mathematics

MDPI

Article

Simultaneous Comparison of Sensitivities and Specificities of Two Diagnostic Tests Adjusting for Discrete Covariates

José Antonio Roldán-Nofuentes

Department of Statistics, School of Medicine, University of Granada, 18016 Granada, Spain; jaroldan@ugr.es

Abstract: Adjusting for covariates is important in the study of the performance of diagnostic tests. In this manuscript, the simultaneous comparison of the sensitivities and specificities of two binary diagnostic tests is studied when discrete covariates are observed in all of the individuals in the sample. Four methods are presented to simultaneously compare the two sensitivities and the two specificities: a global hypothesis test and three other methods based on individual comparisons. The maximum likelihood method was applied to adjust the overall estimators of sensitivities and specificities. Simulation experiments were carried out to study the asymptotic behaviors of the four proposed methods when the covariate is binary, giving general rules of application. The results were applied to a real example.

Keywords: binary diagnostic test; sensitivity; specificity; simultaneous comparison

check for updates

Citation: Roldán-Nofuentes, J.A. Simultaneous Comparison of Sensitivities and Specificities of Two Diagnostic Tests Adjusting for Discrete Covariates. *Mathematics* **2021**, *9*, 2029. https://doi.org/10.3390/math9172029

Academic Editor: Maria Laura Manca

Received: 28 July 2021
Accepted: 20 August 2021
Published: 24 August 2021

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 may be either positive or negative, the diagnostic test is called a binary diagnostic test (BDT). The exercise test for the diagnosis of coronary artery disease is an example of a BDT. The fundamental parameters to measure the effectiveness of a BDT are its sensitivity and the specificity. The sensitivity (Se) is the probability that the BDT result is positive when the individual has the disease, and the specificity (Sp) is the probability that the BDT result is negative when the individual does not have the disease. Both parameters depend only on the intrinsic properties (physical, biological, chemical, etc.) of the BDT. The effectiveness of a BDT is assessed in relation to a gold standard. A gold standard (GS) is a medical test used to objectively diagnose the presence (or absence) of a certain disease. Therefore, a GS is an error-free test. An angiography for diagnosis of coronary artery disease is an example of a GS.

The comparison of the sensitivities (specificities) of two BDTs is an important topic in the study of statistical methods for diagnosis in medicine. The most common type of sample design to compare these parameters is the paired design. The paired design consists of applying the two BDTs to a random sample of n patients whose disease state is known by applying a GS. When the sensitivities and specificities of two BDTs are compared under a paired design, the problem is traditionally solved by conditioning on the disease status and applying a comparison test of two paired binomial proportions (e.g., the McNemar test). Therefore, the comparison of the two sensitivities is made conditioning on the diseased individuals and solving the test $H_0 : Se_1 = Se_2$ vs. $H_1 : Se_1 \neq Se_2$ applying the McNemar test to an α error [1]. Similarly, the comparison of the two specificities is made conditioning on the non-diseased individuals and solving the test $H_0 : Sp_1 = Sp_2$ vs. $H_1 : Sp_1 \neq Sp_2$ by applying the same method. Therefore, sensitivities and specificities are compared independently, by solving the hypothesis tests $H_0 : Se_1 = Se_2$ and $H_0 : Sp_1 = Sp_2$, to the same α error. Roldán-Nofuentes and Sidaty-Regad [2] studied the simultaneous comparison of sensitivities and specificities, and showed that comparing the two sensitivities and the two specificities independently can give rise to global type I errors that greatly exceed the nominal error (and therefore can lead to wrong conclusions).

In clinical practice, when evaluating the effectiveness of a BDT, covariates are frequently observed in all patients in the sample. When the covariate is related to the disease and to the diagnostic test, it is necessary to adjust for covariates [3]. For example, in the diagnosis of coronary disease, smoking is a risk factor for the disease. Because smoking speeds up the heart rate, constricts the main arteries, and can cause disturbances in the rhythm of the heartbeat, if an exercise test is used, adjustment for smoking is needed to properly describe the diagnostic effectiveness of the exercise test. Another topical example is the diagnosis of COVID-19. Lahner et al. [4] studied the diagnosis of this disease in health workers using IgG serology as a diagnostic test (among other tests). Lahner et al. showed that the diagnostic performance of IgG serology is associated with the number of days elapsed (at least 14 or 20 days) after the nasopharyngeal swab. Therefore, adjusting for elapsed days is necessary to evaluate the diagnostic effectiveness of IgG serology. This problem also arises when comparing the effectiveness of two BDTs [3]. Therefore, when two BDTs are compared, it is necessary to eliminate the effect that the covariates have on the estimation of sensitivities and specificities, and on the comparison of these parameters.

This manuscript is an extension of the study by Roldán-Nofuentes and Sidaty-Regad [2], to the situation in which a discrete covariate is observed in all patients in the sample. Therefore, a global hypothesis test was studied to simultaneously compare the sensitivities and specificities of two BDTs when discrete covariates are observed in all patients in the sample. Other alternatives to the global hypothesis test were also studied. Adjusting for covariates in this situation eliminates the effect of covariates in the simultaneous comparison of the two sensitivities and specificities. This problem is approached by applying the maximum likelihood method to the estimation of the parameters and the delta method to the estimation of the variances-covariances. This manuscript is structured as follows. In Section 2, the model to simultaneously compare the sensitivities and specificities of two BDTs in the presence of a discrete covariate is described, in addition to other alternative methods. In Section 3, simulation experiments are carried out to study the sizes and the powers of the methods proposed in Section 2. In Section 4, a function written in R [5] is presented that allows the problem studied in this manuscript to be solved. In Section 5, the results are applied to the diagnosis of coronary heart disease, and in Section 6 the results are discussed.

2. Global Hypothesis Test

The objective is to study the simultaneous comparison of overall sensitivities and overall specificities of the two BDTs, i.e., to solve the global hypothesis test:

$$H_0 : (Se_1 = Se_2 \text{ and } Sp_1 = Sp_2) \text{ vs. } H_1 : (Se_1 \neq Se_2 \text{ and/or } Sp_1 \neq Sp_2) \tag{1}$$

when the two BDTs are applied to all individuals in a sample with a size of n and a discrete covariate is observed in all of them. Therefore, let us consider two BDTs, Test 1 and Test 2, that are applied to all n individuals in a random sample. The disease state (disease present or disease absent) of all of the individuals in the sample is known by applying a GS. Let T_h be the binary random variable that models the result of the hth BDT: $T_h = 1$ when the result of the BDT is positive and $T_h = 0$ when it is negative. Let the binary random variable D that models the result of the GS: $D = 1$ when the individual is diseased and $D = 0$ when the individual is non-diseased. Moreover, let us consider that for all of the n individuals of the sample we observe a vector $\mathbf{X} = (X_1, X_2, \ldots, X_M)$ of a discrete covariate, where X_m is each of the different values or patterns that the covariate can take with $m = 1, \ldots, M$. Let us suppose that the number of individuals that verify $\mathbf{X} = X_m$ is n_m, and therefore $n = \sum_{m=1}^{M} n_m$. Table 1 shows the observed frequencies for $\mathbf{X} = X_m$, where $n_{ijm} = s_{ijm} + r_{ijm}$.

Table 1. Observed frequencies for $\mathbf{X} = X_m$.

	$T_1 = 1$		$T_1 = 0$		Total
	$T_2 = 1$	$T_2 = 0$	$T_2 = 1$	$T_2 = 0$	
$D = 1$	s_{11m}	s_{10m}	s_{01m}	s_{00m}	s_m
$D = 0$	r_{11m}	r_{10m}	r_{01m}	r_{00m}	r_m
Total	n_{11m}	n_{10m}	n_{01m}	n_{00m}	n_m

The sample of n individuals is the product of a multinomial distribution whose probabilities are:

$$\tau_{mij} = P(\mathbf{X} = X_m, D = 1, T_1 = i, T_2 = j)$$

and:

$$\upsilon_{mij} = P(\mathbf{X} = X_m, D = 0, T_1 = i, T_2 = j),$$

with:

$$\sum_{m=1}^{M}\sum_{i,j=0}^{1}\tau_{mij} + \sum_{m=1}^{M}\sum_{i,j=0}^{1}\upsilon_{mij} = 1.$$

From the multinomial distribution sized n and probabilities τ_{mij} and υ_{mij}, $8M - 1$ parameters can be estimated, because in total there are $8M$ probabilities that are subject to $\sum_{m=1}^{M}\sum_{i,j=0}^{1}\tau_{mij} + \sum_{m=1}^{M}\sum_{i,j=0}^{1}\upsilon_{mij} = 1$ (i.e., $\upsilon_{M11} = 1 - \sum_{m=1}^{M}\sum_{i,j=0}^{1}\tau_{mij} - \sum_{\substack{m=1\\(m,i,j)\neq(M,1,1)}}^{M}\sum_{i,j=0}^{1}\upsilon_{mij}$). If the covariate is binary, then 15 parameters can be estimated.

Let $\psi_m = P(\mathbf{X} = X_m)$ be the probability that an individual $\mathbf{X} = X_m$ and $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_M)^T$, with $\sum_{m=1}^{M}\psi_m = 1$. Let ϕ_{ijm} and φ_{ijm} be the probabilities defined as

$$\phi_{ijm} = P(D = 1, T_1 = i, T_2 = j|\mathbf{X} = X_m) \text{ and } \varphi_{ijm} = P(D = 0, T_1 = i, T_2 = j|\mathbf{X} = X_m),$$

then probabilities τ_{mij} and υ_{mij} can be written as:

$$\tau_{mij} = \psi_m\phi_{ijm} \text{ and } \upsilon_{mij} = \psi_m\varphi_{ijm}. \tag{2}$$

The sample of n individuals can be seen as a sample of a mixture of M multinomial independent 2×4 tables. By conditioning on the mth table, i.e., conditioning on $\mathbf{X} = X_m$, and applying the conditional dependence model of Berry et al. [6], it holds that:

$$\begin{gathered}\phi_{ijm} = P(D = 1, T_1 = i, T_2 = j|\mathbf{X} = X_m) = \\ P(D = 1|\mathbf{X} = x_m)\left[P(T_1 = i|\mathbf{X} = X_m, D = 1) \times P(T_2 = j|\mathbf{X} = X_m, D = 1) + \delta_{ij}\varepsilon_{1m}\right] = \\ p_m\left[Se_{1m}^{i}(1 - Se_{1m})^{1-i}Se_{2m}^{j}(1 - Se_{2m})^{1-j} + \delta_{ij}Se_{1m}Se_{2m}(\alpha_{1m} - 1)\right]\end{gathered}$$

and:

$$\begin{gathered}\varphi_{ijm} = P(D = 0, T_1 = i, T_2 = j|\mathbf{X} = X_m) = \\ P(D = 0|\mathbf{X} = x_m)\left[P(T_1 = i|\mathbf{X} = X_m, D = 0) \times P(T_2 = j|\mathbf{X} = X_m, D = 0) + \delta_{ij}\varepsilon_{0m}\right] = \\ q_m\left[Sp_{1m}^{1-i}(1 - Sp_{1m})^{i}Sp_{2m}^{1-j}(1 - Sp_{2m})^{j} + \delta_{ij}(1 - Sp_{1m})(1 - Sp_{2m})(\alpha_{0m} - 1)\right],\end{gathered}$$

where $p_m = P(D = 1|\mathbf{X} = X_m) = \sum_{i,j=0}^{1}\phi_{ijm}$ is the disease prevalence for the individuals with $\mathbf{X} = X_m$, $q_m = 1 - p_m$, $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = -1$ if $i \neq j$, and the parameter α_{1m} (α_{0m}) is the covariance [6] between both BDTs when $D = 1$ $(D = 0)$ and $\mathbf{X} = X_m$. The covariances verify [6] that $1 \leq \alpha_{1m} \leq 1/\max\{Se_{1m}, Se_{2m}\}$ and $1 \leq \alpha_{0m} \leq 1/\max\{(1 - Sp_{1m}), (1 - Sp_{2m})\}$. If $\alpha_{1m} = \alpha_{0m} = 1$, then both BDTs are conditionally independent on the disease when $\mathbf{X} = X_m$, an assumption that is not realistic, so in practice $\alpha_{1m} > 1$ and/or $\alpha_{0m} > 1$.

For the mth table (i.e., $\mathbf{X} = X_m$), let $\boldsymbol{\omega}_m = (\phi_{11m}, \phi_{10m}, \phi_{01m}, \phi_{00m}, \varphi_{11m}, \varphi_{10m}, \varphi_{01m}, \varphi_{00m})^T$ be the vector whose components are the probabilities ϕ_{ijm} and $]\varphi_{ijm}$. Therefore, conditioning on $\mathbf{X} = X_m$, $\boldsymbol{\omega}_m$ is the probability vector of a multinomial distribution. Let $\boldsymbol{\omega} = (\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_M)^T$ be the vector whose components are $\boldsymbol{\omega}_m$. In $\mathbf{X} = X_m$, the sensitivities of the BDTs are:

$$Se_{1m} = P(T_1 = 1 | D = 1, \mathbf{X} = X_m) \text{ and } Se_{2m} = P(T_2 = 1 | D = 1, \mathbf{X} = X_m),$$

and the specificities are:

$$Sp_{1m} = P(T_1 = 0 | D = 0, \mathbf{X} = X_m) \text{ and } Sp_{2m} = P(T_2 = 0 | D = 0, \mathbf{X} = X_m).$$

Let $p = \sum\limits_{m=1}^{M} \psi_m p_m = \sum\limits_{m=1}^{M} \psi_m \left(\sum\limits_{i,j=0}^{1} \phi_{ijm} \right)$ be the overall prevalence and $q = 1 - p$ $= \sum\limits_{m=1}^{M} \psi_m q_m = \sum\limits_{m=1}^{M} \psi_m \left(\sum\limits_{i,j=0}^{1} \varphi_{ijm} \right)$. The overall sensitivity and the overall specificity of each BDT are:

$$Se_h = P(T_h = 1 | D = 1) = \frac{\sum\limits_{m=1}^{M} \psi_m p_m Se_{hm}}{p} \text{ and } Sp_h = P(T_h = 0 | D = 0) = \frac{\sum\limits_{m=1}^{M} \psi_m q_m Sp_{hm}}{q}, \tag{3}$$

With $h = 1, 2$, and where:

$$Se_{1m} = \frac{\phi_{11m} + \phi_{10m}}{p_m} \text{ and } Sp_{1m} = \frac{\varphi_{01m} + \varphi_{00m}}{q_m}$$

are the sensitivity and specificity of Test 1 in $\mathbf{X} = X_m$, and:

$$Se_{2m} = \frac{\phi_{11m} + \phi_{01m}}{p_m} \text{ and } Sp_{2m} = \frac{\varphi_{10m} + \varphi_{00m}}{q_m}$$

are the sensitivity and specificity of Test 2 in $\mathbf{X} = X_m$. The overall sensitivity and the overall specificity of each BDT are written in terms of ψ_m, ϕ_{ijm} and φ_{ijm} as:

$$Se_1 = \frac{\sum\limits_{m=1}^{M} \psi_m (\phi_{11m} + \phi_{10m})}{\sum\limits_{m=1}^{M} \left(\psi_m \sum\limits_{i,j=0}^{1} \phi_{ijm} \right)} \text{ and } Sp_1 = \frac{\sum\limits_{m=1}^{M} \psi_m (\varphi_{00m} + \varphi_{01m})}{\sum\limits_{m=1}^{M} \left(\psi_m \sum\limits_{i,j=0}^{1} \varphi_{ijm} \right)}$$

for Test 1, and:

$$Se_2 = \frac{\sum\limits_{m=1}^{M} \psi_m (\phi_{11m} + \phi_{01m})}{\sum\limits_{m=1}^{M} \left(\psi_m \sum\limits_{i,j=0}^{1} \phi_{ijm} \right)} \text{ and } Sp_2 = \frac{\sum\limits_{m=1}^{M} \psi_m (\varphi_{00m} + \varphi_{10m})}{\sum\limits_{m=1}^{M} \left(\psi_m \sum\limits_{i,j=0}^{1} \varphi_{ijm} \right)}$$

for Test 2.

The parameters of the model are estimated by applying the maximum likelihood method. If the covariate has M patterns then $8M - 1$ parameters must be estimated: $2M$ sensitivities, $2M$ specificities, $2M$ covariances, M prevalences and $M - 1$ probabilities ψ_m (since $\sum\limits_{m=1}^{M} \psi_m = 1$). If the covariate is binary ($M = 2$) then 15 parameters must be estimated:

four sensitivities (Se_{11}, Se_{21}, Se_{12} and Se_{22}), four specificities (Sp_{11}, Sp_{21}, Sp_{12} and Sp_{22}), four covariances (α_{11}, α_{01}, α_{12} and α_{02}), two prevalences (p_1 and p_2) and the probability ψ_1 (since $\psi_2 = 1 - \psi_1$). Therefore, all the parameters of the model can be estimated from the sample of n individuals, since the number of parameters that must be estimated is equal to the number of parameters that can be estimated from the initial multinomial distribution. The log-likelihood function based on n individuals is:

$$l(\boldsymbol{\psi}, \boldsymbol{\omega}) = \sum_{i,j=0}^{1} \sum_{m=1}^{M} x_{ijm} \log\left(\psi_m \phi_{ijm}\right) + \sum_{i,j=0}^{1} \sum_{m=1}^{M} y_{ijm} \log\left(\psi_m \varphi_{ijm}\right).$$

This function can be written as:

$$l(\boldsymbol{\psi}, \boldsymbol{\omega}) = l_1(\boldsymbol{\psi}) + l_2(\boldsymbol{\omega}), \tag{4}$$

where:

$$l_1(\boldsymbol{\psi}) = \sum_{i,j=0}^{1} \sum_{m=1}^{M} n_{ijm} \log(\psi_m) \tag{5}$$

and:

$$l_2(\boldsymbol{\omega}) = \sum_{i,j=0}^{1} \sum_{m=1}^{M} x_{ijm} \log\left(\phi_{ijm}\right) + \sum_{i,j=0}^{1} \sum_{m=1}^{M} y_{ijm} \log\left(\varphi_{ijm}\right). \tag{6}$$

Maximum likelihood estimators of $\boldsymbol{\psi}$ and $\boldsymbol{\omega}$ are easily obtained from Functions (5) and (6), i.e.,

$$\hat{\psi}_m = \frac{n_m}{n},\ \hat{\phi}_{ijm} = \frac{s_{ijm}}{n_m} \text{ and } \hat{\varphi}_{ijm} = \frac{r_{ijm}}{n_m}.$$

The estimators of sensitivities and specificities in $\mathbf{X} = X_m$, the estimator of overall prevalence, and the estimators of overall sensitivities and of overall specificities are easily obtained by substituting the parameters for their estimators into their respective equations. The Fisher information matrix of function (4) is:

$$I(\boldsymbol{\psi}, \boldsymbol{\omega}) = \text{Diag}\{I_1, I_2\},$$

where $I_1 = I(\boldsymbol{\psi})$ and $I_2 = I(\boldsymbol{\omega})$ are the Fisher information matrixes of Functions (5) and (6) respectively, verifying that:

$$I^{-1}(\boldsymbol{\psi}, \boldsymbol{\omega}) = \text{Diag}\left\{I_1^{-1}, I_2^{-1}\right\}$$

and, therefore, the covariances between $\boldsymbol{\psi}$ and $\boldsymbol{\omega}$ are zero. Because vector $\boldsymbol{\psi}$ is the probability vector of a multinomial distribution, the variance-covariance matrix of $\hat{\boldsymbol{\psi}}$ is:

$$\sum\nolimits_{\hat{\boldsymbol{\psi}}} = I_1^{-1} = \left\{Diag(\boldsymbol{\psi}) - \boldsymbol{\psi}\boldsymbol{\psi}^T\right\}/n.$$

The variance-covariance matrix of $\hat{\boldsymbol{\omega}}_m$ is:

$$\sum\nolimits_{\hat{\boldsymbol{\omega}}_m} = \left\{Diag(\boldsymbol{\omega}_m) - \boldsymbol{\omega}_m \boldsymbol{\omega}_m^T\right\}/n_m$$

and the variance-covariance matrix of $\hat{\boldsymbol{\omega}}$ is:

$$\sum\nolimits_{\hat{\boldsymbol{\omega}}} = I_2^{-1} = Diag\left\{\sum\nolimits_{\hat{\boldsymbol{\omega}}_1}, \dots \sum\nolimits_{\hat{\boldsymbol{\omega}}_M}\right\}.$$

The proof can be seen in Appendix A.

Let $\boldsymbol{\theta} = (Se_1, Sp_1, Se_2, Sp_2)^T$ be a vector whose components are the overall sensitivities and the overall specificities; then, by applying the delta method [7], the variance-covariance matrix of $\hat{\boldsymbol{\theta}}$ is:

$$\sum_{\hat{\boldsymbol{\theta}}} = \left(\frac{\partial\boldsymbol{\theta}}{\partial\boldsymbol{\psi}}\right)\sum_{\hat{\boldsymbol{\psi}}}\left(\frac{\partial\boldsymbol{\theta}}{\partial\boldsymbol{\psi}}\right)^T + \left(\frac{\partial\boldsymbol{\theta}}{\partial\boldsymbol{\omega}}\right)\hat{\sum}_{\hat{\boldsymbol{\omega}}}\left(\frac{\partial\boldsymbol{\theta}}{\partial\boldsymbol{\omega}}\right)^T.$$

The estimated variance-covariance matrix $\hat{\sum}_{\hat{\boldsymbol{\theta}}}$ is obtained by substituting into this expression the parameters for their estimators.

The global hypothesis test (1) is equivalent to the hypothesis test:

$$H_0 : \mathbf{A}\boldsymbol{\theta} = 0 \text{ vs. } H_1 : \mathbf{A}\boldsymbol{\theta} \neq 0,$$

where $\mathbf{A}$ is a complete range matrix with the size 2×4, i.e.,

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

By applying the multivariate central limit theorem, it is verified that $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \to N(\mathbf{0}, \sum_{\boldsymbol{\theta}})$ when n is large. Then, the statistic:

$$Q^2 = \hat{\boldsymbol{\theta}}^T \mathbf{A}^T \left(\mathbf{A}\hat{\sum}_{\hat{\boldsymbol{\theta}}}\mathbf{A}^T\right)^{-1} \mathbf{A}\hat{\boldsymbol{\theta}}$$

is distributed according to a Hotelling T-squared distribution. This distribution has 2 and n degrees of freedom, where 2 is the dimension of the vector $\mathbf{A}\hat{\boldsymbol{\theta}}$. When n is large, Q^2 is distributed according to a central chi-squared distribution with 2 degrees of freedom when the null hypothesis is true, i.e.,

$$Q^2 = \hat{\boldsymbol{\eta}}^T \mathbf{A}^T \left(\mathbf{A}\hat{\sum}_{\hat{\boldsymbol{\eta}}}\mathbf{A}^T\right)^{-1} \mathbf{A}\hat{\boldsymbol{\eta}} \xrightarrow[n\to\infty]{} \chi^2_2. \quad (7)$$

To calculate this test statistic, it is necessary to verify that $s_{10m} + s_{01m} + r_{10m} + r_{01m} > 0$.

The global hypothesis test (1) can also be solved from the individual hypothesis test, i.e., $H_0 : Se_1 = Se_2$ and $H_0 : Sp_1 = Sp_2$, each of which are independent of the α error. In this situation, the corresponding test statistics are:

$$z = \frac{\hat{S}e_1 - \hat{S}e_2}{\sqrt{\hat{V}ar(\hat{S}e_1) + \hat{V}ar(\hat{S}e_2) - 2\hat{C}ov(\hat{S}e_1, \hat{S}e_2)}} \quad (8)$$

and:

$$z = \frac{\hat{S}p_1 - \hat{S}p_2}{\sqrt{\hat{V}ar(\hat{S}p_1) + \hat{V}ar(\hat{S}p_2) - 2\hat{C}ov(\hat{S}p_1, \hat{S}p_2)}}. \quad (9)$$

Both test statistics have a normal standard distribution when the sample size n is large. Another method used to solve the global test consists of solving each of the individual tests along with a method of multiple comparisons, such as the Bonferroni method [8] or the Holm method [9]. The Bonferroni and Holm methods are very easy to apply and are based on the p-values of the individual hypothesis tests. In the situation studied here, the Bonferroni method consists of solving each individual hypothesis test with an $\alpha/2$ error. The Holm method is a less conservative method than the Bonferroni method. Let p_1 and p_2 be the p-values obtained in each individual hypothesis test and let us suppose that $p_1 \leq p_2$; then, the Holm method [9] consists of the following two steps:

(1) If $p_1 > \alpha/2$, then none of the two null hypothesis $H_0 : Se_1 = Se_2$ and $H_0 : Sp_1 = Sp_2$ are rejected. If $p_1 \leq \alpha/2$, then the null hypothesis corresponding to that hypothesis test is rejected and we go on to the next step.

(2) If $p_2 > \alpha$, then the corresponding null hypothesis is not rejected. If $p_2 \leq \alpha$, then the null hypothesis is rejected and the process ends.

In this proposed model, it is assumed that a discrete covariate is observed in all of the individuals in the sample. If several discrete covariates are observed, the problem is solved in a similar manner. In this situation, a single discrete covariate is considered, whose number of patterns is the product of the patterns of the observed covariates [10]. For example, if two covariates are observed with two and three patterns, respectively, for example, sex and age group (young, adult, and older), then a covariate that has six patterns is considered (young man, adult man, older man, young woman, adult woman, and older woman).

3. Simulation Experiments

Monte Carlo simulation experiments were carried out to study the sizes and the powers of the four methods described in Section 2: global hypothesis test with $\alpha = 5\%$; individual hypothesis tests each with $\alpha = 5\%$; individual hypothesis tests along with the Bonferroni method and $\alpha = 5\%$; and individual hypothesis tests along with the Holm method and $\alpha = 5\%$. For the global hypothesis test with $\alpha = 5\%$, the global type I error is the error that is committed when the alternative hypothesis is accepted ($Se_1 \neq Se_2$ and/or $Sp_1 \neq Sp_2$) when the null hypothesis is true ($Se_1 = Se_2$ and $Sp_1 = Sp_2$). Regarding the individual hypothesis tests with $\alpha = 5\%$ (with or without a multiple comparison method), the objective is to study the magnitude and behavior of the global type I error and of the global power. The global type I error is the error made when we reject $H_0 : Se_1 = Se_2$ and/or $H_0 : Sp_1 = Sp_2$ when both are true, whether or not each test is with $\alpha = 5\%$ or applies the Bonferroni (or Holm) method. The argument for the global power is similar to this.

These experiments consisted of generating $N = 10,000$ random samples with multinomial distributions with a size of $n = \{50, 100, 200, 500, 1000, 2000\}$, whose probabilities were calculated from Equation (2). It was considered that the discrete covariate $\mathbf{X}$ is binary $(M = 2)$ with patterns X_1 and X_2, such as the presence of a risk factor (Yes or No), family history of the disease (Yes or No), or sex; this situation is very frequent in clinical practice. As values for ψ_1 ($\psi_2 = 1 - \psi_1$), we considered 0.25 and 0.50, and for the prevalence p_m, we considered the values 10%, 25%, and 50%. As values of the sensitivities (Se_{11}, Se_{12}, Se_{21} and Se_{22}) and specificities (Sp_{11}, Sp_{12}, Sp_{21} and Sp_{22}) in each pattern of the covariate, we took the values $\{0.70, 0.80, 0.90\}$. Then, from the values Se_{hm} and Sp_{hm}, we calculated the maximum values of the covariances α_{1m} and α_{0m}, and as values of α_{1m} and α_{0m}, we took intermediate and high values, i.e.,

$$\alpha_{1m} = \frac{f}{Max\{Se_{1m}, Se_{2m}\}} + 1 - f$$

and:

$$\alpha_{0m} = \frac{f}{Max\{(1 - Sp_{1m}), (1 - Sp_{2m})\}} + 1 - f,$$

with $f = \{0.10, 0.50, 0.90\}$. From all of the above values, the overall sensitivities and overall specificities were calculated by applying Equation (3). The simulation experiments were designed in such a manner that, if it is not possible in a sample to estimate a parameter (for example, if $\hat{S}e_{hm} = 0$), then that sample is discarded and another is generated in its place until N samples are obtained.

3.1. *Type I Errors*

Tables 2 and 3 show the type I errors obtained for the four methods proposed in Section 2, considering different scenarios. Table 2 shows some results for $Se_h = 0.90$ and $Sp_h = \{0.70, 0.80\}$ $(Se_h > Sp_h)$, and Table 3 shows some results for $Se_h = \{0.70, 0.80\}$ and $Sp_h = 0.90$ $(Se_h < Sp_h)$.

Table 2. Type I errors (in %) of different methods to simultaneously compare the sensitivities $(Se_h = 0.90)$ and specificities $(Sp_h = \{0.70, 0.80\})$ of two BDTs in the presence of a binary covariate.

$Se_1 = Se_2 = 0.90$, $Sp_1 = Sp_2 = 0.70$, $p_1 = 10\%$, $p_2 = 25\%$, $\psi_1 = 25\%$, $\psi_2 = 75\%$

	$\alpha_{11} = 0.009$ $\alpha_{01} = 0.021$ $\alpha_{12} = 0.009$ $\alpha_{02} = 0.021$				$\alpha_{11} = 0.045$ $\alpha_{01} = 0.105$ $\alpha_{12} = 0.045$ $\alpha_{02} = 0.105$				$\alpha_{11} = 0.081$ $\alpha_{01} = 0.189$ $\alpha_{12} = 0.081$ $\alpha_{02} = 0.189$			
n	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
50	0.34	2.14	1.08	1.08	0.08	0.88	0.34	0.34	0	0	0	0
100	1.28	4.10	2.08	2.08	0.86	2.88	1.42	1.42	0	0.02	0	0
200	1.40	4.52	2.36	2.36	1.20	3.70	1.64	1.64	0.02	0.82	0.16	0.16
500	1.98	4.48	2.60	2.60	1.78	4.66	2.46	2.46	0.58	3.00	1.38	1.38
1000	2.30	5.24	2.74	2.78	1.82	4.82	2.38	2.38	0.88	3.26	1.48	1.48
2000	3.50	7.24	3.58	3.60	2.44	5.26	2.40	2.40	1.20	4.18	1.96	1.96

$Se_1 = Se_2 = 0.90$, $Sp_1 = Sp_2 = 0.70$, $p_1 = 10\%$, $p_2 = 25\%$, $\psi_1 = 50\%$, $\psi_2 = 50\%$

	$\alpha_{11} = 0.009$ $\alpha_{01} = 0.021$ $\alpha_{12} = 0.009$ $\alpha_{02} = 0.021$				$\alpha_{11} = 0.045$ $\alpha_{01} = 0.105$ $\alpha_{12} = 0.045$ $\alpha_{02} = 0.105$				$\alpha_{11} = 0.081$ $\alpha_{01} = 0.189$ $\alpha_{12} = 0.081$ $\alpha_{02} = 0.189$			
n	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
50	0.58	2.42	1.26	1.26	0.10	1.02	0.22	0.22	0	0	0	0
100	1.20	4.08	1.86	1.86	0.50	2.88	1.16	1.16	0	0.04	0	0
200	1.76	4.40	2.22	2.22	1.16	3.64	1.58	1.58	0.04	0.68	0.14	0.14
500	2.44	4.98	2.42	2.42	1.60	4.20	1.82	1.82	0.64	3.08	1.20	1.20
1000	3.22	7.06	3.08	3.10	2.56	5.82	2.98	3.04	1.08	3.66	1.76	1.76
2000	4.20	8.16	4.22	4.28	3.26	7.14	3.26	3.32	1.76	4.48	2.06	2.06

$Se_1 = Se_2 = 0.90$, $Sp_1 = Sp_2 = 0.80$, $p_1 = 10\%$, $p_2 = 25\%$, $\psi_1 = 25\%$, $\psi_2 = 75\%$

	$\alpha_{11} = 0.009$ $\alpha_{01} = 0.016$ $\alpha_{12} = 0.009$ $\alpha_{02} = 0.016$				$\alpha_{11} = 0.045$ $\alpha_{01} = 0.080$ $\alpha_{12} = 0.045$ $\alpha_{02} = 0.080$				$\alpha_{11} = 0.081$ $\alpha_{01} = 0.144$ $\alpha_{12} = 0.081$ $\alpha_{02} = 0.144$			
n	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
50	0.24	1.38	0.58	0.58	0	0.45	0.05	0.05	0	0	0	0
100	0.92	3.30	1.68	1.68	0.40	2.15	0.75	0.75	0	0	0	0
200	1.14	4.24	1.90	1.90	0.75	2.75	1.15	1.15	0	0.30	0.05	0.05
500	1.76	4.68	2.36	2.36	1.40	3.75	1.85	1.85	0.25	2.25	0.70	0.70
1000	2.48	5.08	2.68	2.68	1.55	4.80	2.15	2.15	1.00	3.15	1.65	1.65
2000	3.28	6.22	3.22	3.22	2.80	6.45	3.20	3.20	1.75	3.95	2.35	2.35

$Se_1 = Se_2 = 0.90$, $Sp_1 = Sp_2 = 0.80$, $p_1 = 10\%$, $p_2 = 25\%$, $\psi_1 = 50\%$, $\psi_2 = 50\%$

	$\alpha_{11} = 0.009$ $\alpha_{01} = 0.016$ $\alpha_{12} = 0.009$ $\alpha_{02} = 0.016$				$\alpha_{11} = 0.045$ $\alpha_{01} = 0.080$ $\alpha_{12} = 0.045$ $\alpha_{02} = 0.080$				$\alpha_{11} = 0.081$ $\alpha_{01} = 0.144$ $\alpha_{12} = 0.081$ $\alpha_{02} = 0.144$			
n	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
50	0.32	1.78	0.66	0.66	0.04	0.74	0.22	0.22	0	0	0	0
100	1.22	3.62	1.84	1.84	0.42	2.40	0.84	0.84	0	0	0	0
200	1.52	4.48	2.04	2.04	1.08	3.50	1.46	1.46	0	0.48	0.02	0.02
500	2.40	5.28	2.66	2.66	1.70	4.54	2.22	2.22	0.32	2.02	0.56	0.56
1000	3.22	6.78	3.12	3.12	2.42	5.24	2.32	2.32	0.88	3.52	1.48	1.48
2000	3.74	8.10	3.88	3.88	3.56	7.30	3.80	3.86	1.64	4.52	2.28	2.28

M1: global hypothesis test. M2: individual test with α = 5%. M3: individual tests with Bonferroni method. M4: individual tests with Holm method.

Table 3. Type I errors (in %) of different methods to simultaneously compare the sensitivities ($Se_h = \{0.70, 0.80\}$) and specificities ($Sp_h = 0.90$) of two BDTs in the presence of a binary covariate.

Se_1 = Se_2 = 0.70, Sp_1 = Sp_2 = 0.90, p_1 = 10%, p_2 = 25%, ψ_1 = 50%, ψ_2 = 50%												
	$\alpha_{11} = 0.021$ $\alpha_{01} = 0.009$ $\alpha_{12} = 0.021$ $\alpha_{02} = 0.009$				$\alpha_{11} = 0.105$ $\alpha_{01} = 0.045$ $\alpha_{12} = 0.105$ $\alpha_{02} = 0.045$				$\alpha_{11} = 0.189$ $\alpha_{01} = 0.081$ $\alpha_{12} = 0.189$ $\alpha_{02} = 0.081$			
n	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
50	0.10	0.66	0.20	0.20	0	0.08	0	0	0	0	0	0
100	0.72	2.40	1.06	1.06	0.02	0.68	0.14	0.14	0	0	0	0
200	1.88	4.78	1.72	1.74	0.62	2.60	0.74	0.76	0.02	0.08	0.02	0.02
500	2.68	6.04	2.64	2.70	1.98	4.38	1.96	2.00	0.10	0.68	0.24	0.24
1000	3.88	8.64	4.00	4.04	3.14	7.26	3.30	3.36	0.80	2.54	1.10	1.10
2000	4.54	8.84	4.52	4.58	4.24	8.82	4.08	4.10	1.44	4.06	1.48	1.48
Se_1 = Se_2 = 0.70, Sp_1 = Sp_2 = 0.90, p_1 = 10%, p_2 = 50%, ψ_1 = 25%, ψ_2 = 75%												
	$\alpha_{11} = 0.021$ $\alpha_{01} = 0.009$ $\alpha_{12} = 0.021$ $\alpha_{02} = 0.009$				$\alpha_{11} = 0.105$ $\alpha_{01} = 0.045$ $\alpha_{12} = 0.105$ $\alpha_{02} = 0.045$				$\alpha_{11} = 0.189$ $\alpha_{01} = 0.081$ $\alpha_{12} = 0.189$ $\alpha_{02} = 0.081$			
n	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
50	0	0.22	0.04	0.04	0	0	0	0	0	0	0	0
100	0.26	1.82	0.42	0.42	0.04	0.56	0.06	0.06	0	0	0	0
200	0.88	2.98	1.20	1.20	0.32	1.96	0.60	0.60	0	0	0	0
500	2.26	4.46	1.98	1.98	1.28	4.24	1.76	1.76	0.02	0.36	0.12	0.12
1000	3.18	6.68	3.18	3.22	1.98	4.92	2.08	2.08	0.38	1.72	0.66	0.66
2000	3.78	8.20	3.74	3.86	3.46	6.88	3.22	3.24	1.18	3.74	1.40	1.40
Se_1 = Se_2 = 0.80, Sp_1 = Sp_2 = 0.90, p_1 = 10%, p_2 = 25%, ψ_1 = 25%, ψ_2 = 75%												
	$\alpha_{11} = 0.016$ $\alpha_{01} = 0.009$ $\alpha_{12} = 0.016$ $\alpha_{02} = 0.009$				$\alpha_{11} = 0.080$ $\alpha_{01} = 0.045$ $\alpha_{12} = 0.080$ $\alpha_{02} = 0.045$				$\alpha_{11} = 0.144$ $\alpha_{01} = 0.081$ $\alpha_{12} = 0.144$ $\alpha_{02} = 0.081$			
n	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
50	0.06	0.60	0.14	0.14	0	0.04	0.02	0.02	0	0	0	0
100	0.72	2.66	1.08	1.08	0.06	0.74	0.16	0.16	0	0	0	0
200	1.58	4.50	1.44	1.46	0.62	2.48	0.98	0.98	0	0.02	0	0
500	3.24	6.30	2.84	2.88	1.96	4.62	2.06	2.06	0.06	0.84	0.14	0.14
1000	3.62	8.12	3.66	3.76	2.78	6.58	2.76	2.86	0.54	2.30	0.88	0.88
2000	5.10	9.82	5.06	5.12	4.28	8.08	4.00	4.04	1.60	4.16	1.94	1.94
Se_1 = Se_2 = 0.80, Sp_1 = Sp_2 = 0.90, p_1 = 10%, p_2 = 50%, ψ_1 = 25%, ψ_2 = 75%												
	$\alpha_{11} = 0.016$ $\alpha_{01} = 0.009$ $\alpha_{12} = 0.016$ $\alpha_{02} = 0.009$				$\alpha_{11} = 0.080$ $\alpha_{01} = 0.045$ $\alpha_{12} = 0.080$ $\alpha_{02} = 0.045$				$\alpha_{11} = 0.144$ $\alpha_{01} = 0.081$ $\alpha_{12} = 0.144$ $\alpha_{02} = 0.081$			
n	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
50	0	0.20	0.02	0.02	0	0	0	0	0	0	0	0
100	0.10	1.58	0.18	0.18	0.02	0.32	0.06	0.06	0	0	0	0
200	1.06	3.14	1.52	1.52	0.26	2.02	0.56	0.56	0	0	0	0
500	1.78	4.34	1.92	1.92	1.46	3.84	1.76	1.76	0.00	0.42	0.14	0.14
1000	2.44	6.02	2.62	2.64	2.06	4.64	2.10	2.10	0.26	1.68	0.58	0.58
2000	4.02	8.24	3.94	4.00	2.88	6.42	2.94	2.94	0.82	2.94	1.14	1.14

M1: global hypothesis test. M2: individual test with α = 5%. M3: individual tests with Bonferroni method. M4: individual tests with Holm method.

In the study, it is considered that the type I error exceeds the nominal error when the global type I error is equal to or greater than or than 7%. The covariances α_{1m} and α_{0m} have an important effect on the type I errors of the four methods: type I errors decrease when the values of the covariances increase. From the results, the following general conclusions were obtained:

(a). Global hypothesis test. The type I error of the global hypothesis test is very small when the sample size is small and increases as the sample size increases, until it approaches the nominal error without exceeding it. Therefore, the global hypothesis test is a conservative test when the sample size is small ($n = 50$) or moderate ($n = 100$–200), and its global type I error approaches the nominal error (without exceeding it) when the sample size is large ($n = 500$–1000) or very large ($n = 2000$).

(b). Individual tests with $\alpha = 5\%$. The type I error of the individual tests with $\alpha = 5\%$ is less than the nominal error when the sample size is small and increases as the sample size increases. The type I error can clearly exceed the nominal error when the sample size is large. Therefore, the method based on individual tests with $\alpha = 5\%$ can give false significance when the sample size is large and should not be used.

(c). Individual tests combined with the Bonferroni method. The type I error of the method based on the individual tests combined with the Bonferroni method has a behavior very similar to the type I error of the global hypothesis test, and there is no important difference between both type I errors

(d). Individual tests combined with the Holm method. The type I error of the method based on the individual tests combined with the Holm method is very similar (even the same in many cases) to the type I error of the individual tests combined with the Bonferroni method.

3.2. Powers

Tables 4 and 5 show the powers obtained for the four methods proposed in Section 2, considering different scenarios. The covariances α_{1m} and α_{0m} have an important effect on the powers of the methods: the powers increase when the values of the covariances increase. From the results, the following general conclusions are obtained:

(a). The power of the method based on the individual tests with $\alpha = 5\%$ is greater than the powers of the other methods, due to the fact that its global type I error is also greater than that of the other methods (clearly exceeding the nominal error when the sample size is large).

(b). The power of the method based on individual tests combined with the Bonferroni method and the power of the method based on individual tests combined with the Holm method are practically equal. Therefore, both methods show an asymptotic behavior, in terms of type I error and power, that is practically identical.

(c). In very general terms, the power of the method based on the individual tests combined with Bonferroni (Holm) is slightly greater than the power of the global hypothesis test when the sample size is small or moderate. When the sample size is large or very large, the power of the global hypothesis test is, in very general terms, slightly higher than that of the method based on individual tests with Bonferroni (Holm). In these situations, all of these methods have a very similar type I error.

Table 4. Powers (in %) of different methods to simultaneously compare the sensitivities ($Se_h = \{0.70, 0.90\}$) and specificities ($Sp_h = \{0.70, 0.90\}$) of two BDTs in the presence of a binary covariate.

Se_1 = 0.70, Se_2 = 0.90, Sp_1 = 0.90, Sp_2 = 0.90, p_1 = 10%, p_2 = 25%, ψ_1 = 25%, ψ_2 = 75%												
	$\alpha_{11} = 0.007$ $\alpha_{01} = 0.009$ $\alpha_{12} = 0.007$ $\alpha_{02} = 0.009$				$\alpha_{11} = 0.035$ $\alpha_{01} = 0.045$ $\alpha_{12} = 0.035$ $\alpha_{02} = 0.045$				$\alpha_{11} = 0.063$ $\alpha_{01} = 0.081$ $\alpha_{12} = 0.063$ $\alpha_{02} = 0.081$			
n	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
50	0.04	0.48	0.14	0.14	0	0.10	0	0	0	0	0	0
100	0.54	2.40	0.80	0.80	0.02	0.60	0.12	0.12	0.01	0.01	0.01	0.01
200	1.72	4.12	1.68	1.68	0.86	2.90	0.84	0.84	0.02	0.30	0.02	0.02
500	5.26	8.98	3.78	3.84	4.78	7.92	3.36	3.40	2.04	4.38	1.46	1.48
1000	21.04	31.78	20.48	20.64	23.18	35.16	22.52	22.74	21.26	37.66	22.76	22.90
2000	56.24	70.72	58.08	58.42	66.18	78.72	68.40	68.60	77.82	88.60	81.30	81.48
Se_1 = 0.70, Se_2 = 0.90, Sp_1 = 0.90, Sp_2 = 0.90, p_1 = 10%, p_2 = 50%, ψ_1 = 25%, ψ_2 = 75%												
	$\alpha_{11} = 0.007$ $\alpha_{01} = 0.009$ $\alpha_{12} = 0.007$ $\alpha_{02} = 0.009$				$\alpha_{11} = 0.035$ $\alpha_{01} = 0.045$ $\alpha_{12} = 0.035$ $\alpha_{02} = 0.045$				$\alpha_{11} = 0.063$ $\alpha_{01} = 0.081$ $\alpha_{12} = 0.063$ $\alpha_{02} = 0.081$			
n	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
50	0.02	0.28	0.08	0.08	0	0.04	0	0	0	0	0	0
100	0.18	1.68	0.26	0.26	0.01	0.36	0.01	0.01	0.01	0.01	0.01	0.01
200	1.32	3.40	1.28	1.28	0.56	2.16	0.50	0.51	0.02	0.06	0.02	0.02
500	5.42	9.34	3.94	4.08	4.24	8.28	3.32	3.38	1.62	4.90	1.72	1.72
1000	20.84	31.50	20.76	20.96	21.14	32.68	21.16	21.38	21.4	37.88	22.88	23.16
2000	55.12	69.84	56.84	57.22	65.42	77.58	67.16	67.44	77.52	89.38	81.39	81.46
Se_1 = 0.90, Se_2 = 0.70, Sp_1 = 0.90, Sp_2 = 0.70, p_1 = 10%, p_2 = 25%, ψ_1 = 25%, ψ_2 = 75%												
	$\alpha_{11} = 0.007$ $\alpha_{01} = 0.007$ $\alpha_{12} = 0.007$ $\alpha_{02} = 0.007$				$\alpha_{11} = 0.035$ $\alpha_{01} = 0.035$ $\alpha_{12} = 0.035$ $\alpha_{02} = 0.035$				$\alpha_{11} = 0.063$ $\alpha_{01} = 0.063$ $\alpha_{12} = 0.063$ $\alpha_{02} = 0.063$			
n	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
50	1.70	6.78	3.54	3.54	1.34	5.66	2.96	2.96	1.02	5.62	2.68	2.68
100	8.98	19.34	12.46	12.46	10.26	21.28	13.94	13.94	11.78	24.76	16.06	16.06
200	30.01	42.88	29.54	29.54	33.82	49.96	33.56	33.56	44.80	58.62	44.22	44.22
500	78.62	85.30	77.16	77.44	86.46	90.66	84.30	84.56	92.36	95.48	91.74	91.78
1000	98.68	99.18	98.34	98.46	99.58	99.70	99.16	99.20	99.98	100	99.92	99.94
2000	100	100	100	100	100	100	100	100	100	100	100	100
Se_1 = 0.90, Se_2 = 0.70, Sp_1 = 0.90, Sp_2 = 0.70, p_1 = 10%, p_2 = 25%, ψ_1 = 50%, ψ_2 = 50%												
	$\alpha_{11} = 0.007$ $\alpha_{01} = 0.007$ $\alpha_{12} = 0.007$ $\alpha_{02} = 0.007$				$\alpha_{11} = 0.035$ $\alpha_{01} = 0.035$ $\alpha_{12} = 0.035$ $\alpha_{02} = 0.035$				$\alpha_{11} = 0.063$ $\alpha_{01} = 0.063$ $\alpha_{12} = 0.063$ $\alpha_{02} = 0.063$			
n	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
50	0.44	2.16	0.84	0.84	0.28	1.50	0.56	0.56	0.12	1.42	0.44	0.44
100	1.60	4.78	2.28	2.28	1.16	3.94	1.88	1.88	1.02	3.52	1.48	1.48
200	5.18	10.88	4.46	4.70	4.72	10.08	4.04	4.28	5.18	10.66	4.42	4.48
500	25.26	34.72	22.24	22.42	29.08	39.24	24.80	25.26	32.88	42.38	26.64	27.46
1000	62.78	74.76	59.60	60.24	71.58	81.84	68.50	69.20	85.32	91.18	83.84	84.32
2000	93.44	95.96	92.34	92.62	97.66	98.66	96.86	97.00	99.82	99.94	99.74	99.74

M1: global hypothesis test. M2: individual test with α = 5%. M3: individual tests with Bonferroni method. M4: individual tests with Holm method.

Table 5. Powers (in %) of different methods to simultaneously compare the sensitivities ($Se_h = \{0.80, 0.90\}$) and specificities ($Sp_h = \{0.70, 0.80, 0.90\}$) of two BDTs in the presence of a binary covariate.

Se$_1$ = 0.80, Se$_2$ = 0.90, Sp$_1$ = 0.90, Sp$_2$ = 0.70, p$_1$ = 10%, p$_2$ = 25%, ψ_1 = 25%, ψ_2 = 75%												
	$\alpha_{11} = 0.008$ $\alpha_{01} = 0.007$				$\alpha_{11} = 0.040$ $\alpha_{01} = 0.035$				$\alpha_{11} = 0.072$ $\alpha_{01} = 0.063$			
	$\alpha_{12} = 0.008$ $\alpha_{02} = 0.007$				$\alpha_{12} = 0.040$ $\alpha_{02} = 0.035$				$\alpha_{12} = 0.072$ $\alpha_{02} = 0.063$			
n	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
50	1.94	6.66	3.86	3.86	1.38	6.30	3.16	3.16	1.08	6.04	2.86	2.86
100	8.46	18.80	12.16	12.16	9.78	21.38	13.90	13.90	11.82	24.10	16.82	16.82
200	29.86	43.30	29.68	29.68	35.12	47.86	35.01	35.01	42.76	58.16	42.26	42.26
500	76.62	83.94	75.40	75.40	84.22	89.62	83.84	83.88	92.80	95.20	91.32	91.32
1000	97.78	98.88	97.48	97.54	99.22	99.68	99.06	99.08	99.90	99.98	99.88	99.88
2000	100	100	100	100	100	100	100	100	100	100	100	100
Se$_1$ = 0.80, Se$_2$ = 0.90, Sp$_1$ = 0.90, Sp$_2$ = 0.70, p$_1$ = 10%, p$_2$ = 50%, ψ_1 = 25%, ψ_2 = 75%												
	$\alpha_{11} = 0.008$ $\alpha_{01} = 0.007$				$\alpha_{11} = 0.040$ $\alpha_{01} = 0.035$				$\alpha_{11} = 0.072$ $\alpha_{01} = 0.063$			
	$\alpha_{12} = 0.008$ $\alpha_{02} = 0.007$				$\alpha_{12} = 0.040$ $\alpha_{02} = 0.035$				$\alpha_{12} = 0.072$ $\alpha_{02} = 0.063$			
n	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
50	0.26	2.60	0.86	0.86	0.12	1.86	0.68	0.68	0.16	1.60	0.48	0.48
100	2.58	6.78	3.78	3.78	2.24	6.82	3.52	3.52	1.76	5.86	2.92	2.92
200	8.18	13.64	8.06	8.06	8.38	14.82	8.32	8.32	9.22	16.46	9.16	9.16
500	21.12	30.98	21.04	21.04	26.74	35.70	26.24	26.34	29.56	41.84	29.38	29.38
1000	49.10	58.38	46.00	46.38	58.12	66.60	54.84	55.22	68.78	75.72	65.34	65.58
2000	84.48	88.22	80.62	81.08	91.36	93.70	87.88	88.34	97.02	97.66	95.04	95.54
Se$_1$ = 0.90, Se$_2$ = 0.80, Sp$_1$ = 0.90, Sp$_2$ = 0.80, p$_1$ = 10%, p$_2$ = 25%, ψ_1 = 25%, ψ_2 = 75%												
	$\alpha_{11} = 0.008$ $\alpha_{01} = 0.008$				$\alpha_{11} = 0.040$ $\alpha_{01} = 0.040$				$\alpha_{11} = 0.072$ $\alpha_{01} = 0.072$			
	$\alpha_{12} = 0.008$ $\alpha_{02} = 0.008$				$\alpha_{12} = 0.040$ $\alpha_{02} = 0.040$				$\alpha_{12} = 0.072$ $\alpha_{02} = 0.072$			
n	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
50	0.34	2.18	0.90	0.90	0.30	1.76	0.62	0.62	0.02	0.84	0.24	0.24
100	2.80	7.10	4.18	4.18	2.22	7.58	3.88	3.88	1.44	7.32	2.88	2.88
200	9.90	16.78	9.82	9.82	10.26	19.70	10.16	10.16	14.36	24.56	14.29	14.29
500	28.40	38.24	28.31	28.31	37.68	49.34	37.54	37.60	55.82	66.88	55.57	55.57
1000	60.62	69.44	57.92	58.18	74.08	81.66	71.54	71.90	91.42	95.00	90.46	90.58
2000	91.46	94.30	89.42	89.90	97.76	98.24	96.80	97.00	99.92	100	99.80	99.80
Se$_1$ = 0.90, Se$_2$ = 0.80, Sp$_1$ = 0.90, Sp$_2$ = 0.80, p$_1$ = 10%, p$_2$ = 25%, ψ_1 = 50%, ψ_2 = 50%												
	$\alpha_{11} = 0.008$ $\alpha_{01} = 0.008$				$\alpha_{11} = 0.040$ $\alpha_{01} = 0.040$				$\alpha_{11} = 0.072$ $\alpha_{01} = 0.072$			
	$\alpha_{12} = 0.008$ $\alpha_{02} = 0.008$				$\alpha_{12} = 0.040$ $\alpha_{02} = 0.040$				$\alpha_{12} = 0.072$ $\alpha_{02} = 0.072$			
n	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
50	0.18	1.44	0.46	0.46	0.08	0.86	0.22	0.22	0.02	0.34	0.06	0.06
100	1.10	3.40	1.50	1.50	0.48	2.28	0.90	0.90	0.36	1.90	0.68	0.68
200	2.56	6.02	2.52	2.54	2.37	5.60	2.32	2.32	1.71	4.04	1.58	1.58
500	7.60	12.80	6.64	6.78	7.62	12.32	5.98	6.06	7.68	11.30	5.28	5.36
1000	20.08	31.10	17.96	18.46	24.52	35.70	22.78	23.08	31.70	42.40	27.60	28.28
2000	45.58	57.12	43.34	43.84	56.86	68.70	55.30	55.92	80.24	86.62	79.06	79.52

M1: global hypothesis test. M2: individual test with α = 5%. M3: individual tests with Bonferroni method. M4: individual tests with Holm method.

3.3. Application Rules

Based on the conclusions obtained from the simulation experiments, the following general application rules can be given when simultaneously comparing the accuracies of two BDTs in the presence of a binary covariate:

(a). When the sample size is small or moderate, solve the individual hypothesis tests $H_0 : Se_1 = Se_2$ (Equation (8)) and $H_0 : Sp_1 = Sp_2$ (Equation (9)) combined with the Bonferroni (or Holm) method using an error $\alpha = 5\%$.

(b). When the sample size is large or very large, solve the global test $H_0 : (Se_1 = Se_2$ and $Sp_1 = Sp_2)$ (Equation (7)) using an error $\alpha = 5\%$. If the global hypothesis test is not significant, then the equality of the accuracy of the two BDTs is not rejected. If the global hypothesis test is significant, then the causes of the significance will be investigated via testing $H_0 : Se_1 = Se_2$ and $H_0 : Sp_1 = Sp_2$ by individually applying Equations (8) and (9) combined with the Bonferroni (Holm) method using an error $\alpha = 5\%$. The global hypothesis test is initially applied because it is a somewhat more powerful method than the individual tests combined with the Bonferroni (Holm) method when the sample size is large or very large.

These application rules are given solely based on the sample size n because it is the only parameter of the study whose value was set by the researcher.

4. The "scapbc" Function

A function was written in R [5] that allows simultaneously comparing the accuracies of two BDTs subject to a paired design in the presence of a binary covariate. The function is called "scapbc" (simultaneous accuracy comparison in the presence of a binary covariate) and is executed with the command:

$$\text{scapbc}(s_{111}, s_{101}, s_{011}, s_{001}, r_{111}, r_{101}, r_{011}, r_{001}, s_{112}, s_{102}, s_{012}, s_{002}, r_{112}, r_{102}, r_{012}, r_{002}, \alpha)$$

where $(s_{111}, s_{101}, \ldots, r_{012}, r_{002})$ are the observed frequencies and "α" is the α error. The function checks that the values of the arguments are valid. The function solves the problem by applying the rules given in Section 3.3, by applying the Bonferroni method. The results obtained are recorded in the file "results_scapbc.txt" in the same folder from which the function is run. The "scapbc" function is available as the Supplementary Materials of this manuscript.

5. Example

The results were applied to the diagnosis of coronary artery disease [11]. Weiner et al. [11] applied two BDTs (exercise test and clinical history) and a GS (coronary angiography) to a sample of 2045 patients (1465 men and 580 women). The observed frequencies of the study are shown in Table 6, where the variable T_1 models the result of the exercise test, T_2 models the result of the clinical history, and D models the result of the coronary angiography.

Table 6. Observed frequencies in the study of Weiner et al.

	Men				
	$T_1 = 1$		$T_1 = 0$		Total
	$T_2 = 1$	$T_2 = 0$	$T_2 = 1$	$T_2 = 0$	
$D = 1$	786	29	183	25	1023
$D = 0$	69	46	176	151	442
Total	855	75	359	176	1465
	Women				
	$T_1 = 1$		$T_1 = 0$		Total
	$T_2 = 1$	$T_2 = 0$	$T_2 = 1$	$T_2 = 0$	
$D = 1$	124	4	32	9	169
$D = 0$	81	68	101	161	411
Total	205	72	133	170	580

In this study, the risk of coronary heart disease is 2.4 times higher in men than in women [11]. The estimated value of the odds ratio is 5.63 (95% confidence interval: 4.56 to 6.95). Therefore, sex is a covariate that is related to the disease. In the exercise test, ST segment depression is less sensitive in women than in men, so sex is a covariate that can influence the test result. Therefore, adjusting for sex is necessary to simultaneously compare the two sensitivities and the two specificities. Executing the command

$$\text{scapbc}(786, 29, 183, 25, 69, 46, 176, 151, 124, 4, 32, 9, 81, 68, 101, 161, 0.05),$$

generates the results shown in Table 7.

Table 7. Results obtained in the study of Weiner et al.

Estimates by Sex						
	$\hat{S}e_{1m} \pm SE$	$\hat{S}p_{1m} \pm SE$	$\hat{S}e_{2m} \pm SE$	$\hat{S}p_{2m} \pm SE$	$\hat{p}_m$	$\hat{\psi}_m$
Men	0.797 ± 0.013	0.740 ± 0.021	0.947 ± 0.007	0.446 ± 0.024	69.8%	71.6%
Women	0.757 ± 0.033	0.637 ± 0.024	0.923 ± 0.020	0.557 ± 0.025	29.1%	28.4%

Overall estimates			
	$\hat{S}e_h \pm SE$	$\hat{S}p_h \pm SE$	$\hat{p}$
Exercise test	0.791 ± 0.012	0.691 ± 0.016	58.3%
Clinical history	0.944 ± 0.007	0.499 ± 0.017	

SE: standard error.

Because the sample size is very large, the global hypothesis test is solved (application rules of Section 3.3). The test statistic for the global hypothesis test is $Q^2 = 224.252$ and p-value = 0. Therefore, the null hypothesis (equality of the two sensitivities and of the two specificities) of the global hypothesis test is rejected. To investigate the causes of significance, it is necessary to solve the individual tests and apply the Bonferroni (or Holm) method. The test statistic for $H_0 : Se_1 = Se_2$ vs. $H_1 : Se_1 \neq Se_2$ is 12.265 $(p{-}\text{value} = 0)$, and the test statistic for $H_0 : Sp_1 = Sp_2$ vs. $H_1 : Sp_1 \neq Sp_2$ is 8.593 $(p{-}\text{value} = 0)$. Applying the Bonferroni method with $\alpha = 5\%$, the two null hypotheses are rejected. Therefore, the sensitivity of the clinical history is significantly greater than the sensitivity of the exercise test (95% confidence interval: 0.128 to 0.177), and the specificity of the exercise test is significantly greater than the specificity of the clinical history (95% confidence interval: 0.148 to 0.235). The same conclusions are obtained if the Holm method is applied.

6. Discussion

Comparison of the sensitivities and specificities of two BDTs is a topic of great interest in the study of statistical methods applied to diagnosis and has been the subject of numerous studies in the statistical literature. When two BDTs are compared, it is common to observe discrete covariates in all of the individuals in the sample. In this situation, if the covariates are related to the disease and to either of the two BDTs, then it is necessary to adjust for covariates. This adjustment has the purpose of eliminating the effect of the covariate in the estimation of the global sensitivity and specificity of each BDT, and consequently eliminating its effect in the comparison of the parameters. Therefore, adjustment for covariates is important because the comparison of two diagnostic tests may be biased when an adjustment is not made. This manuscript makes a contribution to this topic, by simultaneously comparing the accuracies of two BDTs by adjusting for discrete covariates. Therefore, in this manuscript the simultaneous comparison of the sensitivities and the specificities of two BDTs was studied when discrete covariates are observed in all of the individuals in the sample. The overall estimators of the sensitivities and specificities were obtained by applying the maximum likelihood method and the variances-covariances were estimated by applying the delta method. In this situation, simultaneous comparison of sensitivities and specificities of two BDTs was resolved by four methods: the global hypothesis test $H_0 : (Se_1 = Se_2 \text{ and } Sp_1 = Sp_2)$ with an α error; individual tests $H_0 : Se_1 = Se_2$ and $H_0 : Sp_1 = Sp_2$, each with an α error; individual tests $H_0 : Se_1 = Se_2$ and $H_0 : Sp_1 = Sp_2$ and application of the Bonferroni method with an α error; and individual tests $H_0 : Se_1 = Se_2$ and $H_0 : Sp_1 = Sp_2$ and application of the Holm method with an α error.

Simulation experiments were carried out to study the behaviors of the different methods when the covariate is binary. The results showed that the method based on the individual tests $H_0 : Se_1 = Se_2$ and $H_0 : Sp_1 = Sp_2$, each with an α error, can give rise to type I errors that far exceed the nominal error, and therefore this method gives rise to too many false significances. Furthermore, the method based on the global hypothesis test has better asymptotic behavior when the sample size is large or very large than the methods based on individual tests and the application of the Bonferroni or Holm methods. However, when the sample size is small or moderate, these latter two methods perform better than the method based on the global hypothesis test. Therefore, based on the results of the simulation experiments, some rules of application of the methods can be given according to the sample size (which is the only value set by the researcher). These rules are: (a) When the sample size is small or moderate, solve the individual hypothesis tests $H_0 : Se_1 = Se_2$ and $H_0 : Sp_1 = Sp_2$ combined with the Bonferroni (or Holm) method with an error $\alpha = 5\%$; (b) When the sample size is large or very large, solve the global test $H_0 : (Se_1 = Se_2 \text{ and } Sp_1 = Sp_2)$ with an error $\alpha = 5\%$. If the global hypothesis test is not significant, then it is not rejected that the two sensitivities are equal and that the two specificities are equal. If the global hypothesis test is significant, then the causes of significance are investigated by solving the individual tests combined with the Bonferroni (Holm) method with an error $\alpha = 5\%$. The method based on the global hypothesis test is very similar to the analysis of variance: first the global test is solved and, if it is significant, then the individual tests are solved and a multiple comparisons method is applied.

Simulation experiments have shown that the covariances between the two BDTs have an important effect on type I errors and powers. Type I errors are greater when the two BDTs are conditionally independent of the disease than when the two BDTs are conditionally dependent on the disease. Regarding the powers, for a fixed sample size, the power of each method is greater when the two BDTs are conditionally dependent on the disease than when they are conditionally independent of the disease. In practice, the only parameter that the researcher can control is the sample size. Therefore, although the effect of the covariances is important, the increase in power can only be achieved by increasing the sample size (the researcher cannot increase the values of the covariances, because these depend on the intrinsic properties of both diagnostic tests).

Simulation experiments have also shown that the global hypothesis test, whose test statistic is a Wald-type test statistic, has a good asymptotic performance in terms of type I error and power. The type I error of the global test is close to the nominal error when the sample size is large or very large. Regarding the power, in general terms and depending on the covariances between the two BDTs, a large sample size is needed for the power to be large. Therefore, the global test performance when the covariate is binary is very similar to that obtained in other studies [2].

The proposed method is based on the fact that the covariate is discrete. A future study should address the problem that occurs when the covariate is quantitative.

Finally, a function was written in R that allows us to solve the problem posed when the covariate is binary. The function is easy to use and provides all of the results so that the researcher can easily solve the problem. The function is available as Supplementary Materials to this manuscript.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/math9172029/s1. The "scapbc" function is a function written in R that allows simultaneous comparison of the accuracies of two BDTs subject to a paired design in the presence of a binary covariate.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The author wants to thank the anonymous referees for their helpful comments that improved the quality of the manuscript.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A

The log-likelihood function $l_2(\boldsymbol{\omega})$ (Equation (6)) can be written as:

$$l_2(\boldsymbol{\omega}) = \sum_{i,j=0}^{1} \sum_{m=1}^{M} x_{ijm} \log(\phi_{ijm}) + \sum_{i,j=0}^{1} \sum_{m=1}^{M} y_{ijm} \log(\varphi_{ijm}) = \sum_{m=1}^{M} l_{2m},$$

where:

$$l_{2m} = \sum_{i,j=0}^{1} \left[x_{ijm} \log(\phi_{ijm}) + y_{ijm} \log(\varphi_{ijm}) \right]$$

is the log-likelihood function in the mth covariate pattern. Then, the Fisher information matrix of function $l_2(\boldsymbol{\omega})$ is:

$$I_2 = Diag\{I_{21}, \ldots, I_{2M}\}$$

and, therefore:

$$\sum\nolimits_{\hat{\boldsymbol{\omega}}} = I_2^{-1} = Diag\left\{\sum\nolimits_{\hat{\boldsymbol{\omega}}_1}, \cdots \sum\nolimits_{\hat{\boldsymbol{\omega}}_M}\right\}.$$

References

1. Zhou, X.H.; Obuchowski, N.A.; McClish, D.K. *Statistical Methods in Diagnostic Medicine*, 2nd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2011.
2. Roldán-Nofuentes, J.A.; Sidaty-Regad, S.B. Recommended methods to compare the accuracy of two binary diagnostic tests subject to a paired design. *J. Stat. Comput. Simul.* **2019**, *89*, 2621–2644. [CrossRef]
3. Janes, H.; Pepe, M.S. Adjusting for Covariates in Studies of Diagnostic, Screening, or Prognostic Markers: An Old Concept in a New Setting. *Am. J. Epidemiol.* **2008**, *168*, 89–97. [CrossRef] [PubMed]
4. Lahner, E.; Dilaghi, E.; Prestigiacomo, C.; Alessio, G.; Marcellini, L.; Simmaco, M.; Santino, I.; Orsi, G.B.; Anibaldi, P.; Marcolongo, A.; et al. Prevalence of SARS-CoV-2 infection in health workers (HWs) and diagnostic test performance: The experience of a teaching Hospital in Central Italy. *Int. J. Environ. Res. Public Health* **2020**, *17*, 4417. [CrossRef] [PubMed]
5. R Core Team. *R: A Language and Environment for Statistical Computing*; R Core Team: Vienna, Austria, 2016; Available online: https://www.R-project.org/ (accessed on 8 July 2021).
6. Berry, G.; Smith, C.; Macaskill, P.; Irwig, L. 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. *Stat. Med.* **2002**, *21*, 853–862. [CrossRef] [PubMed]
7. Agresti, A. *Categorical Data Analysis*, 3rd ed.; Wiley: New York, NY, USA, 2013.
8. Bonferroni, C.E. Teoria statistica delle classi e calcolo delle probabilità. *Pubbl. R Ist. Super. Sci. Econ. Commer. Firenze* **1936**, *8*, 3–62.
9. Holm, S. A simple sequential rejective multiple testing procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.
10. Hosmer, D.W.; Lemeshow, S. *Applied Logistic Regression*, 2nd ed.; Wiley: New York, NY, USA, 2000.
11. Weiner, D.A.; Ryan, T.J.; McCabe, C.H.; Kennedy, J.W.; Schloss, M.; Tristani, F.; Chaitman, B.R.; Fisher, L.D. Correlations among history of angina, ST-segment and prevalence of coronary artery disease in the coronary artery surgery study (CASS). *N. Engl. J. Med.* **1979**, *301*, 230–235. [CrossRef] [PubMed]