Computational methods to simultaneously compare the predictive values of two diagnostic tests with missing data: *EM-SEM* algorithms and multiple imputation

J.A. Roldán-Nofuentes

Biostatistics, School of Medicine, University of Granada, 18016, Spain

Email: jaroldan@ugr.es

This is an Accepted Manuscript of an article published by Taylor & Francis in the Journal of Statistical Computation and Simulation on 2021, available at https://www.tandfonline.com/doi/abs/10.1080/00949655.2021.1926461

Abstract. Predictive values are measures of the clinical accuracy of a binary diagnostic test, and depend on the sensitivity and the specificity of the diagnostic test and on the disease prevalence among the population being studied. This article studies hypothesis tests to simultaneously compare the predictive values of two binary diagnostic tests in the presence of missing data. The hypothesis tests were solved applying two computational methods: the expectation maximization and the supplemented expectation maximization algorithms, and multiple imputation. Simulation experiments were carried out to study the sizes and the powers of the hypothesis tests, giving some general rules of application. Two R programmes were written to apply each method, and they are available as supplementary material for the manuscript. The results were applied to the diagnosis of Alzheimer's disease.

Key words: EM and SEM algorithms, Missing data, Multiple imputation, Partial verification, Predictive values.

Mathematics Subject Classification: 62P10, 6207.

1. Introduction

A diagnostic test is a medical test that is applied to a patient to determine the presence or absence of a certain disease. When the result of a diagnostic test is positive or negative, the diagnostic test is called a binary diagnostic test (*BDT*). The mammography for breast cancer is an example of a *BDT*. The clinical effectiveness of a *BDT* is measured in terms of two parameters: the positive predictive value and the negative predictive value. Positive predictive value (τ) is the probability of a patient having the disease when the result of the *BDT* is positive, and the negative predictive value (υ) is the probability of the patient not having the disease when the result of the *BDT* is negative. The predictive values (*PVs*) depend on the sensitivity (*Se*) and on the specificity (*Sp*) of the *BDT* and on the disease prevalence (*p*) among the population studied, i.e.

$$\tau = \frac{p \times Se}{p \times Se + q \times (1 - Sp)} \quad \text{and} \quad \upsilon = \frac{q \times Sp}{p \times (1 - Se) + q \times Sp}, \tag{1}$$

where $q = 1 - p$. While *Se* and *Sp* quantify how well the *BDT* reflexes the true disease status, the *PVs* quantify the clinical value of the *BDT*, since the patient is more interested in knowing the probability of having or not having the disease given a result of the diagnostic test. The parameters of a *BDT* are estimated in relation to a gold standard (*GS*), which is a medical test which determines without any errors whether or not the patient has the disease. A biopsy for breast cancer is an example of a *GS*.

In clinical practice, the most common sample design to compare the *PVs* of two *BDTs* is paired design [1, 2]. This type of design consists of applying the two *BDTs* to all of the individuals in a sample sized *n* whose disease status is known through the application of a *GS*. The comparison of the *PVs* of two *BDTs* subject to paired design has been the subject of different studies in statistics literature. Leisenring et al [3], Wang et al [4], Kosinski [5] and Tsou [6] have studied asymptotic methods to compare the two positive *PVs* and the negative

PVs independently, i.e. solving the two hypothesis tests $H_0:\tau_1=\tau_2$ and $H_0:\upsilon_1=\upsilon_2$ each one of them to an α error. The Kosinski method has a better asymptotic performance (in terms of type I error and power) than the methods of Leisenring et al and of Wang et al. The method of Tsou leads to the same results as the Kosinski method. Roldán Nofuentes et al [7] studied a global hypothesis test to simultaneously compare the *PVs* of two *BDTs*, i.e. solving the global hypothesis test $H_0:(\tau_1=\tau_2$ and $\upsilon_1=\upsilon_2)$ vs $H_1:(\tau_1\neq\tau_2$ and/or $\upsilon_1\neq\upsilon_2)$, and proposed a method based on chi-squared distribution and multiple comparisons. These authors have demonstrated that the comparison of the positive *PVs* (negative *PVs*) of two *BDTs* subject to a paired design must be carried out simultaneously, solving the global hypothesis test $H_0:(\tau_1=\tau_2$ and $\upsilon_1=\upsilon_2)$. They have also demonstrated that the comparison of the *PVs* is made independently, i.e. solving the tests $H_0:\tau_1=\tau_2$ and $H_0:\upsilon_1=\upsilon_2$ each one of them to an α error, the results may be mistaken. In Appendix *A* these methods are summarized.

When assessing or comparing parameters of *BDTs* it is not common for the *GS* to be applied to all of the individuals in the sample, leading to the problem known as partial disease verification [8, 9]. Therefore, if the *GS* consists of a costly test or one which means some risk for the individual, then it will not be applied to all of the individuals in the sample, and consequently the true disease status is unknown for a subgroup of individuals. When comparing parameters of two *BDTs* in the presence of partial disease verification, it is common to assume that the verification process is missing at random (*MAR*). This assumes that the process to verify the disease status of an individual through the application of the *GS* only conditionally depends on the results of the two *BDTs* and it does not depend on the disease status of the individual. Subject to the *MAR* assumption, there are many different studies in statistics literature which compare parameters of two or more *BDTs* in the presence of partial verification. Zhou [9] studied a hypothesis test to compare the sensitivities (specificities) of two *BDTs* applying the method of maximum likelihood. Roldán Nofuentes

and Luna [10] studied the individual comparison of the *PVs* of two *BDTs* applying the maximum likelihood (*ML*) method. Marín-Jiménez and Roldán-Nofuentes [11] extended the study of Roldán-Nofuentes and Luna [10] to the case of more than two *BDTs* comparing the *PVs* simultaneously applying the *ML* method. Harel and Zhou [12] compared the sensitivities (specificities) of two *BDTs* through confidence intervals applying multiple imputation (*MI*). Roldán-Nofuentes and Luna [13] compared the sensitivities and the specificities independently, as well as the *PVs*, of two *BDTs* applying the expectation maximization (*EM*) algorithm and the supplemented expectation maximization (*SEM*) algorithm.

In this article, hypothesis tests are studied to simultaneously compare the *PVs* of two *BDTs* when in the presence of partial disease verification the missing data mechanism is *MAR*, applying two computational methods: the *EM* and *SEM* algorithms, and *MI*. The *EM* algorithm is a classic method for estimating parameters in the presence of missing data. The advantage of the *EM* algorithm over the *ML* method [11] is that the *EM* algorithm can be applied when some observed frequency is zero, while the *ML* method [11] cannot be applied in this situation. Regarding the *MI*, this method offers better results than the *ML* method when comparing the sensitivities (specificities) of two *BDTs* in the presence of missing data. Therefore, it is convenient to evaluate this method to solve the problem presented here. *MI* cannot be applied when some observed frequency is zero. Therefore, with both types of computational methods we seek to solve the global hypothesis global test to simultaneously compare the positive *PVs* and the negative *PVs* of the two *BDTs*, i.e.

$$\begin{aligned} H_0 &: \tau_1 = \tau_2 \ \text{ and } \ \upsilon_1 = \upsilon_2 \\ H_1 &: \tau_1 \neq \tau_2 \ \text{ and/or } \ \upsilon_1 \neq \upsilon_2. \end{aligned} \tag{2}$$

In Section 2, we solve the global test applying the *EM* and *SEM* algorithms, and in Section 3 the same test is solved applying *MI*. In Section 4, simulation experiments are carried out to study the type I error and the power of the previous tests with each one of the two methods (*EM-SEM* algorithms and *MI*), and some general rules of application are given. In Section 5,

two programmes written in *R* are presented to solve the problem posed by applying each computational method. In Section 6, the results were applied to a real example of the diagnosis of Alzheimer's disease, and in Section 7 the results obtained are discussed.

2. *EM* and *SEM* algorithms

Let us consider two *BDTs* that are applied to all of the individuals in a random sample sized *n*, and let us also consider a *GS* that is only applied to a subgroup of the sample. This situation leads to the observed frequencies in Table 1*a*, where the variable T_h models the result of the *h*th *BDT* ($T_h = 1$ when the result is positive and $T_h = 0$ when it is negative), the variable *V* models the verification process ($V = 1$ when the disease status of an individual is verified with the *GS* and $V = 0$ when it is not), and the variable *D* models the result of the *GS* ($D = 1$ when the individual verified has the disease and $D = 0$ when this individual does not). In Table 1a, a_{ij} is the number of individuals with the disease among whom $T_1 = i$ and $T_2 = j$, b_{ij} is the number of individuals without the disease among whom $T_1 = i$ and $T_2 = j$, and c_{ij} is the number of individuals with an unknown disease status among whom $T_1 = i$ and $T_2 = j$, with $i, j = 0,1$. Let $a = \sum_{i,j=0}^{1} a_{ij}$, $b = \sum_{i,j=0}^{1} b_{ij}$, $c = \sum_{i,j=0}^{1} c_{ij}$, $n_{ij} = a_{ij} + b_{ij} + c_{ij}$ and $n = \sum_{i,j=0}^{1} n_{ij}$. For the *h*th *BDT*, let $Se_h = P(T_h = 1 | D = 1)$, $Sp_h = P(T_h = 0 | D = 0)$, $\tau_h = P(D = 1 | T_h = 1)$ and $\upsilon_h = P(D = 0 | T_h = 0)$, with $h = 1,2$. Let $p = P(D = 1)$ be the disease prevalence and $q = 1 - p = P(D = 0)$. From the expressions (1) each *Se* and *Sp* is written, in terms of the *PVs* and of *p*, as

$$Se_h = \frac{\tau_h(\upsilon_h - q)}{pY_h} \text{ and } Sp_h = \frac{\upsilon_h(\tau_h - p)}{qY_h}, \tag{3}$$

where $Y_h = \tau_h + \upsilon_h - 1$.

==== INSERT TABLE 1 HERE ====

In the presence of partial disease verification, the verification probabilities are designed as $\lambda_{ijk} = P\left(V=1 \middle| T_1=i, T_2=j, D=k\right)$, i.e. λ_{ijk} is the probability of verifying the *GS* the disease status of an individual for whom $T_1=i$, $T_2=j$ and $D=k$, with $i,j,k=0,1$. Assuming that the missing data mechanism is *MAR*, i.e. that the probability of verifying the disease status of an individual only conditionally depends on the results of the two *BDTs* and does not depend on the result of the *GS*, it is verified that $\lambda_{ijk}=\lambda_{ij}=P\left(V=1 \middle| T_1=i, T_2=j\right)$. Supposing the *MAR* assumption, the data from Table 1a are the product of a multinomial distribution sized *n* and the probabilities are written in terms of the *PVs* as

$$\xi_{ij} = P\left(V=1, D=1, T_1=i, T_2=j\right) =$$

$$p\lambda_{ij}\left[\frac{\tau_1^i\left(\upsilon_1-q\right)^i\left(\tau_1+p\right)^{1-i}\left(1-\upsilon_1\right)^{1-i}}{p^i p^{1-i} Y_1^i Y_1^{1-i}} \times \frac{\tau_2^j\left(\upsilon_2-q\right)^j\left(\tau_2+p\right)^{1-j}\left(1-\upsilon_2\right)^{1-j}}{p^j p^{1-j} Y_2^j Y_1^{1-j}} + \delta_{ij}\varepsilon_1\right],$$

$$\psi_{ij} = P\left(V=1, D=0, T_1=i, T_2=j\right) = \tag{4}$$

$$q\lambda_{ij}\left[\frac{\left(1-\tau_1\right)^i\left(\upsilon_1-q\right)^i\left(\tau_1-p\right)^{1-i}\upsilon_1^{1-i}}{q^i q^{1-i} Y_1^i Y_1^{1-i}} \times \frac{\left(1-\tau_2\right)^j\left(\upsilon_2-q\right)^j\left(\tau_2-p\right)^{1-j}\upsilon_2^{1-j}}{q^j q^{1-j} Y_2^j Y_2^{1-j}} + \delta_{ij}\varepsilon_0\right],$$

$$\zeta_{ij} = P\left(V=0, T_1=i, T_2=j\right) = \frac{1-\lambda_{ij}}{\lambda_{ij}}\left(\xi_{ij}+\psi_{ij}\right),$$

where

$$\varepsilon_1 = \frac{\tau_1\tau_2\left(\upsilon_1-q\right)\left(\upsilon_2-q\right)\left(\alpha_1-1\right)}{p^2 Y_1 Y_2} \text{ and } \varepsilon_0 = \frac{\left(1-\tau_1\right)\left(1-\tau_2\right)\left(\upsilon_1-q\right)\left(\upsilon_2-q\right)\left(\alpha_0-1\right)}{q^2 Y_1 Y_2}.$$

with $\delta_{ij}=1$ if $i=j$ and $\delta_{ij}=-1$ if $i\neq j$. Parameters α_1 and α_0 are the covariances [14] between the two *BDTs* when $D=1$ and when $D=0$ respectively, verifying that

$$1 \le \alpha_1 \le \frac{1}{\max\left\{\frac{\tau_1\left(\upsilon_1-q\right)}{pY_1}, \frac{\tau_2\left(\upsilon_2-q\right)}{pY_2}\right\}} \text{ and } 1 \le \alpha_0 \le \frac{1}{\max\left\{\left(1-\frac{\upsilon_1\left(\tau_1-p\right)}{qY_1}\right), \left(1-\frac{\upsilon_2\left(\tau_2-p\right)}{qY_2}\right)\right\}}.$$

If $\alpha_1 = \alpha_0 = 1$, then the two *BDTs* are conditionally independent on the disease, a situation which is not realistic in practice, and therefore it must be verified that $\alpha_1 > 1$ and/or $\alpha_0 > 1$. An *EM* algorithm is then proposed to estimate the *PVs* of the two *BDTs*.

2.1. EM algorithm

In the 3×4 table of observed frequencies, the missing information is the true disease status of the individuals who are not verified with the *GS*, i.e. the missing data is the value of the variable D for the individuals among whom $V = 0$. This information is reconstructed in the *E* step of the algorithm and in the *M* step the values of the maximum likelihood estimators are imputed. Let us assume that that among the c_{ij} individuals who are not verified $(V = 0)$, d_{ij} have the disease and $c_{ij} - d_{ij}$ do not have it, with $i, j = 0,1$. Then the table of observed frequencies can be expressed in the form of a 2×4 table with frequencies $a_{ij} + d_{ij}$ for $D = 1$ and $b_{ij} + c_{ij} - d_{ij}$ for $D = 0$ (Table 1b). Let $\boldsymbol{\theta} = \left(\tau_1, \upsilon_1, \tau_2, \upsilon_2, p, \alpha_1, \alpha_0\right)^T$ be the vector of parameters. From the table of complete data, the log-likelihood function based on *n* individuals is

$$l(\boldsymbol{\theta}) = \sum_{i,j=0}^{1} \left(a_{ij} + d_{ij}\right) \log\left(\phi_{ij}\right) + \sum_{i,j=0}^{1} \left(b_{ij} + c_{ij} - d_{ij}\right) \log\left(\varphi_{ij}\right), \tag{5}$$

where $\phi_{ij} = P\left(T_1 = i, T_2 = i, D = 1\right)$ and $\varphi_{ij} = P\left(T_1 = i, T_2 = i, D = 0\right)$. The expressions of these probabilities in terms of the *PVs* are shown in Appendix *B* of the supplementary material. The components of the vector $\boldsymbol{\theta}$ are going to be estimated applying the *EM* algorithm. Therefore, let us suppose that $d_{ij}^{(k)}$ is the value of d_{ij} in the *k*th iteration of the *EM* algorithm, and $d^{(k)} = \sum_{i,j=0}^{1} d_{ij}^{(k)}$. The values of the *MLEs* in the *k*th iteration are calculated through the following equations

$$\hat{\tau}_1^{(k)} = \frac{1}{n_{10}+n_{11}} \sum_{j=0}^{1} \left(a_{1j} + d_{1j}^{(k)} \right), \quad \hat{\upsilon}_1^{(k)} = \frac{1}{n_{00}+n_{01}} \sum_{j=0}^{1} \left(b_{0j} + c_{0j} - d_{0j}^{(k)} \right),$$

$$\hat{\tau}_2^{(k)} = \frac{1}{n_{01}+n_{11}} \sum_{i=0}^{1} \left(a_{i1} + d_{i1}^{(k)} \right), \quad \hat{\upsilon}_2^{(k)} = \frac{1}{n_{00}+n_{10}} \sum_{i=0}^{1} \left(b_{i0} + c_{i0} - d_{i0}^{(k)} \right),$$

$$\hat{p}^{(k)} = \frac{a + d^{(k)}}{n}, \quad \hat{\alpha}_1^{(k)} = \frac{\left(a + d^{(k)} \right) \left(a_{11} + d_{11}^{(k)} \right)}{\left[\sum_{i=0}^{1} \left(a_{i1} + d_{i1}^{(k)} \right) \right] \left[\sum_{j=0}^{1} \left(a_{1j} + d_{1j}^{(k)} \right) \right]}, \tag{6}$$

$$\hat{\alpha}_0^{(k)} = \frac{\left(b + c - d^{(k)} \right) \left(b_{11} + c_{11} - d_{11}^{(k)} \right)}{\left[\sum_{i=0}^{1} \left(b_{i1} + c_{i1} - d_{i1}^{(k)} \right) \right] \left[\left\{ \sum_{j=0}^{1} \left(b_{1j} + c_{1j} - d_{1j}^{(k)} \right) \right\} \right]}.$$

The estimators in the $(k+1)th$ iteration of the algorithm are calculated applying equations (6) substituting the superindex k with $k+1$, where

$$d_{ij}^{(k+1)} = c_{ij} \frac{\hat{\phi}_{ij}^{(k)}}{\hat{\phi}_{ij}^{(k)} + \hat{\varphi}_{ij}^{(k)}}, \quad i, j = 0,1,$$

when $\hat{\phi}_{ij}^{(k)}$ and $\hat{\varphi}_{ij}^{(k)}$ are the estimators of the probabilities ϕ_{ij} and φ_{ij} in the *k*th iteration of the algorithm, and which are calculated substituting in the expressions of ϕ_{ij} and φ_{ij} (see Appendix *B* of supplementary material) the parameters with their respective estimators obtained in the *k*th iteration. As initial value $d_{ij}^{(0)}$ one can take any value between 0 and u_{ij}. The *EM* algorithm stops when the difference between the values of the log-likelihood functions of two consecutive iterations is lower than a value δ, for example $\delta = 10^{-10}$ or $\delta = 10^{-12}$. If the *EM* algorithm has converged in *K* iterations, we denote through $\hat{\boldsymbol{\theta}} = \left(\hat{\tau}_1, \hat{\upsilon}_1, \hat{\tau}_2, \hat{\upsilon}_2, \hat{p}, \hat{\alpha}_1, \hat{\alpha}_0 \right)^T$ the final estimators obtained. The estimators of the *PVs* obtained by applying the *EM* algorithm converge to the maximum likelihood estimators [10, 11] (proof can be seen in Appendix *C* of the supplementary material). The variances-covariances of $\hat{\boldsymbol{\theta}}$ are then estimated applying the *SEM* algorithm [15].

2.2. SEM algorithm

The estimation of the matrix of asymptotic variances-covariances of $\hat{\boldsymbol{\theta}}$ can be obtained applying the *SEM* algorithm [15], which is a computational method which estimates the variances-covariances matrix of a vector of estimators from the calculations performed in the application of the *EM* algorithm. Let $\sum_{\hat{\boldsymbol{\theta}}}$ be the variances-covariances matrix of $\hat{\boldsymbol{\theta}}$, Dempster et al [16] demonstrated that

$$\sum_{\hat{\boldsymbol{\theta}}} = I_{oc}^{-1}\left(I - DM\right)^{-1}, \tag{7}$$

where I is the identity matrix and $DM = I_{mis}I_{oc}^{-1}$, when I_{oc} is the Fisher information matrix of the complete data and I_{mis} is the Fisher information matrix of the missing data. The *SEM* algorithm consists of three phases: 1) assessment of the matrix I_{oc}^{-1}, 2) assessment of the matrix *DM*, and 3) assessment of the matrix $\sum_{\hat{\boldsymbol{\theta}}}$. The main phase is to calculate the elements of the DM matrix. The following three phases are then analysed.

The first phase consists of assessing the I_{oc}^{-1}. This matrix is the inverse of the Fisher information matrix of the complete data, i.e. $I_{oc} = -\dfrac{\partial^2 l(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}$, where $l(\boldsymbol{\theta})$ is the function (5) and each θ_i is one of the parameters of $\boldsymbol{\theta}$. This matrix is calculated from the last table after the application of the *EM* algorithm, substituting the parameters with their corresponding estimations obtained in the last iteration of the *EM* algorithm. If the *EM* algorithm has converged in *K* iterations, then the frequencies of the last 2×4 table are $a_{ij} + d_{ij}^{(K)}$ for $D=1$ and $b_{ij} + c_{ij} - d_{ij}^{(K)}$ for $D=0$.

The second part of the *SEM* algorithm consists of calculating the *DM* matrix. The elements of this matrix, denoted as r_{ij}, $i, j = 1,...,7$, are obtained applying the following algorithm:

INPUT: $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^{(k)} = \left(\tau_1^{(k)}, \upsilon_1^{(k)}, \tau_2^{(k)}, \upsilon_2^{(k)}, p^{(k)}, \alpha_1^{(k)}, \alpha_0^{(k)}\right)^T$.

Step 1: Calculate $\boldsymbol{\theta}^{(k+1)} = \left(\tau_1^{(k+1)}, \upsilon_1^{(k+1)}, \tau_2^{(k+1)}, \upsilon_2^{(k+1)}, p^{(k+1)}, \alpha_1^{(k+1)}, \alpha_0^{(k+1)}\right)$ applying the *EM* algorithm proposed in Section 2.1.

Step 2: Obtain the vectors $\boldsymbol{\theta}_1^{(k)} = \left(\tau_1^{(k)}, \hat{\upsilon}_1, \hat{\tau}_1, \hat{\upsilon}_2, \hat{p}, \hat{\alpha}_1, \hat{\alpha}_0\right)^T$, $\boldsymbol{\theta}_2^{(k)} = \left(\hat{\tau}_1, \upsilon_1^{(k)}, \hat{\tau}_2, \hat{\upsilon}_2, \hat{p}, \hat{\alpha}_1, \hat{\alpha}_0\right)^T$,..., $\boldsymbol{\theta}_7^{(k)} = \left(\hat{\tau}_1, \hat{\upsilon}_1, \hat{\tau}_2, \hat{\upsilon}_2, \hat{p}, \hat{\alpha}_1, \alpha_0^{(k)}\right)^T$, and for each one of these seven vectors run the first iteration of the *EM* algorithm taking $\boldsymbol{\theta}_i^{(k)}$ as the initial value of $\boldsymbol{\theta}$, and obtain the vectors $\hat{\boldsymbol{\theta}}_1^{*(k+1)}$, $\hat{\boldsymbol{\theta}}_2^{*(k+1)}$,..., $\hat{\boldsymbol{\theta}}_7^{*(k+1)}$.

Step 3: Calculating the elements of the *DM* matrix as

$$r_{ij}^{(k)} = \frac{\hat{\theta}_{ij}^{*(k+1)} - \hat{\theta}_j}{\theta_i^{(k)} - \hat{\theta}_i}, \quad i, j = 1, ..., 7,$$

where $\hat{\theta}_{ij}^{*(k+1)}$ is the *j*th component of vector $\hat{\boldsymbol{\theta}}_i^{*(k+1)}$, $\theta_i^{(k)}$ is the *i*th component of vector $\boldsymbol{\theta}^{(k)}$ and $\hat{\theta}_i$ is the *i*th component of vector $\hat{\boldsymbol{\theta}}$.

OUTPUT: $\boldsymbol{\theta}^{(k+1)}$ and $r_{ij}^{(k)}$, $i, j = 1, ..., 7$.

This algorithm is repeated until $\left|r_{ij}^{(k+1)} - r_{ij}^{(k)}\right| \leq \sqrt{\delta}$ [15]. Consequently, the lower δ is, the lower the numerical errors that are made when calculating the elements of the *DM* matrix, thus making fewer numerical errors in the estimation of the variances-covariances matrix $\sum_{\hat{\boldsymbol{\theta}}}$.

The last phase of the *SEM* algorithm consists of estimating the variances-covariances matrix $\sum_{\hat{\boldsymbol{\theta}}}$ applying equation (7). The estimated variances-covariances matrix is not normally symmetrical due to the numerical errors made in the calculation of *DM* matrix. The assessment of $\hat{\sum}_{\hat{\boldsymbol{\theta}}}$ is performed calculating the matrix $\Delta\hat{\sum}_{\hat{\boldsymbol{\theta}}} = \hat{I}_{oc}^{-1} DM \left(I - DM\right)^{-1}$, a matrix

which represents the increase in the variances-covariances estimated owing to the missing information. The smaller the value of δ is, the more symmetrical the matrix $\Delta\hat{\Sigma}_{\hat{\boldsymbol{\theta}}}$ will be, and therefore the more symmetrical $\hat{\Sigma}_{\hat{\boldsymbol{\theta}}}$ will be. Thus, the problem of the asymmetry of $\hat{\Sigma}_{\hat{\boldsymbol{\theta}}}$ is solved taking a value a very small value of δ [15].

2.3. Global Test

The global hypothesis test (2) to simultaneously compare the *PVs* of the two *BDTs* is equivalent to $H_0 : \mathbf{A}\boldsymbol{\eta} = 0$ vs $H_1 : \mathbf{A}\boldsymbol{\eta} \neq 0$, where $\boldsymbol{\eta} = \left(\tau_1, \upsilon_1, \tau_2, \upsilon_2\right)^T$ and $\mathbf{A}$ is a complete range matrix sized 2×4 whose elements are known constants, i.e.

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix}.$$

Applying the multivariate central limit theorem it is verified that $\sqrt{n}\left(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}\right) \xrightarrow[n\to\infty]{} N\left(\mathbf{0}, \Sigma_{\boldsymbol{\eta}}\right)$ where $\Sigma_{\boldsymbol{\eta}}$ is the variance-covariance matrix of $\boldsymbol{\eta}$. Then, the statistic $Q^2 = \hat{\boldsymbol{\eta}}^T \mathbf{A}^T \left(\mathbf{A}\hat{\Sigma}_{\hat{\boldsymbol{\eta}}} \mathbf{A}^T\right)^{-1} \mathbf{A}\hat{\boldsymbol{\eta}}$ is distributed according to a Hotelling's *T*-squared distribution dimension 2 and *n* degrees of freedom, where 2 is the dimension of the vector $\hat{\boldsymbol{\eta}}$. When *n* is large, the statistic Q^2 is distributed according to a central chi-squared distribution with 2 degrees of freedom when the null hypothesis is true, i.e.

$$Q^2 = \hat{\boldsymbol{\eta}}^T \mathbf{A}^T \left(\mathbf{A}\hat{\Sigma}_{\hat{\boldsymbol{\eta}}} \mathbf{A}^T\right)^{-1} \mathbf{A}\hat{\boldsymbol{\eta}} \xrightarrow[n\to\infty]{} \chi_2^2.$$

The matrix $\hat{\Sigma}_{\hat{\boldsymbol{\eta}}}$ is obtained from the matrix $\hat{\Sigma}_{\hat{\boldsymbol{\theta}}}$ eliminating the rows and columns corresponding to $\hat{p}$, $\hat{\alpha}_1$ and $\hat{\alpha}_0$.

The global hypothesis global test (2) can also be solved from the individual hypothesis test, i.e. $H_0 : \tau_1 = \tau_2$ and $H_0 : \upsilon_1 = \upsilon_2$, each one of them independently to an α error, when the corresponding test statistics are

$$z = \frac{\hat{\vartheta}_1 - \hat{\vartheta}_2}{\sqrt{\hat{V}ar\left(\hat{\vartheta}_1\right) + \hat{V}ar\left(\hat{\vartheta}_2\right) - 2\hat{C}ov\left(\hat{\vartheta}_1, \hat{\vartheta}_2\right)}}$$

both with normal standard distributions when the sample size is large and where $\hat{\vartheta}_i$ is $\hat{\tau}_i$ or $\hat{\upsilon}_i$. Another method to solve the global test consists of solving each one of the individual tests along with a method of multiple comparisons, such as the classic method of Bonferroni [17] or the Holm method [18]. The Bonferroni method consists of solving each individual test to an $\alpha/2$ error, and the Holm method is less conservative than the Bonferroni method.

3. Multiple imputation

Multiple imputation [19, 20, 21] is an alternative method to the *EM* algorithm which is used to solve problems with missing data. Multiple imputation (*MI*) consists of constructing *M* sets of complete data, with $M \geq 2$, obtained replacing the missing data with *M* sets imputed independently. From each complete dataset the parameters are estimated, thereby obtaining *M* estimators of each parameter. Then the *M* estimators of each parameter are combined properly to obtain a global estimator of each parameter and its variance. From these combined values it is possible to obtain confidence intervals for each parameter and also to solve the hypothesis test. In the context of the comparison of parameters of two *BDTs*, Harel and Zhou [12] studied the comparison of the sensitivities (specificities) of two *BDTs* in the presence of missing data *MAR* through confidence intervals applying *MI*.

We then study the simultaneous comparison of the *PVs* applying *MI*. Firstly, the *MICE* method is introduced and the hypothesis test is solved.

3.1. MICE Method

For the imputation of the missing data, we have applied the multiple imputation by chained equations (*MICE*), a method which is also known as fully conditional specification or sequential regression multivariate imputation. The *MICE* method requires us to assume that the missing data are *MAR*, and it has the advantage of being a very flexible method which can be used with binary, ordinal or continuous variables. White et al [22] provided a detailed explanation of the imputation of binary variables with the *MICE* method. Therefore, in our situation we have three random binary variables: T_1, T_2 and *D*. For the variables T_1 and T_2 there are no missing data, since the two *BDTs* have been applied to all of the individuals in a sample. Nevertheless, the variable *D* is missing for a subset of individuals in the sample, since the disease status is unknown for these individuals. Firstly, all missing values are filled in at random. The variable *D* is then regressed on the variables T_1 and T_2 through a logistic regression. The estimation is thus restricted to individuals with observed T_1 and T_2. The missing values in *D* are then replaced by simulated draws from the posterior predictive distribution of *D*. This process is called a cycle, and to stabilize the results, this process is repeated a determined number of times, finally obtaining a set of imputed data. In the situation that we study here, from the 3×4 table (see Table 1: Observed frequencies in the presence of partial verification) *M* 2×4 tables are imputed (see Table 1: Complete data), and from each one of these *M* 2×4 tables, we calculate the estimators of the positive (negative) *PVs* and their variances-covariances. As in the case of the *EM-SEM* algorithms, the comparison of the *PVs* can be solved from the global hypothesis test or from the individual hypothesis tests, each one of them to an α error or applying a multiple comparison method.

3.2. Global Test

The solution of multiparametric hypothesis tests in problems with missing data applying *MI* has been the subject of several studies. Li et al [23] proposed a Wald statistic based on the *F*-distribution to solve a multidimensional hypothesis test, and Li et al [24] solved the hypothesis test combining the *p*-values (or in an equivalent way, the Wald statistics) obtained in the *M* sets of imputed data from the same *F*-distribution. Meng and Rubin [25] solved the multidimensional hypothesis applying the likelihood ratio test.

In the situation studied here, let $x_{ij}^{(m)}$ $\left(y_{ij}^{(m)}\right)$ be the numbers of diseased (non-diseased) individuals among whom *Test* 1 leads to a result *i* and *Test* 2 leads to result *j*, with $i, j = 0,1$, in the *m*th complete dataset ($m = 1,...,M$) obtained by applying the *MICE* method. Let $\hat{\boldsymbol{\eta}}^{(m)} = \left(\hat{\tau}_1^{(m)}, \hat{\upsilon}_1^{(m)}, \hat{\tau}_2^{(m)}, \hat{\upsilon}_2^{(m)}\right)^T$ be the estimator of $\boldsymbol{\eta}$ in the *m*th complete dataset and $\overline{\boldsymbol{\eta}} = \frac{1}{M}\sum_{m=1}^{M} \hat{\boldsymbol{\eta}}^{(m)}$ the overall estimate of $\boldsymbol{\eta}$. In each one of the *M* sets of complete data, we estimate the *PVs* and their variances-covariances. Positive *PVs* are estimated as

$$\hat{\tau}_1^{(m)} = \frac{x_{11}^{(m)} + x_{10}^{(m)}}{x_{11}^{(m)} + x_{10}^{(m)} + y_{11}^{(m)} + y_{10}^{(m)}} \text{ and } \hat{\tau}_2^{(m)} = \frac{x_{11}^{(m)} + x_{01}^{(m)}}{x_{11}^{(m)} + x_{01}^{(m)} + y_{11}^{(m)} + y_{01}^{(m)}},$$

and negative *PVs* as

$$\hat{\upsilon}_1^{(m)} = \frac{y_{01}^{(m)} + y_{00}^{(m)}}{x_{01}^{(m)} + x_{00}^{(m)} + y_{01}^{(m)} + y_{00}^{(m)}} \text{ and } \hat{\upsilon}_2^{(m)} = \frac{y_{10}^{(m)} + y_{00}^{(m)}}{x_{10}^{(m)} + x_{00}^{(m)} + y_{10}^{(m)} + y_{00}^{(m)}}.$$

Appendix *A* shows their variances-covariances. The global hypothesis test (2) can be solved through the methods that are now described.

a) *Wald Test*

Let $\hat{\Sigma}^{(m)}$ be the estimated variances-covariances matrix of $\hat{\boldsymbol{\eta}}^{(m)}$. The matrix $\hat{\Sigma}^{(m)}$ is calculated from the *m*th dataset applying the delta method, and its elements are shown in Appendix *A*.

Let $\bar{\Sigma} = \frac{1}{M}\sum_{m=1}^{M}\hat{\Sigma}^{(m)}$, $\mathbf{B} = \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\boldsymbol{\eta}}^{(m)} - \bar{\boldsymbol{\eta}}\right)\left(\hat{\boldsymbol{\eta}}^{(m)} - \bar{\boldsymbol{\eta}}\right)^{T}$ and $r_1 = \left(1+M^{-1}\right)\text{trace}\left(\mathbf{B}\bar{\Sigma}^{-1}\right)/2$.

The matrix $\bar{\Sigma}$ measures the within imputation variability, the matrix $\mathbf{B}$ measures the between imputation variability and *r* is the estimated average odds ratio of the fractions of missing information. Then, Wald test statistic [23] for the global hypothesis test is

$$F_1 = \frac{\bar{\boldsymbol{\eta}}^{T}\mathbf{A}^{T}\left(\mathbf{A}\bar{\Sigma}\mathbf{A}^{T}\right)^{-1}\mathbf{A}\bar{\boldsymbol{\eta}}}{2\left(1+r_1\right)},$$

whose distribution is one *F* with 2 (the dimension of the vector $\bar{\boldsymbol{\eta}}$) and

$$l = \begin{cases} 4+\left(2M-6\right)\left[1+\dfrac{M-2}{\left(M-1\right)r_1}\right]^2, & \text{if } 2\left(M-1\right)>4 \\ \dfrac{3}{2}\left(M-1\right)\left(1+r_1^{-1}\right)^2, & \text{if } 2\left(M-1\right)\le 4 \end{cases}$$

degrees of freedom.

b) Combination of p-values

The solution of the global test can be made combining the *p*-values obtained in each one of the M complete datasets [24], or what amounts to the same thing, combining the Wald statistics. For the *m*th set of imputed data the Wald test statistic is $F^{(m)} = \left(\hat{\boldsymbol{\eta}}^{(m)}\right)^{T}\mathbf{A}^{T}\left(\mathbf{A}\hat{\Sigma}^{(m)}\mathbf{A}^{T}\right)^{-1}\mathbf{A}\hat{\boldsymbol{\eta}}^{(m)}$. Let $\bar{F} = \frac{1}{M}\sum_{m=1}^{M}F^{(m)}$ and $r_2 = \left(1+M^{-1}\right)\left[\frac{1}{M-1}\sum_{m=1}^{M}\left(\sqrt{F^{(m)}} - \overline{\sqrt{F}}\right)^2\right]$, where $\overline{\sqrt{F}} = \frac{1}{M}\sum_{m=1}^{M}\sqrt{F^{(m)}}$. The quantity $\bar{F}$ is the average Wald statistic and r_2 is $1+M^{-1}$ times the sample variance of $\left\{\sqrt{F^{(m)}},\ m=1,...,M\right\}$.

The statistic for the global hypothesis test is [25]

$$F_2 = \frac{1}{1+r_2}\left(\frac{\bar{F}}{2} - \frac{M+1}{M-1}r_2\right),$$

and the combined p-value is $p-value = P\left(F_{2,l} \geq F_2\right)$ with $l = 2^{-3/M}\left(M-1\right)\left(1+r_2^{-1}\right)^2$.

c) Combined likelihood-ratio tests

A third method to solve the global test is combining likelihood-ratio tests [25]. Let $\mathbf{z}^{(m)} = \left(x_{11}^{(m)}, x_{10}^{(m)}, x_{01}^{(m)}, x_{00}^{(m)}, y_{11}^{(m)}, y_{10}^{(m)}, y_{01}^{(m)}, y_{00}^{(m)}\right)^T$ be the vector of imputed frequencies in the *m*th complete dataset. Let $\phi_{ij}^{(m)}$ and $\varphi_{ij}^{(m)}$ be the probabilities corresponding to each cell of the imputed 2×4 table, whose expressions are similar to those given in Appendix *A* of supplementary material adding the superindex (*m*) to all of the parameters. Let $\boldsymbol{\psi}^{(m)} = \left(\phi_{11}^{(m)}, \phi_{10}^{(m)}, \phi_{01}^{(m)}, \phi_{00}^{(m)}, \varphi_{11}^{(m)}, \varphi_{10}^{(m)}, \varphi_{01}^{(m)}, \varphi_{00}^{(m)}\right)^T$ be the corresponding vector of probabilities. The complete-data log-likelihood function is

$$L^{(m)}\left(\boldsymbol{\psi}^{(m)}; \mathbf{z}^{(m)}\right) \propto \sum_{i,j=0}^{1} x_{ij}^{(m)} \log\left[\phi_{ij}^{(m)}\right] + \sum_{i,j=0}^{1} y_{ij}^{(m)} \log\left[\varphi_{ij}^{(m)}\right].$$

Maximizing this function it holds that $\hat{\phi}_{ij}^{(m)} = x_{ij}^{(m)}/n$ and $\hat{\varphi}_{ij}^{(m)} = y_{ij}^{(m)}/n$, which are the non-restricted estimators of $\boldsymbol{\psi}^{(m)}$ (i.e. $\hat{\boldsymbol{\psi}}^{(m)}$), with $i, j = 0,1$. If in the *m*th set of imputed data the null hypothesis $H_0 : \left(\tau_1^{(m)} = \tau_2^{(m)} \text{ and } \upsilon_1^{(m)} = \upsilon_2^{(m)}\right)$ is true, then it is easy to show that the log-likelihood function is

$$\begin{aligned} L_0^{(m)}\left(\boldsymbol{\psi}_0^{(m)}, \mathbf{z}^{(m)}\right) \propto\ & x_{11}^{(m)} \log\left[\phi_{11}^{*(m)}\right] + \left(x_{10}^{(m)} + x_{01}^{(m)}\right)\log\left[\phi_{10}^{*(m)}\right] + x_{00}^{(m)} \log\left[\phi_{00}^{*(m)}\right] \\ & + y_{11}^{(m)} \log\left[\varphi_{11}^{*(m)}\right] + \left(y_{10}^{(m)} + y_{01}^{(m)}\right)\log\left[\varphi_{10}^{*(m)}\right] + y_{00}^{(m)} \log\left[\varphi_{00}^{*(m)}\right], \end{aligned}$$

where $\phi_{ij}^{*(m)}$ and $\varphi_{ij}^{*(m)}$ are the probabilities subject to the null hypothesis. Maximizing this function it holds that

$$\hat{\phi}_{ij}^{*(m)} = \begin{cases} x_{ii}^{(m)}/n, & \text{if } i = j \\ \left[x_{10}^{(m)} + x_{01}^{(m)}\right]/(2n), & \text{if } i \neq j \end{cases} \text{ and } \hat{\varphi}_{ij}^{*(m)} = \begin{cases} y_{ii}^{(m)}/n, & \text{if } i = j \\ \left[y_{10}^{(m)} + y_{01}^{(m)}\right]/(2n), & \text{if } i \neq j, \end{cases}$$

which are the estimators of $\boldsymbol{\psi}^{(m)}$ subject to the null hypothesis (i.e. $\hat{\boldsymbol{\psi}}_0^{(m)}$), with $i, j = 0,1$. Performing algebraic operations, the likelihood-ratio test statistic for the global test $H_0 : \left(\tau_1^{(m)} = \tau_2^{(m)} \text{ and } \upsilon_1^{(m)} = \upsilon_2^{(m)}\right)$ is

$$F_3^{(m)} = 2\left[L^{(m)}\left(\hat{\boldsymbol{\psi}}^{(m)}; \mathbf{z}^{(m)}\right) - L_0^{(m)}\left(\hat{\boldsymbol{\psi}}_0^{(m)}; \mathbf{z}^{(m)}\right)\right] =$$

$$2\left[x_{10}^{(m)} \log\left(\frac{2x_{10}^{(m)}}{x_{10}^{(m)} + x_{01}^{(m)}} \right) + x_{01}^{(m)} \log\left(\frac{2x_{01}^{(m)}}{x_{10}^{(m)} + x_{01}^{(m)}} \right) + y_{10}^{(m)} \log\left(\frac{2y_{10}^{(m)}}{y_{10}^{(m)} + y_{01}^{(m)}} \right) + y_{01}^{(m)} \log\left(\frac{2y_{01}^{(m)}}{y_{10}^{(m)} + y_{01}^{(m)}} \right)\right].$$

Let $\bar{F}_3 = \frac{1}{M}\sum_{m=1}^{M} F_3^{(m)}$ be the average of likelihood-ratio statistics, and let $\bar{\boldsymbol{\psi}} = \frac{1}{M}\sum_{m=1}^{M} \hat{\boldsymbol{\psi}}^{(m)}$ and $\bar{\boldsymbol{\psi}}_0 = \frac{1}{M}\sum_{m=1}^{M} \hat{\boldsymbol{\psi}}_0^{(m)}$ be the vectors whose components are the measures of the estimators in the *M* sets of imputed data subject to the non-restricted model and subject to the null hypothesis, respectively. Let

$$F_3^* = \frac{2}{M}\sum_{m=1}^{M}\left[L^{(m)}\left(\bar{\boldsymbol{\psi}}; \mathbf{z}^{(m)}\right) - L_0^{(m)}\left(\bar{\boldsymbol{\psi}}_0; \mathbf{z}^{(m)}\right)\right],$$

which is the measure of the likelihood-ratio statistics each one of which is assessed in $\bar{\boldsymbol{\psi}}$ and $\bar{\boldsymbol{\psi}}_0$, and let $r_3 = \frac{M+1}{2(M-1)}\left(\bar{F}_3 - F_3^*\right)$. Finally, the likelihood-ratio test statistic for the global hypothesis test is [25]

$$F_3 = \frac{F_3^*}{2\left(1 + r_3\right)},$$

which is distributed according to a *F*-distribution with 2 and

$$l = \begin{cases} 4 + \left(2M - 6\right)\left[1 + \dfrac{M-2}{(M-1)r_3}\right]^2, & \text{if } 2(M-1) > 4 \\ \dfrac{3}{2}(M-1)\left(1 + r_3^{-1}\right)^2, & \text{if } 2(M-1) \le 4 \end{cases}$$

degrees of freedom.

3.3. Individual tests

As with the *EM-SEM* algorithms, the global hypothesis test can also be solved from the individual tests applying *MI* with methods that compare the positive *PVs* and the negative *PVs* independently or with a method of multiple comparisons. The methods that are going to be considered to individually compare the *PVs* are those of Leisenring et al [3], Wang et al [4] and Kosinski [5]. For the method by Wang et al and for the method by Kosinski, the combination of results is achieved applying the rules of Rubin [19]. For the method by Leisenring et al, the combination of results is achieved calculating the average statistic. The test statistics of the method by Wang et al and of the method by Kosinski are of the type $\hat{\theta}\Big/\sqrt{\hat{V}ar\left(\hat{\theta}\right)}$, and therefore the combination of results is achieved applying the rules of Rubin. Nevertheless, in the case of the method by Leisenring et al, the test statistic to compare the equality of the two positive (negative) *PVs* is not of the type $\hat{\theta}\Big/\sqrt{\hat{V}ar\left(\hat{\theta}\right)}$, and therefore the rules of Rubin cannot be applied.

For the method of Wang et al and for the method of Kosinski the results obtained are then combined applying the rules of Rubin. Firstly, we calculate the overall estimate of the difference between the two positive *PVs*, i.e. $\bar{\tau}=\frac{1}{M}\sum_{m=1}^{M}\hat{\tau}^{(m)}$, where $\hat{\tau}^{(m)}=\hat{\tau}_1^{(m)}-\hat{\tau}_2^{(m)}$. The variance of $\bar{\tau}$ is $\hat{V}ar\left(\bar{\tau}\right)=\bar{V}ar\left(\hat{\tau}\right)+\frac{1}{M+1}B$, where $\bar{V}ar\left(\hat{\tau}\right)=\frac{1}{M}\sum_{m=1}^{M}\hat{V}ar\left(\hat{\tau}^{(m)}\right)$ is the within imputation variance (the average of the complete data variance estimates) and $B=\frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\tau}^{(m)}-\bar{\tau}\right)^2$ is the between imputation variance (the variance of the complete data point estimates). Finally, the test statistic for the test $H_0:\tau_1=\tau_2$ is $\frac{\bar{\tau}}{\sqrt{\hat{V}ar\left(\bar{\tau}\right)}}$, whose

distribution is (Rubin, 1987) a *t*-distribution with $v=(M-1)\left(1+\frac{M}{M+1}\frac{\hat{V}ar(\bar{\tau})}{B}\right)$ degrees of freedom. The comparison of the negative *PVs* is made in a similar way substituting τ with υ.

For the method by Leisenring et al, in the *m*th complete dataset the test statistic for $H_0:\tau_1=\tau_2$ is $z_\tau^{(m)}$ (its expression can be seen in Appendix *A*), whose distribution is a normal standard one when the sample size is large. Then for the Central Limit Theorem, the average of all of the test statistics $\bar{z}_\tau=\frac{1}{M}\sum_{m=1}^{M}z_\tau^{(m)}$ has a normal standard distribution when *M* is large. The process for the test $H_0:\upsilon_1=\upsilon_2$ is similar to the previous one.

4. Simulation experiments

Monte Carlo simulation experiments were carried out to study the type I errors and the powers of the hypothesis tests studied in Sections 2 and 3, as well as the relative biases of the estimators of the *PVs* obtained with both methods. These experiments consisted of generating $N=10000$ random samples of multinomial distributions sized $n=\{50,100,200,500,1000,2000\}$, and whose probabilities were calculated from equations (4). As *PVs* we considered the values $\{0.70,0.75,0.80,0.85,0.90,0.95\}$, which are values that appear quite frequently in clinical practice, and as disease prevalence we took the values $p=\{25\%,50\%,75\%\}$. Once the *PVs* and *p* are set, the *Se* and the *Sp* of each *BDT* were calculated from equations (3). As values of the covariances (α_i) we considered intermediate and high values. Finally, the probabilities of the multinomial distributions were calculated applying equations (4). Therefore, the probabilities of the multinomial distributions were calculated from the *PVs*, and there was no previous setting of the values of sensitivities and specificities of the *BDTs*. The simulation experiments were designed in such a way that in all

of the samples generated it is possible to apply the *EM-SEM* algorithms and *MI*. Therefore, if in a sample a frequency a_{ij} (or b_{ij}) is zero then it is not possible to apply *MI* (it is not possible to apply logistic regression to impute the missing data), then this sample was discarded and another one was generated instead until completing the *N* samples. Regarding the *EM-SEM* algorithms, we established as a stop criterion $\delta = 10^{-12}$ and $\sqrt{\delta} = 10^{-6}$ respectively, and as initial values of the *EM* algorithm the values $d_{ij}^{(0)} = c_{ij}/2$ are used. Regarding *MI*, for each one of the *N* random samples $M = 20$ complete data sets were generated and 100 cycles were performed. In the first phase simulations were made considering $M = 20$ and $M = 50$ and performing 100 and 200 cycles in each case, obtaining very similar results; therefore, to reduce computation time we finally considered $M = 20$ and 100 cycles. For all of the study, we set as the nominal error $\alpha = 5\%$, and considered that a method overwhelms the nominal error or exceeds it too much when its type I error is higher than 7%. The simulation experiments were carried out with *R* [26] and for the *MI* we used the "mice" library [27].

Therefore, in the simulation experiments we studied and compared the type I errors and the powers of sixteen different methods to solve the global hypothesis test (2). Of the sixteen methods, four are based on the *EM-SEM* algorithms and the other twelve on *MI*. The methods based on the *EM-SEM* algorithms are: (a) a global hypothesis global test based on the chi-squared distribution with $\alpha = 5\%$; (b) an individual comparison of the positive *PVs* and the negative *PVs* with $\alpha = 5\%$, Bonferroni and Holm. The twelve methods based on *MI* are: (a) a global hypothesis test applying the Wald method, the combination of *p*-values and the combined likelihood ratio tests, all of them with $\alpha = 5\%$; (b) an individual comparison of the positive *PVs* and the negative *PVs* applying the methods of Leisenring et al, Wang et al and Kosinski, each of them with $\alpha = 5\%$, Bonferroni and Holm.

4.1. EM-SEM algorithms

Table 2 shows some of the results obtained for the type I errors of the different methods applying the *EM-SEM* algorithms. In this Table, *EM-SEM-Global* consists of solving the global hypothesis test, *EM-SEM-Individual* consists of comparing the *PVs* solving the individual hypothesis tests each of them to an error $\alpha = 5\%$, and *EM-SEM-Bonferroni* consists of comparing the *PVs* solving the individual hypothesis tests along with the Bonferroni method to an error $\alpha = 5\%$. The results obtained applying the Holm method to an error $\alpha = 5\%$ are not shown as they are practically the same as those obtained with the Bonferroni method. In general terms, the type I errors of all of the methods increase when the verification probabilities increase, and decrease when the covariances α_i increase. In general terms, all of the methods are very conservative when the sample size is small $(n = 50)$ or moderate $(n = 100 - 200)$. When the sample size is large $(n \geq 500)$, depending on the verification probabilities and on the covariances α_i, the global test has a type I error that fluctuates around the nominal error. On some occasions, above all when the verification probabilities are low or the covariances are high, the type I error of the global test may slightly exceed the nominal error without actually overwhelming it. Method *EM-SEM-Individual* may overwhelm the nominal error, above all when the sample size is large. Method *EM-SEM-Bonferroni* has a type I error whose behaviour is very similar to that of the global test.

==== INSERT TABLE 2 HERE ====

Regarding the powers of these methods, Table 3 shows some results. The power of these methods increases when there is an increase in in the verification probabilities, whereas the

covariances α_i do not have a clear effect upon the power. All of the methods have a very small power when the sample size is small $(n=50)$ or moderate $(n=100-200)$, and it is necessary to have a large sample size (depending on the verification probabilities) so that the power is higher (over 80%). Although there are no clear rules, in general terms the global test is normally more powerful than Method 2. Regarding the method *EM-SEM-Individual*, there are also no clear rules, sometimes the global test is more powerful and on other occasions the method *EM-SEM-Individual* is more powerful (it is a method that easily overwhelms the nominal error), depending on the values that the *PVs* take.

==== INSERT TABLE 3 HERE ====

From the results of the simulation experiments applying the *EM-SEM* algorithms, the method to compare the *PVs* of the two *BDTs* with the best asymptotic behaviour is the global test, since its type I error does not exceed the nominal error too much and, in general terms, it has more power than the method *EM-SEM-Bonferroni* (this is a method whose type I error also does not exceed the nominal error too much). Method *EM-SEM-Individual* may exceed the nominal error too much and, therefore, lead to false significances.

The same previous results are obtained if the global test is solved by applying the *ML* method [10, 11], because the estimators obtained through the *EM-SEM* algorithms converge to the *ML* estimators.

4.2. Multiple Imputation

Table 4 shows the results obtained for the type I errors through *MI* for the same scenarios as Table 2. In this table, *Leisenring-Individual*, *Wang-Individual* and *Kosinski-Individual*, refers to the individual comparison of the *PVs* applying the method of Leisenring et al with $\alpha = 5\%$,

the method of Wang et al with $\alpha = 5\%$ and the method of Kosinski with $\alpha = 5\%$, respectively. Similarly, *Leisenring-Bonferroni*, *Wang-Bonferroni* and *Kosinski-Bonferroni*, refers to the individual comparison of the *PVs* applying the method of Leisenring et al, Wang et al and Kosisnki, respectively, along with the Bonferroni method and $\alpha = 5\%$. The results obtained applying the Holm method are not shown as they are practically identical those obtained with Bonferroni. As with the *EM-SEM* algorithms, the type I errors of all the methods based on *MI* increase when the verification probabilities increase, and decrease when the covariances α_i increase.

Regarding the global tests, the type I error of the test based on the combination of *p*-values is very similar to the type I error of the combined likelihood-ratio tests, both of which fluctuate around the nominal error when the sample size is large. The global test based on the Wald test is very conservative (even when the sample size is large), and its type I error is smaller than that of the other two methods.

The methods based on the individual comparisons to an error $\alpha = 5\%$ (*Leisenring-Individual*, *Wang-Individual* and *Kosinski-Individual*) have type I errors that may exceed the nominal error too much, above all when the sample size is large. Therefore, these methods may lead to an excess of false significances. Regarding the methods based on the individual tests along with Bonferroni, the *Leisenring-Bonferroni* method has a type I error which may exceed the nominal error too much when the sample size is large. *Kosinski-Boferroni* method has a type I error with better fluctuations around the nominal error (with exceeding it too much) than the *Wang-Bonferroni* method, above all when the sample size is large. In general terms, there is no important difference between the type I error of the *Kosinski-Bonferroni* method and the type I error of the global test based on the combination of *p*-values (or combined likelihood-ratio tests), above all when the sample size is large.

==== INSERT TABLE 4 HERE ====

Regarding the powers, Table 5 shows the results for the same scenarios given in Table 3. The power of all of the methods increases when the verification probabilities increase, whereas the covariances α_i do not have a clear effect upon the power.

In general terms, the global test based on the combination of *p*-values is more powerful than the combined likelihood-ratio tests, and the difference is greater when the verification probabilities are low than when they are high. Moreover, both methods are more powerful than the Wald test (as this test is very conservative in relation to the other two).

Comparing the global test based on the combination of *p*-values in relation to the *Leisenring-Individual*, *Wang-Individual* and *Kosinski-Individual* methods, there are no clear rules about their behaviour. Sometimes the global test is more powerful and on other occasions these methods are more powerful (methods which may clearly overwhelm the nominal error), depending on the verification probabilities and on the values that the *PVs* take.

Regarding the *Leisenring-Bonferroni* method, in general terms the global test based on the combination of *p*-values is less powerful, due to the fact that the type I error of the *Leisenring-Bonferroni* method (which may clearly overwhelm the nominal error) is greater than that of the global test (which does not overwhelm the nominal error). Regarding the *Wang-Bonferroni* method and the *Kosinski-Bonferroni* method, there is no important difference between their powers, and these powers are a little lower than those of the global test based on the combination of *p*-values, above all when the sample size is large.

Having analysed the results of the simulation experiments applying *MI*, the method to compare the *PVs* of two *BDTs* with the best asymptotic behaviour is the global test based on the combination of *p*-values, since its type I error does not overwhelm the nominal error and

its power is somewhat higher than that of the other methods which do not overwhelm the nominal error.

==== INSERT TABLE 5 HERE ====

4.3. Relative biases

Table 6 shows some results for the relative biases of the estimators of the *PVs* applying the *EM* algorithm and applying *MI*. The relative biases decrease when the verification probabilities increase whereas the covariances α_i have practically no effect on the estimators obtained applying both methods. In general terms, the difference between the relative biases obtained with both methods is very small, and therefore both methods lead to estimations which on average are very similar to each other.

==== INSERT TABLE 6 HERE ====

4.4. EM-SEM algorithms when some frequency is zero

The *EM-SEM* algorithms can be applied when some frequency a_{ij} or b_{ij} is equal to zero. Simulation experiments have been carried out to study the asymptotic behavior of these algorithms in this situation. These experiments have been designed in a similar way to the previous case, but the samples in which some frequency a_{ij} or b_{ij} is equal to zero have not been eliminated. Table 7 shows the type I errors and the powers for the same scenarios given in Tables 2 and 3. In general terms, the conclusions are the same as those given in Section 4.1.

==== INSERT TABLE 7 HERE ====

4.5. EM-SEM algorithms or MI?

Comparing the type I error of the global test based on the *EM-SEM* algorithms (Table 2) with the type I error of the global test based on the combination of *p*-values (Table 4), In general terms, there is no important difference among them. Regarding the power of both methods (Table 3 and Table 5), in general terms the power of the test based on *MI* (combination of *p*-values) is a little higher than the power of the test based on the *EM* and *SEM* algorithms when the sample is small or moderate; when the sample is large, the power of both methods is very similar. Therefore, from the simulation experiments, it is possible to give the following general rule of application based on the sample size:

a) Apply the *EM-SEM* algorithms or *MI* when the sample size is large.

b) Apply *MI* based on the combination of *p*-values when the sample size is small or moderate.

4.6. Causes of the significance

When the *EM-SEM* algorithms are applied, if the global test is not significant to an α error then we do not reject the homogeneity of the *PVs* of both *BDTs.* If the global hypothesis test is significant to an α error, then the causes of the significance are investigated solving the individual hypothesis tests, i.e. $H_0 : \tau_1 = \tau_2$ and $H_0 : \upsilon_1 = \upsilon_2$, along with the Bonferroni (or Holm) method to an α error. The application of the individual tests along with the Bonferroni (or Holm) method is justified, just as in the simulation experiments, by the fact that this method has a type I error which does not overwhelm the nominal error.

When *MI* is applied, the investigation of the causes of the significance is made in a similar way to in the previous case. We solve the global test based on the combination of *p*-values to an α error and if the test is not significant then we do not reject the homogeneity of the *PVs*. If the global test is significant, then the causes of the investigation are studied solving the

individual tests applying the Kosinski method along with the Bonferroni (or Holm) method to an α error for any sample size, although it is also possible to apply the method of Wang et al along with Bonferroni (or Holm) when the sample size is small or moderate. When the sample size is large, the Kosinski method has a better type I error behaviour than the method of Wang et al, and their powers are very similar. When the sample size is small or moderate, the Kosinski method and the method of Wang et al have very similar type I errors and powers.

5 Empv and Mipv programmes

Two programmes in *R* were written: Empv (*EM* algorithm for *PVs*) and Mipv (Multiple imputation for *PVs*). The Empv programme compares the *PVs* of two *BDTs* in the presence of partial disease verification applying the *EM-SEM* algorithms and the Mipv programme solves the same problem applying *MI* (*MICE* method). Both programmes are available as supplementary material for this article.

The Empv programme is run with the command "empv$(a_{11}, a_{10}, a_{01}, a_{00}, b_{11}, b_{10}, b_{01}, b_{00}, c_{11}, c_{10}, c_{01}, c_{00})$". The programme checks that the values of the frequencies are feasible (for example, that there are no negative values, frequencies with decimals, etc…). By default, the stop criterion of the *EM* algorithm is 10^{-12}, the confidence for the calculation of the intervals is 95% and, as initial values of the *EM* algorithm, the values $d_{ij}^{(0)} = c_{ij}/2$ are used. The programme provides the estimations of the *PVs* and their corresponding standard errors, the inverse Fisher information matrix of the complete data, the *DM* matrix, the estimated variances-covariances matrix of the *PVs*, the test statistics and *p*-values of the global test and of the individual tests, as well as the confidence intervals for the difference of the two positive (negative) *PVs*.

The Mipv programme solves the problem applying the *MICE* method and is run with the command "mipv$(a_{11}, a_{10}, a_{01}, a_{00}, b_{11}, b_{10}, b_{01}, b_{00}, c_{11}, c_{10}, c_{01}, c_{00})$". The programme checks that the values of the frequencies are feasible (for example, that there is no frequency equal to zero, negative values, frequencies with decimals, etc…). By default, 20 complete datasets are generated, for each complete dataset 100 cycles are performed and the confidence for the calculation of the intervals is 95%. The programme provides the estimations of the *PVs* and their corresponding standard errors, the estimated variances-covariances matrix of the *PVs*, the test statistics and *p*-values of the global test and of the individual tests applying the method of Wang et al and the Kosinski method, as well as the confidence intervals for the difference of the two positive (negative) *PVs*.

6. Application

The results obtained were applied to the study by Hall et al [28] on the diagnosis of Alzheimer's disease. Hall et al used two diagnostic tests for the diagnosis of Alzheimer's disease: a new diagnostic test (*Test* 1) based on a cognitive test applied to the patient and in another relative test to another person who knows the patient, and a standard diagnostic test based on a cognitive test (*Test* 2), and as the *GS* they used a clinical assessment (neurological examination, computerized tomography, neuropsychological and laboratory tests…). Table 8 shows the results obtained applying the two diagnostic tests to a sample of 588 patients over 75 years of age, and where the random variable T_1 models the result of *Test* 1, T_2 models the result of *Test* 2 and D models the result of the *GS*. The study by Hall et al corresponds to a two-phase study: in the first phase, the two diagnostic tests were applied to all of the patients, and in the second phase the *GS* was only applied to a subset of patients depending on the results of both diagnostic tests. Consequently, it is assumed that the verification process is

MAR. As the sample size is large, the comparison of the *PVs* can be made applying the *EM-SEM* algorithms and *MI*.

==== INSERT TABLE 8 HERE ====

Table 8 (Results with *EM-SEM* algorithms) shows the estimations of the *PVs* and their standard errors (*st.er.*) obtained applying the Empv programme with the command "empv(31,5,3,1,25,10,19,55,22,6,65,346)". The *EM* algorithm has converged in 186 iterations. The inverse Fisher information matrix of the complete data, the *DM* matrix and the variances-covariances matrix can be seen when the programme is run. The test statistic for the global test is $Q^2 = 30.097$ and the *p*-value is 2.914×10^{-7}, and therefore with $\alpha = 5\%$ we reject the equality of the *PVs*. Solving the individual tests $H_0 : \tau_1 = \tau_2$ and $H_0 : \upsilon_1 = \upsilon_2$ it holds that the respective test statistics are $z = 3.251$ $\left(p\text{-value} = 0.001\right)$ and $z = 0.362$ $\left(p\text{-value} = 0.718\right)$. Applying the Bonferroni (or Holm) method with $\alpha = 5\%$, we reject the equality of the two positive *PVs* and we do not reject the equality of the two negative *PVs*. When the two diagnostic are applied to the population being studied, the positive *PV* of *Test* 1 is significantly higher than that of *Test* 2 (95% confidence interval for the difference: 0.069 to 0.278). If the problem is solved applying the maximum likelihood method [11], the same results are obtained as applying the *EM-SEM* algorithms.

Table 8 (Results with the *MICE* method) shows the estimations obtained applying the Mipv programme with the command "mipv (31,5,3,1,25,10,19,55,22,6,65,346)". The test statistic for the global test applying the method of the combination of *p*-values is $F_2 = 15.974$ and the *p*-value is 1.291×10^{-7}, and therefore with an error $\alpha = 5\%$ we reject the equality of the *PVs* of both *BDTs*. Solving the individual tests $H_0 : \tau_1 = \tau_2$ and $H_0 : \upsilon_1 = \upsilon_2$ applying the

Kosinski method it holds that the respective test statistics are $z = 4.808$ $\left(p\text{-value} = 1.527 \times 10^{-7}\right)$ and $z = 0.747$ $\left(p\text{-value} = 0.455\right)$. Applying the Bonferroni (or Holm) method with $\alpha = 5\%$, we reject the equality of the two positive *PVs* and we do not reject the equality of the two negative *PVs*. When the two diagnostic tests are applied to the population being studied, the positive *PV* of *Test* 1 is significantly greater than that of *Test* 2 (95% confidence interval for the difference: 0.101 to 0.239).

As general conclusions of the simulation experiments, it has been obtained that the *EM-SEM* algorithms and *MI* can be applied when the sample size is large, as in this example. In this example, it can be observed how with both methods, *EM-SEM* algorithms and *MI*, the results obtained are very similar, both in terms of the point estimators and of their variances-covariances. Moreover, the conclusions are the same: we reject the equality between the two positive *PVs* and we do not reject the equality between the two negative *PVs*. Therefore, both methods lead to the same conclusions and both are equally valid.

7. Discussion

This manuscript studies the computational methods to solve this problem in the presence of missing data. The comparison of the two positive *PVs* and of the two negative *PVs* was studied simultaneously applying the *EM-SEM* algorithms and *MI*. With both methods it is required that the missing data be *MAR*, and therefore if the verification process conditionally depends on the disease status then this assumption is not verified and the methods cannot be applied.

Simulation experiments were carried out to study the asymptotic behaviour of the global test of comparison of the *PVs*, and of other alternative methods, both with the *EM-SEM* algorithms and with *MI*, giving some general rules of application based on the sample size. In general terms, *MI* can be applied to any sample size, whereas the *EM-SEM* algorithms require

the sample size to be large. If the global test is not significant to an α error then we do not reject the homogeneity of the *PVs* of both *BDTs*. If the global test is significant then the causes of the significance are investigated comparing the two positive *PVs* and the two negative *PVs* independently and a method of multiple comparisons is applied, such as Bonferroni or Holm, which are very easy methods to apply. This procedure is very similar to an analysis of variance: the global test is solved and if this is significant paired comparisons are carried out and a method of multiple comparisons is applied.

The application of the *EM-SEM* algorithms leads to the same results as the application of the Marín-Jiménez and Roldán-Nofuentes method [11]. These authors simultaneously compared the *PVs* of multiple *BDTs* obtaining the estimators through the maximum likelihood method and estimating the variances-covariances applying the delta method. The advantage of the *EM-SEM* algorithms over the *ML* method is that the former can be applied when some a_{ij} or b_{ij} frequency is equal to zero, while with the *ML* method, the variance-covariance matrix cannot be estimated if some frequency a_{ij} or b_{ij} is equal to zero.

As with the *ML* method, *MI* can only be applied if all frequencies a_{ij} or b_{ij} are greater than zero. If some frequency a_{ij} or b_{ij} is equal to zero then it is not possible to generate complete datasets through logistic regression. Furthermore, the Wald method and the combination of *p*-values require that the fractions of missing information for all components of the parameter vector to be equal. When there are important differences between the fractions of missing information and these are large, there may be an important effect on the size of the test and on the power [23, 24]. Traditionally, Rubin [19] recommended imputing five complete datasets in order to be able to apply *MI*. In the situation analysed in this manuscript, in the initial simulation experiments $M = \{20, 50\}$ compete datasets were considered, and the results were very similar, and therefore it was decided to generate

$M = 20$ complete datasets to save computation time. The simulation experiments demonstrated that the global test based on the combination of *p*-values (which is the test based on *MI* with the best asymptotic behaviour) has a type I error which is very similar to the global test based on the *EM-SEM* algorithms. Regarding power, that of the global test based on the combination of *p*-values is a little higher than that of the global test based on the *EM-SEM* algorithms when the sample is small or moderate, and they are very similar when the sample is large. Therefore, the number of complete datasets was sufficiently large, and did not have any negative effect on the size and the power of the global test.

Disclosure statement

No potential conflict of interest was reported by the author.

References

1. Pepe, M.S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction.* New York: Oxford University Press.
2. Zhou, X.H., Obuchowski, N.A. and McClish, D.K. (2011). *Statistical Methods in Diagnostic Medicine (Second Edition)*. New Jersey: John Wiley & Sons.
3. Leisenring, W., Alonzo, T. and Pepe, M.S. (2000). Comparisons of predictive values of binary medical diagnostic tests for paired designs. Biometrics, 56, 345-351.
4. Wang, W., Davis, C.S. and Soong, S.J. (2006). Comparison of predictive values of two diagnostic tests from the same sample of subjects using weighted least squares. Statistics in Medicine, 25, 2215-2229.
5. Kosinski, A.S. (2013). A weighted generalized score statistic for comparison of predictive values of diagnostic tests. Statistics in Medicine, 32, 964-977.

6. Tsou, T.S. (2018). A new likelihood approach to inference about predictive values of diagnostic tests in paired designs. Statistical Methods in Medical Research, 27, 541-548.
7. Roldán-Nofuentes, J.A., Luna del Castillo, J.D. and Montero-Alonso, M.A. (2012). Global hypothesis test to simultaneously compare the predictive values of two binary diagnostic tests. Computational Statistics and Data Analysis, 56, 1161-1173.
8. Begg, C.B. and Greenes, R.A. (1983). Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics, 39, 207-215.
9. Zhou, X.H. (1998). Comparing accuracies of two screening tests in a two-phase study for dementia. The Journal of the Royal Statistical Society, Series C Applied Statistics, 47, 135-147.
10. Roldán Nofuentes, J.A. and Luna del Castillo, J.D. (2008). The effect of verification bias on the comparison of predictive values of two binary diagnostic tests. Journal of Statistical Planning and Inference, 138, 950-963.
11. Marín-Jiménez, A.E. and Roldán-Nofuentes, J.A. (2014). Global hypothesis test to compare the likelihood ratios of multiple binary diagnostic tests with ignorable missing data. SORT - Statistics and Operations Research Transactions, 38, 305-324.
12. Harel, O. and Zhou, X.H. (2007). Multiple imputation for the comparison of two screening tests in two-phase Alzheimer studies. Statistics in Medicine, 26, 2370-2388.
13. Roldán Nofuentes, J.A. and Luna del Castillo, J.D. (2008). EM algorithm for comparing two binary diagnostic tests when not all the patients are verified. Journal of Statistical Computation and Simulation, 78, 19-35.
14. Berry G., Smith, C., Macaskill, P. and Irwig L. (2002). Analytic methods for comparing two dichotomous screening or diagnostic tests applied to two populations of differing disease prevalence when individuals negative on both tests are unverified. Statistics in Medicine, 21, 853-862.

15. Meng, X. and Rubin, D.B. (1991). Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. Journal of the American Statistical Association, 86, 899-909.

16. Dempster, A., Laird, N. and 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.

17. Bonferroni, C.E. (1936). *Teoria statistica delle classi e calcolo delle probabilità*, Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 8, 3-62.

18. Holm, S. (1979). A simple sequential rejective multiple testing procedure. Scandinavian Journal of Statistics, 6, 65-70.

19. Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

20. Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall.

21. Little, R.J.A. and Rubin, D.B. (2002). *Statistical analysis with missing data (Second Edition)*. New Jersey: Wiley.

22. White, I.R., Royston, P. and Wood, A.M. (2011). Multiple imputation using chained equations: issues and guidance for practice. Statistics in Medicine, 30, 377-399.

23. Li, K.H, Raghunathan, T.E. and Rubin, D.B. (1991). Large-sample significance levels from multiply imputed data using moment-based statistics and an F reference distribution. Journal of the American Statistical Association, 86, 1065-1073.

24. Li, K.H, Meng, X.L., Raghunathan, T.E. and Rubin, D.B.(1991). Significance levels from repeated p-values with multiply-imputed data. Statistica Sinica, 1, 65-92.

25. Meng, X.L., Raghunathan, T.E. and Rubin, D.B. (1992). Performing likelihood ratio tests with multiply-imputed data sets. Biometrika, 79, 103-111.

26. R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

27. van Buuren, S. and Groothuis-Oudshoorn, K. (2011). Mice: multivariate imputation by chained equations in R. Journal of Statistical Software, 45, 1-67.

28. Hall, K.S., Ogunniyi, A.O., Hendrie, H.C., Osuntokun, B.O., Hui, S.L., Musick, B., Rodenberg, C.S., Unverzagt, F.W., Guerje, O., and Baiyewu, O. (1996). A cross-cultural community based study of dementias: methods and performance of survey instrument. International Journal of Methods in Psychiatric Research, 6, 129-142.

Supplementary material of the manuscript:

Computational methods to simultaneously compare the predictive values of two diagnostic tests with missing data: *EM-SEM* algorithms and multiple imputation

Appendix *A*

Let us consider that two *BDTs* are applied to all the individuals in a random sample sized *n*, whose disease status (present or absent) is known through the application of a *GS*. Let x_{ij} $\left(y_{ij}\right)$ be the numbers of diseased (non-diseased) individuals among whom *Test* 1 leads to a result *i* and *Test* 2 leads to result *j*, with $i, j = 0,1$ (0 indicates a negative result and 1 a positive result). We now summarize the methods of Leisenring et al [3], Wang et al [4] and Kosinski [5]. The Tsou method [6] is not considered since it is equivalent to the Kosinski method.

Method of Leisenring et al

Leisenring et al [3] studied the comparison of the positive and negative *PVs* of two binary tests through marginal regression models, and they were able to estimate these models separately or jointly using *GEE* models. Leisenring et al deduced score statistics to compare the positive and negative *PVs* of two binary tests in paired designs. Using the notation from the previous Section, the score statistic for the test $H_0 : \tau_1 = \tau_2$ is

$$z_\tau = \frac{x_{11}\left(1-2\bar{Z}_1\right)+x_{01}\left(1-\bar{Z}_1\right)-x_{10}\bar{Z}_1}{\sqrt{x_{11}\left(1-\bar{D}_1\right)^2\left(1-2\bar{Z}_1\right)^2+x_{01}\left(1-\bar{D}_1\right)^2\left(1-\bar{Z}_1\right)^2+x_{10}\left(1-\bar{D}_1\right)^2\bar{Z}_1^2+y_{11}\bar{D}_1^2\left(1-2\bar{Z}_1\right)^2+y_{01}\bar{D}_1^2\left(1-\bar{Z}_1\right)^2+y_{10}\bar{D}_1^2\bar{Z}_1^2}}$$

and the score statistic to compare the test $H_0 : \upsilon_1 = \upsilon_2$ is

$$z_\upsilon = \frac{y_{00}\left(1-2\bar{Z}_2\right)+y_{10}\left(1-\bar{Z}_2\right)-y_{01}\bar{Z}_2}{\sqrt{y_{00}\left(1-\bar{D}_2\right)^2\left(1-2\bar{Z}_2\right)^2+y_{10}\left(1-\bar{D}_2\right)^2\left(1-\bar{Z}_2\right)^2+y_{01}\left(1-\bar{D}_2\right)^2\bar{Z}_2^2+x_{00}\bar{D}_2^2\left(1-2\bar{Z}_2\right)^2+x_{10}\bar{D}_2^2\left(1-\bar{Z}_2\right)^2+x_{01}\bar{D}_2^2\bar{Z}_2^2}}.$$

Score statistics have a normal distribution when the null hypothesis is true, and where

$$\bar{Z}_1 = \frac{x_{11} + x_{01} + y_{11} + y_{01}}{2x_{11} + x_{01} + x_{10} + 2y_{11} + y_{10} + y_{01}},\ \bar{D}_1 = \frac{2x_{11} + x_{01} + x_{10}}{2x_{11} + x_{01} + x_{10} + 2y_{11} + y_{10} + y_{01}}.$$

$$\bar{Z}_2 = \frac{x_{00} + x_{10} + y_{00} + y_{10}}{2x_{00} + x_{01} + x_{10} + 2y_{00} + y_{01} + y_{10}} \text{ and } \bar{D}_2 = \frac{2y_{00} + y_{01} + y_{10}}{2x_{00} + x_{01} + x_{10} + 2y_{00} + y_{01} + y_{10}}.$$

Method of Wang et al

Wang et al [4] studied the comparison of the *PVs* of two binary tests through a weighted least square method and compared their method to that of Leisenring et al, before recommending the comparison of the *PVs* using the weighted least square method based on the difference between the two positive (negative) *PVs*. The test statistics for $H_0 : \tau_1 = \tau_2$ and $H_0 : \upsilon_1 = \upsilon_2$ are

$$z_\tau = \frac{\hat{\tau}_1 - \hat{\tau}_2}{\sqrt{\hat{Var}(\hat{\tau}_1) + \hat{Var}(\hat{\tau}_2) - 2\hat{Cov}(\hat{\tau}_1, \hat{\tau}_2)}}$$

and

$$z_\upsilon = \frac{\hat{\upsilon}_1 - \hat{\upsilon}_2}{\sqrt{\hat{Var}(\hat{\upsilon}_1) + \hat{Var}(\hat{\upsilon}_2) - 2\hat{Cov}(\hat{\upsilon}_1, \hat{\upsilon}_2)}}$$

respectively, where $\hat{\tau}_1 = \frac{x_{11} + x_{10}}{x_{11} + x_{10} + y_{11} + y_{10}}$, $\hat{\tau}_2 = \frac{x_{11} + x_{01}}{x_{11} + x_{01} + y_{11} + y_{01}}$, $\hat{\upsilon}_1 = \frac{y_{01} + y_{00}}{x_{01} + x_{00} + y_{01} + y_{00}}$ and $\hat{\upsilon}_2 = \frac{y_{10} + y_{00}}{x_{10} + x_{00} + y_{10} + y_{00}}$. Both test statistics follow a standard normal distribution, and the variances are estimated by applying the delta method (the expressions are shown in the method of Roldán-Nofuentes et al [7] which will now be summarized).

Method of Kosinski

Kosinski [5] proposed a weighted generalized score statistic to solve the hypothesis test of comparison of the *PVs*. The weighted generalized score statistic for the test $H_0:\tau_1=\tau_2$ is

$$z_\tau=\frac{\hat{\tau}_1-\hat{\tau}_2}{\sqrt{\left\{\hat{\tau}_p\left(1-\hat{\tau}_p\right)-2C_p^\tau\right\}\left(\frac{1}{n_{10}+n_{11}}+\frac{1}{n_{01}+n_{11}}\right)}},$$

and the weighted generalized score statistic for the test $H_0:\upsilon_1=\upsilon_2$ is

$$z_\upsilon=\frac{\hat{\upsilon}_1-\hat{\upsilon}_2}{\sqrt{\left\{\hat{\upsilon}_p\left(1-\hat{\upsilon}_p\right)-2C_p^\upsilon\right\}\left(\frac{1}{n_{00}+n_{01}}+\frac{1}{n_{00}+n_{10}}\right)}},$$

which has a standard normal distribution when the null hypothesis is true, where $\hat{\tau}_p=\frac{2x_{11}+x_{10}+x_{01}}{2n_{11}+n_{10}+n_{01}}$ and $\hat{\upsilon}_p=\frac{2y_{00}+y_{01}+y_{10}}{2n_{00}+n_{01}+n_{10}}$ are the pooled positive *PV* and pooled negative *PV* respectively, and

$$C_p^\tau=\frac{x_{11}\left(1-\hat{\tau}_p\right)^2+y_{11}\hat{\tau}_p^2}{2n_{11}+n_{10}+n_{01}},\ C_p^\upsilon=\frac{x_{00}\hat{\upsilon}_p^2+y_{00}\left(1-\hat{\upsilon}_p\right)^2}{2n_{00}+n_{01}+n_{10}}$$

and $n_{ij}=x_{ij}+y_{ij}$.

Method of Roldán-Nofuentes et al

Roldán-Nofuentes et al [7] studied the simultaneous comparison of the *PVs* of two *BDTs* subject to a paired design. The simultaneous comparison of the *PVs* of two binary tests consists of solving the hypothesis test

$$H_0:\left(\tau_1=\tau_2\ \text{ and }\ \upsilon_1=\upsilon_2\right)\ \text{vs}\ H_1:\left(\tau_1\neq\tau_2\ \text{ and/or }\ \upsilon_1\neq\upsilon_2\right),$$

Applying the delta method, the estimated variances-covariances of the estimators of the *PVs* are:

$$\hat{V}ar(\hat{\tau}_1) = \frac{(x_{10}+x_{11})(y_{10}+y_{11})}{n(x_{10}+x_{11}+y_{10}+y_{11})^3},\ \hat{V}ar(\hat{\upsilon}_1) = \frac{(x_{00}+x_{01})(y_{00}+y_{01})}{n(x_{00}+x_{01}+y_{00}+y_{01})^3}$$

$$\hat{V}ar(\hat{\tau}_2) = \frac{(x_{01}+x_{11})(y_{01}+y_{11})}{n(x_{01}+x_{11}+y_{01}+y_{11})^3},\ \hat{V}ar(\hat{\upsilon}_2) = \frac{(x_{00}+x_{10})(y_{00}+y_{10})}{n(x_{00}+x_{10}+y_{00}+y_{10})^3},$$

$$\hat{C}ov(\hat{\tau}_1,\hat{\tau}_2) = \frac{x_{01}x_{10}y_{11} + x_{11}\left[y_{01}(y_{10}+y_{11}) + y_{11}(x_{01}+x_{10}+x_{11}+y_{10}+y_{11})\right]}{(x_{01}+x_{11}+y_{01}+y_{11})^2(x_{10}+s_{11}+y_{10}+y_{11})^2},$$

$$\hat{C}ov(\hat{\tau}_1,\hat{\upsilon}_2) = -\frac{x_{00}(x_{10}+x_{11})y_{10} + x_{10}y_{10}(x_{10}+x_{11}+y_{00}+r_{10}) + x_{10}(y_{00}+r_{10})y_{11}}{(x_{00}+x_{10}+y_{00}+y_{10})^2(x_{10}+x_{11}+y_{10}+y_{11})^2},$$

$$\hat{C}ov(\hat{\tau}_2,\hat{\upsilon}_1) = -\frac{x_{00}(x_{01}+x_{11})y_{01} + x_{01}y_{01}(x_{01}+x_{11}+y_{00}+y_{01}) + x_{01}(y_{00}+y_{01})y_{11}}{(x_{00}+x_{01}+y_{00}+y_{01})^2(x_{01}+x_{11}+y_{01}+y_{11})^2},$$

$$\hat{C}ov(\hat{\upsilon}_1,\hat{\upsilon}_2) = \frac{x_{00}(y_{00}+y_{01})y_{10} + y_{00}\left[y_{00}^2 + x_{01}x_{10} + x_{00}(x_{01}+x_{10}+y_{00}+y_{01})\right]}{(x_{00}+x_{01}+y_{00}+y_{01})^2(x_{00}+x_{10}+y_{00}+y_{10})^2},$$

$$\hat{C}ov(\hat{\tau}_1,\hat{\upsilon}_1) = \hat{C}ov(\hat{\tau}_2,\hat{\upsilon}_2) = 0\,.$$

The test statistic for $H_0:(\tau_1=\tau_2$ and $\upsilon_1=\upsilon_2)$ is $Q^2 = \hat{\boldsymbol{\eta}}^T\mathbf{A}^T\left(\mathbf{A}\hat{\Sigma}\mathbf{A}^T\right)^{-1}\mathbf{A}\hat{\boldsymbol{\eta}}$, where $\hat{\boldsymbol{\eta}} = (\hat{\tau}_1,\hat{\upsilon}_1,\hat{\tau}_1,\hat{\upsilon}_2)^T$, $\hat{\Sigma}$ is the estimated variance-covariance matrix of $\hat{\boldsymbol{\eta}}$ and $\mathbf{A}$ is the design matrix, i.e.

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix}.$$

The test statistic Q^2 is distributed asymptotically according to a central chi-square distribution with two degrees of freedom if H_0 is true.

When all of these methods are used applying multiple imputation, all the equations are valid for the *m*th complete dataset, adding superindex (m) to all of the terms of the equations.

Appendix *B*

Probabilities $\phi_{ij} = P(T_1 = i, T_2 = i, D = 1)$ and $\varphi_{ij} = P(T_1 = i, T_2 = i, D = 0)$ are written in terms of the *PVs* as:

$$\phi_{11} = \frac{\alpha_1\tau_1\tau_2(\upsilon_1 - q)(\upsilon_2 - q)}{pY_1Y_2},\ \phi_{10} = \frac{\tau_1(\upsilon_1 - q)\left[pY_2 - \alpha_1\tau_2(\upsilon_2 - q)\right]}{pY_1Y_2},$$

$$\phi_{01} = \frac{\tau_2(\upsilon_2 - q)\left[pY_1 - \alpha_1\tau_1(\upsilon_1 - q)\right]}{pY_1Y_2},$$

$$\phi_{00} = \frac{\tau_1(\upsilon_1 - q)\left\{p\left[1 - \upsilon_2 + \tau_2(\alpha_1 - 1)\right] - \alpha_1\tau_2(1 - \upsilon_2)\right\} - p(p - \tau_2)(1 - \upsilon_2)Y_1}{pY_1Y_2},$$

$$\varphi_{11} = \frac{\alpha_0(1 - \tau_1)(1 - \tau_2)(\upsilon_1 - q)(\upsilon_2 - q)}{qY_1Y_2},\ \varphi_{10} = \frac{(1 - \tau_1)(\upsilon_1 - q)\left[qY_2 - \alpha_0(1 - \tau_2)(\upsilon_2 - q)\right]}{qY_1Y_2},$$

$$\varphi_{01} = \frac{(1 - \tau_2)(\upsilon_2 - q)\left[qY_1 - \alpha_0(1 - \tau_1)(\upsilon_1 - q)\right]}{qY_1Y_2},$$

$$\varphi_{00} = \frac{\alpha_0(1 - \tau_1)(1 - \tau_2)(\upsilon_1 - q)(\upsilon_2 - q) - q^2Y_1Y_2}{qY_1Y_2} +$$

$$\frac{(1 - \tau_2)(\tau_1 - p)\upsilon_1 + \left[p(2\upsilon_1 - 1) - \tau_1(\upsilon_1 - p) - \tau_2Y_1\right]\upsilon_2}{qY_1Y_2}.$$

Appendix *C*

The maximum likelihood estimator of *PVs* in the presence of partial verification of the disease are [10, 11]

$$\hat{\tau}_1 = \frac{1}{n_{10} + n_{11}}\sum_{j=0}^{1}\frac{n_{1j}a_{1j}}{a_{1j} + b_{1j}} \text{ and } \hat{\upsilon}_1 = \frac{1}{n_{00} + n_{01}}\sum_{j=0}^{1}\frac{n_{0j}b_{0j}}{a_{0j} + b_{0j}}$$

for test 1, and

$$\hat{\tau}_2 = \frac{1}{n_{01} + n_{11}}\sum_{i=0}^{1}\frac{n_{i1}a_{i1}}{a_{i1} + b_{i1}} \text{ and } \hat{\upsilon}_2 = \frac{1}{n_{00} + n_{10}}\sum_{i=0}^{1}\frac{n_{i0}b_{i0}}{a_{i0} + b_{i0}}$$

for test 2. From the table of complete data (Table 1b), the log-likelihood function based on *n* individuals is

$$l(\boldsymbol{\theta}) = \sum_{i,j=0}^{1} \left(a_{ij} + d_{ij}\right) \log\left(\phi_{ij}\right) + \sum_{i,j=0}^{1} \left(b_{ij} + c_{ij} - d_{ij}\right) \log\left(\varphi_{ij}\right).$$

From this function it is obtained that $\hat{\phi}_{ij} = \frac{a_{ij} + d_{ij}}{n}$ and $\hat{\varphi}_{ij} = \frac{b_{ij} + c_{ij} - d_{ij}}{n}$. In order to demonstrate that the *EM* algorithm converges to the *ML* estimators, we are going to follow the same steps as Little and Rubin [7]. With the *EM* algorithm, the estimator of τ_1 is calculated as

$$\hat{\tau}_1^{(k+1)} = \frac{1}{n_{10} + n_{11}} \sum_{j=0}^{1} \left(a_{1j} + d_{1j}^{(k+1)}\right) = \frac{1}{n_{10} + n_{11}} \sum_{j=0}^{1} \left(a_{1j} + c_{1j} \frac{\hat{\phi}_{1j}^{(k)}}{\hat{\phi}_{1j}^{(k)} + \hat{\varphi}_{1j}^{(k)}} \right).$$

Therefore it is necessary to show that $\sum_{j=0}^{1} \left(a_{1j} + d_{1j}^{(k+1)}\right)$ converges to $\sum_{j=0}^{1} \frac{n_{1j} a_{1j}}{a_{1j} + b_{1j}}$. Taking

$\hat{\phi}_{1j}^{(k)} = \hat{\phi}_{1j}^{(k+1)} = \hat{\phi}_{1j} = \frac{a_{1j} + d_{1j}}{n}$, $\hat{\varphi}_{1j}^{(k)} = \hat{\varphi}_{1j}^{(k+1)} = \hat{\varphi}_{1j} = \frac{b_{1j} + c_{1j} - d_{1j}}{n}$ and

$d_{1j}^{(k)} = d_{1j}^{(k+1)} = d_{1j} = c_{1j} \frac{\hat{\phi}_{1j}}{\hat{\phi}_{1j} + \hat{\varphi}_{1j}}$, then it is obtained that $d_{1j} = \frac{a_{1j} c_{1j}}{a_{1j} + b_{1j}}$. Then

$$\sum_{j=0}^{1} \left(a_{1j} + d_{1j}^{(k+1)}\right) = \sum_{j=0}^{1} \left(a_{1j} + \frac{a_{1j} c_{1j}}{a_{1j} + b_{1j}} \right) = \sum_{j=0}^{1} \left(a_{1j} + \frac{a_{1j} c_{1j}}{n_{1j} - c_{1j}} \right) = \sum_{j=0}^{1} \frac{n_{1j} a_{1j}}{a_{1j} + b_{1j}}.$$

Therefore, $\hat{\tau}_1^{(k+1)}$ converges to $\hat{\tau}_1$. The convergence of the rest of the estimators of *PVs* is demonstrated in a similar way.