STATISTICS IN MEDICINE
Statist. Med. 2008; **27**:4469–4488
Published online 12 June 2008 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3317

Bayesian latent class models with conditionally dependent diagnostic tests: A case study

Joris Menten[1,2,*,†], Marleen Boelaert[3] and Emmanuel Lesaffre[2,4]

[1]*Clinical Trials Unit, Department of Public Health, Institute of Tropical Medicine, Antwerp B2000, Belgium*
[2]*Biostatistical Centre, Faculty of Medicine, Catholic University of Leuven, Leuven B3000, Belgium*
[3]*Epidemiology and Disease Control Unit, Department of Public Health, Institute of Tropical Medicine, Antwerp B2000, Belgium*
[4]*Department of Biostatistics, Erasmus MC, 3000 CA Rotterdam, The Netherlands*

SUMMARY

In the assessment of the accuracy of diagnostic tests for infectious diseases, the true disease status of the subjects is often unknown due to the lack of a gold standard test. Latent class models with two latent classes, representing diseased and non-diseased subjects, are often used to analyze this type of data. In its basic format, latent class analysis requires the observed outcomes to be statistically independent conditional on the disease status. In most diagnostic settings, this assumption is highly questionable. During the last decade, several methods have been proposed to estimate latent class models with conditional dependence between the test results. A class of flexible fixed and random effects models were described by Dendukuri and Joseph in a Bayesian framework. We illustrate these models using the analysis of a diagnostic study of three field tests and an imperfect reference test for the diagnosis of visceral leishmaniasis. We show that, as observed earlier by Albert and Dodd, different dependence models may result in similar fits to the data while resulting in different inferences. Given this problem, selection of appropriate latent class models should be based on substantive subject matter knowledge. If several clinically plausible models are supported by the data, a sensitivity analysis should be performed by describing the results obtained from different models and using different priors. Copyright © 2008 John Wiley & Sons, Ltd.

KEY WORDS: diagnostic accuracy; latent class models; sensitivity; specificity; visceral leishmaniasis

1. INTRODUCTION

Visceral leishmaniasis (VL), also known as Kala-Azar, is a deadly protozoal disease transmitted by sandflies. It occurs mainly in rural areas of Eastern Africa, Southern Asia, and Latin America

*Correspondence to: Joris Menten, Clinical Trials Unit, Department of Public Health, Institute of Tropical Medicine, Nationalestraat 155, Antwerp B2000, Belgium.
†E-mail: jmenten@itg.be

Contract/grant sponsor: Belgian State—Federal Office for Scientific Technical and Cultural Affairs

Received 15 November 2007
Accepted 4 April 2008

and causes an estimated 500 000 new cases a year. Many patients do not receive proper medical attention, as VL diagnosis and treatment is often only available in tertiary health centers. To reach more patients at the primary-care level, safe and effective drugs and simple and robust diagnostic tests are needed.

The current reference standard for VL diagnosis is demonstration of *Leishmania* parasites in tissue smears or in culture from these tissues. These parasitological techniques show a specificity close to 100 per cent, but their sensitivity is considerably lower and varies depending on the type of tissue aspirate [1]. A prospective, multinational study was designed to evaluate the value of three diagnostic tests that were considered candidates for use in peripheral health services—the freeze-dried version of the direct agglutination test (DAT), the rk39 dipstick test and a urine latex antigen test (KAtex)—in the two most affected regions: East-Africa and the Indian subcontinent. The aim of this study was to estimate the diagnostic accuracy, specifically the test sensitivity and specificity, of these three field tests in order to identify the most appropriate test, if any, for implementation in the countries' treatment programs. In addition, there was interest to determine whether the diagnostic accuracy was consistent across countries or whether there were important differences in sensitivity or specificity of the tests between geographic regions. A meta-analysis of the DAT and rk39 tests [2] indicated that sensitivity of these two tests might be higher and more homogeneous in the studies carried out in South-Asia.

As no perfect reference standard was available in the study, we used Bayesian latent class models (LCMs) to estimate test sensitivities and specificities. We expected test results to be correlated and that the data would violate the conditional independence assumption of standard latent class analysis. On the basis of subject matter knowledge, we defined a restricted list of plausible dependencies between test results and incorporated these test correlations in a Bayesian LCM. We fitted various clinically plausible models for dependencies between the test results using fixed and random effects formulations. Equivalent fixed and random effects models resulted in similar inferences on the parameters of interest. However, some parameter estimates differed considerably between models describing different dependence structures.

In this article, we describe these analyses as a case study in the use of LCMs for the analysis of a diagnostic study with correlated test results. Further, we provide another illustration of the observation of Albert and Dodd [3] that model diagnostics may not allow to distinguish between competing dependence structures that result in different conclusions. This article is organized as follows. Section 2 provides further information on the study design and data structure. In Section 3 we give an overview of different LCMs proposed to model conditional dependence and describe the Bayesian fixed and random effects models proposed by Dendukuri and Joseph [4] in more detail. Section 4 describes issues of model selection and identifiability related to LCMs. In Section 5, we present an analysis of the Sudanese data from the VL study and illustrate the issues related to model selection and identifiability. A discussion follows in Section 6, in which we review implications of our findings on the analysis of diagnostic studies without a perfect reference test and present recommendations for the use of LCMs in this setting.

2. STUDY DESIGN

The multinational VL study was performed in six study sites distributed over five countries (Ethiopia, Kenya, Sudan, India, and Nepal). At each site prospective recruitment was done of all persons that presented with symptoms of VL [5]. Three field tests were performed on all recruited

DOI: 10.1002/sim

subjects: the DAT and the rK39 dipstick test, which are antibody-detection tests performed on a serum sample, and KAtex, an antigen-detection test performed on a urine sample. In addition, microscopic examination of tissue aspirates (parasitology) was performed as a reference test. However, it is well known that parasitology is not a true gold standard. The specificity of parasitology is close to 100 per cent, but sensitivity is expected to vary according to the aspiration site. Sensitivity of spleen aspirates approaches 95 per cent, but bone marrow or lymph node aspirates have much lower sensitivity, namely 60–80 per cent and 50–60 per cent, respectively [1]. Splenic aspiration carries certain risks and requires a high level of clinical and laboratory expertise. It was only performed if the safety of the procedure could be guaranteed (trained personnel, hemoglobin and platelet count within acceptable limits, blood for transfusion, and surgical facility available); otherwise bone marrow or lymph node aspirates were obtained. On the basis of clinical grounds, it was considered unlikely that the tests would be fully independent conditional on disease status. *A priori*, a number of plausible correlations between the tests were identified (Section 4). This diagnostic study illustrates a setting where a perfect reference test is unavailable and available tests are imperfect and correlated.

3. FIXED AND RANDOM EFFECTS LCMs

LCMs are often used to analyze diagnostic studies when a perfect reference test is lacking. In these models, the true disease status of a person is an unobserved, or latent, variable with two mutually exclusive categories, 'diseased' and 'non-diseased'. This unobserved variable determines the probability to test positive or negative to a number of diagnostic tests. In its basic format, LCMs require the observed outcomes to be independent within the categories of the latent class. Recently, extensions of these LCMs are described, which allow for conditional dependence between test results [4, 6, 7]. In the following sections, we describe the basic conditional independence model and the fixed and random effects model extensions that allow for conditional dependence between test results.

3.1. Conditional independence model

Let y_{ij} be the observed binary outcome (0=negative, 1=positive) for the jth test Tj on the ith subject with true disease status d_i (0=not diseased, 1=diseased), where $i=1,\ldots,N$, $j=1,2,\ldots,J$, and y_{ij} is a realization of the binary random variable Y_{ij}. The outcome pattern over all tests for an individual subject i is then a vector $\mathbf{y}_i$ of length J with $\mathbf{y}_i=(y_{i1},y_{i2},\ldots,y_{iJ})^{\mathrm{T}}$.

Results for an individual test are Bernoulli distributed with $P(Y_{ij}=1|D_i=d_i)$, the probability of testing positive on the jth test given an individual's true disease status d_i. The conditional independence assumption can be expressed as

$$P(Y_{i1}=y_1,Y_{i2}=y_2,\ldots,Y_{iJ}=y_J|D_i=d_i)=\prod_{j=1}^{J}P(Y_{ij}=y_j|D_i=d_i) \tag{1}$$

This can be expressed in terms of the test sensitivities and specificities as

$$P(Y_{i1}=y_1,Y_{i2}=y_2,\ldots,Y_{iJ}=y_J|D_i=1)=\prod_{j=1}^{J}S_j^{y_j}(1-S_j)^{(1-y_j)} \tag{2}$$

DOI: 10.1002/sim

$$P(Y_{i1}=y_1, Y_{i2}=y_2, \ldots, Y_{iJ}=y_J | D_i=0) = \prod_{j=1}^{J} C_j^{(1-y_j)}(1-C_j)^{y_j} \quad (3)$$

with $S_j = P(Y_{ij}=1|D_i=1) = P(Y_j=1|D_i=1)$ being the sensitivity of test Tj and $C_j = 1 - P(Y_{ij}=1|D_i=0) = 1 - P(Y_j=1|D_i=0)$ being the specificity of test Tj.

The marginal distribution of the outcomes can then be modeled using a multinomial distribution with class probabilities:

$$\begin{aligned} &P(Y_1=y_1, Y_2=y_2, \ldots, Y_J=y_J) \\ &= \sum_{k=0}^{1} P(D=k) P(Y_1=y_1, Y_2=y_2, \ldots, Y_J=y_J | D=k) \\ &= \sum_{k=0}^{1} P(D=k) \prod_{j=1}^{J} P(Y_j=y_j|D=k) \\ &= \pi \prod_{j=1}^{J} S_j^{y_j}(1-S_j)^{(1-y_j)} + (1-\pi) \prod_{j=1}^{J} C_j^{(1-y_j)}(1-C_j)^{y_j} \end{aligned} \quad (4)$$

with π being the disease prevalence. Estimation of the unknown parameters can proceed by maximum likelihood [8] or Markov chain Monte Carlo (MCMC) approaches [9].

In most diagnostic settings, the assumption of conditional independence is, however, highly questionable. Dependence between diagnostic tests can be induced by subject- and observer-related effects. Subject-specific factors that may induce dependencies between diagnostic tests for infectious diseases include disease severity, immunological response, and pathogen burden in subjects with the disease of interest and the presence of cross-reacting diseases in subjects without the target disease. Observer differences and variations in sample handling and storage may also induce correlations between test results. When assessing diagnostic tests for infectious diseases, the dependencies between the test results are likely to differ between diseased and non-diseased subjects.

If LCMs are used in the analysis of a diagnostic study, the model should allow for dependencies between test results. The specification of the dependencies should be flexible enough to allow correlations between several tests and incorporate different dependence structures across the latent classes. A number of authors have described methods to generalize LCMs over the last 10 years [4, 6, 7], allowing for conditional dependence between the diagnostic tests. In a Bayesian context, Dendukuri and Joseph described two classes of LCMs [4] that allow for a flexible specification of the dependence structure. The main distinction between the two model classes they proposed is whether the test sensitivities and specificities remain constant (fixed effects model) from subject to subject or not (random effects model).

3.2. *Fixed effects LCMs*

In equation (4), conditional dependence can be modeled by introducing covariances between pairs of tests. For example, if we assume that tests T1 and T2 are correlated in diseased subjects only,

DOI: 10.1002/sim

the probability of an outcome pattern $\mathbf{y}$ is

$$
\begin{aligned}
&P(Y_1=y_1, Y_2=y_2, \ldots, Y_J=y_J)\\
&=\sum_{k=0}^{1} P(D=k)P(Y_1=y_1, Y_2=y_2, \ldots, Y_J=y_J|D=k)\\
&=P(D=1)P(Y_1=y_1, Y_2=y_2|D=1)\prod_{j=3}^{J} P(Y_j=y_j|D=1)+P(D=0)\prod_{j=1}^{J} P(Y_j=y_j|D=0)\\
&=\pi(S_1^{y_1}S_2^{y_2}(1-S_1)^{(1-y_1)}(1-S_2)^{(1-y_2)}+(-1)^{(y_1-y_2)}\mathrm{cov}_{12|D=1})\prod_{j=3}^{J} S_j^{y_j}(1-S_j)^{(1-y_j)}\\
&\quad+(1-\pi)\prod_{j=1}^{J} C_j^{(1-y_j)}(1-C_j)^{y_j} \qquad (5)
\end{aligned}
$$

where $\mathrm{cov}_{jj'|D=d_i}$ denotes the pairwise covariance between test results j and j' for subjects with disease status d_i. The corresponding correlation is given by $\rho_{jj'|D=d_i}$, where $\rho_{jj'|D=1}=\mathrm{cov}_{jj'|D=1}/\sqrt{S_jS_{j'}(1-S_j)(1-S_{j'})}$ and $\rho_{jj'|D=0}=\mathrm{cov}_{jj'|D=0}/\sqrt{(1-C_j)(1-C_{j'})C_jC_{j'}}$. Additional pairwise covariances can be entered similarly.

Clearly, models with four or more diagnostic tests result in a complicated notation, especially when including higher-order correlations. A model for four dependent tests containing all two-, three- and four-way correlations, parametrized in terms of conditional probabilities, was described by Berkvens *et al.* [10].

3.3. Random effects LCMs

In the random effects model, the probability of testing positive for test Tj depends not only on the unobserved disease status but also on continuous latent random variables through a regression model [4, 6, 11]. In this model, outcomes for a single test for an individual subject i are Bernoulli distributed with

$$P(Y_{ij}=1|D_i=d_i, \mathbf{Z}_i=\mathbf{z}_i)=\eta^{-1}(\alpha_{jd_i}+\boldsymbol{\beta}_{jd_i}^{\mathrm{T}}\mathbf{z}_i) \qquad (6)$$

where η is a link function. For the link function η both probit [6, 12] and logit [7] links have been proposed. For the probit link, $\eta^{-1}(y)=\Phi(y)$, where Φ represents the cumulative distribution function of the Normal(0, 1) distribution and for the logit link, $\eta^{-1}(y)=1/(1+\mathrm{e}^{-y})$. The regression equation consists of an intercept term α_{jd_i}, a vector of realized values of random effects $\mathbf{z}_i$, and a coefficient vector $\boldsymbol{\beta}_{jd_i}$. The vector of random effects consists of K random variables with $Z_{ki}\backsim\mathrm{N}(0,1)$, with $\mathbf{z}_i=(z_{1i}, z_{2i}, \ldots, z_{Ki})^{\mathrm{T}}$. The coefficient vector $\boldsymbol{\beta}_{jd_i}=(\beta_{j1d_i}, \ldots, \beta_{jKd_i})^{\mathrm{T}}$ describes the dependency of test Tj on the K random effects. The dependence structure of the model is defined by the random effects and coefficient vectors $\boldsymbol{\beta}_{jd_i}$. Tests that share a common random value within each patient will show dependence, conditional on the patient's disease status, without the need to explicitly specify a covariance parameter [4, 6, 11].

For example, if we assume that tests T1 and T2 are correlated in diseased subjects only, there would be a single random effect z_{1i} and the coefficient vectors would be scalars with $\beta_{110}=\cdots=\beta_{J10}=0$, $\beta_{111}=\beta_{211}\equiv\gamma_{12|D=1}$, and $\beta_{311}=\cdots=\beta_{J11}=0$, where the size of $\gamma_{12|D=1}$ indicates the strength of the dependency of T1 and T2 on the random effect in diseased subjects and consequently

DOI: 10.1002/sim

is proportional to the strength of the association between T1 and T2. The subscript of $\gamma_{12|D=1}$ is chosen to describe the correlation induced by the random effect and does not correspond to the subscripts of the $\boldsymbol{\beta}$ vectors.

The random effects can be thought of as unobserved characteristics of the subjects that influence the probability of testing positive for one or more of the diagnostic tests. The model can be generalized through the introduction of observed covariates influencing the diagnostic test results and the use of other probability distributions than the normal for the random effects. Given (a) the disease status of the subject, (b) random effects, and (c) an observed covariate vector, the results of the different tests are then assumed to be independent as in equation (1). The resulting full likelihood of the model over all subjects and estimation methods are described by Dendukuri and Joseph [4] for a Bayesian setting and Qu *et al.* [6] in a frequentist approach.

By allowing the probability of a test result to depend on observed and/or unobserved subject characteristics, we explicitly abandon the equality of the population sensitivity with the probability of testing positive for an individual diseased subject i. To obtain the population-averaged sensitivity and specificity, we need to average over the random effects distributions.

3.4. Comparison of fixed and random effects LCMs and alternative approaches

Similar dependence structures can be parametrized using the two modeling approaches. Pairwise dependencies between two tests can either be modeled by a covariance term in the fixed effects formulation or by a common Gaussian random effect in the random effects formulation. In our setting of multiple diagnostic tests, the most plausible cause of the correlations between test results is through observed or unobserved subject characteristics, for example, immunoresponse or pathogen load. In this case, the fixed effects model is understood as a marginal manifestation of a random effects model. Exceptions would generally be linked to deficiencies in the study conduct, e.g. correlations induced by lack of blinding of test results or improper storage of clinical samples. These fixed and random effects models can cover a wide range of possible dependencies between test results.

Other models have been proposed that need further generalization of this framework. Albert *et al.* [13], for example, replace the Gaussian distribution of the random effects by a finite mixture distribution in which some individuals are unequivocally and correctly classified by all tests, whereas others are subject to diagnostic error. This model was proposed for situations where most severely diseased and healthiest patients are the easiest to classify. The model assumes that there are subjects for whom no diagnostic error is made. In our setting of diagnostic laboratory tests for infectious diseases, this situation is unlikely to happen.

An alternative approach to modeling the dependence between diagnostic tests is the use of LCMs with more than 2 latent classes [14–16], for example, adding a class with subjects with ambiguous disease status. Within each of these classes, conditional independence is then assumed. In our study, we focussed on the 2-class model as it easily provides estimates of test sensitivity and specificity, which is more difficult for the 3-class model [17].

4. MODEL IDENTIFIABILITY AND SELECTION

In general, LCMs are only identifiable if the number of parameters estimated is less than or equal to the number of independent multinomial cell frequencies, which is $2^J - 1$ for J tests.

DOI: 10.1002/sim

The number of parameters to be estimated in an LCM with an unconstrained conditional dependence structure is, however, $2^{J+1}-1$. Consequently, the parameter space needs to be reduced to allow the model to be estimated by placing constraints on either the prevalence, test accuracy parameters, or dependency structure. Conditional independence is such a reduction of the parameter space and allows estimation of a model with at least three tests by setting all two-way and higher associations, conditional on the disease status, to be zero.

Instead of applying a popular, but not necessarily justifiable, assumption as conditional independence, constraints on the parameter space can be based on subject matter knowledge or other external sources of information. This can take the form of deterministic or probabilistic constraints [10]. Setting a model parameter to a particular value, including zero, is an example of a deterministic constraint. On the other hand, specifying a prior distribution for a parameter is a probabilistic constraint. In general, probabilistic constraints are preferred as the resulting inference will better reflect the available knowledge and uncertainty. In a Bayesian approach, prior information, e.g. from previous similar studies or expert opinion, can be used in the specification of the prior distributions of the parameters of interest or nuisance parameters, as covariances. Experts can usually state which pairwise dependencies they find clinically plausible and may even be able to rank dependencies in the order of probability of occurrence. They are, however, rarely able to provide meaningful prior probabilities for the correlations among diagnostic tests or random effect coefficients.

In addition to the lack of model identifiability—due to the estimation of more parameters than available independent data points—it has been reported that (1) some LCMs may be only weakly estimable [9, 18] and (2) different LCMs may provide a similar fit to the data while leading to different study conclusions [3, 13]. Weak estimability occurs when, although the technical conditions for identifiability are met, there is not enough information in the data to estimate all the parameters in the model without relying heavily on the prior distributions of the parameters. Examples of weakly estimable models in the context of LCMs are given by Garrett and Zeger [18] and by Garrett *et al.* [9]. An example of the second phenomenon is given by Albert and Dodd [3], who analyzed Handelmans dentistry data using a general random effects model, a restricted model where a common sensitivity and specificity over raters is assumed, and a finite mixture model where some patients are diagnosed without error, while others may be subject to diagnostic error. All three models provided an acceptable fit to the data and model diagnostics (loglikelihoods, chi-squared values) were very similar across the models. Parameter estimates from the models differed greatly, however. We will illustrate these two issues with another data set where a number of models yield substantially different parameter estimates but result in a similar fit to the data. Combining the dependence structures from two of these models, that each in itself was fully estimable, results in a weakly estimable model.

Model selection criteria can be used to identify models that do not give an acceptable fit to the data at hand. In a frequentist setting, Pearson and likelihood ratio statistics are calculated and compared with a chi-squared distribution [8]. In a Bayesian setting, the deviance information criterion (DIC), Bayes' factors, and Bayesian p-values [10, 19] can be used, as well as graphical assessment of the posterior predictive distributions of the model parameters and marginal counts over the different possible outcome patterns. The posterior predictive plots can indicate which models are clearly incompatible with the data and which parameters are not well identified from a given set of data. The DIC is a generalization of the Akaike information criterion and Bayesian information criteria and estimates the expected deviance of the replicated data. The model with the lowest DIC should show the best out-of-sample prediction and would be the preferred model

DOI: 10.1002/sim

[19, 20]. The Bayesian p-value is defined as the probability that replicated data from a Bayesian model are more extreme than the observed data. A p-value close to 0 indicates lack-of-fit of a selected model [10, 19].

5. APPLICATION TO THE VL DATA

5.1. Study design and data structure

In the multinational VL study, four tests were performed for each subject: the field tests of interest DAT (T1), rk39 (T2), and KAtex (T3), and the imperfect reference test parasitology (T4). All tests were analyzed as binary outcomes (test positive or negative) with standard cut-offs for the semi-quantitative DAT (T1 positive if titer $\geqslant$1:3200) and KAtex test (T3 positive if result +, ++, or +++). We could have modeled the actual, rather than dichotomized, results. This would require either the use of a parametric model for the test outcomes, for example, a proportional odds model, or the estimation of many more probabilities if a fully nonparametric approach was used. As our interest lies in the estimation of diagnostic accuracy of the tests as used and interpreted in field conditions, we chose to work with the dichotomized results using standard cut-offs. This results in a test outcome pattern of the form $\mathbf{y}=(y_1, y_2, y_3, y_4)^{\mathrm{T}}$, where $y_j=1$ if the test result is positive and $y_j=0$ if negative (Table I).

On the basis of clinical information, the following correlations between test results were expected:

1. As both the DAT (T1) and the rk39 (T2) tests are based on antibody detection, the results are likely to be positively correlated in VL subjects. Individuals with a strong immunoresponse are more likely to test positive, whereas immunosuppressed subjects, e.g. the very young or HIV or TB co-infected, are more likely to show false-negative results.
2. The antigen-detection test KAtex (T3) is more likely to test positive in individuals with a high leishmanial parasite load. Similarly, microscopical detection of parasites (T4) is more likely to be successful in subjects with many circulating parasites. Consequently, we expected a positive correlation between KAtex and parasitology results in VL subjects.
3. It has been shown that cases of leishmanial and HIV co-infection show a low immunoresponse combined with a high parasite load [21]. This would lead to a negative correlation in test results between the antibody-detection tests (T1 and T2) on one hand and KAtex (T3) and parasitology (T4) on the other hand in subjects with HIV–leishmanial co-infection.
4. The antibody-detection tests (T1 and T2) can stay positive for months after cure and can also be positive in healthy persons with asymptomatic leishmanial infection. This would lead to a positive correlation between DAT and rk39 in non-VL subjects. However, only patients with the full-blown clinical syndrome of febrile splenomegaly were recruited into the study. The number of asymptomatic leishmania infections presenting with another febrile illness was likely to be low, but not necessarily zero. Consequently, this correlation is less likely to be important.

In the models described in this article, we considered further correlations unlikely to be of importance, although they can be hypothesized. For example, lack of proper blinding could lead to general positive correlations between test results, both in VL and non-VL subjects, as the interpretation of one test may be influenced by the knowledge of the results of another. However, the study

DOI: 10.1002/sim

Table I. Observed and posterior median predicted frequency of test outcome patterns ($N=291$) and posterior mean predicted probability (per cent) of being diseased given the outcome pattern for the different fixed effects model formulations.

Test				Observed	Predicted frequency		Predicted probability (per cent) of being diseased under model					
T1	T2	T3	T4	frequency	Model 0	Model 5	0	1	2	3	4	5
1	1	1	1	51	45	48	100	100	100	100	100	100
1	1	1	0	1	7	3	99	99	100	90	93	100
1	1	0	1	4	8	5	100	100	100	96	98	98
1	1	0	0	16	3	14	44	41	98	3	47	100
1	0	1	1	15	12	14	100	100	100	100	100	100
1	0	1	0	1	2	1	86	86	93	89	90	94
1	0	0	1	2	2	1	93	93	97	95	97	97
1	0	0	0	5	12	7	5	5	61	3	41	63
0	1	1	1	7	7	8	100	100	100	100	100	100
0	1	1	0	4	1	1	72	72	68	66	64	69
0	1	0	1	1	1	1	83	82	83	81	81	83
0	1	0	0	15	23	16	1	1	15	0	8	17
0	0	1	1	1	2	1	99	97	98	99	98	96
0	0	1	0	1	2	2	20	15	11	12	11	8
0	0	0	1	1	2	2	28	21	19	20	19	14
0	0	0	0	166	155	161	0	0	0	0	0	0

protocol required test readers to be blinded to the results of other tests; hence, this dependence is unlikely to occur in the study.

5.2. *Model descriptions*

In addition to a conditional independence model (model 0), we fitted various models incorporating several conditional dependencies between test results (Table II and Appendix A). Models 1–3 incorporate a single pairwise correlation, corresponding to the effects of immunoresponse to VL infection (model 1), parasite load (model 2), and asymptomatic leishmanial infection coinciding with another febrile disease (model 3). Models 4 and 5 incorporate two pairwise correlations. Model 4 includes the effects of immunoresponse to VL infection and to asymptomatic leishmanial infection, whereas model 5 includes the effects of immunoresponse to VL infection and parasite load. Model 6 captures the hypothesized negative correlation between immunoresponse to VL infection on the one hand and parasite load on the other hand in VL subjects. Expert opinion indicated that, *a priori*, models 5 and 6 were clinically most plausible.

For all dependence models apart from model 6, equivalent fixed (Appendix A.1) and random effects (Appendix A.2) formulations were constructed. In model 6, the dependencies across the four tests were induced by a common random effect with positive coefficients for tests 1 and 2, and negative coefficients for tests 3 and 4. Incorporating this structure in the fixed effects formulation would require the specification of a complex model with third- and fourth-order correlation terms. Given the difficulties in interpreting such a model, only the random effects formulation was considered.

As the test accuracy may vary between the five countries included in the study, we modeled the data from each country separately and present in this paper only the data obtained in Sudan

DOI: 10.1002/sim

Table II. Description and model selection criteria for fixed and random effects models in the visceral leishmaniasis diagnostic study.

Model No.	Correlations in diseased subjects	Correlations in non-diseased subjects	Fixed effects formulation					Random effects formulation	
			q	pD	Mean deviance	DIC	Bayesian p-value	Mean deviance	Bayesian p-value
0			9	9.5	109.6	119.2	<0.001	109.8	<0.001
1	T1–T2		10	10.1	109.7	119.8	<0.001	110.6	<0.001
2	T3–T4		10	8.2	69.5	77.7	0.184	69.2	0.124
3		T1–T2	10	8.5	69.2	77.7	0.194	70.2	0.102
4	T3–T4	T1–T2	11	8.4	69.1	77.5	0.197	69.3	0.124
5	T1–T2 and T3–T4		11	8.8	69.4	78.2	0.195	70.0	0.106
6	T1–T2 and T3–T4*		11					70.3	0.102

Note: A low value for the Bayesian p-value indicates lack-of-fit of the proposed model. q, number of parameters in the model; pD, Bayesian measure of model complexity; DIC, deviance information criterion. Deviance for the random effects model was calculated over the margins.
*T1 and T2 negatively correlated with T3 and T4.

($N=291$), see Table I. Similar results were obtained in the other countries [5]. We fitted all LCMs using WinBugs 1.4 (MRC Statistical Unit Cambridge, Cambridge, U.K.) called from R 2.3.1 (R Foundation for Statistical Computing, Vienna, Austria). The analysis programs are available on the Internet at http://med.kuleuven.be/biostat/software/software.htm. Vague priors were specified for all model parameters. We used Beta(1, 1) distributions, equivalent to uniform distributions over the interval [0, 1], as prior for the prevalence and test sensitivities and specificities in the fixed effects formulation. Uniform priors over the feasible range, as determined by the test sensitivities and specificities [4, 22], were specified for test covariances. In the random effects model, we used normal priors with mean μ equal to zero and standard deviation σ equal to 1.69 on the logit scale for the probability of testing positive or negative for a subject with all random effects equal to zero. This prior matches a uniform prior over the interval [0, 1] in the first two moments on the probability scale [23] and consequently results in similar priors as used for the fixed effects model. Using a vague normal prior on the logit scale would result in a prior on the probability scale that strongly favors sensitivities and specificities close to 0 or 1. Vague normal priors ($\mu=0$ and $\sigma=4.5$), constrained to be positive, were specified for the random effect coefficients β_{jkd_i}. The σ of 4.5 for these normal priors was chosen as the highest standard deviation that did not result in convergence problems when fitting the models in WinBugs. In the random effects models, some sensitivities and specificities were constrained to be >50 per cent to avoid label switching. Label switching occurs when the diseased subjects are modeled as non-diseased and *vice versa*, resulting in test sensitivities and specificities of less than 50 per cent. The fixed effects models converged to a unique solution without providing additional constraints.

We monitored convergence of the MCMC algorithm using trace plots and the potential scale reduction $\hat{R}$ [19] calculated in WinBugs. Loosely speaking, the random effects models converged more slowly in WinBugs than the fixed effects models and each individual simulation was of longer duration. Fixed effects models appeared to converge within 20 iterations, whereas random effect models showed good mixing of the chains within 100 iterations. To ensure adequate convergence all results were obtained using two chains of 10 000 iterations, of which we discarded the first 2000 (burn-in).

DOI: 10.1002/sim

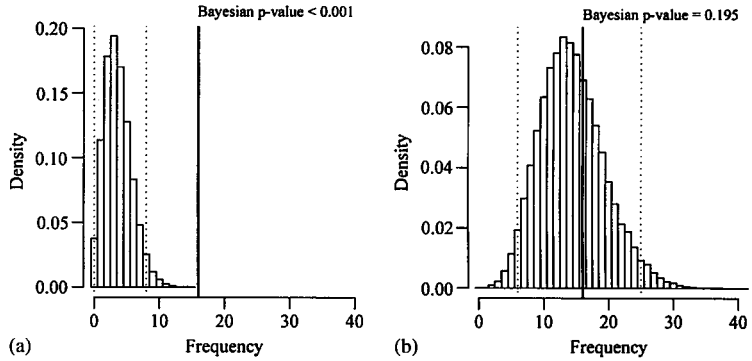

Figure 1. Posterior predictive graph: observed (bold line) and predicted (histogram) frequency of response pattern '1100' for the conditional independence model 0 (a) and fixed effects model 5 (b). Dotted lines indicate the 95 per cent prediction interval for the response pattern frequency.

5.3. *Results*

The conditional independence model (model 0) resulted in an unacceptably bad model fit (Table II). The observed frequency of test pattern '1100' was considerably higher than predicted by the conditional independence model (Figure 1(a) and Table I). Incorporating dependency between the immunoresponse tests (T1 and T2) in diseased subjects (model 1) did not improve the model fit, indicating that false-negative results in these two antibody-detection tests are unrelated. Models 2–6 showed an improved and acceptable fit to the data, as indicated by the posterior predictive distribution (Figure 1(b) and Table I) and the formal model selection criteria (Table II).

Although fitted margins were virtually indistinguishable between models 2–6, the parameter estimates differed greatly. This was observed for both the fixed (Table III) and random (Table V) effects formulations of the models. Similarly, the predicted disease probability for some of the observed outcome patterns showed important differences between the different models. The posterior mean predicted disease probabilities are shown in Table I for the fixed effects formulation; the results for the random effects formulation were similar.

Prevalence estimates (95 per cent credible interval) varied from 29.7 per cent (24.4–35.2) in model 3 to 37.0 per cent (31.0–43.1) in model 2, with similarly large differences in sensitivity and specificity estimates of the different diagnostic tests. Model 4 showed estimates intermediate between model 2 and 3 but showed much larger standard errors for model parameters than any of the other models. The prevalence estimate in model 4 was 33.2 per cent with the 95 per cent credible interval (25.9–41.0) covering the prevalence point estimates of both models 2 and 3.

The high standard errors indicate that model 4 is only weakly estimable, which was confirmed by graphing the posterior distributions of the model parameters. For example, for the specificity of the DAT (C_1) we observed a bimodal distribution with one mode slightly below 90 per cent and the other mode close to 100 per cent. This is due to the fact that the dependence in our data set was induced by the excess of subjects with response pattern '1100' compared with what was predicted under the conditional independence assumption (Figure 1(a)). This dependence can be explained by VL subjects showing negative results on both T3 and T4, indicating a correlation between T3 and T4 in VL subjects ($\rho_{34|D=1}$), or by non-VL subjects showing positive results both on T1 and T2, indicating a correlation between T1 and T2 in non-VL subjects ($\rho_{12|D=0}$). Models

DOI: 10.1002/sim

Table III. Posterior mean and standard error for prevalence (π), test sensitivities (S_i) and specificities (C_i), and covariances ($\rho_{jj'|D=d_i}$) from the fixed effects models 2–5 in the visceral leishmaniasis diagnostic study.

	Model 2		Model 3		Model 4		Model 5	
Parameter	Mean	SE	Mean	SE	Mean	SE	Mean	SE
π	37.0	(3.1)	29.7	(2.7)	33.2	(4.0)	37.0	(3.1)
S_1	85.6	(4.1)	85.4	(4.0)	85.6	(4.1)	85.7	(4.1)
C_1	98.2	(1.4)	89.4	(2.2)	93.6	(3.7)	98.2	(1.4)
S_2	78.1	(4.2)	77.0	(4.5)	77.1	(4.4)	77.9	(4.2)
C_2	91.7	(2.4)	83.8	(2.6)	87.2	(3.7)	91.8	(2.4)
S_3	73.0	(4.7)	90.5	(3.4)	81.3	(7.4)	72.9	(4.7)
C_3	98.2	(1.2)	98.3	(1.2)	98.2	(1.2)	98.3	(1.2)
S_4	75.1	(4.8)	93.0	(3.4)	83.9	(7.8)	74.9	(4.8)
C_4	98.9	(0.8)	98.9	(0.8)	98.9	(0.8)	98.8	(0.9)
$\rho_{12\|D=0}$			0.53	(0.08)	0.41	(0.18)		
$\rho_{12\|D=1}$							−0.06	(0.09)
$\rho_{34\|D=1}$	0.68	(0.09)			0.48	(0.23)	0.68	(0.08)

including only one of these two correlations (models 2 and 3) will correspond to one of the two observed modes. Model 2, including $\rho_{34|D=1}$, corresponds to correlated errors in T3 and T4 in VL subjects. This means that subjects with pattern '1100' are modeled as VL subjects that show, incorrectly, negative results for T3 and T4 and results in low sensitivity estimates of T3 and T4 and high specificity estimates for T1 and T2. Conversely, model 3, including $\rho_{12|D=0}$, corresponds to correlated errors in T1 and T2 in non-VL subjects. This means that subjects with pattern '1100' are modeled as non-VL subjects that are false positives for T1 and T2 and results in low specificity estimates of T1 and T2 and high sensitivity estimates for T3 and T4.

Given the observation of Albert *et al.* [13] and Albert and Dodd [3] that sensitivity, specificity, and prevalence estimates may be biased when the conditional dependence structure between tests in an LCM is misspecified, these results may not be surprising. Models 2 and 3 describe fundamentally different dependence structures with model 2 assuming conditional dependence only in diseased subjects, whereas model 3 assumes conditional dependence only in non-diseased subjects. Model 4, which incorporates conditional dependencies in both diseased and non-diseased, shows results intermediate between models 2 and 3. It should be noted that for some parameters, the estimates are stable over the models considered. From Table III we can see that estimates of the sensitivities of T1 and T2 and of the specificities of T3 and T4 are similar across models. In addition, a number of models (models 2, 5, and 6) show similar results although, in part, they describe different dependence structures.

A priori, $\rho_{34|D=1}$ was presumed to be more important than $\rho_{12|D=0}$. In addition, parameter estimates from models 2 and 5 were more in line with expert expectations. In Sudan, parasitological examination relied only on bone marrow and lymph node aspiration for which a sensitivity of 93.0 per cent would be unexpectedly high. On the basis of these considerations, we selected model 5, the *a priori* most plausible model, for our study conclusions [5]. This *post hoc* reliance on prior expert opinion for model selection (in this case, the expected sensitivity of the parasitological examination) might suggest that the incorporation of probabilistic priors in a more general dependence model, for example, model 4, would lead to an identifiable model. However, when in model 4 different

DOI: 10.1002/sim

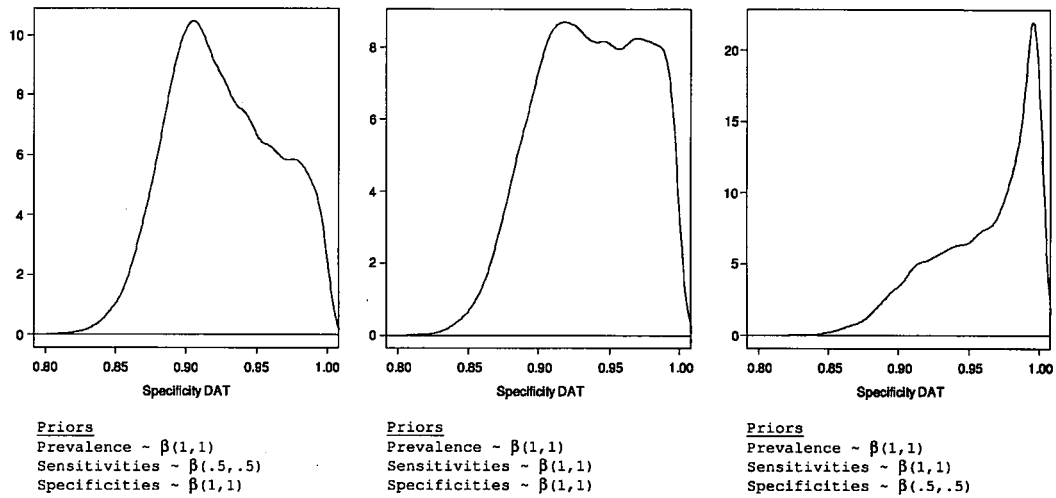

Figure 2. Posterior distribution for the specificity of T1 for model 4 using three different vague priors.

non-informative priors were used, relatively large changes in the posterior distribution of the parameter estimates were observed. As an example, Figure 2 shows the posterior distribution of C_1 when priors for the test sensitivities or specificities are changed from Beta(1, 1) to Beta(0.5, 0.5). These small changes in vague priors resulted in dramatic changes in the posterior distribution of the parameter estimates. Consequently, care should be exercised when providing informative priors for the parameters of interest in models that are only weakly estimable as these priors may exert a larger influence than warranted by the strength of their supporting evidence. In addition, this example illustrates the importance of assessing the full posterior distribution under different priors rather than only the posterior modes or means. Similar caveats hold for maximum likelihood estimation, where the full likelihood function should be assessed to identify local maxima and areas where the likelihood function is relatively flat.

Apart from the *a priori* clinically plausible test dependencies described above (Section 5.1), we considered the four tests to be conditionally independent. This could be considered as strong deterministic constraints on a relatively large number of theoretically possible, but clinically unlikely, pairwise and higher-level covariance terms. Our choice of dependence structures considered was based on subject matter knowledge and expert opinion. Additional dependencies between test results can be hypothesized but were assumed to be zero in our analyses. To assess the effect of this assumption, we studied models with probabilistic rather than deterministic constraints on these additional covariance terms. In Table IV, we show parameter estimates from two models, equivalent to models 2 and 3 in Table III, but with normal priors with $\mu=0$ and $\sigma=0.1$ for all remaining pairwise covariance terms. The apparent fit of the models with these probabilistic constraints, as described by the DIC and Bayesian p-value [10], improved compared with the models with deterministic constraints. The parameter estimates, however, remained similar, apart from a slight increase in standard errors.

Fixed (Table III) and random (Table V) effects formulations resulted in similar inferences on the parameters of interest for all models except the weakly identifiable model 4. As indicated above, the

DOI: 10.1002/sim

Table IV. Posterior mean and standard error for prevalence (π), test sensitivities (S_i) and specificities (C_i) together with model fit criteria from the fixed effects models 2 and 3 with probabilistic constraints (normal priors with $\mu=0$ and $\sigma=0.1$) on additional covariance terms.

	Model 2′		Model 3′	
Parameter	Mean	SE	Mean	SE
π	36.6	(3.5)	29.6	(3.2)
S_1	86.1	(4.9)	86.1	(4.7)
C_1	97.9	(1.7)	89.7	(2.5)
S_2	78.1	(4.3)	76.0	(4.6)
C_2	91.3	(2.9)	83.5	(3.1)
S_3	72.3	(5.1)	89.6	(4.7)
C_3	97.8	(1.3)	98.1	(1.8)
S_4	75.0	(5.2)	92.5	(4.1)
C_4	98.4	(1.1)	98.4	(1.6)
pD	9.5		9.8	
DIC	71.4		72.3	
Bayesian p-value	0.578		0.550	

Table V. Posterior mean and standard error for prevalence (π), test sensitivities (S), and specificities (C) and random effect coefficients ($\gamma_{jj'|D=d_i}$) from the random effects models 2–6 in the visceral leishmaniasis diagnostic study.

	Model 2		Model 3		Model 4		Model 5		Model 6	
Parameter	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
π	37.2	(3.1)	29.2	(2.8)	35.1	(4.3)	37.2	(3.1)	36.5	(3.1)
S_1	85.0	(4.0)	86.3	(4.0)	85.5	(4.0)	84.9	(4.1)	86.2	(3.9)
C_1	97.8	(1.3)	88.2	(2.4)	95.3	(4.1)	97.7	(1.3)	97.6	(1.4)
S_2	78.5	(4.3)	77.2	(4.6)	78.2	(4.2)	78.2	(4.2)	79.1	(4.1)
C_2	92.0	(2.4)	82.3	(2.9)	89.3	(4.5)	91.8	(2.4)	91.5	(2.4)
S_3	69.3	(5.0)	90.9	(3.3)	74.3	(9.6)	69.3	(5.0)	69.8	(5.2)
C_3	98.1	(1.1)	97.7	(1.2)	98.0	(1.1)	98.1	(1.1)	97.9	(1.2)
S_4	71.1	(5.0)	93.8	(3.2)	76.5	(9.9)	71.2	(5.0)	71.9	(5.2)
C_4	98.5	(0.9)	98.4	(0.9)	98.5	(0.9)	98.5	(0.9)	98.5	(0.9)
$\gamma_{12\|D=0}$			2.88	(0.55)	1.20	(1.04)				
$\gamma_{12\|D=1}$							0.44	(0.33)	0.23	(0.20)
$\gamma_{34\|D=1}$	4.38	(1.05)			3.67	(1.83)	4.36	(1.03)	4.53	(1.16)

fixed effects models result in direct estimates of the test sensitivities and specificities as parameters from the model. The random effects models require marginalization to obtain population-averaged sensitivity and specificity estimates.

6. DISCUSSION

In this paper, we described the analysis of a diagnostic phase three-type study with correlated test results [24]. On the basis of the working mechanism of the diagnostic tests, we identified a

DOI: 10.1002/sim

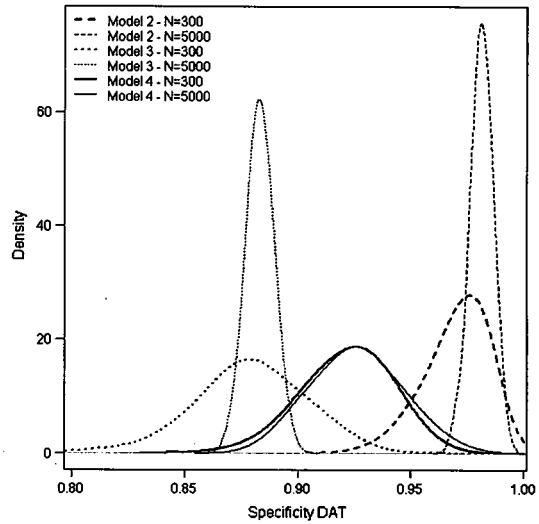

Figure 3. Estimated specificity of DAT (T1) from fixed effects models 2–4 applied on 1000 simulated data sets with $N=300$ and 5000 and parameters obtained from applying model 2 in the visceral leishmaniasis diagnostic study.

restricted number of dependence structures as *a priori* plausible. However, even with a severely restricted dependence structure, the model containing all plausible correlations between tests was not fully identifiable. In this model some parameters were estimable, whereas others were not. Increasing the sample size would not improve estimability of these parameters, as illustrated in Figure 3. Figure 3 shows the estimated specificity of DAT (C_1) obtained from models 2–4 on 1000 simulated data sets. We simulated 500 data sets each with sample sizes 300 and 5000 using the dependence structure assumed in model 2 and parameter estimates obtained from applying model 2 on the VL data. All three models showed an excellent fit to the data with Bayesian p-values close to 0.5. Although increasing the sample size resulted in an increased precision in the estimate of C_1 for models 2 and 3, the increased sample size did not improve the estimability of C_1 in model 4.

Using a Bayesian approach, prior information, e.g. from other studies or expert opinion, can be incorporated in the model to improve identifiability of the model. Unfortunately, introducing prior information on the parameters of interest may influence inference from the study more strongly than warranted by the strength of the prior information and in a non-transparent way. Providing prior information on the dependence structure can lead to an identifiable model in a more intuitive way for the data analysts, but it might be difficult to elicit distributional priors from experts on the test covariances or random effects coefficients. Consequently, when using informative priors in this setting, it is important to provide informative priors for the dependence structure as well as for the parameters of interest and to assess the variability of study conclusions under different priors.

These problems lead some authors to dismiss LCMs as a viable analysis method for the analysis of diagnostic studies. Instead of LCMs, Alonzo and Pepe [25] advocate the use of a combined reference standard (CRS) for the analysis of diagnostic studies. However, if there is genuine uncertainty on the disease status of each subject, the analysis strategy should accurately reflect this. Using a CRS assumes that we can in fact know the true disease status of each subject and

DOI: 10.1002/sim

that, by making a certain combination—which we know *a priori*—of individual tests, this perfect diagnosis can be made for each patient. In contrast, LCMs allow the estimation of test sensitivities and specificities incorporating the true uncertainty in disease status of the patients.

Evidently, if a perfect reference test was available in our study setting, its use would be preferable to the modeling approach we applied to these data. Our study aim, the evaluation of diagnostic accuracy of three novel tests in field settings, required, however, the use of a phase III study design [26] in which consecutive patients were recruited at primary health-care centers. The use of separately selected cases and non-diseased controls, as usual in phases I and II of diagnostic test development, would result in spectrum bias [27]. Given the observed positive correlation of KAtex and parasitological results, the use of parasitologically confirmed cases would bias results in favor of a higher sensitivity estimate of the KAtex test. The use of the more sensitive, although not perfect, aspiration of the spleen as a parasitological reference test was not routinely possible in all primary health-care centers included in our study [5].

LCMs should only be used if one is willing to state all necessary assumptions and to critically assess them. As for all statistical modeling, a purely data-driven approach is unlikely to be useful and the choice of an appropriate LCM should be based on extensive subject knowledge. *A priori*, a list of probable, plausible, and implausible test dependencies and a range of plausible models should be defined. The fit of these models is then assessed using graphical methods and model diagnostics. Models that show important lack of fit are then removed from further consideration. If different plausible models show a similar and acceptable fit to the data and result in significantly different conclusions, the results from the different models should be reported in a sensitivity analysis.

In a Bayesian setting, contextual information can be incorporated in informative priors for some of the model parameters. In this case, reparametrization, e.g. in the form of conditional dependencies [10], may help in obtaining expert opinion. Experts may have opinions on the diagnostic test accuracy and dependence structure, but although the first may be relatively straightforward to quantify, the second may be hard to capture in an informative prior. In this case it is tempting to specify informative priors for test sensitivities, specificities, or disease prevalence, but not for the parameters describing the dependence structure. This may, however, result in an inference that is consistent with prior opinion on diagnostic test accuracy but not on prior opinion on the test dependencies. When using informative priors, an effort should be made to obtain meaningful priors also on the dependence structure of the data. For example, when analyzing a population-based phase III diagnostic study, correlations observed between test results in earlier research phases may be used as priors for the dependence structure.

The issues described in this paper are not unique to LCMs. Similar problems occur in the model-based analysis of incomplete data [28] and in general latent variable models [29]. Molenberghs *et al.* [30] show that models for non-random missing data mechanisms cannot be fully tested using the observed data. They stress the importance of contextual information and subject matter knowledge to determine which models are most plausible. Sensitivity analysis is widely used in missing data analysis to explore the impact of a range of plausible models of study conclusions. These sensitivity analyses may show that some parameters vary considerably between different models, whereas other parameters may be fairly stable. Attempts have been made to summarize the results of these sensitivity analyses in graphical display or in a single summary, the region of uncertainty, which covers both imprecision and ignorance on the parameters of interest [30, 31]. Applied to LCMs, these approaches may lead to further advances in the analysis of diagnostic studies without a gold standard.

DOI: 10.1002/sim

APPENDIX A

Probability models for the 16 possible outcome patterns in the VL study are given below for the conditional independence and six dependence models in both random effects and fixed effects model formulations.

A.1. Definition of fixed effects models

The different fixed effects models are as follows:

$P(Y_1=y_1, Y_2=y_2, Y_3=y_3, Y_4=y_4)=$

Model 0:

$$\pi S_1^{y_1}(1-S_1)^{(1-y_1)}S_2^{y_2}(1-S_2)^{(1-y_2)}S_3^{y_3}(1-S_3)^{(1-y_3)}S_4^{y_4}(1-S_4)^{(1-y_4)}$$
$$+(1-\pi)C_1^{(1-y_1)}(1-C_1)^{y_1}C_2^{(1-y_2)}(1-C_2)^{y_2}C_3^{(1-y_3)}(1-C_3)^{y_3}C_4^{(1-y_4)}(1-C_4)^{y_4}$$

Model 1:

$$\pi(S_1^{y_1}(1-S_1)^{(1-y_1)}S_2^{y_2}(1-S_2)^{(1-y_2)}+(-1)^{(y_1-y_2)}\mathrm{cov}_{12|D=1})S_3^{y_3}(1-S_3)^{(1-y_3)}S_4^{y_4}(1-S_4)^{(1-y_4)}$$
$$+(1-\pi)C_1^{(1-y_1)}(1-C_1)^{y_1}C_2^{(1-y_2)}(1-C_2)^{y_2}C_3^{(1-y_3)}(1-C_3)^{y_3}C_4^{(1-y_4)}(1-C_4)^{y_4}$$

Model 2:

$$\pi S_1^{y_1}(1-S_1)^{(1-y_1)}S_2^{y_2}(1-S_2)^{(1-y_2)}(S_3^{y_3}(1-S_3)^{(1-y_3)}S_4^{y_4}(1-S_4)^{(1-y_4)}+(-1)^{(y_3-y_4)}\mathrm{cov}_{34|D=1})$$
$$+(1-\pi)C_1^{(1-y_1)}(1-C_1)^{y_1}C_2^{(1-y_2)}(1-C_2)^{y_2}C_3^{(1-y_3)}(1-C_3)^{y_3}C_4^{(1-y_4)}(1-C_4)^{y_4}$$

Model 3:

$$\pi S_1^{y_1}(1-S_1)^{(1-y_1)}S_2^{y_2}(1-S_2)^{(1-y_2)}S_3^{y_3}(1-S_3)^{(1-y_3)}S_4^{y_4}(1-S_4)^{(1-y_4)}$$
$$+(1-\pi)(C_1^{(1-y_1)}(1-C_1)^{y_1}C_2^{(1-y_2)}(1-C_2)^{y_2}$$
$$+(-1)^{(y_1-y_2)}\mathrm{cov}_{12|D=0})C_3^{(1-y_3)}(1-C_3)^{y_3}C_4^{(1-y_4)}(1-C_4)^{y_4}$$

Model 4:

$$\pi S_1^{y_1}(1-S_1)^{(1-y_1)}S_2^{y_2}(1-S_2)^{(1-y_2)}(S_3^{y_3}(1-S_3)^{(1-y_3)}S_4^{y_4}(1-S_4)^{(1-y_4)}$$
$$+(-1)^{(y_3-y_4)}\mathrm{cov}_{34|D=1})+(1-\pi)(C_1^{(1-y_1)}(1-C_1)^{y_1}C_2^{(1-y_2)}(1-C_2)^{y_2}$$
$$+(-1)^{(y_1-y_2)}\mathrm{cov}_{12|D=0})C_3^{(1-y_3)}(1-C_3)^{y_3}C_4^{(1-y_4)}(1-C_4)^{y_4}$$

Model 5:

$$\pi(S_1^{y_1}(1-S_1)^{(1-y_1)}S_2^{y_2}(1-S_2)^{(1-y_2)}+(-1)^{(y_1-y_2)}\mathrm{cov}_{12|D=1})$$
$$\times(S_3^{y_3}(1-S_3)^{(1-y_3)}S_4^{y_4}(1-S_4)^{(1-y_4)}+(-1)^{(y_3-y_4)}\mathrm{cov}_{34|D=1})$$
$$+(1-\pi)C_1^{(1-y_1)}(1-C_1)^{y_1}C_2^{(1-y_2)}(1-C_2)^{y_2}C_3^{(1-y_3)}(1-C_3)^{y_3}C_4^{(1-y_4)}(1-C_4)^{y_4}$$

DOI: 10.1002/sim

A.2. Definition of random effects models

In the random effects formulation, the probability of an outcome pattern for individual patients, conditional on their disease status and random effects value, is given by $P(Y_{i1}=y_1, Y_{i2}=y_2, Y_{i3}=y_3, Y_{i4}=y_4|D_i=d_i, \mathbf{Z}_i=\mathbf{z}_i)=\prod_{j=1}^{4} P(Y_{ij}=y_j|D_i=d_i, \mathbf{Z}_i=\mathbf{z}_i)$, where $P(Y_{ij}=1|D_i=d_i, \mathbf{Z}_i=\mathbf{z}_i)$ as given in the table below (see also equation (6)):

Model	Test	Diseased subjects	Non-diseased subjects
0	T1	$\eta^{-1}(\alpha_{11})$	$\eta^{-1}(\alpha_{10})$
	T2	$\eta^{-1}(\alpha_{21})$	$\eta^{-1}(\alpha_{20})$
	T3	$\eta^{-1}(\alpha_{31})$	$\eta^{-1}(\alpha_{30})$
	T4	$\eta^{-1}(\alpha_{41})$	$\eta^{-1}(\alpha_{40})$
1	T1	$\eta^{-1}(\alpha_{11}+\gamma_{12\|D=1}z_i)$	$\eta^{-1}(\alpha_{10})$
	T2	$\eta^{-1}(\alpha_{21}+\gamma_{12\|D=1}z_i)$	$\eta^{-1}(\alpha_{20})$
	T3	$\eta^{-1}(\alpha_{31})$	$\eta^{-1}(\alpha_{30})$
	T4	$\eta^{-1}(\alpha_{41})$	$\eta^{-1}(\alpha_{40})$
2	T1	$\eta^{-1}(\alpha_{11})$	$\eta^{-1}(\alpha_{10})$
	T2	$\eta^{-1}(\alpha_{21})$	$\eta^{-1}(\alpha_{20})$
	T3	$\eta^{-1}(\alpha_{31}+\gamma_{34\|D=1}z_i)$	$\eta^{-1}(\alpha_{30})$
	T4	$\eta^{-1}(\alpha_{41}+\gamma_{34\|D=1}z_i)$	$\eta^{-1}(\alpha_{40})$
3	T1	$\eta^{-1}(\alpha_{11})$	$\eta^{-1}(\alpha_{10}+\gamma_{12\|D=0}z_i)$
	T2	$\eta^{-1}(\alpha_{21})$	$\eta^{-1}(\alpha_{20}+\gamma_{12\|D=0}z_i)$
	T3	$\eta^{-1}(\alpha_{31})$	$\eta^{-1}(\alpha_{30})$
	T4	$\eta^{-1}(\alpha_{41})$	$\eta^{-1}(\alpha_{40})$
4	T1	$\eta^{-1}(\alpha_{11})$	$\eta^{-1}(\alpha_{10}+\gamma_{12\|D=0}z_{2i})$
	T2	$\eta^{-1}(\alpha_{21})$	$\eta^{-1}(\alpha_{20}+\gamma_{12\|D=0}z_{2i})$
	T3	$\eta^{-1}(\alpha_{31}+\gamma_{34\|D=1}z_{1i})$	$\eta^{-1}(\alpha_{30})$
	T4	$\eta^{-1}(\alpha_{41}+\gamma_{34\|D=1}z_{1i})$	$\eta^{-1}(\alpha_{40})$
5	T1	$\eta^{-1}(\alpha_{11}+\gamma_{12\|D=1}z_{1i})$	$\eta^{-1}(\alpha_{10})$
	T2	$\eta^{-1}(\alpha_{21}+\gamma_{12\|D=1}z_{1i})$	$\eta^{-1}(\alpha_{20})$
	T3	$\eta^{-1}(\alpha_{31}+\gamma_{34\|D=1}z_{2i})$	$\eta^{-1}(\alpha_{30})$
	T4	$\eta^{-1}(\alpha_{41}+\gamma_{34\|D=1}z_{2i})$	$\eta^{-1}(\alpha_{40})$
6	T1	$\eta^{-1}(\alpha_{11}+\gamma_{12\|D=1}z_i)$	$\eta^{-1}(\alpha_{10})$
	T2	$\eta^{-1}(\alpha_{21}+\gamma_{12\|D=1}z_i)$	$\eta^{-1}(\alpha_{20})$
	T3	$\eta^{-1}(\alpha_{31}-\gamma_{34\|D=1}z_i)$	$\eta^{-1}(\alpha_{30})$
	T4	$\eta^{-1}(\alpha_{41}-\gamma_{34\|D=1}z_i)$	$\eta^{-1}(\alpha_{40})$

With z_i, z_{1i}, z_{2i} being the Gaussian random effects with mean 0 and standard deviation 1 and $\eta^{-1}(y)=1/(1+e^{-y})$.

DOI: 10.1002/sim

ACKNOWLEDGEMENTS

The authors would like to thank Abraham Aseffa, Sayda El-Safi, Asrat Hailu, Jane Mbui, Maowia Mukhtar, Suman Rijal, Shyam Sundar, Monique Wasunna, and Rosanna Peeling for study conduct and data collection and Dirk Berkvens and Niko Speybroek for their helpful comments. The third author acknowledges the partial support from the Interuniversity Attraction Poles Programs P5/24 and P6/03—Belgian State—Federal Office for Scientific Technical and Cultural Affairs.

REFERENCES

1. Zijlstra EE, Ali MS, el Hassan AM, el Toum IA, Satti M, Ghalib HW, Kager PA. Kala-azar: a comparative study of parasitological methods and the direct agglutination test in diagnosis. *Transactions of the Royal Society of Tropical Medicine and Hygiene* 1992; **86**(5):505–507.
2. Chappuis F, Rijal S, Soto A, Menten J, Boelaert M. A meta-analysis of the diagnostic performance of the direct agglutination test and rk39 dipstick for visceral leishmaniasis. *British Medical Journal* 2006; **333**(7571):723–7261. DOI: 10.1136/bmj.38917.503056.7C.
3. Albert PS, Dodd LE. A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics* 2004; **60**(2):427–435. DOI: 10.1111/j.0006-341X.2004.00187.x.
4. Dendukuri N, Joseph L. Bayesian approaches to modeling the conditional dependence between multiple tests. *Biometrics* 2001; **57**:158–167. DOI: 10.1111/j.0006-341X.2001.00158.x.
5. Boelaert M, El Safi S, Hailu A, Mukhtar M, Rijal S, Sundar S, Wasunna M, Aseffa A, Mbui J, Menten J, Desjeux P, Peeling RW. Diagnostic tests for kala-azar management at primary care level: a head-on comparison of the freeze-dried DAT, rk39 strip test, and KAtex in a multi-centre study in East-Africa and the Indian subcontinent. *Transactions of the Royal Society of Tropical Medicine and Hygiene* 2008; **102**(1):32–40. DOI: 10.1016/j.trstmh.2007.09.003.
6. Qu Y, Tan M, Kutner MH. Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics* 1996; **52**:797–810. DOI: 10.2307/2533043.
7. Qu Y, Hadgu A. A model for evaluating sensitivity and specificity for correlated diagnostic tests in efficacy studies with an imperfect reference test. *Journal of the American Statistical Association* 1998; **93**:920–928.
8. McCutcheon AL. *Latent Class Analysis*. Quantitative Applications in the Social Sciences Series, vol. 64. Sage Publications: Thousand Oaks, CA, 1987.
9. Garrett ES, Eaton WW, Zeger SL. Methods for evaluating the performance of diagnostic tests in the absence of a gold standard: a latent class model approach. *Statistics in Medicine* 2002; **21**:1289–1307. DOI: 10.1002/sim.1105.
10. Berkvens D, Speybroeck N, Praet N, Adel A, Lesaffre E. Estimating disease prevalence in a bayesian framework using probabilistic constraints. *Epidemiology* 2006; **17**(2):145–153. DOI: 10.1097/01.ede.0000198422.64801.8d.
11. Hadgu A, Qu Y. A biomedical application of latent class models with random effects. *Journal of the Royal Statistical Society Series C—Applied Statistics* 1998; **47**:603–616. DOI: 10.1111/1467-9876.00131.
12. Goetghebeur E, Liinev J, Boelaert M, der Stuyft PV. Diagnostic test analyses in search of their gold standard: latent class analyses with random effects. *Statistical Methods in Medical Research* 2000; **9**:231–248. DOI: 10.1177/096228020000900304.
13. Albert PS, McShane LM, Shih JH, Network TUCIBT. Latent class modeling approaches for assessing diagnostic error without a gold standard: with applications to p53 immunohistochemical assay in bladder tumors. *Biometrics* 2001; **57**(2):610–619. DOI: 10.1111/j.0006-341X.2001.00610.x.
14. Goodman LA. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* 1974; **61**(2):215–231. DOI: 10.2307/2334349.
15. Formann AK. Measurement errors in caries diagnosis: some further latent class models. *Biometrics* 1994; **50**(3):865–875. DOI: 10.2307/2532801.
16. Boelaert M, Aoun K, Liinev J, Goetghebeur E, Stuyft PVD. The potential of latent class analysis in diagnostic test validation for canine leishmania infantum infection. *Epidemiology and Infection* 1999; **123**(3):499–506. DOI: 10.1017/S0950268899003040.
17. Yang I, Becker MP. Latent variable modeling of diagnostic accuracy. *Biometrics* 1997; **53**:948–958. DOI: 10.2307/2533555.
18. Garrett ES, Zeger SL. Latent class model diagnosis. *Biometrics* 2000; **56**(4):1055–1067. DOI: 10.1111/j.0006-341X.2000.01055.x.

19. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis* (2nd edn). Chapman & Hall, CRC: Boca Raton, FL, U.S.A., 2004.
20. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit (with Discussion and Rejoinder). *Journal of the Royal Statistical Society, Series B* 2002; **64**:583–639. DOI: 10.1111/1467-9868.00353.
21. Deniau M, Canavate C, Faraut-Gambarelli F, Marty P. The biological diagnosis of leishmaniasis in HIV-infected patients. *Annals of Tropical Medicine and Parasitology* 2003; **97**(Suppl. 1):115–133. DOI: 10.1179/000349803225002598.
22. Black MA, Craig BA. Estimating disease prevalence in the absence of a gold standard. *Statistics in Medicine* 2002; **21**(18):2653–2669. DOI: 10.1002/sim.1178.
23. Agresti A, Hitchcock DB. Bayesian inference for categorical data analysis: a survey. *Technical Report*, University of Florida, FL, U.S.A., 2005.
24. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press: Oxford, U.K., 2003.
25. Alonzo TA, Pepe MS. Using a combination of reference tests to assess the accuracy of a new diagnostic test. *Statistics in Medicine* 1999; **18**(22):2987–3003.
26. Pepe MS. Evaluating technologies for classification and prediction in medicine. *Statistics in Medicine* 2005; **24**(24):3687–3696. DOI: 10.1002/sim.2431.
27. Knottnerus JA, van Weel C, Muris JWM. Evidence base of clinical diagnosis—evaluation of diagnostic procedures. *British Medical Journal* 2002; **324**(7335):477–480. DOI: 10.1136/bmj.324.7335.477.
28. Molenberghs G, Kenward MG. *Missing Data in Clinical Studies*. Wiley: Hoboken, NJ, U.S.A., 2007.
29. Skrondal A, Rabe-Hesketh AR. *Generalized Latent Variable Modeling—Multilevel, Longitudinal and Structural Equation Models*. Chapman & Hall, CRC: Boca Raton, FL, U.S.A., 2004.
30. Molenberghs G, Goetghebeur EJT, Lipsitz SR, Kenward MG. Nonrandom missingness in categorical data: strengths and limitations. *American Statistician* 1999; **53**(2):110–118.
31. Molenberghs G, Kenward MG, Goetghebeur E. Sensitivity analysis for incomplete contingency tables: the Slovenian plebiscite case. *Journal of the Royal Statistical Society Series C—Applied Statistics* 2001; **50**:15–29. DOI: 10.1111/1467-9876.00217.