Comparison of the predictive values of multiple binary diagnostic tests

Roldán Nofuentes, José Antonio
University of Granada, School of Medicine, Biostatistics
Avenida de Madrid s/n
Granada, 18076, Spain
E-mail: jaroldan@ugr.es

Luna del Castillo, Juan de Dios
University of Granada, School of Medicine, Biostatistics
Avenida de Madrid s/n
Granada, 18076, Spain
E-mail: jdluna@ugr.es

Montero Alonso, Miguel Ángel
University of Granada, School of Medicine, Biostatistics
Avenida de Madrid s/n
Granada, 18076, Spain
E-mail: mmontero@ugr.es

1. Introduction

Sensitivity and specificity are the classic parameters to assess the accuracy of a binary diagnostic test in relation to a gold standard. Sensitivity (Se) is the probability of the diagnostic test being positive when the subject is diseased, and specificity (Sp) is the probability of the diagnostic test being negative when the subject is not diseased. Other parameters to assess the accuracy of a binary diagnostic test are positive and negative predictive values. The positive predictive value (PPV) is the probability of a patient being diseased when the test result is positive, and the negative predictive value (NPV) is the probability of a patient not being diseased when the test result is negative. The predictive values (PVs) represent the accuracy of the diagnostic test when it is applied to a cohort of subjects, and they are measures of the clinical accuracy of the diagnostic test. The predictive values depend on the sensitivity and the specificity of the diagnostic test and on the disease prevalence, and are easily calculated applying Bayes' Theorem, i.e.

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

where p is the disease prevalence.

In the study of the statistical methods for diagnosis, one of the most interesting topics is the comparison of the accuracy of two binary diagnostic tests in relation to the same gold standard. In paired designs, the estimation and comparison of the positive (negative) predictive values has been the subject of several studies (Bennett, 1972 and 1985; Leisenring et al, 2000; Moskowitz and Pepe, 2004 and 2006; Wang et al, 2006). In this work we study a global hypothesis test to simultaneously compare the positive and negative predictive values of multiple binary diagnostic tests when the binary tests and the gold standard are applied to all of the subjects in a random sample.

2. The model

Let us consider *J* binary diagnostic tests $(J \geq 3)$ whose performance is compared in relation to the same gold standard. Let T_j be the random binary variable that models the result of *j*th binary diagnostic test, in such a way that $T_j = 1$ when the result of the binary test is positive and $T_j = 0$ when the result is negative. Let *D* be the random binary variable that models the result of the gold standard, in such a way that $D = 1$ when the subject is diseased and $D = 0$ when the subject is non-diseased. Let $Se_j = P(T_j = 1 | D = 1)$ be the sensitivity of the *j*th diagnostic test and $Sp_j = P(T_j = 0 | D = 0)$ be the specificity, $p = P(D = 1)$ the disease prevalence, $PPV_j = P(D = 1 | T_j = 1)$ the positive predictive value of the *j*th diagnostic test and $NPV_j = P(D = 0 | T_j = 0)$ the negative predictive value, with $j = 1,..., J$. When the *J* binary tests and the gold standard are applied to all of the subjects in a random sample sized *n*, $s_{i_1,...,i_J}$ is the number of diseased subjects in which $T_1 = i_1$, $T_2 = i_2$,..., $T_J = i_J$, and $r_{i_1,...,i_J}$ is the number of non-diseased subjects in which $T_1 = i_1$, $T_2 = i_2$,..., $T_J = i_J$, with $i_j = 0,1$ and $j = 1,..., J$. Let

$$s = \sum_{i_1,...,i_J=0}^{1} s_{i_1,...,i_J} \text{ and } r = \sum_{i_1,...,i_J=0}^{1} r_{i_1,...,i_J}$$

be the total number of diseased subjects and the total number of non-diseased subjects respectively, with $n = s + r$. The observed data are the result of a multinomial distribution whose probabilities are given by

$$p_{i_1,...,i_J} = P(D = 1, T_1 = i_1, T_2 = i_2, ..., T_J = i_J) \text{ and } q_{i_1,...,i_J} = P(D = 0, T_1 = i_1, T_2 = i_2, ..., T_J = i_J).$$

Let $\boldsymbol{\pi} = (p_{1,...,1}, ..., p_{0,...,0}, q_{1,...,1}..., q_{0,...,0})^T$ be a vector sized 2^{J+1} whose components are the previous probabilities and let $\boldsymbol{\eta} = (PPV_1, ..., PPV_J, NPV_1, ..., NPV_J)^T$ be a vector sized *2J* whose components are the positive and negative predictive values of each one of the *J* diagnostic tests. In terms of the probabilities of the vector $\boldsymbol{\pi}$, the predictive values of the *j*th binary test are given by the expressions

$$PPV_j = \frac{\sum_{\substack{i_1,...,i_J=0 \\ i_j=1}}^{1} p_{i_1,...,i_J}}{\sum_{\substack{i_1,...,i_J=0 \\ i_j=1}}^{1} p_{i_1,...,i_J} + \sum_{\substack{i_1,...,i_J=0 \\ i_j=1}}^{1} q_{i_1,...,i_J}} \text{ and } NPV_j = \frac{\sum_{\substack{i_1,...,i_J=0 \\ i_j=0}}^{1} q_{i_1,...,i_J}}{\sum_{\substack{i_1,...,i_J=0 \\ i_j=0}}^{1} p_{i_1,...,i_J} + \sum_{\substack{i_1,...,i_J=0 \\ i_j=0}}^{1} q_{i_1,...,i_J}}.$$

As the probabilities of the vector $\boldsymbol{\pi}$ are the probabilities of a multinomial distribution, their maximum likelihood estimators are

$$\hat{p}_{i_1,...,i_J} = s_{i_1,...,i_J} / n \text{ and } \hat{q}_{i_1,...,i_J} = r_{i_1,...,i_J} / n,$$

so that the maximum likelihood estimators of the predictive values of the *j*th diagnostic test are

$$\widehat{PPV}_j = \frac{\sum_{\substack{i_1,\ldots,i_J=0 \\ i_j=1}}^{1} s_{i_1,\ldots,i_J}}{\sum_{\substack{i_1,\ldots,i_J=0 \\ i_j=1}}^{1} s_{i_1,\ldots,i_J} + \sum_{\substack{i_1,\ldots,i_J=0 \\ i_j=1}}^{1} r_{i_1,\ldots,i_J}} \quad \text{and} \quad \widehat{NPV}_j = \frac{\sum_{\substack{i_1,\ldots,i_J=0 \\ i_j=0}}^{1} r_{i_1,\ldots,i_J}}{\sum_{\substack{i_1,\ldots,i_J=0 \\ i_j=0}}^{1} s_{i_1,\ldots,i_J} + \sum_{\substack{i_1,\ldots,i_J=0 \\ i_j=0}}^{1} r_{i_1,\ldots,i_J}}.$$

Applying the delta method, the asymptotic variance-covariance matrix of the vector $\hat{\boldsymbol{\eta}}$ is

$$\Sigma_{\hat{\boldsymbol{\eta}}} = \left(\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\pi}}\right) \Sigma_{\hat{\boldsymbol{\pi}}} \left(\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\pi}}\right)^T,$$

where $\Sigma_{\hat{\boldsymbol{\pi}}}$ is the variance-covariance matrix of $\hat{\boldsymbol{\pi}}$, i.e.

$$\Sigma_{\hat{\boldsymbol{\pi}}} = \left\{\text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T\right\} / n.$$

The global hypothesis test to simultaneously compare the positive and negative and predictive values of the J binary diagnostic tests is

$$H_0 : \boldsymbol{\varphi}\boldsymbol{\eta} = \mathbf{0} \text{ vs } H_1 : \boldsymbol{\varphi}\boldsymbol{\eta} \neq \mathbf{0},$$

where $\boldsymbol{\varphi}$ is a complete range matrix size $2(J-1) \times 2J$ whose elements are known constants. For $J = 3$

$$\boldsymbol{\varphi} = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix},$$

and for $J = 4$

$$\boldsymbol{\varphi} = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix}.$$

Finally, the statistic for the global hypothesis test is

$$Q^2 = \hat{\boldsymbol{\eta}}^T \boldsymbol{\varphi}^T \left(\boldsymbol{\varphi} \hat{\Sigma}_{\hat{\boldsymbol{\theta}}} \boldsymbol{\varphi}^T\right)^{-1} \boldsymbol{\varphi} \hat{\boldsymbol{\eta}},$$

which is asymptotically distributed according to a central chi-square distribution with $2(J-1)$ degrees of freedom.

An alternative method to solve the global hypothesis test consists of comparing the positive and negative predictive values applying the method proposed by Leisenring et al (2000) (or the method of Wang et al (2006)) and applying a multiple comparison method, e.g. the method proposed by Holm (1979) or the method proposed by Hochberg (1988) which are less conservative than the method of multiple comparison proposed by Bonferroni (1935).

3. Simulation study

We have carried out Monte Carlo simulation experiments to study the type I error and the power of the global hypothesis test when comparing the predictive values of three binary diagnostic tests, and we have compared the type I error (power) with the type I error (power) when the global hypothesis test is solved making the comparisons of the *PVs* applying the method of Leisenring et al (2000) and the method of Wang et al (2006), both if the marginal hypothesis tests

$$\left(H_0 : PPV_i = PPV_j \text{ vs } H_1 : PPV_i \neq PPV_j\right) \text{ and } \left(H_0 : NPV_i = NPV_j \text{ vs } H_1 : NPV_i \neq NPV_j\right)$$

are carried out to an error rate α, with the method proposed by Holm (1979) or with the method proposed by Hochberg (1988). These experiments consisted of the generation of 5000 random samples of multinomial distributions with different sizes and the probabilities of the multinomial distribution were generated using the dependence model of Torrance-Rynard and Walter (1997). As the nominal error we also took $\alpha = 5\%$ .In Table 1 we show some of the results obtained for the type I error when

$$\left(PPV_1 = PPV_2 = PPV_3 = 0.80 \text{ , } NPV_1 = NPV_2 = NPV_3 = 0.70\right),$$

and in Table 2 we show some of the results obtained for the power when

$$\left(PPV_1 = 0.80,\ NPV_1 = 0.75,\ PPV_2 = 0.85,\ NPV_2 = 0.80,\ PPV_3 = 0.90,\ NPV_3 = 0.85\right),$$

and for high values of dependence factors.

Table 1. Type I errors of the hypothesis tests.

n	Global	Method of Wang et al			Method of Leisenring et al		
	test	$\alpha = 5\%$	Holm	Hochberg	$\alpha = 5\%$	Holm	Hochberg
50	0.0198	0.0882	0.0056	0.0060	0.0788	0.0054	0.0056
100	0.0400	0.1522	0.0224	0.0250	0.1362	0.0178	0.0188
200	0.0636	0.2016	0.0376	0.0412	0.1844	0.0284	0.0310
300	0.0644	0.2178	0.0494	0.0504	0.2070	0.0360	0.0396
400	0.0597	0.2212	0.0468	0.0498	0.2132	0.0442	0.0464
500	0.0578	0.2228	0.0460	0.0486	0.2160	0.0432	0.0466
1000	0.0520	0.2008	0.0396	0.0428	0.1972	0.0386	0.0412
2000	0.0532	0.2032	0.0438	0.0482	0.2014	0.0428	0.0474

Table 2. Powers of the hypothesis tests.

n	Global	Method of Wang et al			Method of Leisenring et al		
	test	$\alpha = 5\%$	Holm	Hochberg	$\alpha = 5\%$	Holm	Hochberg
50	0.3010	0.7240	0.2714	0.2860	0.7214	0.2664	0.2790
100	0.8090	0.9630	0.8136	0.8172	0.9632	0.8126	0.8172
200	0.9942	0.9988	0.9940	0.9944	0.9988	0.9940	0.9944
300	1	1	1	1	1	1	1
400	1	1	1	1	1	1	1
500	1	1	1	1	1	1	1
1000	1	1	1	1	1	1	1
2000	1	1	1	1	1	1	1

From the results of simulation experiments it holds that, in general terms, the type I error of the global hypothesis test fluctuates around the nominal error of 5% particularly for $n \geq 400$. Regarding the method proposed by Wang et al (Leisenring et al) to an error rate $\alpha = 5\%$, the type I error clearly overwhelms the nominal error, so those method may lead to erroneous results. As for the method proposed by Wang et al (Leisenring et al) along with the Holm's method or Hochberg's method, the type I error is almost always

slightly lower than the nominal error. With respect to the power, the power of the global hypothesis test increases when the dependence factors increase, and the prevalence has little effect on the power of the global hypothesis test. In general terms, with samples sized $n \geq 200$ the power of the global hypothesis test is very high (higher than 90%). Similar conclusions are reached when the global test is solved applying the method proposed by Wang et al (Leisenring et al) along with Holm's method or Hochberg's method. If we use the method proposed by Wang et al (Leisenring et al) to an error rate $\alpha = 5\%$, the power is higher due to the fact that this method has a type I error that clearly overwhelms the nominal error.

4. Conclusions

The positive predictive value and the negative predictive value of a binary diagnostic test are, along with sensitivity and specificity, fundamental parameters to assess and compare the classificatory accuracy of binary diagnostic tests. For the same binary test, the positive and negative predictive value depend on the sensitivity and specificity and on the disease prevalence, and therefore this dependence must be considered when comparing the predictive values of binary tests. In this study, we have proposed a global hypothesis test to simultaneously compare the positive and negative predictive value of multiple binary diagnostic tests when the binary tests and the gold standard are applied to all of the subjects in a random sample. The global hypothesis test is based on the chi-square distribution, estimating the variance-covariance matrix of the *PVs* through the delta method. Based on the results of the simulation experiments, we propose a method to compare the *PVs* of multiple binary diagnostic tests in paired designs: 1) Solving the global hypothesis test based on the chi-square distribution to an error rate of α; 2) If the global hypothesis test is significant to an error rate of α, the study of the causes of the significance must be carried out solving the marginal hypothesis tests (through the method of Leisenring el al (2000) or the method of Wang et al (2006)) along with the multiple comparison method (Holm's method or Hochberg's method) to the same error rate of global hypothesis test.

REFERENCES

Bennett, B.M. 1972. On comparison of sensitivity, specificity and predictive value of a number of diagnostic procedures. Biometrics 28, 793-800.

Bennett, B.M. 1985. On tests for equality of predictive values for t diagnostic procedures. Statistics in Medicine 4, 535-539.

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

Hochberg, Y. 1988. A sharper Bonferroni procedure for multiple tests of significance. Biometrika 75, 800-802.

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

Leisenring, W., Alonzo, T. and Pepe, M.S. 2000. Comparisons of predictive values of binary medical diagnostic tests for paired designs. Biometrics 56, 345-351.

Moskowitz, C.S. and Pepe, M.S. 2006. Comparing the predictive values of diagnostic tests: sample size and analysis for paired study designs. Clinical Trials 3, 272-279.

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.

Wang, W., Davis, C.S. and Soong, S.J. 2006. Comparison of predictive values of two diagnostic tests from the same sample of subjects using weighted least squares. Statistics in Medicine 25, 2215-2229.