Estimating Disease Prevalence in a Bayesian Framework Using Probabilistic Constraints

Dirk Berkvens, Niko Speybroeck,* Nicolas Praet,* Amel Adel,† and Emmanuel Lesaffre‡*

Abstract: Studies sometimes estimate the prevalence of a disease from the results of one or more diagnostic tests that are applied to individuals of unknown disease status. This approach invariably means that, in the absence of a gold standard and without external constraints, more parameters must be estimated than the data permit. Two assumptions are regularly made in the literature, namely that the test characteristics (sensitivity and specificity) are constant over populations and the tests are conditionally independent given the true disease status. These assumptions have been criticized recently as being unrealistic. Nevertheless, to estimate the prevalence, some restrictions on the parameter estimates need to be imposed. We consider 2 types of restrictions: deterministic and probabilistic restrictions, the latter arising in a Bayesian framework when expert knowledge is available. Furthermore, we consider 2 possible parameterizations allowing incorporation of these restrictions. The first is an extension of the approach of Gardner et al and Dendukuri and Joseph to more than 2 diagnostic tests and assuming conditional dependence. We argue that this system of equations is difficult to combine with expert opinions. The second approach, based on conditional probabilities, looks more promising, and we develop this approach in a Bayesian context. To evaluate the combination of data with the (deterministic and probabilistic) constraints, we apply the recently developed Deviance Information Criterion and effective number of parameters estimated (p_D) together with an appropriate Bayesian *P* value. We illustrate our approach using data collected in a study on the prevalence of porcine cysticercosis with verification from external data.

(*Epidemiology* 2006;17: 145–153)

Diagnostic tests form an essential part of all disciplines of epidemiology, providing an estimate of the true prevalence of the disease, infection, or condition.

Submitted 27 June 2003; accepted 27 September 2005.

From the *Institute of Tropical Medicine, Department of Animal Health, Antwerp, Belgium; †Ecole Nationale Vétérinaire, Département Clinique, El Harach, Alger, Algeria; and ‡Biostatistical Centre, Catholic University of Leuven, Leuven, Belgium.

Supplemental material for this article is available with the online version of the journal at www.epidem.com; click on "Article Plus."

Correspondence: Dirk Berkvens, Institute of Tropical Medicine, Department of Animal Health, Nationalestraat 155, B-2000 Antwerp, Belgium. E-mail: dberkvens@itg.be.

ISSN: 1044-3983/06/1702-0145
DOI: 10.1097/01.ede.0000198422.64801.8d

Suppose that D^+ (D^-) indicates that a subject is diseased (disease-free) and T^+ (T^-) indicates a positive (negative) result on a diagnostic test T. In the presence of a gold standard, the number of diseased subjects (n_{D^+}) and disease-free subjects (n_{D^-}) are known (Table 1). A gold standard can be a diagnostic test with both test sensitivity and test specificity equal to one, or (for example) an experiment in which a proportion of the subjects are artificially infected. The columns "diseased" and "disease-free" in Table 1 represent this situation and constitute the so-called full table, ie, the table in which the distinction between the 2 infection status categories can be made.

From Table 1, sensitivity (Se) and specificity (Sp) of the test are estimable by $n_{T^+|D^+}/n_{D^+}$ and $n_{T^-|D^-}/n_{D^-}$, respectively. On the other hand, in a field observation only the probability of a positive test result can be directly estimated, ie, $P(T^+) = n_{T^+}/n$ (the apparent prevalence). The column "Total" in Table 1 is actually the marginal or collapsed table over the diseased and disease-free subjects and represents this situation. When Se and Sp are known, the true prevalence $P(D^+)$ can be estimated using the following expression[1]:

$$P(D^+) = \frac{P(T^+) + Sp - 1}{Se + Sp - 1} \quad (1)$$

Unfortunately, Se and Sp are rarely known exactly and must be estimated from data. Hence, we need to take into account the sampling variability with which the prevalence is estimated, which could be done using the approach of Rogan and Gladen.[1]

Some traditional textbooks on diagnostic testing still refer to the test sensitivity and specificity as values that are intrinsic to the diagnostic test, ie, constant and universally applicable.[2,3] Our own experience (and that of others) indicates that both test sensitivity and specificity vary with external factors.[4–8] Consequently, test sensitivity and specificity, as traditionally defined, are purely theoretical concepts determined in the population used to validate the test. Therefore, when using a diagnostic test in the population of interest, the characteristics of that population must be used to get an improved estimate of Se and Sp.[7] Observe, however, that assumptions of constancy of Se and Sp over different populations is still being made.[9]

For a long time, it was assumed that 2 (or more) diagnostic tests are conditionally independent on the disease status,[10–12] for example, $P(T_1^+ \cap T_2^+ \,|\, D^+) = P(T_1^+ \,|\, D^+)\, P(T_2^+ \,|\, D^+)$. When the 2 diagnostic tests have a similar biologic basis, as is often the case, the conditional independence assumption is

TABLE 1. Two-by-Two Contingency Table When Testing n Subjects for Disease D With One Diagnostic Test T

	Diseased	Disease-Free	Total
+ Test result	$n_{T^+\|D^+}$	$n_{T^+\|D^+}$	n_{T^+}
− Test result	$n_{T^-\|D^+}$	$n_{T^-\|D^-}$	n_{T^-}
Total	n_{D^+}	n_{D^-}	n

D^+ (D^-) indicates that the subject is (is not) diseased; T^+ (T^-), a positive (negative) result with test T.

untenable. Toft et al[13] review the possible pitfalls when using the Hui-Walter paradigm in real life, particularly the problems encountered when trying to stratify the population into 2 or more subpopulations with different true prevalence but constant test characteristics.

When these 2 simplifying assumptions cannot be made, estimation of the true prevalence either becomes impossible or requires extra information added to the estimation process. Indeed, when h tests are applied to each individual, $2^{h+1} - 1$ parameters must be estimated. These parameters are the true prevalence (one parameter), the test sensitivities (h parameters), the test specificities (h parameters), and

$$2\sum_{i=2}^{h}\binom{h}{i} = 2(2^h - h - 1) = 2^{h+1} - 2h - 2$$

parameters describing the dependence of the h tests given the true disease status of the subject. Yet only $2^h - 1$ parameters can be estimated, because only data from the collapsed table (over disease status) are available. Consequently, the true prevalence of the disease cannot be estimated if no constraints are imposed on the parameters. The most popular constraint has been to assume conditional independence.

Table 2 shows the maximum number of parameters that can be estimated and the number of parameters that need to be estimated as a function of the number of diagnostic tests, as well as the number of parameters to be estimated given conditional independence of the tests.

TABLE 2. Maximum Number of Estimable Parameters and Number of Parameters to Be Estimated in the Absence of Conditional Independence and Under Conditional Independence as a Function of the Number of Tests per Subject

Number of Tests	Maximum Number of Estimable Parameters	Parameters to be Estimated Under Conditional Dependence	Parameters to Be Estimated Under Conditional Independence
1	1	3	3
2	3	7	5
3	7	15	7
4	15	31	9
5	31	63	11
h	$2^h - 1$	$2^{h+1} - 1$	$2h + 1$

In particular, Table 2 indicates that, under conditional independence, parameters can be estimated for $h \geq 3$, whereas for $h \geq 4$, the number of estimable parameters actually exceeds the number of parameters to estimate.

Estimating the true prevalence thus becomes a matter of adding constraints on the parameters. These constraints must come from external sources, eg, previous similar studies, expert opinion, and so on. Hence, the estimated true prevalence and test characteristics will be the result of a combination of the data (test results) and the external information on these test characteristics, which is the best that can be obtained. Consequently, several authors have suggested a Bayesian approach to incorporate this external information by specifying prior distributions on the parameters obtained from eliciting the opinion of experts.[14,15] Most often, prior knowledge on sensitivity and specificity is incorporated. Unfortunately, in practice, experts often do not have and cannot have (see, for example, nonconstant sensitivity and specificity) a clearcut opinion on these test characteristics. As a result, the experts' opinions will often be in conflict with the actually observed data. Of course, the Bayesian framework allows more diffuse prior distributions, but this will, in our context, often render the parameters inestimable. In this article, we show that, if possible, prior information on conditional probabilities is easier to specify.

To verify whether the prior information is in conflict with the test results, the recently developed deviance information criterion (DIC)[16] and an appropriate Bayesian *P* value can be used.[17] To quantify the impact of the constraints, the effective number of estimated parameters (p_D) of the model can be calculated.[16]

In the next section, we discuss 2 parameterizations to model conditional dependence. We then distinguish between deterministic and probabilistic constraints and show that the number of parameters effectively estimated (p_D) can be used to quantify the effect of these constraints on the number of effectively estimated parameters. In the next section, we indicate that DIC and an appropriate Bayesian *P* value can pinpoint a conflict between the prior information and the test results. We then examine the behavior of DIC, p_D, and the Bayesian *P* value using a theoretical dataset. Finally, we apply one of the models developed here to field data. A discussion of our approach and the results follows in the last section.

Markov Chain Monte Carlo (MCMC) estimations were carried out in WinBUGS 1.4.[18] Additional calculations were performed in R[19] making extensive use of the "bugs" function[17] posted on the web.[20] The software developed for the evaluation of DIC, p_D, and the Bayesian *P* value can be downloaded.[21]

MODELING CONDITIONAL DEPENDENCE BETWEEN TESTS THROUGH CONDITIONAL PROBABILITIES

For the situations in which 2 diagnostic tests are applied to all subjects, Gardner et al[22] and Dendukuri and Joseph[23]

calculated the probabilities of the different outcomes as a function of test sensitivities, test specificities, and covariances. Furthermore, these authors suggest combining prior information on these parameters with the test results in a Bayesian manner. Their results can be expanded to more than 2 tests.[24] However, the prior distributions for the covariances (ie, generalized beta distributions) are quite difficult to elicit from experts, because they cannot be related to real-life situations. Although not well recognized in the literature, this is equally true for the sensitivity parameters, the reason being that the sensitivity of a diagnostic test needs to be determined in experimental conditions (and hence also quite distinct from real-life settings) on a small number of subjects. In contrast, the specificity of a test can be determined somewhat more easily in a population that is known to be disease-free.

Eliciting information from experts on the conditional performance of one test given the results of another test could be much easier in certain cases. For instance, a question such as "What is the probability that a subject tests positively in test 2 given that the subject is diseased and has tested positively in test 1?" relates the characteristics of 2 tests applied on the same subject. This can be easier to answer because the experts usually have one or more so-called reference tests (very often with a very high specificity) and know the performance of other tests in relation to the reference test in the infected and uninfected subpopulations.

Model (2), given by

$$P(T_1^{i_1} \cap \cdots \cap T_h^{i_h}) = P(D^+)\prod_{t=1}^{h}[(1-i_t) - (-1)^{i_t}P\left(T_t^+ \mid D^+ \bigcap_{t' \mid t>1}^{t-1} T_{t'}^{i_{t'}}\right)] + [1-P(D^+)]\prod_{t=1}^{h}[i_t + (-1)^{i_t}P\left(T_t^- \mid D^- \bigcap_{t' \mid t>1}^{t-1} T_{t'}^{i_{t'}}\right)], \quad (2)$$

expresses the cell probabilities of the collapsed $2^{(h+1)}$ table (hence of a 2^h table) in terms of the prevalence of the disease, the sensitivity and specificity of the first test, and conditional probabilities. In Appendix A1.1 (available with the online version of this article), the different conditional probabilities are listed in a hierarchical fashion: parameters θ_1–θ_3 are used when only a single test is applied, θ_1–θ_7 are used for 2 tests, θ_1–θ_{15} for 3 tests, and θ_1–θ_{31} for 4 tests. In Appendix A1.2 (available with the online version of this article), expressions are given to calculate the prevalence and the test characteristics from the parameters defined in A1.1. Finally, in Appendix A1.3 (available with the online version of this article), the equations are given to calculate the cell probabilities of the different test result combinations when h = 4. When fewer than 4 tests are used, the probabilities can be extracted from these equations by dropping excess terms, eg, $P(111) = \theta_1\,\theta_2\,\theta_4\,\theta_8 + (1-\theta_1)(1-\theta_3)(1-\theta_7)(1-\theta_{15})$.

DETERMINISTIC VERSUS PROBABILISTIC CONSTRAINTS AND THE USE OF p_D

Constraints on the parameters need to be imposed to estimate the prevalence and the test characteristics using equation 2. We classify these constraints into 2 types: deterministic and probabilistic. Setting Se (or Sp) to a particular value is an example of a deterministic constraint, as is the assumption of conditional independence. Specifying a prior distribution for a parameter or for a function of parameters (like a contrast) is an example of a probabilistic constraint in a Bayesian setting.

In a frequentist context, *m* independent deterministic constraints reduce the number of parameters to estimate exactly by *m*. For instance, when using 2 tests (h = 2), the assumption of conditional independence between the tests reduces the number of parameters to be estimated from 7 to 5 (see Table 2). When fixing the specificity of one test to say one, the number of parameters to estimate is further reduced by one. In a Bayesian context, things are more difficult because it is not immediately clear what impact a probabilistic constraint has on the number of parameters to estimate. In this context, Spiegelhalter et al[16] proposed to measure the effective number of estimated parameters in a fitted statistical model by p_D. This measure is not an integer any more, even for a deterministic constraint, because it is calculated as the difference of the posterior mean of the deviance and the deviance evaluated in the posterior mean. More details are given in the next section.

MEASURING THE DISCORDANCE OF THE PRIOR INFORMATION WITH THE OBSERVED TEST RESULTS

As described in the introduction, experts have difficulty expressing their prior knowledge in quantitative terms (sensitivity and specificity). Our experience shows that often the prior information is in conflict with the actual observed data. In the context of diagnostic testing, this is evidently a crucial handicap. Several authors have addressed this problem in the statistical literature,[25] but it is not immediately clear how the proposed measures for discordance can be implemented in our context. Here, 2 measures are proposed. The first one is based on a Bayesian goodness-of-fit test leading to a Bayesian *P* value. The second one uses the recently introduced deviance information criterion (DIC).[16] Both measures are reviewed here in the context of analyzing collapsed tables of diagnostic test data in a Bayesian manner. Although not absolutely necessary, we assume that Bayesian estimation is done through MCMC sampling and reference is made to the WinBUGS software. A detailed account of the computation of both measures is given in Appendix 2 (available with the online version of this article).

BEHAVIOR OF DEVIANCE INFORMATION CRITERION, p_D, AND BAYESIAN *P* VALUE

Deviance Information Criterion and p_D

In this section, we discuss the performance of DIC and p_D in the context of a possibly overspecified multinomial

model. That is, we look at the behavior of DIC and p_D when $q > (k - 1)$ and we focus on model (2). When $q \leq (k - 1)$, we expect $p_D \approx k - 1 - q$. Unfortunately, this will not necessarily be the case for model (2) because this model is not log-concave in its parameters. Things become worse when $q > (k - 1)$ because then the log-likelihood must be flat around the maximum likelihood estimate if no constraints have been imposed. However, if the multinomial model is parameterized in its multinomial probabilities, ie, in π_i ($i = 1, \ldots, k - 1$), then for all cases, the log-likelihood will be concave in its parameters. Consequently, we suggest evaluating DIC and p_D always in the posterior mean of π_i ($i = 1, \ldots, k - 1$). However, there is one remaining problem, namely that p_D (if based on the multinomial probabilities) is always smaller than $k - 1$ regardless of whether the model has been overspecified. To have an idea of when the model has been overspecified, we suggest calculating p_D also using the posterior means of its parameters, ie, for model (2) on the posterior means of the parameters θ_1 to θ_{31} for $h = 4$. Empiric evidence shows that without sufficient constraints in that case, p_D is negative, resulting in a diagnostic tool that can indicate whether all our parameters are estimable.

To exemplify our reasoning in the previous paragraph, we take the case of $h = 1$, which is when there is only one diagnostic test and the multinomial model contains only 2 cells, ie, $\pi_1 = P(T_1^+)$ and $\pi_2 = P(T_1^-)$. In this case, $\pi_1 = \theta_1\theta_2 + (1 - \theta_1)(1 - \theta_3)$ and $\pi_2 = \theta_1(1 - \theta_2) + (1 - \theta_1)\theta_3$. The log-likelihood is not concave in θ_1, θ_2, and θ_3 but clearly it is in π_1 (we can neglect π_2 because it is $1 - \pi_1$). Without any constraints on θ_1, θ_2, and θ_3, the multinomial parameter π_1 will vary freely, thus $p_D \approx 1$ if based on the posterior mean of π_1. However, experience showed that p_D becomes negative when based on θ_1, θ_2, and θ_3. When putting constraints on θ_1, θ_2, and θ_3, nothing will change if these constraints do not put a constraint on the multinomial parameter π_1, and so p_D will stay around 1. Only when the constraints on θ_1, θ_2, and θ_3 affect the mobility of the multinomial parameter, p_D (based on π_1) will shrink. On the other hand, p_D based on θ_1, θ_2, and θ_3 will be negative if the constraints were not sufficient to constraint π_1. A comparison of the 2 p_D-values will immediately reveal whether the parameters θ_1, θ_2, and θ_3 are estimable.

From a practical point of view, we can conclude in general:

- DIC and p_D should be evaluated in the posterior mean of the multinomial probabilities and in the posterior mean of the parameters of the model. In WinBUGS language, the latter are called "parent nodes." Thus, we need 2 evaluations of DIC and p_D, one within WinBUGS and one outside WinBUGS;
- Only when the 2 p_D-values are smaller or equal to $2^h - 1$ is there hope that the prevalence of the disease can be estimated; and
- Models with a high value for DIC indicate a bad model in a Bayesian sense, meaning that either the model (likelihood) part is badly specified or the prior distributions are not compatible with the data. Consequently, when comparing different prior knowledge combined with the same likelihood, prior knowledge that is in conflict with the observed data is reflected in a high value for DIC.

Bayesian *P* Value

When the model has been overspecified, the Bayesian *P* value (as defined in our approach) will be around 0.50. The reason for this is that the posterior probability for the multinomial probabilities will be flat. However, as shown in Appendix 3 (available with the online version of this article), this test quantity is a useful indicator for the actual model fit because the Bayesian *P* value tends to zero if there is a good model fit and to one if the fit is poor.

Modeling Exercise

We now examine the behavior of DIC, p_D, and the Bayesian *P* value using theoretical frequencies.

The prevalence of the disease is taken equal to 0.5. Furthermore, we assume 2 diagnostic tests T_1 and T_2 ($h = 2$), both with specificity equal to 1, ie, with no false-positive results. The sensitivity of T_1 equals 0.60 and the sensitivity of test T_2 equals 0.70, but there is conditional dependence, ie, in terms of the parameters in Appendix 1, θ_4 and θ_5 are not equal. In summary: $\theta_1 = 0.50$, $\theta_2 = 0.60$, $\theta_3 = 1$, $\theta_4 = 0.90$, $\theta_5 = 0.40$, $\theta_6 = 1$, and $\theta_7 = 1$. This yields the following theoretical probabilities for the 2^2 collapsed contingency table: P(00) = 0.62, P(01) = 0.08, P(10) = 0.03, and P(11) = 0.27. For a study of $N = 1000$, the expected cell frequencies are therefore $r_1 = 620$, $r_2 = 80$, $r_3 = 30$, and $r_4 = 270$ and the expected number of diseased subjects is equal to $N_{D+} = 500$. We test the following models on this dataset:

- M1: no prior constraints;
- M2: specificity of $T_1 = 1$, specificity of $T_2 = 1$;
- M3: specificity of $T_1 = 1$, specificity of $T_2 = 1$, sensitivity of T_1 constrained uniformly to interval [0.5, 0.7] and the sensitivity of T_2 constrained by a uniform prior on θ_4 to interval [0.8, 1] and a uniform prior on θ_5 to interval [0.3, 0.5];
- M4: specificity of $T_1 = 1$, specificity of $T_2 = 1$, the sensitivity of T_1 severely constrained uniformly to interval [0.5999, 0.6001] and the sensitivity of T_2 severely constrained by a uniform prior on θ_4 to interval [0.8999, 0.9001] and a uniform prior on θ_5 to interval [0.3999, 0.4001];
- M5: constraints on specificity and sensitivity of T_1 and T_2 = 1 as in M4. Additionally, the prevalence is severely constrained by a uniform prior on θ_1 to interval [0.4999, 0.5001];
- M6: specificity of $T_1 = 1$, specificity of $T_2 = 1$, sensitivity of T_1 wrongly constrained by a uniform prior to interval [0.8, 1]; and
- M7: specificity of $T_1 = 1$, specificity of $T_2 = 1$, the sensitivity on T_1 wrongly constrained by a uniform prior to interval [0.8, 1] and a wrongly positive conditional sensitivity of T_2 by a uniform prior to interval [0.2, 0.4].

In the next section, these models are applied to the 2^2 contingency table of the expected frequencies. This exercise further exemplifies our reasoning in previous sections.

RESULTS AND DISCUSSION

The results of applying models M1 to M7 are summarized in Table 3. Note that DIC and p_D calculated from the

TABLE 3. Results of the Different Models Using the Theoretical Data Presented in the Text

Model	Bayesian *P* Value	Parent Nodes		Multinomial		Prev	Test 1		Test 2	
		DIC	p_D	DIC	p_D		Se	Sp	Se	Sp
1	0.4916	−90.177	−111.609	24.283	2.936	0.5253	0.3229	1	0.3957	1
2	0.4930	8.303	−13.065	24.342	2.952	0.5688	0.5732	1	0.6680	1
3	0.4793	24.176	2.873	24.279	2.917	0.5058	0.5956	1	0.6939	1
4	0.1852	20.407	0.990	20.471	1.021	0.5006	0.6000	1	0.7000	1
5	0.0004	18.426	0.000	18.445	0.007	0.5000	0.6000	1	0.7000	1
6	0.7000	25.524	2.351	25.468	2.338	0.3850	0.8106	1	0.9049	1
7	1.0000	355.977	1.406	355.899	1.382	0.3848	0.8103	1	0.5006	1

For each model, the posterior mean of the parameters are given. The column "Parent Nodes" indicates that the calculations were done within WinBUGS and are based on the parameters θ_1 to θ_7 in Appendix 1. The column "Multinomial" indicates that the calculations were done outside WinBUGS and are based on the multinomial probabilities.

multinomial probabilities for models M1, M2, and M3 differ only by random MCMC sampling variation.

In models M1 and M2, the constraints are not sufficient to estimate the parameters θ_1 to θ_7 of Appendix 1. This is reflected by negative p_D-values estimated from the parent nodes. Observe that p_D as calculated from the multinomial probabilities is practically equal to 3, the true value. Furthermore, for both models, the Bayesian *P* value is about 0.5, indicating no particular problem. Clearly, the prevalence of the disease is overestimated for both models. The constraint imposed on model M3 brings the parent-node p_D close to 3, indicating that now all parameters are estimable. The prevalence is well estimated now, and the estimated sensitivities are close to their true values. In models M4 and M5, the constraints are made more stringent, but in the correct manner. Model M5 has the lowest DIC value of the 2, with the lowest p_D-value almost equal to zero. This implies that parameters are set to their correct values. Indeed, the Bayesian *P* value indicates a nearly perfect but nonstochastic model. Furthermore, the prevalence and the sensitivities are basically equal to their true values. In models M6 and M7, enough constraints have been put on the parameters, because for each model, the 2 corresponding p_D-values are almost equal to each other. However, the Bayesian *P* values indicate badly fitted models, which is also reflected in a badly estimated prevalence and sensitivities. (Of course, this would not be recognized in practice by the user.)

APPLICATION OF MODEL (2) TO FIELD DATA

The Problem and the Data

Porcine cysticercosis is a major problem in many countries, causing a debilitating and potentially lethal zoonosis.[26,27] Relatively accurate estimates of prevalence of cysticercae in fattening pigs are essential to appraise the risk for human infection. Several diagnostic tests are used, but none is a gold standard and exact information about test sensitivity and specificity is unavailable. A total of 868 traditionally kept pigs, offered for sale on a market near Lusaka (Zambia), were tested with the following 4 diagnostic tests: palpation of the tongue (TONG), visual inspection of the carcass (VISUAL), an antigen enzyme-linked immunosorbent assay (Ag-ELISA), and an antibody enzyme-linked immunosorbent assay (Ab-ELISA). A summary of the results is shown in Table 4.[28]

The data in Table 4 were used to estimate the prevalence and the test characteristics under equation (2) and assuming a variety of expert opinions.

Prior Information

"Expert" opinion in the broadest possible sense was used to specify prior information on the diagnostic test characteristics. In this section, we call a model the combination of equation (2) with a particular set of deterministic and probabilistic (prior information) constraints. Some of the models were constructed from general principles only. For

TABLE 4. Test Results of 868 Traditional Zambian Pigs Subjected to 4 Diagnostic Tests

TONG	VISUAL	Ag-ELISA	Ab-ELISA	Number of Pigs
0	0	0	0	326
0	0	0	1	42
0	0	1	0	281
0	0	1	1	95
0	1	0	0	0
0	1	0	1	0
0	1	1	0	5
0	1	1	1	4
1	0	0	0	1
1	0	0	1	0
1	0	1	0	2
1	0	1	1	0
1	1	0	0	2
1	1	0	1	1
1	1	1	0	35
1	1	1	1	74

0 indicates negative test result; 1, positive test result; TONG, tongue palpation; VISUAL, visual carcass inspection; Ag-ELISA, antigen ELISA; Ab-ELISA, antibody ELISA.

instance, in model M1, the "expert" opinion states that both test sensitivity and specificity can take any value between zero and one and that the 4 tests are mutually conditionally independent. For the other models, proper expert opinion was used. This expert opinion was obtained from helminthologists at the Institute of Tropical Medicine (Antwerp) and at Ghent University. They provided upper and lower limits for the various test sensitivity and specificity values. From biologic principles, they also concluded that the tests TONG and VISUAL are not independent in a truly infected population. A positive test result for TONG is nearly always accompanied by a positive result for VISUAL, whereas a negative TONG test nearly invariably means a negative VISUAL test.

The prior distributions for sensitivity and specificity are taken here as uniform distributions (beta[1, 1] truncated on the interval $[a, b]$, with a being the under limit and b the upper limit as specified by the experts). These uniform distributions can be replaced by beta distributions (beta[α, β], where α and β are determined such that, say, 95% of the probability mass is located in $[a, b]$).

TABLE 5. Parameters to Be Estimated in the 7 Models That Were Constructed From the Available 'Expert' Opinion

	M1*	M2†	M3‡	M4	M5	M6	M7
θ_1	0–1	0–1	0–1	0–1	0–1	0–1	0–1
θ_2	0–1	0–1	0.3–0.7	0–1	0–1	0–1	0–1
θ_3	0–1	1	1	0–1	1	1	1
θ_4	0–1	0–1	0.8–1	0–1	0–1	0–1	0.9–1
θ_5	$= \theta_4$	$= \theta_4$	$= \theta_4$	0–1	0–1	0–1	0–0.1
θ_6	0–1	1	1	0–1	1	1	1
θ_7	$= \theta_6$	—	—	0–1	—	—	—
θ_8	0–1	0–1	0–1	0–1	0–1	0–1	0–1
θ_9	$= \theta_8$	$= \theta_8$	$= \theta_8$	0–1	0–1	0–1	0–1
θ_{10}	$= \theta_8$	$= \theta_8$	$= \theta_8$	0–1	0–1	0–1	0–1
θ_{11}	$= \theta_8$	$= \theta_8$	$= \theta_8$	0–1	0–1	0–1	0–1
θ_{12}	0–1	0–1	0.95–1	0–1	0–1	0.9–1	0.9–1
θ_{13}	$= \theta_{12}$	—	—	0–1	—	—	—
θ_{14}	$= \theta_{12}$	—	—	0–1	—	—	—
θ_{15}	$= \theta_{12}$	—	—	0–1	—	—	—
θ_{16}	0–1	0–1	0.92–1	0–1	0–1	0–1	0–1
θ_{17}	$= \theta_{16}$	$= \theta_{16}$	$= \theta_{16}$	0–1	0–1	0–1	0–1
θ_{18}	$= \theta_{16}$	$= \theta_{16}$	$= \theta_{16}$	0–1	0–1	0–1	0–1
θ_{19}	$= \theta_{16}$	$= \theta_{16}$	$= \theta_{16}$	0–1	0–1	0–1	0–1
θ_{20}	$= \theta_{16}$	$= \theta_{16}$	$= \theta_{16}$	0–1	0–1	0–1	0–1
θ_{21}	$= \theta_{16}$	$= \theta_{16}$	$= \theta_{16}$	0–1	0–1	0–1	0–1
θ_{22}	$= \theta_{16}$	$= \theta_{16}$	$= \theta_{16}$	0–1	0–1	0–1	0–1
θ_{23}	$= \theta_{16}$	$= \theta_{16}$	$= \theta_{16}$	0–1	0–1	0–1	0–1
θ_{24}	0–1	0–1	0.98–1	0–1	0–1	0.9–1	0.9–1
θ_{25}	$= \theta_{24}$	—	—	0–1	0–1	0–1	0–1
θ_{26}	$= \theta_{24}$	—	—	0–1	—	—	—
θ_{27}	$= \theta_{24}$	—	—	0–1	—	—	—
θ_{28}	$= \theta_{24}$	—	—	0–1	—	—	—
θ_{29}	$= \theta_{24}$	—	—	0–1	—	—	—
θ_{30}	$= \theta_{24}$	—	—	0–1	—	—	—
θ_{31}	$= \theta_{24}$	—	—	0–1	—	—	—

*Equivalent to	TONG	VISUAL	Ag-ELISA	Ab-ELISA
Sensitivity	0–1	0–1	0–1	0–1
Specificity	0–1	0–1	0–1	0–1
†Equivalent to	TONG	VISUAL	Ag-ELISA	Ab-ELISA
Sensitivity	0–1	0–1	0–1	0–1
Specificity	1	1	0–1	0–1
‡Equivalent to	TONG	VISUAL	Ag-ELISA	Ab-ELISA
Sensitivity	0.3–0.7	0.8–1	0–1	0.92–1
Specificity	11	1	0.95–1	0.98–1

$\theta_1 \ldots \theta_{31}$, see Appendix 1 for the parameter definition; a-b denotes that a is the lower limit and b is the upper limit of the parameter interval; = [a], value equal to parameter a in brackets; —, not to be estimated.

TABLE 6. Deviance Information Criterion (DIC), Effective Number of Parameters Estimated (p_D), and Bayesian *P* Value (P) for the Models That Converged

Model	DIC	p_D	*P*
M2	97.1	6.52	1.00
M3	925.1	2.89	1.00
M7	70.3	9.86	0.48

Models

Table 5 lists the parameters to be estimated in each of the 7 models (M1 to M7) using all 4 tests that were constructed using the available "expert" opinion together with the limits that were applied to each parameter. The starting model (M1) assumes conditional independence of the 4 tests and no prior information on any of the diagnostic test characteristics (ie, test sensitivities and specificities have uniform prior distributions on ([0, 1]). The model M2 still assumes conditional independence and fixes the specificity of TONG test and the VISUAL test to one, but no other probability constraints were added. The deterministic constraints on model M1 imply that there we are estimating 9 parameters when 15 can be estimated. For model M2, we are estimating 7 parameters with again 15 estimable parameters.

Model M3 again assumes conditional independence, but now probabilistic constraints (inspired by the experts' opinions) apply. At face value, there are still 7 parameters to be estimated, but the probability constraints imply probabilistic relationships among the parameters and hence fewer parameters need to be estimated. The actual number of parameters estimated in the model should be reflected in the value of p_D.

The remaining models all considered conditional dependence. When no constraints are applied, 31 parameters need to be estimated, whereas only 15 parameters are estimable in the collapsed table (see model M4). Putting the TONG specificity and the VISUAL specificity both to one (model M5) reduces the number of parameters to be estimated to 19: conditional probabilities θ_3 and θ_6 become one and all parameters, appearing behind $(1-\theta_3)$ and $(1-\theta_6)$, no longer need to be estimated (ie, θ_7, θ_{13}, θ_{14}, θ_{15}, θ_{26}, θ_{27}, θ_{28}, θ_{29}, θ_{30}, θ_{31}).

The number of parameters to be estimated was further reduced by constraining both θ_{12} and θ_{24} to $[0.9-1]$ (model M6), constraints that are moderate by most standards (a specificity equal to 0.90 is considered a low specificity). Finally, the conditional probabilities θ_4 and θ_5 were constrained to, respectively $[0.9-1]$ and $[0-0.1]$ (model M7). The constraints applied in models M6 and M7 are of probabilistic nature and hence imply that the actual number of parameters to be estimated lies below 19. Model M6 has between 17 and 19 parameters to be estimated. Conditional probabilities θ_4 and θ_5, which are constrained in model M7, reflect the expert opinion that the visual carcass inspection result is highly associated with the result of the tongue palpation. If the 2 tests are made identical ($\theta_4 = 1$ and $\theta_5 = 0$), the minimum number of parameters to be estimated becomes 6 (assuming 3 independent tests with specificity of one test equal to one) and the actual number of parameters to be estimated lies between 6 and 19. The listing for model M7 can be downloaded.[21]

RESULTS

As we expected, not all models converged. Table 6 shows the value of DIC, p_D, and the Bayesian *P* value for each converged model. Table 7 shows the posterior means together with the 95% credibility intervals of the prevalence and the test characteristics of the 4 tests.

Model M1 did not converge in WinBUGS, which is not surprising given that symmetry yields several possible solutions depending on the starting conditions: replacing sensitivity by the complement of specificity, specificity by the complement of sensitivity, and prevalence by its own complement yields a symmetric solution (and there is thus an inherent problem of identifiability). Indeed, constraining the prevalence to either $[0-0.5]$ or $[0.5-1]$ results in convergence and estimates for all parameters (DIC = 63.3, $p_D = 0.3$). Model M2 converged and yielded estimates for all parameters. The expert opinion used in model M3 did not improve the model fit. On the contrary, DIC increased from 97 to 945 and the Bayesian *P* value stayed at 1.0. The Bayesian *P* values for models M2 and M3 near 1.0 suggest a lack-of-fit, indicating that conditional independence test does not hold. Models M4, M5, and M6 did not converge, probably because they were overparameterized, which implies that the constraints were not strict enough to yield identifiable models. Model M7 converged and yielded the minimum DIC and an acceptable Bayesian *P* value of 0.48 (the Bayesian *P* value tended to zero when strict constraints were applied).

Table 6 shows the effective number of parameters estimated. For model M7, $p_D = 9.86$. This illustrates that

TABLE 7. Posterior Mean for the Prevalence and the Test Characteristics Together With the 95% Credibility Interval (in parentheses) for the 3 Models That Converged

		TONG		VISUAL		Ag-ELISA		Ab-ELISA	
Model	Prev	Se	Sp	Se	Sp	Se	Sp	Se	Sp
M2	0.144 (0.12–0.17)	0.918 (0.86–0.96)	1.000	0.965 (0.93–0.99)	1.000	0.961 (0.92–0.99)	0.495 (0.46–0.53)	0.635 (0.55–0.72)	0.815 (0.79–0.84)
M3	0.246 (0.22–0.28)	0.540 (0.47–0.61)	1.000	0.803 (0.80–0.81)	1.000	0.973 (0.95–0.99)	0.900 (0.90–0.901)	0.903 (0.90–0.91)	0.952 (0.95–0.96)
M7	0.642 (0.54–0.91)	0.210 (0.14–0.26)	1.000	0.221 (0.15–0.27)	1.000	0.867 (0.62–0.98)	0.947 (0.90–0.997)	0.358 (0.26–0.41)	0.917 (0.85–0.99)

TABLE 8. Test Characteristic Estimates in a Group of 65 Pigs Dissected Experimentally After Slaughter

	TONG		VISUAL		Ag-ELISA		Ab-ELISA	
	No.*	Se or Sp (95% CI)	No.*	Se or Sp (95% CI)	No.*	Se or Sp (95% CI)	No.*	Se or Sp (95% CI)
Sensitivity	05/31	0.16 (0.05–0.34)	12/31	0.39 (0.22–0.58)	20/31	0.65 (0.45–0.81)	14/31	0.45 (0.27–0.64)
Specificity	34/34	1.00 (0.91–1.00)	34/34	1.00 (0.91–1.00)	31/34	0.91 (0.76–0.98)	30/34	0.88 (0.73–0.97)

*For sensitivity, number of animals tested positive/number of infected animals; for specificity, number of animals tested negative/number of disease-free animals.
CI indicates confidence interval.

the 6 constraints (deterministic and probabilistic) on the 20 parameters to estimate have more effect than one might initially think. Indeed, model M7 is based on model (2), which is parameterized in a hierarchical manner with conditional probabilities. Constraints on lower-order conditional probabilities must have an effect on higher-order conditional probabilities.

Taking into account conditional dependence between the various diagnostic tests considerably reduces the estimated test sensitivity of both tongue palpation and visual carcass inspection and, most importantly, results in a much higher estimate of the true prevalence (Table 7).

External Model Validation

Additional data became available later, allowing external validation of the selected model. Namely, an additional 65 pigs were subjected to the 4 tests and completely dissected out on slaughter (gold standard), permitting the ascertainment of the true infection status and thus allowing estimation of the true prevalence as well as the test characteristics. The true prevalence was estimated as 0.48 (31/65) and the estimates of the test characteristics are shown in Table 8.

Clearly, model M7 (Table 7) resulted in parameter estimates that are reasonably close to those obtained from the experimental dissections (Table 8).

DISCUSSION

Analysis of data generated by the application of one or more diagnostic tests in a specified population invariably entails "overfitting" of the data. The number of parameters that have to be estimated always exceeds the number that can be estimated. This can be resolved only by simplifying the model (deterministic constraints) or through the inclusion of expert opinion (probabilistic constraints). In the latter case, only a Bayesian approach can incorporate that information. Observe that the Bayesian approach is slowly becoming accepted by the medical community. Indeed, everyday practice is a reflection of the Bayesian philosophy. When a test is used within a certain population, it is implicitly assumed that the values of sensitivity and specificity, as supplied by the manufacturer of the test kit, apply to the population studied; this prior knowledge of the test characteristics is given so much credence that the test results are no longer needed to estimate Se and Sp, allowing estimation of the true prevalence.

The model developed on the basis of conditional probabilities allows formalization of this expert opinion, whatever form it might take. Anything from genuine information acquired through high-quality data to a personal opinion can be quantified and fed as a prior belief probability distribution into the model. Whether it is easy to specify a prior opinion on a conditional probability will depend on the actual tests involved, but we argue that it is practically impossible to give reliable prior information on the sensitivity of a diagnostic test. The user can monitor the effect of this prior belief on the results, and it may be easier for the user to appreciate the fact that the actual interpretation of the test results is conditional on the prior opinion. The effect of imposing deterministic or probabilistic constraints is reflected in the value of p_D and can thus be evaluated.

Our approach is in sharp contrast to the approach of Pouillot et al[29] in which conditional independence is accepted when a specific test shows no indication against this assumption. However, not much is known about the power of this test. Instead, we suggest working under the assumption of conditional dependence and applying a sensitivity analysis on the estimation of the prevalence and the test characteristics by varying the prior distributions.

The results of the different scenarios applied to the present example clearly show that the estimate of the infection prevalence depends on the model chosen, and that widely varying estimates can be obtained. It is important that users understand this and realize that the expert opinion has a great impact on the final estimation of the prevalence. However, as the simulation and the real-life study show, DIC, p_D, and the Bayesian *P* value are useful in the process of selecting a model. We must, however, warn the user that the information in the collapsed table over the disease groups contains inherently little information on the prevalence and the test characteristics. Finally, the present example shows that "classic" testing with one or more tests, assuming constancy of test parameters and independence of tests, may grossly underestimate true prevalence and thus, in our case, the seriousness of the zoonosis.

REFERENCES

1. Rogan WJ, Gladen B. Estimating prevalence from the results of a screening test. *Am J Epidemiol*. 1978;107:71–76.
2. Thrusfield D. *Veterinary Epidemiology*. Oxford University Press; 1995.
3. Toma B, Bénet J-J, Dufour B, et al. *Glossaire d'Épidémiologie Animale*. Éditions du Point Vétérinaire; 1991.
4. Billiouw M, Vercruysse J, Marcotty T, et al. *Theileria parva* epidemics: a case study in eastern Zambia. *Vet Parasitol*. 2002;107:51–63.
5. Billiouw M, Brandt J, Vercruysse J, et al. Evaluation of the indirect fluorescent antibody test as a diagnostic tool for East Coast fever in eastern Zambia. *Vet Parasitol*. 2005;127:189–198.

6. Saegerman C, De Waele L, Gilson D, et al. Evaluation of three serum i-ELISAs using monoclonal antibodies and protein G as peroxidase conjugate for the diagnosis of bovine brucellosis. *Vet Microbiol*. 2004; 100:91–105.
7. Greiner M, Gardner IA. Epidemiological issues in the validation of veterinary diagnostic tests. *Prev Vet Med*. 2000;45:3–22.
8. Begg CB. Biases in the assessment of diagnostic tests. *Stat Med*. 1987;6:411–423.
9. Georgiadis M, Johnson W, Gardner I, et al. Correlation-adjusted estimation of sensitivity and specificity of two diagnostic tests. *Appl Stat*. 2003;52:63–76.
10. Hui SL, Walter SD. Estimating the error rates of diagnostic tests. *Biometrics*. 1980;36:167–171.
11. Joseph L, Gyorkos TW, Coupal L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *Am J Epidemiol*. 1995;141:263–273.
12. Johnson WO, Gastwirth JL, Pearson LM. Screening without a 'gold standard': the Hui-Walter paradigm revisited. *Am J Epidemiol*. 2000; 153:921–924.
13. Toft N, Jørgensen E, Højsgaard S. Diagnosing diagnostic tests: evaluating the assumptions underlying the estimation of sensitivity and specificity in the absence of a gold standard. *Prev Vet Med*. 2005;68: 19–33.
14. Enoe C, Georgiadis MP, Johnson WO. Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true diseases state is unknown. *Prev Vet Med*. 2000;45:61–81.
15. Branscum AJ, Gardner IA, Johnson WO. Estimation of diagnostic-test sensitivity and specificity through Bayesian modelling. *Prev Vet Med*. 2005;68:145–163.
16. Spiegelhalter DJ, Best NG, Carlin BP, et al. Bayesian measures of model complexity and fit (with discussion). *J R Stat Soc* [*Ser B*]. 2002;64:583–640.
17. Gelman A, Carlin JB, Stern HS, et al. *Bayesian Data Analysis*, 2nd ed. Chapman & Hall/CRC; 2003.
18. Spiegelhalter DJ, Thomas A, Best NG, et al. WinBUGS Version 1.4 User Manual. MRC Biostatistics Unit; 2003. Available at: http://www.mrc-bsu.cam.ac.uk/bugs/.
19. The R Project for Statistical Computing. Available at: http://www.r-project.org/.
20. bugs.R: functions for running WinBugs from R. Available at: http://www.stat.columbia.edu/~gelman/bugsR/.
21. R and WinBUGS program listings. Available at: http://www.itg.be/itg/uploads/animalhealth/diagtest.pdf.
22. Gardner IA, Stryhn H, Lind P, et al. Conditional dependence between tests affects the diagnosis and surveillance of animal diseases. *Prev Vet Med*. 2000;45:107–122.
23. Dendukuri N, Joseph L. Bayesian approaches to modelling the conditional dependence between multiple diagnostic tests. *Biometrics*. 2001; 57:208–217.
24. Adel A, Berkvens D. Modelling conditional dependence between multiple diagnostic tests, using co-variances between test results. Internal Technical Report, ITM; April 16, 2002. Available at: http://www.itg.be/itg/uploads/animalhealth/internal_document.pdf.
25. Young K, Pettit L. Measuring discordancy between prior and data. *J R Stat Soc* [*Ser B*]. 1996;58:679–689.
26. Garcia HH, Del Brutto OH. *Taenia solium* cysticercosis. *Infect Dis Clin North Am*. 2000;14:97–119.
27. Phiri IK, Dorny P, Gabriel S, et al. The prevalence of porcine cysticercosis in eastern and southern provinces of Zambia. *Vet Parasitol*. 2002;108:31–39.
28. Dorny P, Phiri IK, Vercruysse J, et al. A Bayesian approach for estimating values for prevalence and diagnostic test characteristics of porcine cysticercosis. *Int J Parasitol*. 2004;34:569–576.
29. Pouillot R, Gerbier G, Gardner IA. 'TAGS,' a program for the evaluation of test accuracy in the absence of a gold standard. *Prev Vet Med*. 2002;53:67–81.