AN APPROXIMATE METHOD TO ASSESS A BINARY DIAGNOSTIC TEST WITHOUT THE GOLD STANDARD IN THE PRESENCE OF IGNORABLE MISSING DATA

José Antonio Roldán Nofuentes[1], Pedro Femia Marzo & Juan de Dios Luna del Castillo
Unit of Biostatistics, Department of Statistics, School of Medicine, University of Granada, 18071, Spain

RESUMEN

La evaluación de un nuevo test diagnóstico binario requiere conocer el verdadero estado de la enfermedad de todos lo pacientes en la muestra mediante la aplicación de un gold estándar. En la práctica clínica con frecuencia no se dispone de un gold estándar y es necesario utilizar un test de referencia para evaluar el nuevo test diagnóstico. Asimismo, con frecuencia el test de referencia no se aplica a todos los pacientes de la muestra. En este artículo se propone un algoritmo basado en el algoritmo *EM* para imputar los estimadores máximo verosímiles de la sensibilidad y especificidad de un nuevo test diagnóstico binario con respecto a un test de referencia con exactitud conocida cuando no todos los pacientes son verificados con el test de referencia. Se aplica el algoritmo *SEM* para estimar los errores estándares de los estimadores. Los resultados se han aplicado al diagnóstico del *VIH*.

PALABRAS CLAVE: Algoritmos EM y SEM, Test de Referencia, Sensibilidad, Especificidad.

ABSTRACT

The assessment of a new binary diagnostic test has traditionally required knowledge of the disease status in all of the patients in the sample via the application of a gold standard. In clinical practice, there is often no gold standard to diagnose the disease and it is necessary to use a reference test to evaluate a new diagnostic test. Furthermore, the reference test is often not applied to all of the patients in the sample. In this article, an algorithm is proposed based on the *EM* algorithm to impute the maximum likelihood estimators of sensitivity and specificity of a new binary diagnostic test in relation to a reference test with known accuracy when not all of the patients are verified with the reference test. The *SEM* algorithm is applied in order to estimate the standard errors of the estimators. The results have been applied to the diagnosis of *HIV*.

KEY WORDS: *EM* and *SEM* Algorithms, Reference Test, Sensitivity, Specificity.

1. INTRODUCTION

A diagnostic method is a test which is applied to a patient in order to obtain a provisional diagnosis of the presence or absence of a given disease. When the result of a diagnostic test is binary, the accuracy of the test is measured in terms of its sensitivity and specificity. Sensitivity is the probability of a diseased patient giving a positive test result and specificity is the probability of a non-diseased patient giving a negative test result. Traditionally, the evaluation of a new binary diagnostic test requires knowledge of the real state of the disease (whether present or absent) in each patient through the application of a gold standard. In practice, the gold standard is not always applied to all of the subjects in the sample, which leads to the so-called problem of partial verification of the disease [1, 2].

In some clinical situations, there is not always a gold standard to evaluate the disease status of a patient and, therefore, it is not possible to estimate the sensitivity and specificity of the new test through the classic method. In these situations, a reference test is used to evaluate the accuracy of the new diagnostic test which is to be studied. Several authors have studied the estimation of the accuracy of a diagnostic test in relation to a reference test by supposing that the two tests are conditionally independent on the disease [3, 4, 5]. The assumption of conditional independence between the two diagnostic tests does not seem reasonable in many applications [6, 7]. Others authors [8, 9, 10, 11] have studied the evaluation of a diagnostic test in relation to a reference test when the two diagnostic tests are conditionally dependent on the disease.

Moreover, if the reference test is costly or risky for the patient, it is not applied to all of the patients in the sample, causing a problem quite similar to that of partial disease verification.

The objective of this study is to estimate the sensitivity and the specificity of a new binary diagnostic test in relation to a reference test with known sensitivity and specificity when this test is not applied is not applied to all of the patients in the sample. In Section 2, an algorithm is proposed based on the *EM* algorithm in order to impute the values of the maximum likelihood estimators of sensitivity and specificity of a new diagnostic test. In Section 3, the values of the standard errors of the estimators obtained in Section 2 are imputed through the application of the *SEM* algorithm [12]. In Section 4, simulation experiments are carried out in order

[1] jaroldan@ugr.es

to study the asymptotic coverage of the confidence intervals of sensitivity and specificity. In Section 5, the results obtained are applied to an example, and in Section 6 we discuss our findings.

2. MAXIMUM LIKELIHOOD ESTIMATORS

Let binary diagnostic test 1 be the new test and let diagnostic test 2 be the reference test. The random variable T models the result of test 1, such that $T=1$ when it is positive, indicating the presence of the disease, and $T=0$ when it is negative, indicating the absence of the disease; in a similar way, the random variable R models the result of the reference test, $R=1$ when it is positive and $R=0$ when it is negative; the variable V models the process of verification, $V=1$ when the patient is verified though the reference test and $V=0$ when the patient is not verified; and, lastly, the variable D models the true disease status in each patient, such that $D=1$ when the patient is diseased and $D=0$ when the patient is non-diseased.

If the diagnostic test is applied to all of the subjects in a random sample of n size and the reference test is only applied to a part of the sample, Table 1 is obtained. Let $\pi_1 = P(T=1|D=1)$ and $\nu_1 = P(T=0|D=0)$ be the sensitivity and the specificity of the new test, $\pi_2 = P(R=1|D=1)$ and $\nu_2 = P(R=0|D=0)$ the sensitivity and the specificity of the reference test (we assume that $0<\pi_2<1$ and $0<\nu_2<1$), $p = P\ \ D=1$ the disease prevalence and

$$\lambda_{ijk} = P(V=1|T=i,R=j,D=k), i,j,k=0,1$$

the probability of selecting a subject with the results $T=i$, $R=j$ and $D=k$ to apply the reference test to him or her.

		$T=1$	$T=0$
$V=1$			
	$R=1$	s_{11}	s_{10}
	$R=0$	s_{01}	s_{00}
$V=0$		u_1	u_0
	Total	n_1	n_0

Table 1. Cross-classification of test results.

In general, it can be assumed that the two diagnostic tests are conditionally dependent on the disease [6], that is to say

$$P(T=i,R=j|D=k) = P(T=i|D=k)P(R=j|D=k) + \delta_{ij}\varepsilon_k \tag{1}$$

where $\delta_{ij}=1$ when $i=j$ and $\delta_{ij}=-1$ when $i \neq j$, and ε_k represents the conditional dependence between the two diagnostic tests $(\varepsilon_k > 0)$: ε_1 is the covariance when $D=1$ and ε_0 is the covariance when $D=0$. It is verified [6] that $\varepsilon_k \leq \vartheta_1(1-\vartheta_2)$ when $\vartheta_2 > \vartheta_1$ and $\varepsilon_k \leq \vartheta_2(1-\vartheta_1)$ when $\vartheta_1 > \vartheta_2$, when ϑ is the sensitivity or the specificity. If $\varepsilon_1 = \varepsilon_0 = 0$, expression (1) is equivalent to supposing that the two diagnostic tests are conditionally independent on the disease.

If the missing data mechanism is ignorable, which implies that the verification process is missing at random (MAR), it holds that

$$P(V|T,R,D) = P(V|T), \tag{2}$$

i.e. the verification process with the reference test only depends on the result of the new test and not on the reference test nor on the disease status. Under assumption MAR this problem of estimation can be solved. Under assumption (2), the data from Table 1 are obtained from a multinomial distribution with the probabilities given in Table 2 when

$$\lambda_1 = P(V=1|T=1) = \textstyle\sum_{j,k=0}^{1} \lambda_{1jk} \text{ and } \lambda_0 = P(V=1|T=0) = \sum_{j,k=0}^{1} \lambda_{0jk} \tag{3}$$

	$T=1$	$T=0$
$V=1$		
$R=1$	$1-p\ \left[\ 1-\nu_1\ \ 1-\nu_2\ \ +\varepsilon_0\right]\lambda_1$ $+$ $p\ \ \pi_1\pi_2+\varepsilon_1\ \ \lambda_1$	$1-p\ \left[\nu_1\ 1-\nu_2\ \ -\varepsilon_0\right]\lambda_0$ $+$ $p\left[\ 1-\pi_1\ \ \pi_2-\varepsilon_1\right]\lambda_0$
$R=0$	$1-p\ \left[\ 1-\nu_1\ \ \nu_2-\varepsilon_0\right]\lambda_1$ $+$ $p\left[\pi_1\ 1-\pi_2\ \ -\varepsilon_0\right]\lambda_1$	$1-p\ \ \ \nu_1\nu_2+\varepsilon_0\ \ \lambda_0$ $+$ $p\left[\ 1-\pi_1\ \ 1-\pi_2\ \ +\varepsilon_1\right]\lambda_0$
$V=0$	$\left[\ 1-p\ \ \ 1-\nu_1\ \ +p\pi_1\right]\ 1-\lambda_1$	$\left[\ 1-p\ \ \nu_1+p\ \ 1-\pi_1\ \ \right]\ 1-\lambda_0$

Table 2. Probabilities of the multinomial distribution.

The logarithm of likelihood function of the data in Table 1 is

$$\begin{aligned} l \propto & (s_{11}+s_{01}) log\lambda_1 + (s_{10}+s_{00}) log\lambda_0 + s_{11} log[(1-p)\{(1-v_1)(1-v_2)+\varepsilon_0\}+p\{\pi_1\pi_2+\varepsilon_1\}] \\ & + s_{10} log[(1-p)\{v_1(1-v_2)-\varepsilon_0\}+p\{(1-\pi_1)\pi_2-\varepsilon_1\}] \\ & + s_{01} log[(1-p)\{(1-v_1)v_2-\varepsilon_0\}+p\{\pi_1(1-\pi_2)-\varepsilon_1\}] \\ & + s_{00} log[(1-p)\{v_1v_2+\varepsilon_0\}+p\{(1-\pi_1)(1-\pi_2)+\varepsilon_1\}] \\ & + u_1 log[(1-p)(1-v_1)+p\pi_1]+u_0 log[(1-p)v_1+p(1-\pi_1)]+u_1\log(1-\lambda_1) \\ & + u_0\log(1-\lambda_0) \end{aligned} \tag{4}$$

If the sensitivity and the specificity of the reference test $(\pi_2 \text{ and } v_2)$ and the covariances ε_0 and ε_1 are known, the maximum likelihood estimators (*MLEs*) of the sensitivity and the specificity of the new test, of the disease prevalence and of λ_1 and λ_0 are obtained maximizing function (4). Thus, the *MLEs* of λ_1 and λ_0 are

$$\hat{\lambda}_1=\frac{s_{11}+s_{01}}{n_1} \text{ and } \hat{\lambda}_0=\frac{s_{10}+s_{00}}{n_0}. \tag{5}$$

If the parameters π_2, ν_2, ε_0 and ε_1 are known, the values of the *MLEs* of π_1, ν_1 and p are obtained numerically solving a system of grade 5 nonlinear equations which does not depend on parameters λ_1 and λ_0, or by using the *EM* algorithm.

2.1. EM Algorithm

The *EM* algorithm [13] is a technique which permits the determination of the *MLEs* of parametric models when not all of the data is observed. The *EM* algorithm involves two stages: Step *E* and Step *M*. If there is a model for all of the data of Y with a density function $f(Y|\theta)$ when $\theta=(\theta_1,\ldots,\theta_d)$ is a vector of unknown parameters, the complete information Y can be written as $Y=(Y_{obs},Y_{mis})$, where Y_{obs} represents the observed part of Y and Y_{mis} is the missing part. The *EM* algorithm imputes the value of θ, $\hat{\theta}$, which maximizes $f(Y_{obs}|\theta)$, that is to say, the *MLE* of θ based on the observed data Y_{obs}. The algorithm starts from an initial value $\theta^{(0)}$, if $\theta^{(t)}$ is the estimator of θ in the t-th iteration, the iteration (*t+1*) of the *EM* algorithm is as follows:

Step *E*. To obtain the expectation of the logarithm of the likelihood function of the complete data if θ is $\theta^{(t)}$:

$$Q(\theta|\theta^{(t)}) = \int logf(Y|\theta)f(Y_{mis}|Y_{obs}, \theta = \theta^{(t)})dY_{mis} \quad . \tag{6}$$

Step *M*. To determine $\theta^{(t+1)}$*:* maximizing the expectation of the logarithm of the likelihood function:

$$Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta \quad |\theta^{(t)}), \forall\theta \qquad .(7)$$

The *EM* algorithm implicitly defines an application $\theta \rightarrow M(\theta)$such that

$$\theta^{(t+1)} = M(\theta^{(t)}),\ t = 0,1,\ldots \tag{8}$$

If $\theta^{(t)}$ converges to $\hat{\theta}$ and $M(\theta)$ is continuous, then $\hat{\theta}$ verifies that

$$\hat{\theta} = M(\hat{\theta}). \tag{9}$$

Developing $M(\theta^{(t)})$ in Taylor series around $\hat{\theta}$ it holds that

$$\theta^{(t+1)} - \hat{\theta} \approx (\theta^{(t)} - \hat{\theta})DM, \tag{10}$$

when

$$DM = \left(\frac{\partial M_j(\theta)}{\partial \theta_i}\right)_{\theta=\hat{\theta}} \tag{11}$$

is the Jacobian matrix of $M(\theta)=(M_1(\theta),\ldots,M_d(\theta))$ evaluated in $\theta = \hat{\theta}$. This matrix *DM* has a special relevance in the imputation of the asymptotic variance-covariance matrix of the $MLE\ \hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_d)MLE$ of $\theta = (\theta_1, \ldots, \theta_d)$.

2.2. Application of the EM Algorithm

In the situation which is analyzed here, the missing information is the disease status of each patient (*D*). This information is reconstructed in Step *E* of the algorithm and in Step *M* the *MLEs* are imputed from the data reconstructed in the previous step. The following proposition describes the *EM* algorithm.

Proposition 1. If π_2, ν_2, ε_1 and ε_0 are known, let $\pi_1^{(t)}, v_1^{(t)}$ and $p^{(t)}$ be current values of parameter π_1, ν_1 and *p* after *t* iterations in the *EM* algorithm. The next iteration estimates are given by

$$\pi_1^{(t+1)} = \frac{\sum_{j=0}^{1} \alpha_{1j1}^{(t)} + \beta_{11}^{(t)}}{\sum_{j,k=0}^{1} \alpha_{1jk}^{(t)} + \sum_{j=0}^{1} \quad \beta_{1j}^{(t)}}, v_1^{(t+1)} = \frac{\sum_{j=0}^{1} \alpha_{0j0}^{(t)} + \beta_{00}^{(t)}}{\sum_{j,k=0}^{1} \alpha_{0jk}^{(t)} + \sum_{j=0}^{1} \quad \beta_{0j}^{(t)}} \text{ and } p^{(t+1)} = \frac{\sum_{j=0}^{1} \alpha_{1jk}^{(t)} + \sum_{j=0}^{1} \quad \beta_{1j}^{(t)}}{n} \quad , \tag{12}$$

where

$\alpha_{1jk}^{(t)} = s_{jk}P(D = 1|T = j, V = 1, R = k, \pi_1 = \pi_1^{(t)}, v_1 = v_1^{(t)}p = p^{(t)}, \hat{\lambda}_1, \hat{\lambda}_0),$
$\alpha_{0jk}^{(t)} = s_{jk}P(D = 0|T = j, V = 1, R = k, \pi_1 = \pi_1^{(t)}, v_1 = v_1^{(t)}p = p^{(t)}, \hat{\lambda}_1, \hat{\lambda}_0) = s_{jk} - \alpha_{1jk}^{(t)},$
$\beta_{1j}^{(t)} = u_jP(D = 1|T = j, V = 0, \pi_1 = \pi_1^{(t)}, v_1 = v_1^{(t)}p = p^{(t)}, \hat{\lambda}_1, \hat{\lambda}_0),$
and

$\beta_{0j}^{(t)} = u_jP(D = 0|T = j, V = 0, \pi_1 = \pi_1^{(t)}, v_1 = v_1^{(t)}p = p^{(t)}, \hat{\lambda}_1, \hat{\lambda}_0) = u_j - \beta_{1j}^{(t)}$

We iterate this process until the estimates converge. The convergent values $\hat{\pi}_1$, $\hat{\nu}_1$ and $\hat{p}$ are the *MLEs* for π_1, ν_1 and *p*. For a proof, see Appendix A.

2.3. A new algorithm

The application of the algorithm described in the Section 2.2 requires knowledge of the sensitivity and the specificity of the reference test (π_2 and ν_2) and the covariances ε_0 and ε_1. If the accuracy of the reference test is known but not the covariances, the proposed *EM* algorithm cannot be applied. However, the covariances are limited [6], $\varepsilon_k \in (0, \vartheta_1(1-\vartheta_2))\ if\ \vartheta_1 < \vartheta_2\ or\ \varepsilon_k \in (0, \vartheta_2(1-\vartheta_1))\ if\ \vartheta_2 < \vartheta_1$, where ϑ is the sensitivity or the specificity, and therefore this information can be used to approximate the values of the covariances. In order to solve this problem, we propose the following algorithm based on the *EM* algorithm:

Step 0: $\pi_1^{(0)}, \nu_1^{(0)}, p^{(0)}, \varepsilon_1^{(0)} = \frac{\pi_1^{(0)}(1-\pi_2)}{2}\ and\ \varepsilon_0^{(0)} = \frac{\nu_1^{(0)}(1-\nu_2)}{2}$.

Repeat Steps 1 and 2 until the *EM* algorithm converges.

Step 1: Apply an iteration of the *EM* algorithm and obtain $\pi_1^{(t+1)}, \nu_1^{(t+1)}$ and $p^{(t+1)}$.

Step 2: Calculate $\varepsilon_1^{(t+1)} = \frac{\pi_1^{(t+1)}(1-\pi_2)}{2}\ if\ \pi_1^{(t+1)} < \pi_2$ or $\varepsilon_1^{(t+1)} = \frac{\pi_2\left(1-\pi_1^{(t+1)}\right)}{2}\ if\ \pi_2 < \pi_1^{(t+1)}$ and $\varepsilon_0^{(t+1)} = \frac{\nu_1^{(t+1)}(1-\nu_2)}{2}\ if\ \nu_1^{(t+1)} < \nu_2$ or $\varepsilon_0^{(t+1)} = \frac{\nu_2\left(1-\nu_1^{(t+1)}\right)}{2}\ if\ \nu_2 < \nu_1^{(t+1)}$.

Therefore, this algorithm is based on the approximation of each ε_k covariance to the average point of the interval *(0, ϑ_i(1-ϑ_j)* and the application of the *EM* algorithm. The convergence of the algorithm is given by the convergence of the *EM* algorithm. The initial values of sensitivity, specificity and prevalence can be any values between 0 and 1, and the method always converges to the same solution as has been observed in the simulation experiments described in Section 5. Likewise, we have verified that in order that the method converges, it is necessary that Youden's index [14] of the reference test is bigger than zero *(π_2+ν_2-1>0)*. With respect to the stop criteria, it is recommended to use the very small values, for example 10^{-20} or 10^{-24}. Higher values do not affect the solutions, although they can affect the estimation of the variance and covariance matrix. With respect to accuracy of reference test, we assume that $0 < \pi_2 < 1$ and $0 < \nu_2 < 1$, else the function (23) cannot be evaluated. If $\pi_2 = 1$ and $\nu_2 = 1$, then the reference test is a gold standard, and this problem is the verification bias problem [1, 2].

After obtaining the values of the *MLE*s of sensitivity and specificity of the diagnostic tests under evaluation and of the disease prevalence, it is necessary to obtain the corresponding standard errors.

3. VARIANCE-COVARIANCE MATRIX

The estimation of the asymptotic variance-covariance matrix of $\hat{\pi}_1$, $\hat{\nu}_1$ and $\hat{p}$ can be obtained through the application of the *SEM* algorithm (*Supplemented EM*) [12].

3.1. Missing Information Principle and SEM Algorithm

The Missing Information Principle establishes that the information observed is equal to the complete information minus the missing information, which in terms of Fisher information functions is expressed as

$$I_0\left(\hat{\theta}|Y_{obs}\right) = I_{oc} - I_{mis}, \tag{13}$$

when

$$I_{mis} = E\left[-\frac{\partial^2 \log f\,(Y_{mis}|Y_{obs},\theta)}{\partial\theta^2}|Y_{obs},\theta\right]_{\theta=\hat{\theta}} \quad , \tag{14}$$

where Y_{mis} is the missing information, Y_{obs} the observed information, θ el parameter vector and $\hat{\theta}$ the *MLE*. Equation (13) can be rewritten as

$$I_0\left(\hat{\theta}|Y_{obs}\right) = (I \quad - I_{mis} I_{oc}^{-1})I_{00} \tag{15}$$

Dempster et al [13] demonstrated that

$$I_{mis} I_{oc}^{-1} = DM , \tag{16}$$

when *DM* is the defined matrix in (11). Substituting expression (16) for (15), the asymptotic covariance matrix, obtained as the inverse matrix of (15), is

$$\Sigma = I_{oc}^{-1}(I - DM)^{-1} = I_{oc}^{-1} + I_{oc}^{-1}DM(I - DM)^{-1} = I_{oc}^{-1} + \Delta\Sigma \qquad .(17)$$

The *SEM* algorithm developed by Meng and Rubin [12], is a numerical procedure based on the *EM* algorithm to approximate the variance-covariance matrix of the estimator $\hat{\theta}$ of a parameter vector θ. The *SEM* algorithm consists of three parts: (1) the evaluation of matrix I_{oc}^{-1}, (2) the evaluation of matrix *DM*, and (3) the evaluation of matrix $\sum$ (see Meng and Rubin [13]). The main characteristic of the *SEM* algorithm consists of the imputation of the elements (r_{ij}) of matrix DM. The process is as follows.

Let $\theta = (\theta_1, \dots, \theta_d)$ be a parameter vector, $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_d)$ its *MLE* obtained through the application of the *EM* algorithm or some other procedure and r_{ij} the (*i,j*) th element of the matrix DM. Let $\theta^{(t)}(i)$ be defined as $\theta^{(t)}(i) = (\hat{\theta}_1, \dots, \hat{\theta}_{i-1}\hat{\theta}_i^{(t)}\hat{\theta}_{i+1}, \hat{\theta}_d)$, that is to say, only the *i*th component in $\theta^{(t)}(i)$ is active in the sense that the rest of the components are fixed at their *MLE*s. By definition

$$r_{ij} = \frac{\partial M_j(\hat{\theta})}{\partial \theta_i} = \lim_{\theta_i \to \hat{\theta}_i} \frac{M_j(\hat{\theta}_1, \dots, \hat{\theta}_{i-1}, \theta_i, \hat{\theta}_{i+1}, \hat{\theta}_d) - M_j(\hat{\theta})}{\theta_i - \hat{\theta}_i} = \lim_{t \to \infty} \frac{M_j(\theta^{(t)}(i)) - \hat{\theta}_j}{\theta_i - \hat{\theta}_i} = \lim_{t \to \infty} r_{ij}^{(t)} \quad (18)$$

As $M(\theta)$is implicitly defined by the results of the *EM* algorithm, all of the r_{ij} values can be calculated using the *EM* algorithm. The steps for computing $r_{ij}^{(t)}$ are:

```
INPUT: $\hat{\theta}$ and $\theta^{(t)}$.
Step 1: Obtain $\theta^{(t+1)}$ by executing the EM algorithm.
Repeat Steps 2 and 3 for $i = 1, \dots, d$.
    Step 2: Calculate $\theta^{(t)}(i)$ and consider it to be the estimator of $\theta$, execute an iteration of the EM algorithm to obtain $\tilde{\theta}^{(t+1)}(i)$.
    Step 3: Calculate the ratios
        $r_{ij}^{(t)} = \frac{\theta_j^{(J+1)}(i) - \hat{\theta}_j}{\theta_t^{(t)} - \hat{\theta}_i}, j = 1, \dots, d.$   (19)
OUTPUT: $\tilde{\theta}^{(t+1)}$y $r_{ij}^{(t)}, i, j = 1, \dots, d$.
```

The value of each r_{ij} is obtained when the sequence $r_{ij}^{(t^*)}, r_{ij}^{(t^*+1)}$,. is stabilized for some value of t^*. In this process, the value t^* can be different for each r_{ij} element of the DM matrix. In practice, the stop criteria for the calculation of r_{ij} is that $\left| r_{ij}^{(t^*+1)} - r_{ij}^{(t^*)} \right|$ should be smaller than the square root of the stop criteria used in the application of the *EM* algorithm.

3.2. Application of the SEM Algorithm

The application of the *SEM* algorithm requires firstly the evaluation of the matrix I_{oc}^{-1}. This matrix is the inverse matrix to the Fisher information matrix of the complete data and is calculated from the last table obtained in the application of the algorithm described in Section 2.3 (algorithm 2.3). The elements of this matrix are shown in Appendix B.

Having obtained the matrix I_{oc}^{-1} and the *MLE*s $\hat{\pi}_1$, $\hat{v}_1$ and $\hat{p}$, and taking as values of the covariances ε_1 and ε_0 the last values obtained by the algorithm 2.3, the second part of the *SEM* algorithm consists of the evaluation of the matrix *DM*. The elements r_{ij}, $i, j = 1, 2, 3$, of the matrix *DM* are obtained by the iterative application of the following steps of the algorithm:

```
INPUT: $\hat{\theta} = (\hat{\pi}_1, \hat{v}_1, \hat{p})$ and $\theta^{(t)} = (\pi_1^{(t)}, v_1^{(t)}, p^{(t)})$.
    Step 1: Calculate $\theta^{(t+1)} = (\pi_1^{(t+1)}, v_1^{(t+1)}, p^{(t+1)})$ applying the algorithm 2.3.
```

Step 2: Obtain $\theta_1{}^{(t)} = \left(\pi_1^{(t)}, \hat{v}_1, \hat{p}\right), \theta_2{}^{(t)} = \left(\hat{\pi}_1, v_1^{(t)}, \hat{p}\right)$ and $\theta_3{}^{(t)} = \left(\hat{\pi}_1, \hat{v}_1, p^{(t)}\right)$. For each one of them, execute the first iteration of the algorithm 2.3 considering $\theta_i{}^{(t)}$ to be the initial value of θ and obtain the values $\tilde{\theta}_1^{(t+1)}, \tilde{\theta}_2^{(t+1)}$ and $\tilde{\theta}_3^{(t+1)}$.

Step 3: Calculate the ratios

$$r_{ij}^{(t)} = \frac{\tilde{\theta}_j^{(t+1)}(i) - \hat{\theta}_j}{\theta_i^{(t)} - \hat{\theta}_i}, i, j = 1,2,3. \tag{20}$$

OUTPUT: $\theta^{(t+1)}$and $r_{ij}^{\ t}$, $i, j = 1, 2, 3$.

Therefore, having imputed the *MLEs* of π_1, ν_1 and p through the algorithm 2.3, the initial values of $\theta^{(0)} = \left(\pi_1^{(0)}, v_1^{(0)}, p^{(0)}\right)$ are taken, the same ones, for example, as in the application of the algorithm 2.3, and an iteration of the algorithm 2.3 is executed obtaining $\theta^{(1)} = \left(\pi_1^{(1)}, v_1^{(1)}, p^{(1)}\right)$. Next, the vectors

$$\theta^{\ (0)}(1) = \left(\pi_1^{(0)}, \hat{v}_1, \hat{p}\right), \theta^{\ (0)}(2) = \left(\hat{\pi}_1, v_1^{(0)}, \hat{p}\right) \text{ and } \theta_3{}^{(0)} = \left(\hat{\pi}_1, \hat{v}_1, p^{(0)}\right)$$

are constructed, and for each one of them an iteration of the algorithm 2.3 is executed considering each $\theta^{\ (0)}(i)$ as the initial value which is taken on applying this algorithm, obtaining $\tilde{\theta}^{\ (1)}(1) = \left(\tilde{\pi}_1^{(1)}(1), \tilde{v}_1^{(1)}(1), \tilde{p}_1^{(1)}(1)\right), \tilde{\theta}^{\ (1)}(2) = \left(\tilde{\pi}_1^{(1)}(2), \tilde{v}_1^{(1)}(2), \tilde{p}_1^{(1)}(2)\right)$ and $\tilde{\theta}^{\ (1)}(3) = \left(\tilde{\pi}_1^{(1)}(3), \tilde{v}_1^{(1)}(3), \tilde{p}_1^{(1)}(3)\right)$

. Next, the $r_{ij}^{(0)}$ ratios are calculated:

$$r_{11}^{(0)} = \frac{\tilde{\pi}_1^{(1)}(1) - \hat{\pi}_1}{\pi_1^{(0)} - \hat{\pi}_1}, r_{12}^{(0)} = \frac{\tilde{v}_1^{(1)}(1) - \hat{v}_1}{\pi_1^{(0)} - \hat{\pi}_1}, r_{13}^{(0)} = \frac{\tilde{p}_1^{(1)}(1) - \hat{p}}{\pi_1^{(0)} - \hat{\pi}_i}$$
$$r_{21}^{(0)} = \frac{\tilde{\pi}_1^{(1)}(2) - \hat{\pi}_1}{v_1^{(0)} - \hat{v}_1}, r_{22}^{(0)} = \frac{\tilde{v}_1^{(1)}(2) - \hat{v}_1}{v_1^{(0)} - \hat{v}_1}, r_{23}^{(0)} = \frac{\tilde{p}_1^{(1)}(2) - \hat{p}}{v_1^{(0)} - \hat{v}_1}$$

$$r_{31}^{(0)} = \frac{\tilde{\pi}_1^{(1)}(3) - \hat{\pi}_1}{p^{(0)} - \hat{p}}, r_{32}^{(0)} = \frac{\tilde{v}_1^{(1)}(3) - \hat{v}_1}{p^{(0)} - \hat{p}}, r_{33}^{(0)} = \frac{\tilde{p}_1^{(1)}(3) - \hat{p}}{p^{(0)} - \hat{p}} \tag{21}$$

and the process is repeated until $\left|r_{ij}^{(t+1)} - r_{ij}^{(t)}\right| \leq \delta$, when δ is the value of the stop criteria of the algorithm 2.3. Next, the asymptotic variance-covariance matrix of $\hat{\pi}_1$, $\hat{\nu}_1$ and $\hat{p}$ is estimated applying equation (17).

4. SIMULATION STUDY

The analysis of the asymptotic coverage of the confidence intervals obtained through the application of the algorithm 2.3 and *SEM* algorithm has been carried out through a Monte Carlo study which consisted of the generation of 2000 random multinomial samples, with the probabilities given in Table 2, sized 1000, 2000, 3000, 4000 and 5000. As values of sensitivity and specificity of two diagnostic tests we have taken $(\pi_1 = 0{,}85,\ v_1 = 0.75, \pi_2 = 0.98, v_2 = 0.95)$ and $(\pi_1 = 0{,}80,\ v_1 = 0.70, \pi_2 = 0.95, v_2 = 0.90)$, as they are values which appear quite frequently in clinical practice. In the first case, $\varepsilon_1 \leq 0.017$ and $\varepsilon_0 \leq 0.0375$, and in the second case $\varepsilon_1 \leq 0.04$ and $\varepsilon_0 \leq 0.07$. As values of disease prevalence we have taken 10%, 20%, 30%, 40% and 50%; and as values of verification probabilities we have taken $(\lambda_1 = 0{,}75,\ \lambda_0 = 0.10)$ and $(\lambda_1 = 0{,}95,\ \lambda_0 = 0.25)$, which in practice can be considered extreme verification probabilities for the sample sizes analysed. For each one of the samples generated, the algorithm 2.3 and the *SEM* algorithm have been applied, and for each 2000 samples of the same size generated from the same multinomial distribution the percentage of the 95% confidence intervals $(\hat{\theta} \mp 1.96\hat{\sigma}_{\hat{\theta}})$ which contain the value of sensitivity and specificity respectively.

In Tables 3 and 4, some of the results obtained for $(\pi_1 = 0{,}85,\ \nu_1 = 0.75, \pi_2 = 0.98, \nu_2 = 0.95)$ taking different values of ε_1 and ε_0. From these results, it is obtained that, in general terms, the coverage of the confidence intervals of sensitivity and specificity grow with an increase in the prevalence of the disease and/or with an increase in the probabilities verification. With respect to conditional dependence $(\varepsilon_1\ ,\ \varepsilon_0)$, this has no clear effect upon the coverage of the confidence intervals. In general terms, when the probabilities of verification are low, depending on the prevalence of the disease it is necessary to use samples of between 1000 and 5000 patients so that the confidence intervals of sensitivity and specificity have a coverage of 95%. In general terms, a 10% increase in disease prevalence implies a decrease of 1000 patients in the size of the sample in order to obtain a coverage of 95%. When the probabilities of verification are high, with samples of between 1000 and 2000 patients the confidence intervals of sensitivity and specificity have a coverage of 95%. Similar results have been obtained for $(\pi_1 = 0{,}80,\ \nu_1 = 0.70, \pi_2 = 0.95, \nu_2 = 0.90)$.

$\pi_1 = 0.85\ \ \nu_1 = 0.75\ \ \pi_2 = 0.98\ \ \nu_2 = 0.95\ \ p = 0.10\ \ \lambda_1 = 0.75\ \ \lambda_0 = 0.10$					
n	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$	$\varepsilon_1 = 0.004250$ $\varepsilon_0 = 0.009375$	$\varepsilon_1 = 0.00850$ $\varepsilon_0 = 0.01875$	$\varepsilon_1 = 0.012750$ $\varepsilon_0 = 0.028125$	$\varepsilon_1 = 0.0165$ $\varepsilon_0 = 0.0370$
1000	0.7205	0.8100	0.7570	0.6230	0.3655
2000	0.8865	0.8440	0.8245	0.8570	0.8640
3000	0.9225	0.9010	0.9030	0.9020	0.9155
4000	0.9480	0.9135	0.9305	0.9355	0.9340
5000	0.9480	0.9425	0.9470	0.9505	0.9425
$\pi_1 = 0.85\ \ \nu_1 = 0.75\ \ \pi_2 = 0.98\ \ \nu_2 = 0.95\ \ p = 0.10\ \ \lambda_1 = 0.95\ \ \lambda_0 = 0.25$					
n	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$	$\varepsilon_1 = 0.004250$ $\varepsilon_0 = 0.009375$	$\varepsilon_1 = 0.00850$ $\varepsilon_0 = 0.01875$	$\varepsilon_1 = 0.012750$ $\varepsilon_0 = 0.028125$	$\varepsilon_1 = 0.0165$ $\varepsilon_0 = 0.0370$
1000	0.5375	0.8115	0.8840	0.6635	0.2810
2000	0.9515	0.9485	0.9540	0.9485	0.9535
3000	0.9695	0.9685	0.9685	0.9610	0.9645
4000	0.9685	0.9575	0.9555	0.9610	0.9720
5000	0.9625	0.9580	0.9610	0.9605	0.9525
$\pi_1 = 0.85\ \ \nu_1 = 0.75\ \ \pi_2 = 0.98\ \ \nu_2 = 0.95\ \ p = 0.50\ \ \lambda_1 = 0.75\ \ \lambda_0 = 0.10$					
n	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$	$\varepsilon_1 = 0.004250$ $\varepsilon_0 = 0.009375$	$\varepsilon_1 = 0.00850$ $\varepsilon_0 = 0.01875$	$\varepsilon_1 = 0.012750$ $\varepsilon_0 = 0.028125$	$\varepsilon_1 = 0.0165$ $\varepsilon_0 = 0.0370$
1000	0.9505	0.9420	0.9400	0.9345	0.9500
2000	0.9525	0.9440	0.9495	0.9415	0.9420
3000	0.9505	0.9470	0.9585	0.9635	0.9590
4000	0.9525	0.9540	0.9570	0.9540	0.9525
5000	0.9500	0.9575	0.9580	0.9630	0.9565
$\pi_1 = 0.85\ \ \nu_1 = 0.75\ \ \pi_2 = 0.98\ \ \nu_2 = 0.95\ \ p = 0.50\ \ \lambda_1 = 0.95\ \ \lambda_0 = 0.25$					
n	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$	$\varepsilon_1 = 0.004250$ $\varepsilon_0 = 0.009375$	$\varepsilon_1 = 0.00850$ $\varepsilon_0 = 0.01875$	$\varepsilon_1 = 0.012750$ $\varepsilon_0 = 0.028125$	$\varepsilon_1 = 0.0165$ $\varepsilon_0 = 0.0370$
1000	0.9570	0.9580	0.9375	0.9575	0.9525
2000	0.9485	0.9540	0.9520	0.9550	0.9450
3000	0.9580	0.9535	0.9500	0.9525	0.9550
4000	0.9525	0.9550	0.9490	0.9550	0.9480

5000	0.9525	0.9530	0.9575	0.9575	0.9445

Table 3. Coverage of the 95% confidence interval of sensitivity.

$\pi_1 = 0.85 \quad \nu_1 = 0.75 \quad \pi_2 = 0.98 \quad \nu_2 = 0.95 \quad p = 0.10 \quad \lambda_1 = 0.75 \quad \lambda_0 = 0.10$					
n	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$	$\varepsilon_1 = 0.004250$ $\varepsilon_0 = 0.009375$	$\varepsilon_1 = 0.00850$ $\varepsilon_0 = 0.01875$	$\varepsilon_1 = 0.012750$ $\varepsilon_0 = 0.028125$	$\varepsilon_1 = 0.0165$ $\varepsilon_0 = 0.0370$
1000	0.7915	0.9390	0.9530	0.9765	0.9710
2000	0.9665	0.9530	0.9470	0.9520	0.9505
3000	0.9650	0.9515	0.9565	0.9570	0.9535
4000	0.9625	0.9465	0.9480	0.9595	0.9545
5000	0.9640	0.9530	0.9610	0.9645	0.9545
$\pi_1 = 0.85 \quad \nu_1 = 0.75 \quad \pi_2 = 0.98 \quad \nu_2 = 0.95 \quad p = 0.10 \quad \lambda_1 = 0.95 \quad \lambda_0 = 0.25$					
n	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$	$\varepsilon_1 = 0.004250$ $\varepsilon_0 = 0.009375$	$\varepsilon_1 = 0.00850$ $\varepsilon_0 = 0.01875$	$\varepsilon_1 = 0.012750$ $\varepsilon_0 = 0.028125$	$\varepsilon_1 = 0.0165$ $\varepsilon_0 = 0.0370$
1000	0.7545	0.9450	0.9635	0.9670	0.9460
2000	0.9565	0.9590	0.9555	0.9570	0.9595
3000	0.9565	0.9555	0.9545	0.9580	0.9560
4000	0.9635	0.9575	0.9600	0.9560	0.9540
5000	0.9625	0.9545	0.9650	0.9510	0.9535
$\pi_1 = 0.85 \quad \nu_1 = 0.75 \quad \pi_2 = 0.98 \quad \nu_2 = 0.95 \quad p = 0.50 \quad \lambda_1 = 0.75 \quad \lambda_0 = 0.10$					
n	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$	$\varepsilon_1 = 0.004250$ $\varepsilon_0 = 0.009375$	$\varepsilon_1 = 0.00850$ $\varepsilon_0 = 0.01875$	$\varepsilon_1 = 0.012750$ $\varepsilon_0 = 0.028125$	$\varepsilon_1 = 0.0165$ $\varepsilon_0 = 0.0370$
1000	0.9635	0.9485	0.9470	0.9430	0.9510
2000	0.9540	0.9520	0.9540	0.9460	0.9460
3000	0.9510	0.9445	0.9590	0.9560	0.9555
4000	0.9605	0.9530	0.9545	0.9470	0.9525
5000	0.9560	0.9545	0.9585	0.9570	0.9600
$\pi_1 = 0.85 \quad \nu_1 = 0.75 \quad \pi_2 = 0.98 \quad \nu_2 = 0.95 \quad p = 0.50 \quad \lambda_1 = 0.95 \quad \lambda_0 = 0.25$					
n	$\varepsilon_1 = 0$ $\varepsilon_0 = 0$	$\varepsilon_1 = 0.004250$ $\varepsilon_0 = 0.009375$	$\varepsilon_1 = 0.00850$ $\varepsilon_0 = 0.01875$	$\varepsilon_1 = 0.012750$ $\varepsilon_0 = 0.028125$	$\varepsilon_1 = 0.0165$ $\varepsilon_0 = 0.0370$
1000	0.9460	0.9480	0.9480	0.9545	0.9420
2000	0.9605	0.9490	0.9420	0.9580	0.9465
3000	0.9570	0.9525	0.9515	0.9560	0.9450
4000	0.9590	0.9565	0.9495	0.9465	0.9420
5000	0.9600	0.9460	0.9500	0.9585	0.9460

Table 4. Coverage of the 95% confidence interval of specificity.

5. APPLICATION

The results of Sections 2 and 3 have been applied to the diagnosis of *HIV* using the *ELISA* test as a reference test. Sensitivity and specificity of the *ELISA* test are 0.98 and 0.93 approximately. The diagnosis of this disease can also be made through the *p24* antigen test. The objective is to evaluate the *p24* antigen test taking as a reference test the *ELISA*. In Table 5, the results shown are those obtained when applying the *p24* antigen test to a sample of 2150 individuals and the *ELISA* test to a part of this sample.

Using the *EM* algorithm and taking (*0,5, 0.5, 0.5*) as initial values and 10^{-24} as the value of the stop criteria, the estimated sensitivity and specificity of the *p24* test are $\hat{\pi}_1 = 71.13\%$ and $\hat{\nu}_1 = 93.88\%$ respectively. Applying the *SEM* algorithm, standard errors of sensitivity and specificity are

$\hat{\theta}_{\hat{\pi}_1} = 0.075$ and $\hat{\theta}_{\hat{\nu}_1} = 0.004$ $\hat{\sigma}_{\hat{\pi}_1} = 0.075$ and $\hat{\sigma}_{\hat{\nu}_1} = 0.004$, and the respective 95% confidence intervals are (0.5635, 0.8591) and (0.9307, 0,9468).

		$T=1$	$T=0$
$V=1$			
	$R=1$	203	11
	$R=0$	46	163
$V=0$		13	1714
	Total	262	1888

Table 5. Data on HIV.

6. CONCLUDING REMARKS

In this article, an algorithm is proposed to estimate the sensitivity and the specificity of a new binary diagnostic test in relation to a reference test with known accuracy when the reference test is not applied to all of the patients in the sample. The algorithm is based on the approximation of covariances between the two diagnostic tests and on the application of the *EM* algorithm. Moreover, the *SEM* algorithm developed by Meng and Rubin (1991) is applied to impute the standard errors of the estimators of sensitivity and specificity. Simulation experiments have been carried out in order to study the asymptotic coverage of the confidence intervals of the estimators obtained, analysing the effect upon them of verification probabilities, conditional dependence between the two diagnostic tests and the prevalence of the disease. In general terms, when the probabilities of verification are low, depending on the prevalence of the disease it is necessary to use samples of between 1000 and 5000 patients so that the confidence intervals of sensitivity and specificity have a coverage of 95%. When the probabilities of verification are high, it is necessary to use samples of between 1000 and 2000 patients so that the confidence intervals of sensitivity and specificity have a coverage of 95%. The high number of samples is undoubtedly due to the sources of missing information in the problem analysed. On the one hand, the disease status in each patient is unknown, and on the other hand the reference test is not applied to all of the subjects, which has a clear repercussion on the size of the samples.

Acknowledgements: This research was supported by the Spanish Ministry of Science, Grant Number MTM2009-08886. The authors would like to thank Prof. María del Mar Rueda and Prof. Carlos Bouza for her invitation to write this article for this book.

REFERENCES

[1] BEGG, C. B. & GREENES, R. A. (1983). Assessment of diagnostic tests when disease verification is subject to selection bias. **Biometrics,** 39, 207-215.

[2] ZHOU, X. H. (1993). Maximum likelihood estimators of sensitivity and specificity corrected for verification bias. **Communication in Statistics – Theory and Methods**, 22, 3177-3198.

[3] GART, J. J. & BUCK, A. A. (1966). Comparison of a screening test and a reference test in epidemiologic studies. II. A probabilistic model for the comparison of diagnostic tests. **American Journal of Epidemiology**, 83, 593-602.

[4] GREENBERG, R. A. & JEKEL, J. F. (1969). Some problems in the determination of the false positive and false negative rates of tuberculin tests. **American Review of Respiratory Disease**, 100, 645-650.

[5] HUI, S. L. & WALTER, S. D. (1980). Estimating the error rates of diagnostic tests. **Biometrics**, 36, 167-171.

[6] VACEK, P. (1985). The effect of conditional dependence on the evaluation of diagnostic tests. **Biometrics**, 41, 959-968.

[7] TORRANCE-RYNARD, V. L. & WALTER, S. D. (1997). Effects of dependent errors in the assessment of diagnostic test performance. **Statistics in Medicine**, 16, 2157-2175.

[8] QU, Y., TANG, M. & KUTNER, M. H. (1996). Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. **Biometrics**, 52, 797-810.

[9] ESPELAND, M. A., PLATT, O. S. & GALLAGHER, D. (1989). Joint estimation of incidence and diagnostic error rates from irregular longitudinal data. **Journal of the American Statistical Association**, 84, 972-979.

[10] YANG, I. & BECKER, M. P. (1997). Latent variable modeling of diagnostic accuracy. **Biometrics**, 53, 948-958.

[11] SINCLAIR, M. D. & GASTWIRTH, J. L. (1996). On procedure for evaluating the effectiveness of reinterview survey methods: application to labor force data. **Journal of the American Statistical Association**, 91, 961-969.

[12] MENG, X. L. & RUBIN, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. **Journal of the American Statistical Association**, 86, 899-909.

[13] DEMPSTER, A. P., LAIRD, N. M., & RUBIN, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). **Journal of the Royal Statistical Society**, Series B, 39, 1-38.

[14] YOUDEN, W.J. (1950). Index for rating diagnostic tests. **Cancer**, 3, 32-35.

Appendix A: EM algorithm

If the true state of the disease in all patients is known, the logarithm of the likelihood function of the complete data is

$$l^{'} = \sum_{j,k=0}^{1} \quad a_{1jk} \, log\left(p_{1jk}\right) + \sum_{j=0}^{1} \quad b_{1j} log\left(q_{1j}\right) + \sum_{j,k=0}^{1} \quad a_{0jk} \, log\left(p_{0jk}\right) + \sum_{j=0}^{1} \quad b_{0j} \, log\left(q_{0j}\right)$$

(22)

when a_{ijk} is the number of patients in whom $D = i, T = j, V = 1, R = k$, b_{ij} is the number of patients in whom *(D=i, T=j, V=0), p_{ijk}=P(D=i, T=j, V=1,R=k)* and *q_{ij}=P(D=i, T=j, V=0).*

Let $\hat{\pi}_1^{(t)}, \hat{v}_1^{(t)}$ and $\hat{p}_1^{(t)}$ be current values of parameter π_1, ν_1 and *p* after *t* iterations in the *EM* algorithm. The *E* step of *M* gives the expected value of the complete-data log-likelihood conditional on the observed data and the current values of the parameters. From (22) we obtain the conditional expectation as

$$Q(\pi_1, v_1, p, \lambda_1, \lambda_0) = \sum_{j,k=0}^{1} \alpha_{1jk}^{(t)} \quad log\left(p_{1jk}\right) + \sum_{j=0}^{1} \quad \beta_{1j}^{(t)} log\left(q_{1j}\right) + \sum_{j,k=0}^{1} \quad \alpha_{0jk}^{(t)} log\left(p_{0jk}\right) + \sum_{j=0}^{1} \quad \beta_{0j} \, log\left(q_{0j}\right) \tag{23}$$

where

$$\alpha_{1jk}^{(t)} = s_{jk} P\left(D = 1 \middle| T = j, V = 1, R = k, \pi_1 = \pi_1^{(t)}, v_1 = v_1^{(t)}, p = p^{(t)}, \hat{\lambda}_1, \hat{\lambda}_0\right)$$

$$\alpha_{0jk}^{(t)} = s_{jk} P\left(D = 0 \middle| T = j, V = 1, R = k, \pi_1 = \pi_1^{(t)}, v_1 = v_1^{(t)}, p = p^{(t)}, \hat{\lambda}_1, \hat{\lambda}_0\right) = s_{jk} - \alpha_{1jk}^{(t)}$$

$$\beta_{1j}^{(t)} = u_j P\left(D = 1 \middle| T = j, V = 0, \pi_1 = \pi_1^{(t)}, v_1 = v_1^{(t)}, p = p^{(t)}, \hat{\lambda}_1, \hat{\lambda}_0\right)$$

and

$$\beta_{0j}^{(t)} = u_j P\left(D = 0 \middle| T = j, V = 0, \pi_1 = \pi_1^{(t)}, v_1 = v_1^{(t)}, p = p^{(t)}, \hat{\lambda}_1, \hat{\lambda}_0\right) = u_j - \beta_{1j}^{(t)}$$

.

The conditioned probabilities are calculated using (1) and (2). The *M* step of *EM* gives new estimators $\pi_1^{(t+1)}, v_1^{(t+1)}$ and $p^{(t+1)}$ for π_1, ν_1 and *p* by maximizing the conditional expectation $Q(\pi_1, v_1, p, \lambda_1, \lambda_0)$. These new estimators have the explicit forms given in Proposition 1.

Appendix B: Fisher information matrix

The Fisher information matrix of the complete data, I_{oc} , is the information matrix corresponding to the logarithm of likelihood function (22). This matrix is estimated through the Fisher information matrix corresponding to function (23), and this is a diagonal matrix whose elements are:

$$E\left(-\frac{\partial^2 l^{'}}{\partial \pi_1^2}\right)=\frac{\alpha_{111}^{(T)}\pi_2^2}{\{\pi_1\pi_2+\varepsilon_1\}^2}+\frac{\alpha_{121}^{(T)}(1-\pi_2)^2}{\{(\pi_1(1-\pi_2)-\varepsilon_1\}^2}+\frac{\beta_{11}^{(T)}}{\pi_1{}^2}+\frac{\alpha_{112}^{(T)}\pi_2^2}{\{(1-\pi_1)\pi_2-\varepsilon_1\}^2}+\frac{\alpha_{122}^{(T)}(1-\pi_2)^2}{\{(1-\pi_1)(1-\pi_2)+\varepsilon_1\}^2}+$$

$$\frac{\beta_{10}^{(T)}}{(1-\pi_1)^2}, E\left(-\frac{\partial^2 l^{'}}{\partial \pi_1\ \partial v_1}\right)=0$$

$$E\left(-\frac{\partial^2 l^{'}}{\partial p^2}\right)=\frac{\sum_{j,k=0}^{1}\quad \alpha_{1jk}^{(T)}+\sum_{j=0}^{1}\beta_{1j}^{(T)}}{p^2}+\frac{\sum_{j,k=0}^{1}\quad \alpha_{0jk}^{(T)}++\sum_{j=0}^{1}\beta_{0j}^{(T)}}{(1-p)^2}, E\left(-\frac{\partial^2 l^{'}}{\partial v_1\ \partial p}\right)=0$$

where each value $\alpha_{ijk}^{(T)}$ and $\beta_{ij}^{(T)}$ has been obtained in the last iteration of the method 2.2. On rare occasions, the matrix I_{oc}^{-1} can be badly conditioned and this usually happens when with small disease prevalence the sample is not big enough, and the column of subjects with a negative test from the table of diseased subjects obtained in the last iteration of the *EM* algorithm is equal to zero. The problem is solved by increasing the size of the sample or by increasing the level of verification for subjects with a negative test.