REVSTAT – Statistical Journal
Volume 15, Number 1, January 2017, 45–64

COMPARISON OF THE PREDICTIVE VALUES OF MULTIPLE BINARY DIAGNOSTIC TESTS IN THE PRESENCE OF IGNORABLE MISSING DATA

Authors: Ana Eugenia Marín-Jiménez
– Statistics (Biostatistics), Faculty of Education, Economy and Technology, University of Granada, Ceuta, Spain
anamarin@ugr.es

José Antonio Roldán-Nofuentes
– Statistics (Biostatistics), School of Medicine, University of Granada, Granada, Spain
jaroldan@ugr.es

Received: August 2013 Revised: April 2015 Accepted: May 2015

Abstract:

- The comparison of the predictive values of binary diagnostic tests is an important topic in the study of statistical methods applied to medical diagnosis. In this article, we study a global hypothesis test to simultaneously compare the predictive values of multiple binary diagnostic tests in the presence of ignorable missing data. The global hypothesis test deduced is based on the chi-squared distribution. Simulation experiments were carried out to study the type I error probability and the power of global hypothesis test and of other alternative methods when comparing the predictive values of two and three binary diagnostic tests respectively.

Key-Words:

- *global hypothesis test; predictive values; multiple comparisons; chi-squared distribution; ignorable missing data.*

AMS Subject Classification:

- 62P10, 62H15.

1. INTRODUCTION

A diagnostic test is a medical test that is applied to an individual in order to determine the presence or absence of a disease. When the diagnostic test can only give two possible results (positive, indicating the provisional presence of the disease, or negative, indicating the provisional absence of the disease) the diagnostic test is called a binary diagnostic test (BDT) and it is used very frequently in clinical practice. A stress test for the diagnosis of coronary disease or a mammogram to diagnose breast cancer are two examples of *BDTs*. The most common parameters to assess the accuracy or performance of a BDT are sensitivity (Se) and specificity (Sp). Other commonly used parameters to assess the performance of a BDT are predictive values (PVs). The positive predictive value (PPV) of a BDT is the probability of an individual having the disease given that the result of the BDT is positive and the negative predictive value (NPV) is the probability of an individual not having the disease given that the result of the BDT is negative. The predictive values (PVs) are a measure of clinical accuracy of the BDT, and they depend on the sensitivity and the specificity of the BDT and on the prevalence of the disease (p). Applying Bayes Theorem, the PVs are calculated as

$$PPV = \frac{p \times Se}{p \times Se + (1-p) \times (1-Sp)}$$

(1.1) and

$$NPV = \frac{(1-p) \times Sp}{p \times (1-Se) + (1-p) \times Sp} .$$

In the study of statistical methods for the diagnosis of diseases, comparison of the accuracy or the performance of two diagnostic tests is a topic of particular importance. In paired designs (*i.e.* when the two *BDTs* and the gold standard are applied to all of the individuals in a random sample), comparison of the PVs of two *BDTs* in relation to the same gold standard has been the subject of several studies in the statistical literature [1, 2, 3, 4]. In all of them, the comparison of the two $PPVs$ and the comparison of the two $NPVs$ is carried out independently. Roldán-Nofuentes *et al.* [5] showed that the PVs of two (or more) *BDTs* are correlated and they studied a global hypothesis test based on the chi-squared distribution to simultaneously compare the PVs of two or more *BDTs* in relation to the same gold standard. In all of these studies, the disease status of all of the patients is known, as well as the results of the two diagnostic tests. This situation is also known as 'complete verification' (because the gold standard is applied to all of the individuals in the sample). Poleto *et al.* [6] studied the comparison of the predictive values of two *BDTs* when for some individuals we do not know the results of one of the two *BDTs*. Furthermore, in clinical practice it is common for the gold standard not to be applied to all of the individuals in the sample, thus leading to the problem known as partial disease verification [7, 8, 9].

Therefore, the disease status (if the disease is present or absent) is unknown for a subset of individuals in the sample. In this situation, Roldán-Nofuentes *et al.* [10, 11] studied the comparison of the *PPVs* and of the *NPVs* of two *BDTs*. Nevertheless, in these two studies they did not consider the dependence that exists between the *PVs* of the diagnostic tests. This is the essence of our article, to study a global hypothesis test that allows us to jointly compare the *PVs* of two (or more) *BDTs* in the presence of ignorable missing data. In this article, we study a global hypothesis test to simultaneously compare the predictive values of two or more *BDTs* when, in the presence of partial disease verification, the missing data mechanism is ignorable. In Section 2, we propose a global hypothesis test, and other alternative methods, to simultaneously compare the *PVs* of multiple *BDTs*. In Section 3, Monte Carlo simulation experiments are carried out in order to study the type I error probability and the power of the global hypothesis test (and of the alternative methods) when comparing the *PVs* of two and of three *BDTs* respectively. In Section 4, the method proposed is applied to two examples, and in Section 5 the results obtained are discussed.

2. THE MODEL

Let us consider J *BDTs* ($J \geqslant 2$) that are applied independently to the same random sample of size n extracted from a population that has a determined prevalence of the disease (p). Moreover, let us consider that the gold standard has not been applied to all of the individuals in the random sample. In this situation, the J diagnostic tests are applied to all of the individuals in the sample whilst the gold standard is only applied to a subset of them. Therefore, the results of the J diagnostic tests are known by all of the individuals in the sample, whereas the result of the gold standard (*i.e.* the disease status) is only unknown to a subset of them. Let T_j, V and D be the random binary variables defined as: T_j which models the result of the j-th *BDT* ($j = 1, ..., J$), so that $T_j = 1$ when the test result is positive and $T_j = 0$ when the result is negative; V models the verification process, $V = 1$ when the individual is verified with the gold standard and $V = 0$ when the individual is not verified; and D models the result of the gold standard, $D = 1$ when the individual has the disease and $D = 0$ when the individual does not. Let $Se_j = P(T_j = 1 \mid D = 1)$, $Sp_j = P(T_j = 0 \mid D = 0)$, $PPV_j = P(D = 1 \mid T_j = 1)$ and $NPV_j = P(D = 0 \mid T_j = 0)$ be the sensitivity, the specificity, the positive predictive value and the negative predictive value of the j-th *BDT* respectively. Let the observed frequencies be: $s_{i_1,...,i_J}$ is the number of patients verified in which $T_1 = i_1$, $T_2 = i_2$, ..., $T_J = i_J$ and $D = 1$; $r_{i_1,...,i_J}$ is the number of patients verified in which $T_1 = i_1$, $T_2 = i_2$, ..., $T_J = i_J$ and $D = 0$; and $u_{i_1,...,i_J}$ is the number of patients not verified in which $T_1 = i_1$, $T_2 = i_2$, ..., $T_J = i_J$ with $i_j = 0, 1$ and $j = 1, ..., J$. Let $n_{i_1,...,i_J} = s_{i_1,...,i_J} + r_{i_1,...,i_J} + u_{i_1,...,i_J}$ and

$n = \sum_{i_1,...,i_J=0}^{1} n_{i_1,...,i_J}$. As only a subset of individuals in the sample have their disease status verified with the gold standard, the verification probabilities ($\lambda_{k,i_1,...,i_J}$) are defined as the probability of selecting an individual for whom $D = k$, $T_1 = i_1$, $T_2 = i_2$, ..., $T_J = i_J$ with $k, i_j = 0, 1$, $j = 1, ..., J$, to verify his or her disease status *i.e.*

$$\lambda_{k,i_1,...,i_J} = P\Big(V = 1 \mid D = k, T_1 = i_1, T_2 = i_2, ..., T_J = i_J\Big).$$

Assuming that the verification process with the gold standard only depends on the results of the J *BDTs* and does not depend on the disease status, then the missing data mechanism is missing at random (MAR) [12]. Assuming also that the parameters of the data model and the parameters of the missingness mechanism are distinct, then the missing data mechanism is ignorable [13]. Under this model, the verification probabilities are

$$\lambda_{k,i_1,...,i_J} = \lambda_{i_1,...,i_J} = P\Big(V = 1 \mid T_1 = i_1, T_2 = i_2, ..., T_J = i_J\Big),$$

and all of the parameters can be estimated applying the maximum likelihood method.

2.1. Maximum likelihood estimators of the *PVs*

As the J *BDTs* are applied to all of the n individuals in the random sample and the gold standard is only applied to a subset of them, the frequencies observed $r_{i_1,...,i_J}$, $s_{i_1,...,i_J}$ and $u_{i_1,...,i_J}$ with $i_j = 0, 1$ and $j = 1, ..., J$, which can be written in the form of a 3×2^J table in which the sample of size n has been set, are the realization of a multinomial distribution whose probabilities are

$$\phi_{i_1,...,i_J} = P\Big(V = 1, D = 1, T_1 = i_1, T_2 = i_2, ..., T_J = i_J\Big),$$

$$\varphi_{i_1,...,i_J} = P\Big(V = 1, D = 0, T_1 = i_1, T_2 = i_2, ..., T_J = i_J\Big)$$

and

$$\gamma_{i_1,...,i_J} = P\Big(V = 0, T_1 = i_1, T_2 = i_2, ..., T_J = i_J\Big).$$

Let $\boldsymbol{\omega} = (\phi_{1,...,1}, ..., \phi_{0,...,0}, \varphi_{1,...,1}, ..., \varphi_{0,...,0}, \gamma_{1,...,1}, ..., \gamma_{0,...,0})^{\mathrm{T}}$ be a vector sized $\left(3 \cdot 2^J\right)$ whose components are the probabilities of multinomial distribution and $\eta_{i_1,...,i_J} = \phi_{i_1,...,i_J} + \varphi_{i_1,...,i_J} + \gamma_{i_1,...,i_J}$. Assuming that the missing data mechanism is ignorable, the *PVs* of the j-th *BDT* are written in terms of the parameters of

the vector $\boldsymbol{\omega}$ and of the verification probabilities as

$$PPV_j = \frac{\sum\limits_{i_1,\ldots,i_J=0;\, i_j=1}^{1} \phi_{i_1,\ldots,i_J}\, \lambda_{i_1,\ldots,i_J}^{-1}}{\sum\limits_{i_1,\ldots,i_J=0;\, i_j=1}^{1} \eta_{i_1,\ldots,i_J}}$$

(2.1) and

$$NPV_j = \frac{\sum\limits_{i_1,\ldots,i_J=0;\, i_j=0}^{1} \varphi_{i_1,\ldots,i_J}\, \lambda_{i_1,\ldots,i_J}^{-1}}{\sum\limits_{i_1,\ldots,i_J=0;\, i_j=0}^{1} \eta_{i_1,\ldots,i_J}} ,$$

where $\lambda_{i_1,\ldots,i_J} = (\phi_{i_1,\ldots,i_J} + \varphi_{i_1,\ldots,i_J})/\eta_{i_1,\ldots,i_J}$ are the verification probabilities. Therefore, in equations (2.1) we can observe the dependence of the *PVs* of the verification process subject to the MAR assumption. In this situation the logarithm of the likelihood function is

$$\begin{aligned} l = & \sum_{i_1,\ldots,i_J=0}^{1} s_{i_1,\ldots,i_J} \log\left(\phi_{i_1,\ldots,i_J}\right) + \sum_{i_1,\ldots,i_J=0}^{1} r_{i_1,\ldots,i_J} \log\left(\varphi_{i_1,\ldots,i_J}\right) \\ & + \sum_{i_1,\ldots,i_J=0}^{1} u_{i_1,\ldots,i_J} \log\left(\gamma_{i_1,\ldots,i_J}\right) , \end{aligned}$$

so that maximizing this function, the maximum likelihood estimators (MLEs) of $\phi_{i_1,\ldots,i_J}$, $\varphi_{i_1,\ldots,i_J}$ and $\gamma_{i_1,\ldots,i_J}$ are the estimators of multinomial proportions [14], *i.e.*

$$\hat{\phi}_{i_1,\ldots,i_J} = \frac{s_{i_1,\ldots,i_J}}{n} , \quad \hat{\varphi}_{i_1,\ldots,i_J} = \frac{r_{i_1,\ldots,i_J}}{n} \quad \text{and} \quad \hat{\gamma}_{i_1,\ldots,i_J} = \frac{u_{i_1,\ldots,i_J}}{n} , \tag{2.2}$$

and the *MLE* of $\eta_{i_1,\ldots,i_J}$ is $\hat{\eta}_{i_1,\ldots,i_J} = n_{i_1,\ldots,i_J}/n$. Substituting in equations (2.1) the parameters with their respective *MLEs* given in equations (2.2), the *MLEs* of the *PVs* of the j-th *BDT* are

$$\widehat{PPV_j} = \frac{\sum\limits_{i_1,\ldots i_J=0;\, i_j=1}^{1} \frac{s_{i_1,\ldots i_J}\, n_{i_1,\ldots i_J}}{s_{i_1,\ldots i_J} + r_{i_1,\ldots i_J}}}{\sum\limits_{i_1,\ldots i_J=0;\, i_j=1}^{1} n_{i_1,\ldots i_J}}$$

and

$$\widehat{NPV_j} = \frac{\sum\limits_{i_1,\ldots i_J=0;\, i_j=0}^{1} \frac{r_{i_1,\ldots i_J}\, n_{i_1,\ldots i_J}}{s_{i_1,\ldots i_J} + r_{i_1,\ldots i_J}}}{\sum\limits_{i_1,\ldots i_J=0;\, i_j=0}^{1} n_{i_1,\ldots i_J}} .$$

Once we have obtained the *MLEs* of the *PVs* of the J *BDTs*, we then estimate their variances-covariances.

2.2. Estimation of the variances-covariances of the *PVs*

As the vector $\boldsymbol{\omega}$ is the vector of probabilities of a multinomial distribution, the variance-covariance matrix of $\hat{\boldsymbol{\omega}}$ is $\sum_{\hat{\omega}} = \left\{\text{diag}\left(\boldsymbol{\omega}\right) - \boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\omega}\right\}/n$. Let $\boldsymbol{\tau} = \left(PPV_1, ..., PPV_J, NPV_1, ..., NPV_J\right)^{\mathrm{T}}$ be a vector sized $2J$ whose components are the *PVs* of the J *BDTs*, and let $\hat{\boldsymbol{\tau}}$ be the *MLE* of $\boldsymbol{\tau}$. As $\boldsymbol{\tau}$ is a function of the components of the vector $\boldsymbol{\omega}$, applying the delta method [15] the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\tau}}$ is

$$\sum_{\hat{\boldsymbol{\tau}}} = \left(\frac{\partial \boldsymbol{\tau}}{\partial \boldsymbol{\omega}}\right) \sum_{\hat{\omega}} \left(\frac{\partial \boldsymbol{\tau}}{\partial \boldsymbol{\omega}}\right)^{\mathrm{T}}.$$

Substituting in the previous expression each parameter with its corresponding *MLE*, we obtain the estimated asymptotic variances-covariances of the estimators of the *PVs* of the J *BDTs*.

Moreover, the asymptotic variances-covariances of $\hat{\boldsymbol{\tau}}$ can also be estimated through bootstrap [16], generating, from the random sample of size n, B samples with replacement and from these B samples asymptotic variance-covariance matrix of $\hat{\boldsymbol{\tau}}$ is estimated.

Once we have obtained the *MLEs* of the *PVs* and their estimated asymptotic variances-covariances, it is possible to solve the global hypothesis test to simultaneously compare the *PVs* of the J *BDTs*.

2.3. Global hypothesis test

The global hypothesis test to simultaneously compare the *PVs* of J *BDTs* is

$$H_0\colon \quad PPV_1 = PPV_2 = \cdots = PPV_J \quad \text{and} \quad NPV_1 = NPV_2 = \cdots = NPV_J\,,$$

$$H_1\colon \quad \text{at least one equality is not true}\,,$$

which is equivalent to the hypothesis test

$$H_0\colon\ \mathbf{A}\boldsymbol{\tau} = 0 \quad \text{vs} \quad H_1\colon\ \mathbf{A}\boldsymbol{\tau} \neq 0\,, \tag{2.3}$$

where $\mathbf{A}$ is a full rank matrix sized $2\left(J-1\right) \times 2J$ whose elements are known constants. For two *BDTs* $(J = 2)$ the matrix $\mathbf{A}$ is $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \otimes \begin{pmatrix} 1 & -1 \end{pmatrix}$, and for $J = 3$ this matrix is $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \otimes \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix}$, where $\otimes$ is the Kronecker product. As the vector $\hat{\boldsymbol{\tau}}$ is asymptotically distributed according to a normal multivariate

distribution, *i.e.* $\hat{\boldsymbol{\tau}} \xrightarrow[n\to\infty]{} N\left(\boldsymbol{\tau}, \boldsymbol{\Sigma}_{\boldsymbol{\tau}}\right)$, the Wald statistic for the global hypothesis test (2.3) is

$$Q^2 = \hat{\boldsymbol{\tau}}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\left(\mathbf{A}\hat{\sum}_{\hat{\boldsymbol{\tau}}}\mathbf{A}^{\mathrm{T}}\right)^{-1}\mathbf{A}\hat{\boldsymbol{\tau}}\,, \tag{2.4}$$

which is asymptotically distributed according to a chi-squared distribution with $2\left(J-1\right)$ degrees of freedom when the null hypothesis is true.

If the global hypothesis test is solved applying bootstrap, the statistic for the global test is similar to that given in the expression (2.4), substituting $\hat{\boldsymbol{\tau}}$ with the bootstrap estimator of $\boldsymbol{\tau}$ and $\hat{\sum}_{\hat{\boldsymbol{\tau}}}$ with the variance-covariance matrix estimated through bootstrap.

Other alternative methods will now be proposed to solve the global hypothesis test (2.3).

2.4. Alternative methods

The method proposed in the previous Section to solve the global hypothesis test (2.3) is based on the chi-squared distribution. The following are some alternative methods to solve this hypothesis test:

Method 1. Consists of solving the $J\left(J-1\right)$ marginal hypothesis tests given by

$$H_0\colon\ PPV_k = PPV_l \quad \text{vs} \quad H_1\colon\ PPV_k \neq PPV_l$$

and

$$H_0\colon\ NPV_k = NPV_l \text{ vs } H_1\colon\ NPV_k \neq NPV_l$$

with $k, l = 1, ..., J$ and $k \neq l$, each one to an error rate of $\alpha/\{J\left(J-1\right)\}$, *i.e.* applying the Bonferroni method [17], where the statistics is

$$z = \frac{\widehat{PV_k} - \widehat{PV_l}}{\sqrt{\widehat{\mathrm{Var}}\left(\widehat{PV_k}\right) + \widehat{\mathrm{Var}}\left(\widehat{PV_l}\right) - 2\,\widehat{\mathrm{Cov}}\left(\widehat{PV_k}, \widehat{PV_l}\right)}} \longrightarrow N\left(0,1\right)\,, \tag{2.5}$$

and where PV is PPV or NPV respectively.

Method 2. Consists of solving the $J\left(J-1\right)$ marginal hypothesis tests and applying the multiple comparison method of Holm [18] to a global error rate of α.

Method 3. Consists of solving the $J\left(J-1\right)$ marginal hypothesis tests and applying the multiple comparison method of Hochberg [19] to a global error rate of α.

These methods are very easy to apply from the p-values calculated in the $J\,(J-1)$ marginal hypothesis tests. The Bonferroni method is a classic method of post hoc comparison, and the Holm and Hochberg methods are less conservative post hoc methods than the Bonferroni method. Furthermore all of the aforementioned methods can be applied both if the *PVs* and their variances-covariances are estimated through the maximum likelihood method and the delta method respectively, or if they are estimated through the bootstrap method.

3. SIMULATION EXPERIMENTS

Monte Carlo simulation experiments were carried out to study the type I error probability and the power of the global hypothesis proposed in Section 2.3 and of the alternative methods proposed in Section 2.4, when comparing the *PVs* of two and of three *BDT*s respectively, and both if the variance-covariance matrix is estimated through the delta method and if it is estimated through the bootstrap method. These simulation experiments were designed in a similar way to those carried out by Roldan-Nofuentes *et al.* [5], and consisted of the generation of 5000 random samples with multinomial distributions sized 50, 100, 200, 500, 1000, 2000 and 5000. For all of the study $\alpha = 5\%$ was set. All of the random samples were generated in such a way that in all of them it was possible to estimate the *PVs* and their variances-covariances. In the case of bootstrap, for each random sample 2000 samples with replacement were generated and from these $\hat{\hat{\boldsymbol{\tau}}}$ and $\hat{\sum}_{\hat{\hat{\boldsymbol{\tau}}}}$ were calculated. All of the random samples were generated from the *PVs* and the prevalence, without setting the values of sensitivity and specificity of each *BDT* in the following way:

1. As *PVs* we took the values $\{0.60, 0.65, ..., 0.90, 0.95\}$, which are quite common values in clinical practice, and as values of the disease prevalence we took the values $\{0.05, 0.10, ..., 0.90, 0.95\}$.
2. Once the *PVs* and the disease prevalence were set, the sensitivity and the specificity of each diagnostic test were calculated from equations (1.1), and then the maximum values of the dependence factors between the two *BDTs* were obtained from the values of the sensitivity and specificity of each diagnostic test applying the model of Vacek [20] for two *BDTs* and applying the model of Torrance-Rynard and Walter [21] for three *BDTs*. In Appendix A both models are summarized.
3. For two *BDTs*, as verification probabilities we took the values

$$\left(\lambda_{11} = 0.70,\ \lambda_{10} = \lambda_{01} = 0.40,\ \lambda_{00} = 0.10\right)$$

and

$$\left(\lambda_{11} = 0.95,\ \lambda_{10} = \lambda_{01} = 0.60,\ \lambda_{00} = 0.30\right),$$

which can be considered a scenario with low verification and a scenario

with high verification respectively. For three *BDTs*, we took the values

$$\begin{aligned}&\big(\lambda_{111}=0.70,\ \lambda_{110}=0.40,\ \lambda_{101}=0.40,\ \lambda_{100}=0.25,\\&\ \lambda_{011}=0.40,\ \lambda_{010}=0.25,\ \lambda_{001}=0.25,\ \lambda_{000}=0.05\big)\end{aligned}$$

and

$$\begin{aligned}&\big(\lambda_{111}=1,\ \lambda_{110}=0.80,\ \lambda_{101}=0.80,\ \lambda_{100}=0.40,\\&\ \lambda_{011}=0.80,\ \lambda_{010}=0.40,\ \lambda_{001}=0.40,\ \lambda_{000}=0.20\big)\,,\end{aligned}$$

which can also be considered as scenarios with low and high verification.

4. In the case of two *BDTs*, the probabilities of the multinomial distributions were calculated from the equations of the model of Vacek [20] (Appendix A). In the case of three *BDTs*, the probabilities of the multinomial distributions were calculated from the model of Torrance-Rynard and Walter [21] (Appendix A).

3.1. Two *BDTs*

In Table 1 we show the results obtained for the type I errors probabilities and the powers when comparing the *PVs* of two *BDTs*, for different values of the *PVs* and for intermediate and high dependence factors, when the *PVs* are estimated through maximum likelihood and the variance-covariance matrix is estimated through the delta method (other tables with results from the simulation experiments can be requested from the authors). Regarding the type I error probability, the global hypothesis test has a type I error probability which, in general terms, fluctuates around the nominal error of 5% especially when $n \geqslant 500$. In some cases, especially when $n \leqslant 200$ and the verification probabilities are low and/or the dependence factors are high, the type I error probability may overwhelm the nominal error. This may be due to the fact that the samples are not large enough, and therefore some frequencies of the multinomial distribution which are equal to zero, and the variance-covariance matrix are not well represented. Regarding Methods 1, 2 and 3 (Bonferroni, Holm and Hochberg), the type I error probability of each one of them performs in a similar way to that of the global test, although it is usually somewhat lower than the nominal error (especially for $n \geqslant 2000$).

Regarding the power, in general it is necessary to have samples of between 500 and 1000 individuals (depending on the verification probabilities) so that the power of the global hypothesis test is high (higher than 80% or 90%). The power of the global hypothesis test increases when there is an increase in the verification probabilities; whereas the increase in the dependence factors does not have a clear effect on the power of the global hypothesis test (sometimes it increases and sometimes it decreases). Regarding Methods 1, 2 and 3, their respective powers perform in a similar way to that of the global test, although the power of each one of them is slightly lower than that of the global test.

Table 1: Type I errors probabilities and powers when comparing the *PVs* of two *BDTs* (M1: Method 1, M2: Method 2, M3: Method 3).

TYPE I ERRORS PROBABILITIES								
$PPV_1 = PPV_2 = 0.80, \quad NPV_1 = NPV_2 = 0.90$ $p = 20\%$ $Se_1 = Se_2 = 0.5714, \quad Sp_1 = Sp_2 = 0.9643$								
$\lambda_{11} = 0.70, \quad \lambda_{10} = 0.40, \quad \lambda_{01} = 0.40, \quad \lambda_{00} = 0.10$								
n	$\varepsilon_1 = 0.075, \quad \varepsilon_0 = 0.01$				$\varepsilon_1 = 0.15, \quad \varepsilon_0 = 0.02$			
	Global test	**M1**	**M2**	**M3**	**Global test**	**M1**	**M2**	**M3**
50	0.007	0.007	0.007	0.007	0.003	0.003	0.003	0.003
100	0.039	0.035	0.035	0.039	0.044	0.039	0.039	0.044
200	0.070	0.069	0.069	0.069	0.068	0.067	0.067	0.068
500	0.066	0.065	0.065	0.066	0.069	0.068	0.068	0.069
1000	0.059	0.058	0.058	0.059	0.054	0.050	0.050	0.051
2000	0.056	0.055	0.055	0.055	0.049	0.039	0.039	0.042
5000	0.048	0.042	0.042	0.044	0.054	0.048	0.048	0.049
$\lambda_{11} = 0.95, \quad \lambda_{10} = 0.60, \quad \lambda_{01} = 0.60, \quad \lambda_{00} = 0.30$								
n	$\varepsilon_1 = 0.075, \quad \varepsilon_0 = 0.01$				$\varepsilon_1 = 0.15, \quad \varepsilon_0 = 0.02$			
	Global test	**M1**	**M2**	**M3**	**Global test**	**M1**	**M2**	**M3**
50	0.022	0.021	0.021	0.021	0.011	0.011	0.011	0.011
100	0.062	0.058	0.058	0.062	0.067	0.066	0.066	0.067
200	0.068	0.067	0.067	0.068	0.069	0.068	0.068	0.068
500	0.055	0.056	0.056	0.057	0.058	0.057	0.057	0.057
1000	0.058	0.057	0.057	0.058	0.059	0.057	0.057	0.058
2000	0.054	0.052	0.052	0.053	0.048	0.043	0.043	0.044
5000	0.049	0.044	0.044	0.047	0.052	0.046	0.046	0.047

POWERS								
$PPV_1 = 0.85, \quad PPV_2 = 0.90, \quad NPV_1 = 0.75, \quad NPV_2 = 0.80$ $p = 75\%$ $Se_1 = 0.9444, \quad Sp_1 = 0.50, \quad Se_2 = 0.9429, \quad Sp_2 = 0.6857$								
$\lambda_{11} = 0.70, \quad \lambda_{10} = 0.40, \quad \lambda_{01} = 0.40, \quad \lambda_{00} = 0.10$								
n	$\varepsilon_1 = 0.02, \quad \varepsilon_0 = 0.06$				$\varepsilon_1 = 0.04, \quad \varepsilon_0 = 0.12$			
	Global test	**M1**	**M2**	**M3**	**Global test**	**M1**	**M2**	**M3**
50	0.004	0.004	0.004	0.004	0.006	0.006	0.006	0.006
100	0.090	0.089	0.089	0.090	0.065	0.066	0.066	0.068
200	0.307	0.264	0.026	0.266	0.336	0.333	0.333	0.335
500	0.770	0.667	0.667	0.668	0.913	0.863	0.863	0.863
1000	0.984	0.940	0.940	0.940	0.999	0.998	0.998	0.998
2000	1	0.998	0.998	0.998	1	1	1	1
5000	1	1	1	1	1	1	1	1
$\lambda_{11} = 0.95, \quad \lambda_{10} = 0.60, \quad \lambda_{01} = 0.60, \quad \lambda_{00} = 0.30$								
n	$\varepsilon_1 = 0.02, \quad \varepsilon_0 = 0.06$				$\varepsilon_1 = 0.04, \quad \varepsilon_0 = 0.12$			
	Global test	**M1**	**M2**	**M3**	**Global test**	**M1**	**M2**	**M3**
50	0.051	0.050	0.050	0.050	0.032	0.031	0.031	0.032
100	0.198	0.195	0.195	0.201	0.175	0.185	0.185	0.192
200	0.431	0.388	0.388	0.392	0.553	0.540	0.540	0.543
500	0.833	0.780	0.780	0.782	0.965	0.949	0.949	0.949
1000	0.990	0.978	0.978	0.978	1	1	1	1
2000	1	1	1	1	1	1	1	1
5000	1	1	1	1	1	1	1	1

Regarding the solution of the global test applying the bootstrap method, the results obtained are almost identical to those obtained through the method of maximum likelihood and the delta method. Therefore, in terms of the type I error probability and the power there is practically no difference between solving the global hypothesis test through the maximum likelihood method and the bootstrap method, although the bootstrap requires a greater computational effort.

3.2. Three *BDTs*

In Table 2 we show some of the results obtained for the type I error probability and the power when comparing the *PVs* of three *BDTs*, also for different values of the *PVs* and for intermediate and high dependence factors, when the *PVs* are estimated through maximum likelihood and the variance-covariance matrix is estimated through the delta method (other tables with results from the simulation experiments can be requested from the authors). For three *BDTs* we have not considered sample sizes smaller than 100, since with smaller samples there are too many frequencies equal to 0 (above all when the prevalence is low and/or the verification probabilities are low) and it is not possible to calculate the estimators or the variances-covariances. In general terms, the conclusions reached are similar to those obtained for two *BDTs*, although for the global test and for methods 1, 2 and 3 it is necessary to have larger sample sizes so that the type I error probability fluctuates around the nominal error.

With regard to the power of each method, this increases with an increase in the verification probabilities, and decreases when there is an increase in the values of the dependence factors. In very general terms, when the verification probabilities are low it is necessary to have samples of between 500 and 1000 individuals so that the power of the global test is higher than 80% or 90% (depending on the values of the dependence factors), although in some situations (high dependence factors) it is necessary to have very large samples ($n \geqslant 5000$) in order to reach this power. Regarding Methods 1, 2 and 3, in general terms there is no important difference in power in relation to the global hypothesis test, especially when $n \geqslant 500$, whilst for smaller sample sizes the global test is somewhat more powerful than for the other three methods.

3.3. Conclusions

From the analysis of the results obtained in the simulation experiments one may conclude that the global hypothesis test based on the chi-squared distribution displays the performance of an asymptotic hypothesis test (from a certain sample size onwards, its type I error probability fluctuates around the nominal error).

Table 2: Type I errors probabilities and powers when comparing the *PVs* of three *BDTs* (M1: Method 1, M2: Method 2, M3: Method 3).

TYPE I ERRORS PROBABILITIES

$PPV_1 = PPV_2 = PPV_3 = 0.85$, $NPV_1 = NPV_2 = NPV_3 = 0.80$
$p = 25\%$
$e_1 = Se_2 = Se_3 = 0.2615$, $Sp_1 = Sp_2 = Sp_3 = 0.9846$

$\lambda_{111} = 0.70$, $\lambda_{110} = 0.40$, $\lambda_{101} = 0.40$, $\lambda_{100} = 0.25$
$\lambda_{011} = 0.40$, $\lambda_{010} = 0.25$, $\lambda_{001} = 0.25$, $\lambda_{000} = 0.05$

n	$\delta = 0.0075$, $\varepsilon = 0.0001$				$\delta = 0.015$, $\varepsilon = 0.0002$			
	Global test	**M1**	**M2**	**M3**	**Global test**	**M1**	**M2**	**M3**
100	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
200	0.005	0.003	0.003	0.003	0.005	0.004	0.004	0.004
500	0.041	0.026	0.026	0.028	0.045	0.030	0.030	0.032
1000	0.058	0.041	0.041	0.045	0.059	0.040	0.040	0.045
2000	0.059	0.043	0.043	0.050	0.059	0.058	0.058	0.059
5000	0.053	0.044	0.044	0.048	0.054	0.037	0.037	0.042

$\lambda_{111} = 1$, $\lambda_{110} = 0.80$, $\lambda_{101} = 0.80$, $\lambda_{100} = 0.40$
$\lambda_{011} = 0.80$, $\lambda_{010} = 0.40$, $\lambda_{001} = 0.40$, $\lambda_{000} = 0.20$

n	$\delta = 0.0075$, $\varepsilon = 0.0001$				$\delta = 0.015$, $\varepsilon = 0.0002$			
	Global test	**M1**	**M2**	**M3**	**Global test**	**M1**	**M2**	**M3**
100	0.007	0.005	0.005	0.005	0.006	0.004	0.004	0.005
200	0.043	0.031	0.031	0.032	0.043	0.029	0.029	0.031
500	0.050	0.037	0.038	0.041	0.056	0.039	0.039	0.041
1000	0.057	0.045	0.045	0.047	0.057	0.042	0.042	0.047
2000	0.058	0.044	0.044	0.046	0.059	0.047	0.047	0.052
5000	0.054	0.037	0.037	0.041	0.052	0.041	0.041	0.044

POWERS

$PPV_1 = 0.70$, $PPV_2 = 0.80$, $PPV_3 = 0.90$, $NPV_1 = 0.75$, $NPV_2 = 0.85$, $NPV_3 = 0.95$
$p = 60\%$
$Se_1 = 0.9074$, $Sp_1 = 0.4167$, $Se_2 = 0.9231$, $Sp_2 = 0.6538$, $Se_3 = 0.9706$, $Sp_3 = 0.8382$

$\lambda_{111} = 0.70$, $\lambda_{110} = 0.40$, $\lambda_{101} = 0.40$, $\lambda_{100} = 0.25$
$\lambda_{011} = 0.40$, $\lambda_{010} = 0.25$, $\lambda_{001} = 0.25$, $\lambda_{000} = 0.05$

n	$\delta = 0.001$, $\varepsilon = 0.01$				$\delta = 0.002$, $\varepsilon = 0.02$			
	Global test	**M1**	**M2**	**M3**	**Global test**	**M1**	**M2**	**M3**
100	0.042	0.041	0.041	0.041	0.034	0.033	0.033	0.034
200	0.261	0.260	0.260	0.260	0.219	0.218	0.218	0.218
500	0.823	0.823	0.823	0.823	0.741	0.740	0.740	0.740
1000	0.986	0.986	0.986	0.986	0.950	0.950	0.950	0.950
2000	1	1	1	1	1	1	1	1
5000	1	1	1	1	1	1	1	1

$\lambda_{111} = 1$, $\lambda_{110} = 0.80$, $\lambda_{101} = 0.80$, $\lambda_{100} = 0.40$
$\lambda_{011} = 0.80$, $\lambda_{010} = 0.40$, $\lambda_{001} = 0.40$, $\lambda_{000} = 0.20$

n	$\delta = 0.001$, $\varepsilon = 0.01$				$\delta = 0.002$, $\varepsilon = 0.02$			
	Global test	**M1**	**M2**	**M3**	**Global test**	**M1**	**M2**	**M3**
100	0.316	0.315	0.315	0.315	0.276	0.275	0.275	0.275
200	0.724	0.724	0.724	0.724	0.652	0.651	0.651	0.651
500	0.971	0.970	0.970	0.971	0.910	0.909	0.909	0.910
1000	1	1	1	1	0.995	0.995	0.995	0.995
2000	1	1	1	1	1	1	1	1
5000	1	1	1	1	1	1	1	1

In general terms, its type I error probability fluctuates around the nominal error (especially for $n \geqslant 500$) and it is necessary to have large samples ($n \geqslant 500$) so that the power is greater than 80%. From the results obtained in the simulation experiments carried out, the global hypothesis test

$$H_0\colon\ PPV_1 = PPV_2 = \cdots = PPV_J \quad \text{and} \quad NPV_1 = NPV_2 = \cdots = NPV_J\ ,$$
$$H_1\colon\ \text{at least one equality is not true}\ ,$$

can be solved through the following procedure:

1. Solving the global hypothesis test based on the chi-squared distribution to a global error rate of α using the statistics given by equations (2.4) or bootstrap method.

2. If the global hypothesis test is not significant, then one cannot reject the homogeneity of the J *PPVs* and of the J *NPVs*. If the global hypothesis test is significant to an error rate of α, in order to investigate the causes of the significance the following marginal hypothesis tests are solved

$$H_0\colon\ PPV_i = PPV_j \quad \text{vs} \quad H_1\colon\ PPV_i \neq PPV_j$$
and
$$H_0\colon\ NPV_i = NPV_j \quad \text{vs} \quad H_1\colon\ NPV_i \neq NPV_j$$

using the statistics given by equation (2.5), and applying some of the methods of multiple comparison used (Bonferroni, Holm or Hochberg) to an error rate of α.

4. EXAMPLE

The results obtained in Section 2 and the procedure given in Section 3.3 were applied to the diagnosis of coronary stenosis. Coronary stenosis is a disease that consists of the obstruction of the coronary artery and its diagnosis can be made through a dobutamine echocardiogram, a stress echocardiogram or a *CT* scan, and as the gold standard a coronary angiography is used. Coronary angiography may cause different reactions in patients (thrombosis, heart attacks, infections, even death) and therefore not all patients are verified with the gold standard. In Table 3 (Study of coronary stenosis), we show the results obtained by applying the dobutamine echocardiogram (variable T_1), the stress echocardiogram (variable T_2) and the CT scan (variable T_2) to a sample of 2455 males over 45 years of age and by only applying the coronary angiography (variable D) to a subset of these individuals. This study was carried out in two phases: firstly, the three *BDTs* were applied to all of the individuals in the sample, and secondly the gold standard was applied to a subset of these individuals depending on only the results of the three diagnostic tests. This data are part of a study carried out at the University Hospital in Granada (Spain). In this example, one can assume that

the missing data mechanism is ignorable, and therefore the results from Section 3 can be applied. The values of the estimators of the *PVs* are $\widehat{PPV_1} = 0.742$, $\widehat{PPV_2} = 0.622$, $\widehat{PPV_3} = 0.805$, $\widehat{NPV_1} = 0.933$, $\widehat{NPV_2} = 0.850$, $\widehat{NPV_3} = 0.952$, and applying the delta method, the estimated asymptotic variance-covariance matrix is

$$\hat{\sum}_{\hat{\boldsymbol{\tau}}} = \begin{pmatrix} 0.000234 & 0.000108 & 0.000086 & 0 & -0.000063 & -0.000038 \\ 0.000108 & 0.000258 & 0.000106 & -0.000035 & 0 & -0.000025 \\ 0.000086 & 0.000106 & 0.0000239 & -0.000059 & -0.000069 & 0 \\ 0 & -0.000034 & -0.000059 & 0.000114 & 0.000080 & 0.000045 \\ 0.000063 & 0 & 0.000069 & 0.000080 & 0.000169 & 0.000064 \\ 0.000038 & 0.000025 & 0 & 0.000045 & 0.000064 & 0.000085 \end{pmatrix}.$$

Applying equation (2.4) it holds that $Q^2 = 145.103$ ($p - value = 0$), and therefore we reject the equality of the three *PPVs* and of the three *NPVs*. In order to investigate the causes of the significance, the marginal hypothesis tests ($H_0 \colon PPV_i = PPV_j$ and $H_0 \colon NPV_i = NPV_j$) are solved. In Table 3 (Marginal hypothesis tests), we show the results obtained for each one of the six hypothesis tests that compare the *PVs*. Applying the Bonferroni method, the Holm method or the Hochberg method, it holds that the three *PPVs* are different, and that the *PPV* of the *CT* scan is the largest, followed by that of the dobutamine echocardiogram and, finally, that of the stress echocardiogram.

Table 3: Data from the study of coronary stenosis and marginal hypothesis tests.

Study of coronary stenosis									
	$T_1 = 1$				$T_1 = 0$				
	$T_2 = 1$		$T_2 = 0$		$T_2 = 1$		$T_2 = 0$		
	$T_3 = 1$	$T_3 = 0$	$T_3 = 1$	$T_3 = 0$	$T_3 = 1$	$T_3 = 0$	$T_3 = 1$	$T_3 = 0$	Total
V = 1									
D = 1	457	30	84	5	34	0	7	1	618
D = 0	41	23	5	61	16	86	32	95	359
V = 0	92	31	85	120	42	195	88	825	1478
Total	590	84	174	186	92	281	127	921	2455

Marginal hypothesis tests		
Hypothesis test	z	Two sided p-value
$H_0 \colon PPV_1 = PPV_2$ vs $H_1 \colon PPV_1 \neq PPV_2$	3.61	0.003
$H_0 \colon PPV_1 = PPV_3$ vs $H_1 \colon PPV_1 \neq PPV_3$	7.20	6.06×10^{-13}
$H_0 \colon PPV_2 = PPV_3$ vs $H_1 \colon PPV_2 \neq PPV_3$	10.79	0
$H_0 \colon NPV_1 = NPV_2$ vs $H_1 \colon NPV_1 \neq NPV_2$	7.46	8.37×10^{-14}
$H_0 \colon NPV_1 = NPV_3$ vs $H_1 \colon NPV_1 \neq NPV_3$	1.76	0.078
$H_0 \colon NPV_2 = NPV_3$ vs $H_1 \colon NPV_2 \neq NPV_3$	8.99	0

Regarding the *NPVs*, no significant differences were found between the *NPVs* of the dobutamine echocardiogram and of the *CT* scan, whilst the *NPV* of the dobutamine echocardiogram is significantly lower than the *NPVs* of the other two *BDTs*.

5. DISCUSSION

Different studies have examined the problem of the comparison of the *PVs* of two or more *BDTs* when the diagnostic tests and the gold standard are applied to all of the individuals in a random sample. These models cannot be applied when a subset of individuals in the random sample have not had their disease status verified through the application of the gold standard, since the results obtained may be affected by the verification bias. In this article, we have studied a global hypothesis test to simultaneously compare the *PVs* of two or more *BDTs* when for a subset of individuals in the sample the disease status (either present or absent) is unknown. The global hypothesis test is based on the chi-squared distribution, and can be solved through the method of maximum likelihood and the delta method (equation (2.4) or through the bootstrap method, although the latter requires a greater computational effort. In terms of the type I error probability, both methods lead to very similar results, and the type I error probability fluctuates around the nominal error especially for $n \geqslant 500$. Other alternative methods to solve the global hypothesis test have been studied. The method based on the marginal comparisons of the *PPVs* (*NPVs*) to an error rate of $\alpha = 5\%$ leads to a type I error probability that clearly overwhelms the nominal error, and therefore this method may give rise to erroneous results. The methods based on marginal comparisons applying the corrections of Bonferroni, Holm and Hochberg respectively give rise to a type I error probability that fluctuates around the nominal error especially for $n \geqslant 500$. In terms of power, the global hypothesis test based on the chi-squared distribution (equation (2.4) or bootstrap method) is a little more powerful than the methods based on the corrections of Bonferroni, Holm and Hochberg respectively. Therefore, from the results of the simulation experiments carried out, the following method is proposed to compare the *PVs* of J *BDTs* in the presence of ignorable missing data: 1) Apply the global hypothesis test based on the chi-squared distribution to an error rate of α (equations (2.4) or bootstrap method); 2) If the global hypothesis test is significant to an error rate of α, investigate the causes of the significance solving the marginal hypothesis tests $H_0\colon PPV_i = PPV_j$ and $H_0\colon NPV_i = NPV_j$ along with a method of multiple comparisons (Bonferroni, Holm or Hochberg). This procedure is similar to that used in an analysis of variance. Firstly, the global test is solved and then a method of multiple comparisons is applied.

An alternative method to that proposed in Section 2 consists of solving the global test applying the Wilks method. Similar simulation experiments to those described in Section have demonstrated that the type I error probability and the power of this method are very similar to those obtained with the Wald method (equation (2.4)).

If all of the individuals are verified with the gold standard, and therefore all of the frequencies $u_{i_1,\ldots,i_J}$ are equal to 0, the method proposed by Roldán-Nofuentes *et al.* [5] is a particular case of the scenario analyzed in this study. Therefore, the simultaneous comparison of the *PVs* of two (or more) *BDTs* in paired designs is a particular case of the scenario analyzed in this article.

APPENDIX A

In the case of two *BDTs*, the probabilities of the multinomial distribution were calculated applying the model of conditional dependence of Vacek [20], and their expressions are

$$\begin{aligned}
\phi_{ij} &= \lambda_{ij} p \left\{ Se_1^i (1-Se_1)^{1-i} Se_2^j (1-Se_2)^{1-j} + \delta_{ij}\varepsilon_1 \right\}, \\
\varphi_{ij} &= \lambda_{ij} (1-p) \left\{ Sp_1^{1-i} (1-Sp_1)^i Sp_2^{1-j} (1-Sp_2)^j + \delta_{ij}\varepsilon_0 \right\}, \\
\gamma_{ij} &= (1-\lambda_{ij}) p \left\{ Se_1^i (1-Se_1)^{1-i} Se_2^j (1-Se_2)^{1-j} + \delta_{ij}\varepsilon_1 \right\} \\
&\quad + (1-\lambda_{ij}) (1-p) \left\{ Sp_1^{1-i} (1-Sp_1)^i Sp_2^{1-j} (1-Sp_2)^j + \delta_{ij}\varepsilon_0 \right\},
\end{aligned}$$

where $\delta_{ij} = 1$ when $i = j$ and $\delta_{ij} = -1$, and ε_i is the dependence factor (covariance) between the two *BDTs* when $D = i$. In clinical practice, the two *BDTs* are usually conditionally dependent on the disease, and it is verified [20] that $0 < \varepsilon_1 < Se_1 (1 - Se_2)$ when $Se_2 > Se_1$ and $0 < \varepsilon_1 < Se_2 (1 - Se_1)$ when $Se_1 > Se_2$, and in the same way, $0 < \varepsilon_0 < Sp_1 (1 - Sp_2)$ when $Sp_2 > Sp_1$ and $0 < \varepsilon_0 < Sp_2 (1 - Sp_1)$ when $Sp_1 > Sp_2$. If the two *BDTs* are conditionally independent on the disease then $\varepsilon_1 = \varepsilon_0 = 0$.

In the case of three *BDTs*, the probabilities of the multinomial distributions were calculated applying the model of Torrance-Rynard and Walter [21]:

$$\begin{aligned}
&P\Big(V=1,\, D=1,\, T_1=i_1,\, T_2=i_2,\, T_3=i_3\Big) = \\
&\qquad = p\, \lambda_{i_1 i_2 i_3} \left\{ \prod_{j=1}^{3} Se_j^{i_j} (1-Se_j)^{1-i_j} + \sum_{j,k,j<k}^{3} (-1)^{|i_j - i_k|} \delta_{jk} \right\},
\end{aligned}$$

$$\begin{aligned}
&P\Big(V=1,\, D=0,\, T_1=i_1,\, T_2=i_2,\, T_3=i_3\Big) = \\
&\qquad = (1-p)\, \lambda_{i_1 i_2 i_3} \left\{ \prod_{j=1}^{3} {Sp_j}^{1-i_j} (1-Sp_j)^{i_j} + \sum_{j,k,j<k}^{3} (-1)^{|i_j - i_k|} \varepsilon_{jk} \right\}
\end{aligned}$$

and

$$\begin{aligned}
&P\Big(V=0,\, T_1=i_1,\, T_2=i_2,\, T_3=i_3\Big) = \\
&\quad = p\,(1-\lambda_{i_1 i_2 i_3}) \left\{ \prod_{j=1}^{3} {Se_j}^{i_j} (1-Se_j)^{1-i_j} + \sum_{j,k,j<k}^{3} (-1)^{|i_j - i_k|} \delta_{jk} \right\} \\
&\qquad + (1-p)\,(1-\lambda_{i_1 i_2 i_3}) \left\{ \prod_{j=1}^{3} {Sp_j}^{1-i_j} (1-Sp_j)^{i_j} + \sum_{j,k,j<k}^{3} (-1)^{|i_j - i_k|} \varepsilon_{jk} \right\},
\end{aligned}$$

with $i_j = 0, 1$, $i_k = 0, 1$ and $j, k = 1, 2, 3$, and where δ_{jk} (ε_{jk}) is the factor of dependence between the j-th *BDT* and k-th *BDT* when $D = 1$ ($D = 0$).

The factors of dependence δ_{jk} and ε_{jk} verify restrictions that depend on the values of sensitivity and specificity of the three *BDTs*. In order to simplify the simulation experiments, it has been considered that $\delta_{ij} = \delta$ and $\varepsilon_{ij} = \varepsilon$, so that the factors of dependence verify the following restrictions:

$$\delta \leq \text{Min}\Big\{(1-Se_1)(1-Se_2)Se_3,\ (1-Se_1)Se_2(1-Se_3),\ Se_1(1-Se_2)(1-Se_3)\Big\}$$

and

$$\varepsilon \leq \text{Min}\Big\{(1-Sp_1)(1-Sp_2)Sp_3,\ (1-Sp_1)Sp_2(1-Sp_3),\ Sp_1(1-Sp_2)(1-Sp_3)\Big\}.$$

In clinical practice, factors δ_{jk} and/or ε_{jk} are greater than zero, so that the *BDTs* are conditionally dependent on the disease status. When $\delta_{jk} = \varepsilon_{jk} = 0$ the three BDTs are conditionally independent on the disease status.

ACKNOWLEDGMENTS

This research was supported by the Spanish Ministry of Economy, Grant Number MTM2012-35591. We thank the two referees, the Associate Editor and the Editor (M. Ivette Gomes) of REVSTAT for their helpful comments that improved the quality of the paper.

REFERENCES

[1] Bennet, B. M. (1972). On comparison of sensitivity, specificity and predictive value of a number of diagnostic procedures, *Biometrics*, **28**, 793–800.

[2] Bennet, B. M. (1985). On tests for equality of predictive values for t diagnostic procedures, *Statistics in Medicine*, **4**, 525–539.

[3] Leisenring, W.; Alonzo, T. and Pepe, M. S. (2000). Comparisons of predictive values of binary medical diagnostic tests for paired designs, *Biometrics*, **56**, 354–351.

[4] Wnag, 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] 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, Special issue "Computational Statistics for Clinical Research"* , **56**, 1161–1173.

[6] Poleto, F. Z.; Singer, J. M. and Paulino, C. D. (2011). Comparing diagnostic tests with missing data, *Journal of Applied Statistics*, **38**, 1207–1222.

[7] Begg, C. B. and Greenes, R. A. (1983). Assessment of diagnostic tests when disease verification is subject to selection bias, *Biometrics*, **39**, 207–215.

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

[9] Roldán Nofuentes, J. A. and Luna del Castillo, J. D. (2007). The effect of verification bias in the naive estimators of accuracy of a binary diagnostic test, *Communications in Statistics — Simulation and Computation*, **36**, 959–972.

[10] Roldán Nofuentes, J. A. and Luna del Castillo, J. D. (2008a). EM algorithm for comparing two binary diagnostic tests when not all the patients are verified, *Journal of Statistical Computation and Simulation*, **78**, 19–35.

[11] Roldán Nofuentes, J. A. and Luna del Castillo, J. D. (2008b). The effect of verification bias on the comparison of predictive values of two binary diagnostic tests, *Journal of Statistical Planning and Inference*, **138**, 959–963.

[12] Rubin, D. B. (1976). Inference and missing data, *Biometrika*, **4**, 73–89.

[13] Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*, Chapman and Hall/CRC, USA.

[14] Bishop, Y. M.; Fienberg, S. E. and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*, The MIT Press.

[15] Agresti, A. (2002). *Categorical Data Analysis*, John Wiley and Sons, New York.

[16] Efron, B. and Tibshirsni, R. J. (1993). *An Introduction to the Bootstrap*, Chapman and Hall, New York.

[17] Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilitá, *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] Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance, *Biometrika*, **75**, 800–802.

[20] Vacek, P. (1985). The effect of conditional dependence on the evaluation of diagnostic tests, *Biometrics*, **23**, 959–968.

[21] Torrance-Rynard, V. L. and Walter, S. D. (1997). Effects of dependent errors in the assessment of diagnostic test performance, *Statistics in Medicine*, **16**, 2157–2175.