k=1}^{K} P(T_{i1}, \ldots, T_{iP} | L_k) P(L_k) = \sum_{k=1}^{K} P(L_k) \prod_{p=1}^{P} P(T_{ip} | L_k) \tag{1}$$

assuming conditional independence within latent classes. The model in (1) is thus mathematically equivalent to a latent class model with K latent classes. The likelihood of the observed data is given by: $L \propto \prod_{i=1}^{N} P(T_i)$. When $K = 2$ the model in (1) reduces to the TLCM.

The probability of a positive result on the pth test conditional on latent class k can be expressed as $P(T_{ip} = 1 | L_k) = \Phi(a_{pk})$, where a_{pk} is the unknown parameter to be estimated and Φ is the normal cumulative probability distribution function. As explained below, this formulation allows us to expand the model to add covariates or random effects. We estimated the model in (1) using a Bayesian approach. The form of the prior distributions was as follows:

$$\begin{aligned} P(L_1), \ldots, P(L_K) &\sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K) \\ a_{pk} &\sim \text{Normal}(\mu_{0a_{pk}}, \sigma^2_{0a_{pk}}) \end{aligned} \tag{2}$$

The number of unknown parameters to be estimated in this model is $KP + K - 1$. We discuss the choice of the prior parameters in greater detail in Sections 2.2 and 4.1. Using a Gibbs sampling approach, we can sample from the full-conditional distributions of each parameter to obtain a sample from the joint posterior distribution. The full-conditional distributions are given in the Appendix.

Using simple algebra, we can derive various parameters of interest based on the probabilities in (1). For example, the sensitivity of the pth test with respect to the latent variable l_j is given by

$$P(T_{ip} = 1 | l_j = 1) = \frac{\sum_{k:l_j=1} P(T_{ip} = 1 | L_k) P(L_k)}{\sum_{k:l_j=1} P(L_k)}$$

Similarly, the sensitivity of the pth test with respect to the true disease status is given by

$$P(T_{ip} = 1 | D = 1) = \frac{\sum_{k:D=1} P(T_{ip} = 1 | L_k) P(L_k)}{\sum_{k:D=1} P(L_k)}$$

Similarly, we can define the specificity with respect to a latent variable ($P(T_{ip} = 0 | l_j = 0)$), the specificity with respect to the true disease status ($P(T_{ip} = 0 | D = 0)$), the true disease prevalence ($P(D = 1)$), the prevalence of each latent variable ($P(l_j = 1)$), and the sensitivity and specificity of each latent variable ($P(l_j = 1 | D = 1)$ and $P(l_j = 0 | D = 0)$). Further, we can obtain the positive

and negative predictive values of each test with respect to D or the l_j's as functions of these parameters.

A particular case of the model in (1) is when tests of type j depend only on the jth latent variable l_j which in turn depends on D, i.e.

$$P(T_{ip}|L_k) = P(T_{ip}|l_j) \tag{3}$$

In the *C. trachomatis* example, this would mean that the NAATs depend on the DNA status alone. In this case, the model has fewer unknown parameters: $2P + K - 1$.

2.1. Addition of random effects

The basic MLVMs in (1) or (3) can be extended by the addition of random effects or covariates to further adjust for any heterogeneity between test results within the latent classes L_k. The conditional probability of a positive test result as a function of Q random effects and M covariates may be defined as

$$P(T_{ip} = 1|L_k, r_{iq}, z_{im}) = \Phi\left(a_{pk} + \sum_{q=1}^{Q} b_{pkq} r_{iq} + \sum_{m=1}^{M} c_{pkm} z_{im}\right) \tag{4}$$

where a_{pk}, b_{pkq}, and c_{pkm} are unknown parameters, $r_{iq} \sim \mathrm{N}(0, 1)$ are random effects, and z_{im} are known covariates. This model is more general than the random effects models previously described in the literature [17, 18, 25] in that it allows for multiple random effects to model dependence between different groups of tests. The number of random effects Q can be determined by prior knowledge of the nature of the conditional dependence between the tests or by examining model fit statistics such as the posterior predictive check described in Section 3.1. To simplify the description, we consider a model with a single covariate and single random effect, i.e. one where the conditional probability is $P(T_{ip} = 1|L_k, r_{i1}, z_{i1}) = \Phi(a_{pk} + b_{pk1} r_{i1} + c_{pk1} z_{i1})$.

The following forms of prior distributions can be used for the additional parameters:

$$\begin{aligned} b_{pk1} &\sim \mathrm{Uniform}(L_{0b_{pk1}}, U_{0b_{pk1}}) \\ c_{pk1} &\sim \mathrm{Normal}(\mu_{0c_{pk1}}, \sigma^2_{0c_{pk1}}) \end{aligned} \tag{5}$$

Once again we can use a Gibbs sampler to obtain a sample from the joint posterior distribution of the parameters. The full-conditional distributions are given in the Appendix. Constraining the value of $a_{pk} + b_{pk1} r_{i1} + c_{pk1} z_{i1}$ to lie between -8 and 8 improves the convergence of the Gibbs sampler, while allowing the conditional probability to span the range from 0 to 1.

For the extended model in (4) the sensitivity and specificity of each test can be defined conditional on the covariate and averaged across the random effect (r_{i1}). For example,

$$\begin{aligned} P(T_{ip} = 1|D = d, z_{i1}) &= \sum_{k:D=d} P(L_k) \int_{-\infty}^{\infty} P(T_{ip} = 1|L_k, r_{i1}, z_{i1}) \phi(r_{i1})\, \mathrm{d}r_{i1} \\ &= \sum_{k:D=d} P(L_k) \Phi\left(\frac{a_{pk} + c_{pk1} z_{i1}}{\sqrt{1 + b^2_{pk1}}}\right) \end{aligned}$$

where $\phi(\cdot)$ is the standard normal probability density function.

2.1.1. Another formulation of the model with random effects. Another formulation of the model in (4) allows for the random effect to be different within each latent class, i.e. $r_{i1}=r_{i1k}, r_{i1k}\sim \mathrm{N}(0, b_{pk}^2)$ [25]. The form of the conditional probability would be: $P(T_{ip}=1|L_k, z_{i1}, r_{i1k})=\Phi(a_{pk}+r_{i1k}+c_{pk1}z_{i1})$. This model is perhaps intuitively more appealing, allowing for the source of the dependence to be different in each latent class. However, the likelihood is identical to the model in (4). Moreover, the Gibbs sampler is more complicated in this case because at a given iteration of the Gibbs sampler, an individual is classified into only one of the latent classes, resulting in the sampling of the random effect associated with that class only. For example, in a model with $K=3$, if the ith subject is classified as $k=1$ at a certain iteration then only the random effect r_{i11} will be sampled, but not r_{i12} and r_{i13}. This means that at the next iteration, the probability that the ith subject belongs to latent class k cannot be estimated, as it depends on all three random effects:

$$P(L_k|T_i, a_{pk}, b_{pk1}, c_{pk1}, z_{i1}, r_{i11}, r_{i12}, r_{i13})$$

$$=\frac{P(L_k)\prod_{p=1}^{P}\Phi(a_{pk}+r_{i1k}+c_{pk1}z_{i1})^{T_{ip}}\Phi(-a_{pk}-r_{i1k}-c_{pk1}z_{i1})^{1-T_{ip}}}{\sum_{k=1}^{3}P(L_k)\prod_{p=1}^{P}\Phi(a_{pk}+r_{i1k}+c_{pk1}z_{i1})^{T_{ip}}\Phi(-a_{pk}-r_{i1k}-c_{pk1}z_{i1})^{1-T_{ip}}}$$

A solution to this problem is to use a reversible jump procedure with proposal distributions for the random effects associated with the latent classes other than the one in which a subject has been classified [26].

2.2. Identifiability of the model

2.2.1. Necessary and sufficient conditions for identifiability. One commonly encountered issue in latent class analysis is model identifiability. With results from P dichotomous diagnostic tests, we have 2^P-1 degrees of freedom. A necessary condition for identifiability is that 2^P-1 is greater than or equal to the number of parameters in the model. Further, a sufficient condition for local identifiability is that the Jacobian of the transformation from the multinomial probability vector, of length 2^J, to the vector of parameters being estimated be of full rank (i.e. have a rank equal to the number of parameters to be estimated) [27]. This condition is often evaluated numerically using maximum likelihood estimates of the parameters [28]. A drawback of this approach is that one cannot distinguish if the condition is satisfied only empirically, i.e. only in the case of the data at hand. To ensure that the local identifiability was not limited to a given data set we evaluated the Jacobian several times substituting random values for the parameter estimates. When the model is identifiable, non-informative prior distributions can be used for all parameters. When the model is not identifiable, it may still be possible to obtain a solution by constraining some parameters [27] or by using informative prior distributions for some parameters [3, 18, 29, 30] as explained below. In either case, permutation invariance (label switching) remains an issue with $K!$ solutions being possible for a latent class model with K classes [31].

2.2.2. Values of non-informative prior distribution parameters. We recommend using a standard normal prior with $\mu_{0a_{pk}}=0$ and $\sigma^2_{0a_{pk}}=1$ for the a_{pk} parameters. In the absence of random effects this induces a U(0, 1) prior (a widely used non-informative prior) over the conditional probabilities $P(T_{ip}|L_k)$. Using very large values of $\sigma^2_{0a_{pk}}$ as in [25] will greatly increase the prior probability of $\Phi(a_{pk})$ being close to 0 or 1. Apart from being an unrealistic prior, this causes problems for

the convergence of the Gibbs sampler when the model includes random effects. Similar values can be used for the normal prior distributions over the c_{pkm} parameters. For the b_{pkq} parameters we recommend using U(0, 5) prior distributions. By using a lower limit of 0 for the uniform distribution, we essentially force the correlation between test results within a latent class to be positive. To evaluate the sensitivity of the results to the prior distributions, we allowed the prior standard deviations for the a_{pk} parameters to vary between 1 and 5 and the upper limit of the uniform prior distribution over the b_{pkq} parameters to vary between 5 and 10.

2.2.3. Values of informative prior distribution parameters. It may be possible to find prior information on some parameters in consultation with the literature or experts. For example, the range of values of the sensitivity and specificity of a widely used diagnostic test may be well known. This information can be used to determine the parameters of its prior distribution [18]. When the model is not identifiable, informative prior information is necessary to obtain a meaningful solution. The minimum number of parameters for which informative priors need to be used is the difference between the number of parameters and the number of degrees of freedom. For such problems the prior distribution can continue to impact the posterior distribution even when the sample size is infinite [32]. Thus, it is important that the prior distributions accurately reflect prior knowledge.

3. MODEL CHECKING AND MODEL COMPARISON

3.1. Posterior predictive checks

Following the approach described in Gelman *et al.* [33], we define a statistic based on a sample from the posterior predictive distribution to evaluate model fit. Several statistics for testing for conditional independence and the number of latent classes have been described [25, 34–36]. The method described below is more informative about the nature of the dependence between each pair of tests within each latent class.

Once the convergence of the Gibbs sampler is ascertained, a value is drawn from the posterior predictive distribution, T^*_{ip}, corresponding to each observed test result. The expected percentage agreement between a pair of tests within the kth latent class is

$$\alpha^*_{pp'|k} = \frac{\sum_{i=1}^{N}(T^*_{ip}T^*_{ip'} + (1-T^*_{ip})(1-T^*_{ip'}))I(d_{ik}=k)}{\sum_{i=1}^{N} I(d_{ik}=k)}$$

where d_{ik} is a categorical variable indicating the latent class k into which the ith subject was classified and $I(\cdot)$ is the indicator function. Similarly, we can define $\alpha_{pp'|k}$, the observed agreement between tests p and p'. Finally, we calculate $P(\alpha^*_{pp'|k} > \alpha_{pp'|k})$. When these probabilities are very close to 0 or 1, say less than 0.05 or greater than 0.95, they are taken to indicate that the model may be inappropriate.

3.2. Bayesian Information Criterion

Competing models were compared using the Bayesian Information Criterion (BIC) [37], which is an approximation of the marginal likelihood of each model. For a given model, M, the BIC is a

function of the posterior mode ($\hat{\theta}$) and the number of parameters in the model (s) as follows:

$$\text{BIC} = -2\log L(\hat{\theta}|\text{data}, M) + s\log(N)$$

where N is the sample size. The approximation is appropriate when non-informative prior distributions are used and the sample size is large. The preferred model is the one with the lowest value of the criterion. The posterior mode was estimated using the EM algorithm as described in [17].

4. ANALYSIS OF *C. TRACHOMATIS* DATA

4.1. Likelihood and prior distributions

For the *C. trachomatis* example introduced earlier, the LCR, PCR, DNAP, and culture tests were denoted as $p = 1$, 2, 3, and 4, respectively. The NAAT tests (i.e. LCR and PCR) were assumed to depend on the latent variable l_1 (the true DNA status), while the DNAP and culture tests were assumed to depend directly on the disease status, D. We had $P = 4$ diagnostic tests and therefore $2^4 - 1 = 15$ degrees of freedom (see Figure 1). We assumed that the latent class $l_1 = 0, D = 1$ (i.e. a class where patients are truly diseased but DNA is not present) is a structural zero, resulting in $K = 3$ latent classes: $L_1 = \{l_1 = 1, D = 1\}$, $L_2 = \{l_1 = 1, D = 0\}$, and $L_3 = \{l_1 = 0, D = 0\}$.

We first fit the constrained version of the MLVM described in (3) (denoted by MLVM-Basic). Based on the results of the posterior predictive check of the MLVM-Basic model, we considered an extension of this model (which we denote as MLVM-RE) that included random effects to allow for a dependence between the LCR and PCR tests within the class $L_1 = \{l_1 = 1, D = 1\}$. The MLVM-RE model was defined using the constraints $b_{111} = b_{211} \neq 0$, $b_{1k1} = b_{2k1} = 0, k = 2, 3$, and $b_{pk1} = 0$, $p = 3, 4, k = 1, 2, 3$, resulting in a total of 11 unknown parameters. Both models satisfied necessary and sufficient criteria for identifiability. The following prior distributions were used:

$$P(L_1), P(L_2), P(L_3) \propto \text{Dirichlet}(1, 1, 1)$$

$$a_{pk} \sim \text{Normal}(0, 1)$$

$$b_{pk1} \sim \text{Uniform}(0, 5)$$

By setting the lower bound of the prior distribution of b_{pk1} to 0, we forced a positive dependence between LCR and PCR within latent class L_1.

For comparison, we also fit the TLCM with latent classes labeled as $L_1 = \{D = 1\}$ and $L_2 = \{D = 0\}$ (denoted by TLCM-Basic). From the results of the posterior predictive check for this model, we found that there was a correlation between LCR and PCR and between DNAP and culture. We therefore extended the model by adding two random effects, one for each correlated pair (the model is denoted by TLCM-RE). The TLCM-RE model was defined using the following constraints: $b_{111} = b_{211} \neq 0$ and $b_{312} = b_{412} \neq 0$. Non-informative prior distributions such as those used for the MLVM models were used here as well.

To evaluate the performance of our methods in the presence of non-identifiability, we considered the MLVM model when no constraints are placed on any of the conditional probabilities (1). This model has 14 parameters but the rank of the Jacobian is only 13 indicating that an informative prior distribution is needed for at least one parameter. The specificity of the cell culture test is believed to be between 98 and 100 per cent [3]. This prior information was incorporated into the

model as a truncated normal prior $N(\mu_{a_{0pk}}=0, \sigma^2_{a_{0pk}}=1)I(-5,-2.05)$ prior for the a_{42} and a_{43} parameters.

The Gibbs sampler achieved convergence typically within 500 iterations. However, to be conservative we dropped the first 2000 iterations and retained the next 3000 to obtain summary statistics. Convergence was evaluated using trace plots and the Gelman–Rubin statistic based on five randomly selected sets of starting values [33]. Once it appeared that a model had converged, we selected the sign of the starting values for the a_{pk} parameters such that the sensitivity and specificity were greater than 0.5 to avoid the label-switching problem. We calculated median, 2.5, and 97.5 per cent quantiles to summarize the distribution of each parameter of interest. We also fit all identifiable models using the EM algorithm and obtained results similar to those presented in Section 4.3.

4.2. Parameters of interest

Samples from the posterior densities of the probabilities below were used to draw inferences about them. We drop the notation of 'conditional on the data' for simplicity. For each of the MLVMs, the sensitivity of the tests with respect to D was obtained directly from the model, $P(T_{ip}=1|L_1)$. The specificity with respect to D was obtained by simple algebra as follows:

$$P(T_{ip}=0|D=0)=\frac{\sum_{k=2}^{3} P(T_{ip}=0|L_k)P(L_k)}{\sum_{k=2}^{3} P(L_k)} \tag{6}$$

Similarly, the sensitivity of each test with respect to the DNA status can be obtained as

$$P(T_{ip}=1|l_1=1)=\frac{\sum_{k=1}^{2} P(T_{ip}=1|L_k)P(L_k)}{\sum_{k=1}^{2} P(L_k)} \tag{7}$$

while the specificity with respect to DNA was obtained directly from the model as $P(T_{ip}=0|L_3)$. The sensitivity and specificity with respect to disease under the TLCM models were obtained in a similar manner.

We also estimated the positive predictive values of each combination of test results (i.e. each profile) with respect to the disease and DNA status:

$$P(D=1|T_{i1}=t_1, T_{i2}=t_2, T_{i3}=t_3, T_{i4}=t_4)$$

$$=\frac{P(L_1)\prod_{p=1}^{P} P(T_{ip}|L_1)}{\sum_{k=1}^{3} P(L_k)\prod_{p=1}^{P} P(T_{ip}|L_k)}$$

$$P(DNA=1|T_{i1}=t_1, T_{i2}=t_2, T_{i3}=t_3, T_{i4}=t_4)$$

$$=\frac{\sum_{k=1}^{2} P(L_k)\prod_{p=1}^{P} P(T_{ip}|L_k)}{\sum_{k=1}^{3} P(L_k)\prod_{p=1}^{P} P(T_{ip}|L_k)}$$

The prevalence of the disease was estimated as $P(L_1)$ and the prevalence of DNA was estimated as $\sum_{k=1}^{2} P(L_k)$.

4.3. Results

The observed and predicted frequencies for each test profile under the different models are summarized in Table I. The lowest value of the BIC was obtained for the MLVM-RE model. Since the BIC for this model is at least 10 points lower than the other models considered, there is 'strong' evidence in favor of it based on the criteria discussed by Raftery [37]. There was only a slight difference in the BIC between the TLCM-RE and MLVM-Basic models. Thus, relying on the BIC alone, we would not be able to distinguish between the TLCM-RE and MLVM-Basic models, even though the prevalence, sensitivities, specificities, and predictive values based on these models would lead to different conclusions. Further, relying on the BIC would not allow us determine that the TLCM-RE model is inappropriate according to our predetermined biological model in Figure 1.

Tables II and III list the expected and observed agreement between pairs of tests under the different models, and the probability that the expected agreement exceeded the observed agreement between pairs of tests within latent classes. From these tables we find that MLVM-RE is the only model for which none of the probabilities is less than 0.05 or greater than 0.95. We also find that the TLCM-Basic model does not adjust for the conditional dependence between the NAATs and between the non-NAATs. The MLVM-Basic model improves upon the TLCM-Basic by explaining the conditional dependence between the non-NAATs. However, there is still some

Table I. Posterior median predicted frequency of each combination of test results and model fit statistics for the different models.

Profile					TLCM		MLVM	
LCR	PCR	DNAP	Culture	Observed frequency (per cent)	Basic	RE	Basic	RE
					Median predicted frequency			
1	1	1	1	210 (5.9)	174	218	197	204
1	1	1	0	12 (0.3)	40	10	9	11
1	1	0	1	44 (1.2)	78	50	46	52
1	1	0	0	59 (1.7)	18	58	60	59
1	0	1	1	32 (0.9)	34	26	42	29
1	0	1	0	0 (0)	8	1	2	2
1	0	0	1	10 (0.3)	15	6	10	8
1	0	0	0	39 (1.1)	39	37	40	40
0	1	1	1	14 (0.4)	22	11	27	14
0	1	1	0	0 (0)	5	1	1	1
0	1	0	1	8 (0.2)	10	3	7	4
0	1	0	0	27 (0.8)	28	27	28	29
0	0	1	1	18 (0.5)	4	25	6	24
0	0	1	0	11 (0.3)	13	12	12	12
0	0	0	1	24 (0.7)	26	24	25	25
0	0	0	0	3043 (85.7)	3037	3039	3037	3035
Log-likelihood					−2571	−2506	−2506	−2493
Number of parameters					9	11	10	11
BIC					5216	5102	5094	5076

LCR, ligase chain reaction; PCR, polymerase chain reaction; DNAP, DNA probe test; TLCM: two-latent class model; MLVM, multiple latent variable model; RE, random effects; BIC, Bayesian Information Criterion.

Table II. Posterior medians of expected and observed probability of agreement within latent classes based on the TLCM models.

	$D=1$			$D=0$		
	$\alpha^*_{pp'\|k}$	$\alpha_{pp'\|k}$	$P(\alpha^*_{pp'\|k} > \alpha_{pp'\|k})$	$\alpha^*_{pp'\|k}$	$\alpha_{pp'\|k}$	$P(\alpha^*_{pp'\|k} > \alpha_{pp'\|k})$
TLCM-Basic						
LCR, PCR	0.762	0.833	**0.002**	0.979	0.981	0.341
LCR, DNAP	0.652	0.643	0.590	0.985	0.985	0.428
LCR, culture	0.744	0.723	0.746	0.981	0.981	0.430
PCR, DNAP	0.630	0.607	0.756	0.988	0.988	0.440
PCR, culture	0.711	0.678	0.865	0.984	0.984	0.426
DNAP, culture	0.621	0.816	**0.000**	0.989	0.989	0.414
TLCM-RE						
LCR, PCR	0.865	0.829	**0.970**	0.981	0.982	0.401
LCR, DNAP	0.637	0.646	0.329	0.986	0.987	0.404
LCR, culture	0.723	0.712	0.685	0.984	0.984	0.409
PCR, DNAP	0.619	0.618	0.500	0.988	0.989	0.405
PCR, culture	0.693	0.675	0.768	0.986	0.987	0.375
DNAP, culture	0.819	0.812	0.630	0.991	0.991	0.401

$\alpha^*_{pp'|k}$, expected agreement; $\alpha_{pp'|k}$, observed agreement; LCR, ligase chain reaction; PCR, polymerase chain reaction; DNAP, DNA probe test; TLCM, two-latent class model; RE, random effects.

correlation remaining between the NAATs in the latent category $L_1 = \{l_1 = 1, D = 1\}$. This seems to be adjusted for sufficiently by MLVM-RE.

The prevalences of disease and DNA status estimated from the different models are given in Table IV. The prevalence of disease based on the TLCM models was estimated at around 12 per cent. The prevalence of the three latent categories according to the MLVM models was estimated as $P(L_1, L_2, L_3) = (10, 2, 88$ per cent). Thus, the MLVM models appear to result in the division of the 12 per cent of individuals classified as disease positive under the TLCM into the two groups, which we labeled as (DNA $= 1$, $D = 1$) and (DNA $= 1$, $D = 0$). Table V provides the median positive predictive value for each profile. From Table V we see that the group of individuals that were most likely to be reclassified are those with the profile (LCR $= 1$, PCR $= 1$, DNAP $= 0$, culture $= 0$).

From Figure 2 we find that under the basic TLCM model, LCR and PCR have a higher probability of testing positive among individuals classified as $D = 1$. From Figure 3, we see that under the MLVM-Basic model, culture has the highest probability of testing positive among individuals classified as $D = 1$, DNA $= 1$. Further, among individuals classified as (DNA $= 1$, $D = 0$), the LCR and PCR tests have a high probability of a positive result, whereas the DNAP and culture tests have a nearly zero probability of a positive result. Thus, the MLVM is consistent with the nature and intended use of the tests under consideration, i.e. it recognizes that NAATs are attempting to detect the presence of live or dead DNA. On the other hand, the TLCM does not distinguish the difference between the presence of DNA and the presence of active disease.

Tables VI and VII list the sensitivities and specificities with respect to disease and DNA, respectively. Under the MLVM models, the specificity of LCR and PCR for detecting disease is lower than the specificity of culture and DNAP for detecting disease (Table VI). The specificity of LCR and PCR for detecting DNA is higher than that for detecting disease. The sensitivity of DNAP and culture tests is lower for detecting DNA than that for detecting disease. The addition of

Table III. Posterior medians of expected and observed probability of agreement within latent classes based on the MLVM models.

	$D=1, \text{DNA}=1$			$D=0, \text{DNA}=1$			$D=0, \text{DNA}=0$		
	$\alpha^*_{pp'\|k}$	$\alpha_{pp'\|k}$	$P(\alpha^*_{pp'\|k} > \alpha_{pp'\|k})$	$\alpha^*_{pp'\|k}$	$\alpha_{pp'\|k}$	$P(\alpha^*_{pp'\|k} > \alpha_{pp'\|k})$	$\alpha^*_{pp'\|k}$	$\alpha_{pp'\|k}$	$P(\alpha^*_{pp'\|k} > \alpha_{pp'\|k})$
MLVM-Basic									
LCR, PCR	0.745	0.817	**0.006**	0.744	0.738	0.500	0.986	0.986	0.465
LCR, DNAP	0.737	0.750	0.328	0.121	0.131	0.397	0.988	0.988	0.419
LCR, culture	0.846	0.843	0.526	0.123	0.133	0.404	0.984	0.984	0.413
PCR, DNAP	0.701	0.705	0.442	0.179	0.177	0.469	0.990	0.990	0.454
PCR, culture	0.793	0.786	0.573	0.183	0.181	0.472	0.987	0.987	0.445
DNAP, culture	0.784	0.788	0.433	0.988	0.988	0.324	0.988	0.989	0.391
MLVM-RE									
LCR, PCR	0.841	0.822	0.801	0.893	0.893	0.464	0.979	0.980	0.344
LCR, DNAP	0.713	0.750	0.116	0.045	0.045	0.381	0.985	0.985	0.398
LCR, culture	0.823	0.829	0.381	0.048	0.047	0.400	0.982	0.983	0.357
PCR, DNAP	0.681	0.706	0.197	0.063	0.063	0.435	0.988	0.989	0.315
PCR, culture	0.774	0.774	0.469	0.066	0.065	0.448	0.985	0.986	0.283
DNAP, culture	0.770	0.774	0.428	1.000	1.000	0.303	0.990	0.991	0.328

$\alpha^*_{pp'|k}$, expected agreement; $\alpha_{pp'|k}$, observed agreement; LCR, ligase chain reaction; PCR, polymerase chain reaction; DNAP, DNA probe test; MLVM, multiple latent variable model; RE, random effects.

Published in 2008 by John Wiley & Sons, Ltd.
Statist. Med. 2009; **28**:441–461
DOI: 10.1002/sim

Table IV. Posterior median prevalence (and 95 per cent credible interval) of disease (D) and DNA status.

	TLCM	MLVM		
Model	$D=1$	$D=1,\ \text{DNA}=1$	$D=0,\ \text{DNA}=1$	$D=0,\ \text{DNA}=0$
Basic	0.117 (0.113, 0.121)	—	—	—
RE	0.122 (0.119, 0.127)	—	—	—
Basic	—	0.099 (0.096, 0.103)	0.023 (0.021, 0.025)	0.878 (0.874, 0.882)
RE	—	0.101 (0.092, 0.112)	0.017 (0.013, 0.023)	0.881 (0.870, 0.892)

TLCM, two-latent class model; MLVM, multiple latent variable model; RE, random effects.

random effects to adjust for the conditional dependence between (LCR, PCR) pair did not change the sensitivity and specificity estimates by much.

For the non-identifiable model, where we assumed a uniform informative prior for culture specificity (98–100 per cent), we obtained results similar to the MLVM-Basic model.

5. SUMMARY

In this article we proposed a latent class model for diagnostic test studies that accounts for the nature of the diagnostic test by allowing tests based on the same biological mechanism to be dependent conditional on the latent disease status. We further described how conditional dependence within latent classes can be modeled via the addition of one or more random effects and/or covariates. We described a Bayesian approach for model estimation, and a posterior predictive check to evaluate model fit. To our knowledge, estimation of a latent class model with random effects has not been successfully implemented so far when using non-informative prior distributions. We have developed a library in the R statistical package to implement the models and methods described in this article. The model we describe effectively results in increasing the number of latent classes to more than two. However, we emphasize that our approach is different from earlier work suggesting addition of latent classes to adjust for conditional dependence. The article by Albert *et al.* [23] suggested the addition of unequivocally positive and negative classes, where individuals were all correctly classified by all tests. The article by Espeland and Handelman [22] suggested use of three latent classes: unequivocally positive and negative and undiagnosable. Thus, both these approaches do not allow for latent variables besides the binary disease status.

Accurately specifying the nature of the dependence between the diagnostic tests is important as it has been shown that misspecification could alter estimates of sensitivity and specificity [15, 16, 38]. Further, it is possible that models specifying different dependence structures fit the data equally well but give very different estimates of sensitivity and specificity. Common model selection statistics, such as the log-likelihood and χ^2 test statistic, may not be able to distinguish between such models [38]. Therefore, it is important to caution that while the MLVM-RE model was the best model among those considered, it is possible that there may be other models that fit equally well and lead to different conclusions. We believe that latent class model selection should take into consideration the biology of the problem and not rely solely on mathematical/statistical concerns. Ideally, we should proceed by first defining a suitable substantive model describing the latent variables, the resulting latent classes, and the dependence structure between the tests. The

Table V. Posterior median positive predictive values for each combination of test results based on the different models.

Profile					TLCM's $P(D=1\|\text{Profile})$		MLVM's $P(D=1\|\text{Profile})$		MLVM's $P(\text{DNA}=1\|\text{Profile})$	
LCR	PCR	DNAP	Culture	Frequency (per cent)	Basic	RE	Basic	RE	Basic	RE
1	1	1	1	210 (5.9)	1	1	1	1	1	1
1	1	1	0	12 (0.3)	1	1	0.98	0.98	1	1
1	1	0	1	44 (1.2)	1	1	0.99	0.99	1	1
1	1	0	0	59 (1.7)	0.98	1	0.03	0.04	0.99	0.99
1	0	1	1	32 (0.9)	1	1	1	1	1	1
1	0	1	0	0 (0)	0.98	0.92	0.93	0.91	0.95	0.92
1	0	0	1	10 (0.3)	0.98	0.97	0.97	0.97	0.98	0.97
1	0	0	0	39 (1.1)	0.09	0.18	0.01	0.01	0.32	0.10
0	1	1	1	14 (0.4)	1	1	1	1	1	1
0	1	1	0	0 (0)	0.98	0.87	0.92	0.88	0.94	0.89
0	1	0	1	8 (0.2)	0.98	0.95	0.97	0.95	0.98	0.95
0	1	0	0	27 (0.8)	0.08	0.11	0.01	0.01	0.29	0.05
0	0	1	1	18 (0.5)	0.98	1	0.98	1	0.98	1
0	0	1	0	11 (0.3)	0.07	0.10	0.02	0.10	0.02	0.10
0	0	0	1	24 (0.7)	0.07	0.26	0.05	0.24	0.05	0.24
0	0	0	0	3043 (85.7)	0	0	0	0	0	0

LCR, ligase chain reaction; PCR, polymerase chain reaction; DNAP, DNA probe test; TLCM, two-latent class model; MLVM, multiple latent variable model; RE, random effects.

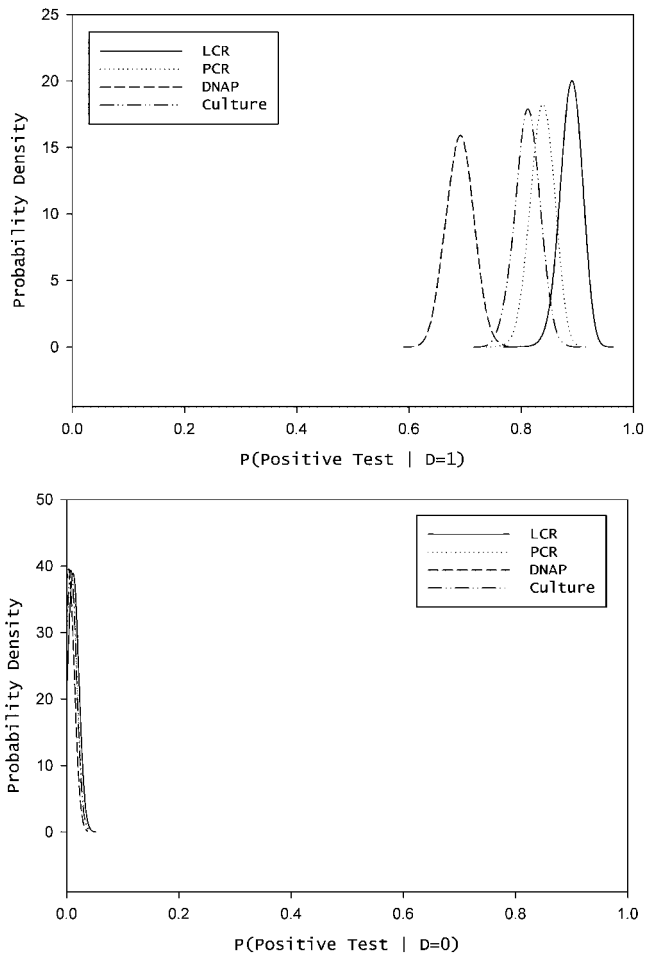

Figure 2. Posterior distributions of a positive test result in each latent class based on the TLCM model.

observed data can then be fit to the predetermined model. Finally, posterior predictive checks, of the type we have described, can be used to determine whether the model agrees with the observed data.

Diagnostic test evaluation in the absence of a gold standard continues to present a substantial problem for diagnostic test evaluation. While latent class models appear to provide a possible solution, they have been criticized for the lack of transparency in how the latent 'true' disease status is determined [39, 40]. We believe that this criticism also applies to a clinical diagnosis, as there is typically no transparent algorithm by which a clinician arrives at a diagnosis using multiple sources of information such as the patient's disease history, physical symptoms, and diagnostic test results. To address this criticism, in part, we have presented the positive and negative predictive values for each latent variable in each test profile.

In our example we found that the selected model classified individuals who were positive on both NAAT tests but negative on the non-NAAT tests as having a lower probability of being

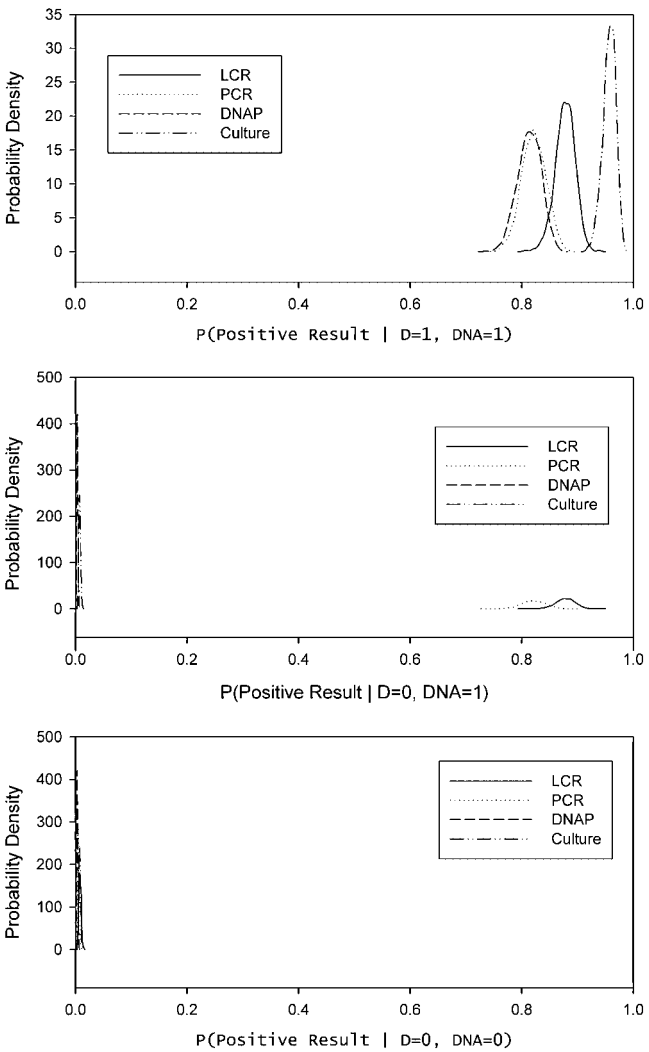

Figure 3. Posterior distributions of the probability of a positive result in each latent class based on the MLVM model.

disease positive than being DNA positive. Based on the current CDC guidelines, a positive result on LCR or PCR would be considered presumptive evidence of treatment that must be confirmed. It is therefore not impossible that following the current guidelines the women in this group would be classified as false positives. The TLCM models classify women in this group as true positives. This results in an estimate of nearly 100 per cent for the specificity of the LCR and PCR tests, which is not in keeping with our knowledge of the biology of the problem. We would like to point out to readers that as in the case of any statistical modeling exercise, there is no way to prove that the MLVM is in fact the correct model. We also acknowledge that in the case of latent class models, in particular, the labeling of the latent classes is subjective. We merely offer that our

Table VI. Posterior median sensitivity and specificity (and 95 per cent credible interval) of the four tests for detecting disease status.

	TLCM-Basic	TLCM-RE
Specificity		
LCR	0.988 (0.987, 0.990)	0.990 (0.989, 0.991)
PCR	0.992 (0.990, 0.993)	0.992 (0.991, 0.993)
DNAP	0.996 (0.995, 0.997)	0.997 (0.996, 0.997)
Culture	0.992 (0.991, 0.993)	0.994 (0.993, 0.995)
Sensitivity		
LCR	0.890 (0.879, 0.901)	0.87 (0.85, 0.88)
PCR	0.838 (0.825, 0.850)	0.81 (0.80, 0.83)
DNAP	0.692 (0.676, 0.707)	0.68 (0.66, 0.69)
Culture	0.813 (0.799, 0.826)	0.80 (0.78, 0.81)
	MLVM-Basic	MLVM-RE
Specificity		
LCR	0.969 (0.967, 0.971)	0.969 (0.963, 0.975)
PCR	0.973 (0.971, 0.975)	0.973 (0.967, 0.978)
DNAP	0.996 (0.995, 0.997)	0.997 (0.994, 0.998)
Culture	0.992 (0.991, 0.994)	0.994 (0.990, 0.997)
Sensitivity		
LCR	0.88 (0.87, 0.89)	0.859 (0.817, 0.897)
PCR	0.82 (0.81, 0.84)	0.803 (0.757, 0.844)
DNAP	0.81 (0.80, 0.83)	0.798 (0.753, 0.843)
Culture	0.96 (0.95, 0.96)	0.952 (0.921, 0.973)

LCR, ligase chain reaction; PCR, polymerase chain reaction; DNAP, DNA probe test; TLCM, two-latent class model; MLVM, multiple latent variable model; RE, random effects.

Table VII. Posterior median sensitivity and specificity (and 95 per cent credible interval) of the four tests for detecting DNA status.

	MLVM-Basic	MLVM-RE
Specificity		
LCR	0.991 (0.990, 0.993)	0.988 (0.984, 0.992)
PCR	0.994 (0.993, 0.996)	0.991 (0.987, 0.994)
DNAP	0.996 (0.996, 0.997)	0.997 (0.994, 0.998)
Culture	0.992 (0.991, 0.993)	0.994 (0.990, 0.997)
Sensitivity		
LCR	0.879 (0.866, 0.892)	0.876 (0.837, 0.910)
PCR	0.825 (0.811, 0.839)	0.824 (0.779, 0.862)
DNAP	0.661 (0.643, 0.678)	0.680 (0.634, 0.725)
Culture	0.778 (0.762, 0.793)	0.811 (0.768, 0.850)

LCR, ligase chain reaction; PCR, polymerase chain reaction; DNAP, DNA probe test; MLVM, multiple latent variable model; RE, random effects.

approach is closer to reality than the widely followed approach of assuming that all diagnostic tests for a certain disease are measuring the same underlying latent variable. It is reasonable to believe that results from additional diagnostic tests would provide a compromise between the TLCM and MLVM models, dividing this group into true positives and false positives.

One of the main findings of this article is that NAATs are less specific than non-NAATs for detection of disease (Table VI) based on our selected model. The NAATs' specificity for detection of disease is decreased because these tests can pick up infections that are no longer active or they can give positive results due to systematic or stochastic contamination in the laboratory. In typical U.S. chlamydia and gonorrhea screening sites, the prevalence of disease is less than 4 per cent. If indeed the specificity of LCR for detection of disease is as low as 96.9 per cent (Table VI) and its sensitivity is 88.2 per cent, then as many as 46 per cent of women with an LCR-positive result would be false positives.

APPENDIX A: FULL-CONDITIONAL DISTRIBUTIONS FOR MLVM

Below we describe the full-conditional distributions for the MLVM-RE model with a single random effect for modeling the conditional dependence in latent class $k=1$ between tests $p=1$ and 2. In the absence of random effects, Steps 2 and 3 below can be dropped and the terms r_{i1} can be set to 0. In the presence of constraints the parameters of the full-conditional distributions need to be modified. Corresponding to the ith individual, we introduce the latent vector $\mathbf{d}_i=(d_{i1},\ldots,d_{iK})$, where $d_{ik}=1$ if the ith individual is classified in latent class k and $d_{ik}=0$ otherwise. Using the approach described by Albert and Chib [41] for Bayesian analysis of probit models, we introduce latent variables $\boldsymbol{\eta}_i=(\eta_{i1},\ldots,\eta_{iP})$ for the ith subject corresponding to the value of each diagnostic test result on a continuous scale.

1. Latent class membership: The probability that the ith subject belongs to latent class L_k is given by

$$\zeta_{ik}=P(L_k|T_i,a_{pk},b_{pk1},r_{i1})=\frac{P(L_k)\prod_{p=1}^{P}P(T_{ip}|L_k)}{\sum_{j=1}^{K}P(L_j)\prod_{p=1}^{P}P(T_{ip}|L_k)}$$

   The full-conditional distribution of the vector $\mathbf{d}_i=(d_{i1},\ldots,d_{iK})$ is given by

$$\mathbf{d}_i\sim\text{Multinomial}(1,(\zeta_{i1},\ldots,\zeta_{iK}))$$

   where Multinomial$(n,(\zeta_{i1},\ldots,\zeta_{iK}))$ is the notation for a multinomial probability distribution with sample size n and probability vector $(\zeta_{i1},\ldots,\zeta_{iK})$.
2. Random effects: For individuals in latent class $k=1$ the full-conditional distribution is

$$r_{i1}|T_i,\mathbf{d}_i,\boldsymbol{\eta}_i,a_{pk},b_{pk1}\sim\text{N}(\mu_r,\sigma_r^2)$$

   where

$$\sigma_r^2=\frac{1}{\sum_{k=1}^{K}d_{ik}\sum_{p=1}^{P}b_{pk1}^2+1}\quad\text{and}\quad\mu_r=\sigma_r^2\sum_{k=1}^{K}d_{ik}\sum_{p=1}^{P}(\eta_{ip}-a_{pk})b_{pk1}$$

In the MLVM-RE model discussed in this article we had placed the constraint $b_{111}=b_{211}$ implying that $\sigma_r^2=(2b_{111}^2+1)^{-1}$. For individuals in latent classes $k=2$ or 3 the full-conditional distribution is $r_{i1}\sim \mathrm{N}(0,1)$.

3. Co-efficients of the random effects: The parameters b_{p11} follow a normal distribution truncated by the lower and upper bounds of the uniform prior on b_{p11}:

$$b_{pk1}|a_{pk}, r_{i1}, \mathbf{d}_i, L_{0b_{pk1}}, U_{0b_{pk1}} \sim \mathrm{N}(\mu_{b_{pk1}}, \sigma^2_{b_{pk1}}), I(L_{0b_{pk1}}, U_{0b_{pk1}}), \quad p=1,2$$

where

$$\sigma^2_{b_{pk1}} = \frac{1}{\sum_{i=1}^{N}\sum_{l=1}^{K}\sum_{j=1}^{P} r_{i1}^2 d_{il} I(b_{pk1}=b_{jl1})} \quad \text{and}$$

$$\mu_{b_{pk1}} = \sigma^2_{b_{p11}} \sum_{i=1}^{N}\sum_{l=1}^{K} d_{il} r_{i1}(\eta_{ij}-a_{jl}) I(b_{pk1}=b_{jl1})$$

Under the constraint $b_{111}=b_{211}$, these parameters are modified as follows:

$$\sigma^2_{b_{111}} = \frac{1}{(2\sum_i r_{i1}^2 d_{i1})} \quad \text{and} \quad \mu_{b_{111}} = \sigma^{-2}_{b_{111}} \sum_i r_{i1} d_{i1}\{(\eta_{i1}-a_{11})+(\eta_{i2}-a_{21})\}$$

4. Probability of each latent class:

$$P(L_1),\ldots,P(L_K)|\alpha_1,\ldots,\alpha_K,\mathbf{d}_i \sim \text{Dirichlet}\left(\alpha_1+\sum_{i=1}^{N} d_{i1},\ldots,\alpha_K+\sum_{i=1}^{N} d_{iK}\right)$$

5. Latent diagnostic test result on a continuous scale: These variables follow a truncated normal distribution: $\eta_{ip}\sim \mathrm{N}(\sum_{k=1}^{K}(a_{pk}+b_{pk1}r_{i1})d_{ik}, 1)$. The distribution is truncated between $[0,8]$ and $[-8,0]$ depending on whether $T_{ip}=0$ or 1.
6. The a_{pk} parameters:

$$a_{pk}|\mu_{a_{0pk}}, \sigma^2_{a_{0pk}}, \mathbf{d}_i, b_{pk1}, r_{i1}, \boldsymbol{\eta}_i \sim \mathrm{N}(\mu_{a_{pk}}, \sigma^2_{a_{pk}})$$

where

$$\sigma^2_{a_{pk}} = \left(\sum_{i=1}^{N}\sum_{l=1}^{K} d_{il} \sum_{j=1}^{P} I(a_{pk}\equiv a_{jl}) + \frac{1}{\sigma^2_{a_{0pk}}}\right)^{-1}$$

$$\mu_{a_{pk}} = \sigma^2_{a_{pk}}\left(\frac{\mu_{a_{0pk}}}{\sigma^2_{a_{0pk}}} + \sum_{i=1}^{N}\sum_{l=1}^{K} d_{il} \sum_{j=1}^{P} I(a_{pk}\equiv a_{jl})(\eta_{ip}-b_{pk1}r_{i1})\right)$$

ACKNOWLEDGEMENTS

This project received support from the Natural Sciences and Engineering Research Council of Canada and the CDC Research Participation Program. The first author holds a salary award from the Fonds de la Recherche en Santé du Québec.

Disclaimer: The views expressed in this article are those of the authors and do not necessarily reflect the views or policies of the CDC or the U.S. Public Health Service.

REFERENCES

1. Walter SD, Irwig LM. Estimation of error rates, disease prevalence, and relative risk misclassified data: a review. *Journal of Clinical Epidemiology* 1988; **41**:923–937.
2. Black CM, Marrazzo J, Johnson RE, Hook III EW, Jones RB, Green TA, Schachter J, Stamm WE, Bolan G, St Louis ME, Martin DH. Head-to-head multicenter comparison of dna probe and nucleic acid amplification tests for *Chlamydia trachomatis* infection in women performed with an improved reference standard. *Journal of Clinical Microbiology* 2002; **40**:3757–3763.
3. Hadgu A, Dendukuri N, Hilden J. Evaluation of nucleic acid amplification tests in the absence of a perfect gold-standard test: a review of the statistical and epidemiologic issues. *Epidemiology* 2005; **16**:604–612.
4. Black CM. Current methods of laboratory diagnosis of *Chlamydia trachomatis* infections. *Clinical Microbiology Reviews* 1997; **10**:160–184.
5. Schachter J. In defense of discrepant analysis. *Journal of Clinical Epidemiology* 2001; **54**:211–212.
6. Hadgu A. The discrepancy in discrepant analysis. *Lancet* 1996; **348**:592–593.
7. Hilden J. Discrepant analysis—or behavior? *Lancet* 1997; **350**:902.
8. Miller WC. Bias in discrepant analysis: when two wrongs don't make it right. *Journal of Clinical Epidemiology* 1998; **51**:219–231.
9. Persing DH, Tenover FC, Versalovic J, Tang Y, Unger ER, Relman DA, White TJ (eds). *Molecular Microbiology*. ASM Press: Washington, DC, 2004.
10. Golden MR, Whittington WL, Handsfield HH, Hughes JP, Stamm WE, Hogben M, Clark A, Malinski C, Helmers JR, Thomas KK, Holmes KK. Effect of expedited treatment of sex partners on recurrent or persistent gonorrhea or chlamydial infections. *New England Journal of Medicine* 2005; **352**:676–685.
11. Rogers SM, Miller WC, Turner CF, Ellen J, Zenilman J, Rothman R, Villarroel MA, Al-Tayyib A, Leone P, Gaydos C, Ganapathi L, Hobbs M, Kanouse D. Concordance of *Chlamydia trachomatis* infections within sexual partnerships. *Sexually Transmitted Infections* 2008; **84**:23–28.
12. Ferrero DV, Meyers HN, Schultz DE, Willis SA. Performance of the gen-probe amplified *Chlamydia trachomatis* assay in detecting *Chlamydia trachomatis* in endocervical and urine specimens from women and urethral and urine specimens from men attending sexually transmitted disease and family planning clinics. *Journal of Clinical Microbiology* 1998; **36**:3230–3233.
13. Van Der Pol B, Ferrero DV, Buck-Barrington L, Hook III E, Lenderman C, Quinn T, Gaydos CA, Lovchik J, Schachter J, Moncada J, Hall G, Tuohy MJ, Jones RB. Multicenter evaluation of the bdprobetec et system for detection of *Chlamydia trachomatis* and *Neisseria gonorrhoeae* in urine specimens, female endocervical swabs, and male urethral swabs. *Journal of Clinical Microbiology* 2001; **39**:1008–1016.
14. Leber AL, Hall GS, LeBar WD. *Cumitech 44*, *Nucleic Acid Amplification Tests for Detection of Chlamydia trachomatis and Neisseria gonorrhoeae*. ASM Press: Washington, DC, 2006.
15. Vacek PM. The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics* 1985; **41**: 959–968.
16. Torrance-Rynard VL, Walter SD. Effects of dependent errors in the assessment of diagnostic test performance. *Statistics in Medicine* 1997; **16**:2157–2175.
17. Qu Y, Tan M, Kutner MH. Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics* 1996; **52**:797–810.
18. Dendukuri N, Joseph L. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics* 2001; **57**:158–167.
19. Hadgu A, Qu Y. A biomedical application of latent class models with random effects. *Applied Statistics* 1998; **47**:603–616.
20. Georgiadis MP, Johnson WO, Singh R, Gardner IA. Correlation-adjusted estimation of sensitivity and specificity of two diagnostic tests. *Applied Statistics* 2003; **52**:63–76.
21. Hanson TE, Johnson WO, Gardner IA. Hierarchical models for the estimation of disease prevalence and the sensitivity and specificity of dependent tests in the absence of a gold-standard. *Journal of Agricultural*, *Biological and Environmental Statistics* 2003; **8**:223–239.
22. Espeland MA, Handelman SL. Using latent class models to characterize and assess relative error in discrete measurements. *Biometrics* 1989; **45**:587–599.
23. Albert PS, McShane LM, Shih JM, U.S. National Cancer Institute (NCI) Bladder Tumor Marker Network. Latent class modeling approaches for assessing diagnostic error without a gold standard: with applications to p53 immunohistochemical assays in bladder tumors. *Biometrics* 2001; **57**:610–619.

Published in 2008 by John Wiley & Sons, Ltd. *Statist. Med.* 2009; **28**:441–461
DOI: 10.1002/sim

24. Yang I, Becker MP. Latent variable modeling of diagnostic accuracy. *Biometrics* 1997; **53**:948–958.
25. Qu Y, Hadgu A. A Bayesian approach to latent class analysis via data augmentation. *Proceedings of the Annual Meeting of the American Statistical Association*. Section on Bayesian Statistical Science, Dallas, 1998; 54–57.
26. Pauler DK, Laird NM. Non-linear hierarchical models for monitoring compliance. *Statistics in Medicine* 2002; **21**:219–229.
27. Goodman LA. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* 1974; **61**:215–231.
28. Formann AK. Latent class model diagnosis from a frequentist point of view. *Biometrics* 2003; **59**:189–196.
29. Joseph L, Gyorkos T, Coupal L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *American Journal of Epidemiology* 1995; **141**:263–272.
30. Greenland S. Multiple-bias modelling for analysis of observational data. *Journal of the Royal Statistical Society*, *Series A* 2005; **168**:267–306.
31. Chung H, Loken E, Schafer JL. Difficulties in drawing inferences with finite-mixture models: a simple example with a simple solution. *American Statistician* 2004; **58**:152–158.
32. Dendukuri N, Rahme E, Belisle P, Joseph L. Bayesian sample size determination for prevalence and diagnostic test studies in the absence of a gold standard test. *Biometrics* 2004; **60**:388–397.
33. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis* (1st edn). Chapman & Hall: London, 1995.
34. Berkhof J, Van Mechelen I, Gelman A. A Bayesian approach to the selection and testing of mixture models. *Statistica Sinica* 2003; **13**:423–442.
35. Garrett ES, Zeger SL. Latent class model diagnosis. *Biometrics* 2000; **56**:1055–1067.
36. Miglioretti DL. Latent transition regression for mixed outcomes. *Biometrics* 2003; **59**:710–720.
37. Raftery AE. Bayesian model selection in social research. *Sociological Methodology* 1995; **25**:111–163.
38. Albert PS, Dodd LE. A cautionary note on robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics* 2004; **60**:427–435.
39. Alonzo TA, Pepe MS. Using a combination of reference tests to assess the accuracy of a new diagnostic test. *Statistics in Medicine* 1999; **22**:2987–3003.
40. Pepe MS, Janes H. Insights into latent class analysis of diagnostic test performance. *Biostatistics* 1996; **16**: 404–411.
41. Albert JH, Chib S. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* 1993; **88**:669–679.