Journal of Statistical Planning and Inference 141 (2011) 3578–3594

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Asymptotic hypothesis test to compare likelihood ratios of multiple diagnostic tests in unpaired designs

Jan Luts [a,b,*], José Antonio Roldán Nofuentes [c], Juan de Dios Luna del Castillo [c], Sabine Van Huffel [a,b]

[a] Department of Electrical Engineering (ESAT), Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium
[b] IBBT-K.U.Leuven Future Health Department, Leuven, Belgium
[c] Biostatistics, School of Medicine, University of Granada, Granada 18071, Spain

ARTICLE INFO

Article history:
Received 5 January 2011
Received in revised form
11 May 2011
Accepted 12 May 2011
Available online 19 May 2011

Keywords:
Binary diagnostic test
Multi-level diagnostic test
Likelihood ratio
Unpaired design
Sensitivity
Specificity

ABSTRACT

The accuracy of a binary diagnostic test is usually measured in terms of its sensitivity and its specificity. Other measures of the performance of a diagnostic test are the positive and negative likelihood ratios, which quantify the increase in knowledge about the presence of the disease through the application of a diagnostic test, and which depend on the sensitivity and specificity of the diagnostic test. In this article, we construct an asymptotic hypothesis test to simultaneously compare the positive and negative likelihood ratios of two or more diagnostic tests in unpaired designs. The hypothesis test is based on the logarithmic transformation of the likelihood ratios and on the chi-square distribution. Simulation experiments have been carried out to study the type I error and the power of the constructed hypothesis test when comparing two and three binary diagnostic tests. The method has been extended to the case of multiple multi-level diagnostic tests.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

The decision about which of two or more competing diagnostic tests to use generally depends on their diagnostic ability, which is often expressed in terms of sensitivity and specificity. Sensitivity refers to the probability of a positive test result when the patient is diseased, while specificity indicates the probability of a negative test result when the patient does not have the disease. Additionally, (positive and negative) predictive values and likelihood ratios can be used to characterize the diagnostic ability of a test (Pepe, 2003; Sackett et al., 1985). Likelihood ratios quantify the change from the prior (i.e. before testing) odds of disease to the posterior (i.e. after testing) odds of disease. The positive (negative) likelihood ratio, which combines the sensitivity and the specificity, represents the increase (decrease) in pre-test odds when the test does (not) diagnose the disease. These measures have the advantage to be very intuitive and to be independent of disease prevalence. Interestingly, likelihood ratios can be defined for dichotomous tests, as well as tests with multiple levels. Since a likelihood ratio is algebraically identical to a risk ratio, confidence intervals can be calculated

* Corresponding author at: Department of Electrical Engineering (ESAT), Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium. Tel.: +32 16 321065; fax: +32 16 321970.
E-mail addresses: jan.luts@esat.kuleuven.be (J. Luts), jaroldan@ugr.es (J.A. Roldán Nofuentes), jdluna@ugr.es (J.D. Luna del Castillo), sabine.vanhuffel@esat.kuleuven.be (S. Van Huffel).

0378-3758/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2011.05.010

in a straightforward way (Simel et al., 1991). There exists a relationship between comparisons based on sensitivity/specificity, predictive values and likelihood ratios. A test with better sensitivity and specificity is also better with respect to the dimensions of the likelihood ratio scale and the predictive value scale. There exists a reciprocal relationship with the predictive value scale, but not with the likelihood ratio scale (Pepe, 2003).

There are a number of studies in the statistical literature that already addressed the issue of comparing likelihood ratios to decide about the usage of competing diagnostic tests. A graphical method was proposed to incorporate the likelihood ratios in a plot of 1—specificity versus sensitivity (similar to receiver operator characteristic curves), thereby allowing comparison of two or more diagnostic tests (Biggerstaff, 2000). However, this approach assumed that the sensitivity and specificity were observed without error so that no measures of variability were included in the graph. Although Biggerstaff (2000) suggested the incorporation of confidence intervals, no such formal comparison of likelihood ratios was made. Alternatively, a regression approach was demonstrated to compare likelihood ratios in a paired design, i.e. each of the cases was analyzed with both diagnostic tests (Leisenring and Pepe, 1998). Such a regression approach allowed evaluating the effects of covariates on the likelihood ratios. Comparisons between the tests could be made by modeling a covariate for the type of test under study. Although this method was originally designed for dichotomous tests, the authors suggested an extension towards continuous tests by including extra covariates in the model. Comparisons of likelihood ratios were also studied in the context of assessing the gain in diagnostic ability when combining two (dichotomous) diagnostic tests as compared to its (dichotomous) component test (Macaskill et al., 2002). A graphical method, similar to the approach by Biggerstaff (2000), was proposed. Asymptotic variance formulae were derived for the difference between logarithms of the likelihood ratios in a paired design using the delta method. Confidence intervals for the ratio of likelihood ratios were obtained in order to choose between a combined or single component test. Finally, separate hypothesis tests and a joint hypothesis test for simultaneous comparison of positive and negative likelihood ratios were constructed for a paired design (Roldán Nofuentes and Luna del Castillo, 2007). The method, which employed the logarithmic transformation and the delta method, was shown to be extensible to the case where there are tests with multi-level results and where there are more than two diagnostic tests to compare.

In contrast to these previous studies, this article focuses on the simultaneous comparison of two or more dichotomous/multi-level diagnostic tests in an unpaired design. This setting applies when the diagnostic tests are applied to different sets of cases. Although paired designs might be preferable because of their efficiency, they are not always practical in a real (clinical) environment because of ethical considerations. For instance, performing a diagnostic test might be time consuming, it can impose an extra burden to the patient or there might be interference between the diagnostic tests (Pepe, 2003). While the choice between diagnostic tests is usually based on multiple criteria measuring clinical effectiveness (e.g. financial costs, hospitalization time, etc.), this study does not attempt to address such factors and exclusively focuses on the comparison of the diagnostic ability of the tests. Furthermore, sometimes, the categorical diagnostic test outcome is derived from some underlying continuous measure. If available, this continuous measure should be used for the comparison of diagnostic ability. However, in case only categorical diagnostic test values are available, the proposed methodology can be used. To this end, we construct a joint hypothesis test for simultaneous testing of likelihood ratios in an unpaired design. This joint hypothesis test allows taking the correlation between the positive and the negative likelihood ratio into account. The proposed method is based on a logarithmic transformation of the likelihood ratios since these measures have an asymmetrical distribution. In contrast to existing literature, this article proposes a test to directly compare two or more competing (binary or multi-level) tests, each applied to a different set of cases, and analyzes its type I error and the power.

In Section 2 we study the likelihood ratios in an unpaired design and develop a procedure for joint hypothesis testing in case of multiple binary diagnostic tests. Section 3 outlines the simulation experiments that have been performed to study the type I error and the power of the global test. The results of this analysis are summarized and compared with the results from three other approaches to compare the likelihood ratios. Section 4 proposes a simultaneous test for comparing likelihood ratios of multiple multi-level diagnostic tests. The new hypothesis test is illustrated on real-life examples in Section 5. Finally, in Section 6, we discuss the conclusions of our study.

2. Hypothesis tests for binary diagnostic tests

Let T_j be the binary random variable that models the result of the jth diagnostic test, with $j=1,\ldots,J$, so that $T_j=0$ when the result of the jth test is negative and $T_j=1$ when the test is positive. Let D be the binary random variable that models the result of the gold standard, so that $D=0$ when the individual is not diseased and $D=1$ when the individual is diseased. Let us consider that the jth diagnostic test is applied to all of the individuals in a random sample sized n_j, while all of the J samples are independent of each other; furthermore, the gold standard is applied to all of the individuals in all of the samples. Let p_j be the disease prevalence (i.e. $P(D=1)$) in the population from which the jth sample has been taken. Let $Se_j=P(T_j=1|D=1)$ and $Sp_j=P(T_j=0|D=0)$ be the sensitivity and the specificity of the jth diagnostic test, respectively. When the result of the jth diagnostic test is positive, the positive likelihood ratio is $LR_j^+ = Se_j/(1-Sp_j)$; and when the result of the diagnostic test is negative the negative likelihood ratio is $LR_j^- = (1-Se_j)/Sp_j$. Table 1 summarizes the results obtained by applying the jth diagnostic test and the gold standard to a random sample sized n_j.

Table 1
Observed frequencies and probabilities for jth binary test.

	$T_j=1$	$T_j=0$
Observed frequencies		
$D=1$	s_{1j1}	s_{1j0}
$D=0$	s_{0j1}	s_{0j0}
Probabilities		
$D=1$	$p_{1j1}=p_jSe_j$	$p_{1j0}=p_j(1-Se_j)$
$D=0$	$p_{0j1}=(1-p_j)(1-Sp_j)$	$p_{0j0}=(1-p_j)Sp_j$

For the jth diagnostic test let the probabilities $p_{ijk}=P(D=i, T_j=k)$, with $i, k=0, 1$ and $j=1, \ldots, J$, so that $\sum_{i,k=0}^{1} p_{ijk}=1$, and let $\boldsymbol{\omega}_j=(p_{0j0}, p_{0j1}, p_{1j0}, p_{1j1})^T$. The logarithms of the likelihood ratios of the jth diagnostic test are written in terms of the probabilities of the vector $\boldsymbol{\omega}_j$ as

$$\log(LR_j^+)=\log\left\{\frac{p_{1j1}(p_{0j0}+p_{0j1})}{p_{0j1}(p_{1j0}+p_{1j1})}\right\} \text{ and } \log(LR_j^-)=\log\left\{\frac{p_{1j0}(p_{0j0}+p_{0j1})}{p_{0j0}(p_{1j0}+p_{1j1})}\right\}. \tag{1}$$

As the elements of each vector $\boldsymbol{\omega}_j$ are the probabilities of multinomial distributions, the maximum likelihood estimators of the probabilities p_{ijk} are $\hat{p}_{ijk}=s_{ijk}/n_j$ and the variance-covariance matrix of the vector $\hat{\boldsymbol{\omega}}_j$ is

$$\sum_{\hat{\boldsymbol{\omega}}_j}=\{diag(\boldsymbol{\omega}_j)-\boldsymbol{\omega}_j\boldsymbol{\omega}_j^T\}/n_j.$$

Substituting in Eq. (1) each parameter with its maximum likelihood estimator, the maximum likelihood estimators of the likelihood ratios of the ith diagnostic test are

$$\log(\widehat{LR}_j^{+})=\log\left\{\frac{s_{1j1}(s_{0j0}+s_{0j1})}{s_{0j1}(s_{1j0}+s_{1j1})}\right\} \text{ and } \log(\widehat{LR}_j^{-})=\log\left\{\frac{s_{1j0}(s_{0j0}+s_{0j1})}{s_{0j0}(s_{1j0}+s_{1j1})}\right\}.$$

For each jth diagnostic test let the vector $\log(\mathbf{LR}_j)=(\log(LR_j^+), \log(LR_j^-))^T$. Applying the delta method, the variance-covariance matrix of $\log(\widehat{\mathbf{LR}}_j)$ is

$$\sum_{\widehat{\log(\mathbf{LR}_j)}}=\left(\frac{\partial\log(\mathbf{LR}_j)}{\partial\boldsymbol{\omega}_j}\right)\sum_{\hat{\boldsymbol{\omega}}_j}\left(\frac{\partial\log(\mathbf{LR}_j)}{\partial\boldsymbol{\omega}_j}\right)^T.$$

Carrying out matrix operations and substituting each parameter with its maximum likelihood estimator, the estimated variances–covariances of the estimators of the logarithms of the likelihood ratios of the jth diagnostic test become

$$\widehat{Var}(\log(\widehat{LR}_j^{+}))=\frac{1-\widehat{Se}_j}{s_{1j1}}+\frac{\widehat{Sp}_j}{s_{0j1}},$$

$$\widehat{Var}(\log(\widehat{LR}_j^{-}))=\frac{\widehat{Se}_j}{s_{1j0}}+\frac{1-\widehat{Sp}_j}{s_{0j0}},$$

and

$$\widehat{Cov}(\log(\widehat{LR}_j^{+}), \log(\widehat{LR}_j^{-}))=-\frac{n_j}{(s_{0j0}+s_{0j1})(s_{1j0}+s_{1j1})}.$$

Therefore, the logarithms of the estimators of the positive and negative likelihood ratios are correlated and, as a consequence, the comparison of the likelihood ratios of two diagnostic tests should be made simultaneously. We next derive an asymptotic hypothesis test to simultaneously compare the positive and negative likelihood ratios of J binary diagnostic tests with independent samples.

Let $\log(\mathbf{LR})=\left(\log(LR_1^+),\log(LR_1^-),\ldots,\log(LR_J^+),\log(LR_J^-)\right)^T$ be a vector of size $2J$ whose components are the logarithms of likelihood ratios of the J diagnostic tests. As the diagnostic tests are applied to samples which are independent from each other, the likelihood ratios (and their logarithms) of the diagnostic tests are independent and therefore the estimated

variance-covariance matrix of the vector $\log(\widehat{\mathbf{LR}})$ is

$$\hat{\sum_{\log(\widehat{\mathbf{LR}})}} = \begin{pmatrix} \hat{\sum_{\log(\widehat{\mathbf{LR}}_1)}} & & & 0 \\ & \ddots & & \\ & & \hat{\sum_{\log(\widehat{\mathbf{LR}}_j)}} & \\ & & & \ddots \\ 0 & & & \hat{\sum_{\log(\widehat{\mathbf{LR}}_J)}} \end{pmatrix},$$

and applying the multivariate central limit theorem it is verified that

$$\sqrt{n}(\log(\widehat{\mathbf{LR}})-\log(\mathbf{LR})) \underset{n\to\infty}{\longrightarrow} N\left(\mathbf{0}, \sum_{\log(\mathbf{LR})}\right),$$

where $n=\sum_j n_j$.

The joint hypothesis test to simultaneously compare the positive and negative likelihood ratios of the J binary diagnostic tests is

$$H_0 : \log(LR_1^+) = \ldots = \log(LR_J^+) \text{ and } \log(LR_1^-) = \ldots = \log(LR_J^-)$$
$$H_1 : \text{at least one equality is not true,}$$

which is equivalent to the hypothesis test

$$H_0 : \boldsymbol{\varphi}\log(\mathbf{LR}) = \mathbf{0}$$
$$H_1 : \boldsymbol{\varphi}\log(\mathbf{LR}) \neq \mathbf{0}, \tag{2}$$

where $\boldsymbol{\varphi}$ is a full rank matrix whose dimensions are $2(J-1)\times 2J$ and whose values are known constants. For example, for $J=3$

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

and for $J=4$

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

Finally, the statistic for the joint hypothesis test in Eq. (2) is

$$Q^2 = \log(\widehat{\mathbf{LR}})^T \boldsymbol{\varphi}^T \left(\boldsymbol{\varphi} \hat{\sum_{\log(\widehat{\mathbf{LR}})}} \boldsymbol{\varphi}^T\right)^{-1} \boldsymbol{\varphi}\log(\widehat{\mathbf{LR}}) \underset{n\to\infty}{\longrightarrow} \chi^2_{2(J-1)}. \tag{3}$$

Since the vector $\log(\widehat{\mathbf{LR}})$ is asymptotically distributed according to a multivariate normal distribution, the statistic Q^2 is asymptotically distributed according to a central chi-square distribution with $2(J-1)$ degrees of freedom (i.e. the number of rows of the matrix $\boldsymbol{\varphi}$) when the null hypothesis is true.

The method proposed to simultaneously compare the likelihood ratios of the J diagnostic tests is based on the estimation of the variance-covariance matrix by applying the delta method. An alternative approach consists of estimating this matrix using bootstrap methods. Consequently, for each one of the J samples H bootstrap data sets are generated and from them the variances–covariances of the likelihood ratios are estimated. Note that we assume for both approaches that there are no zero cell counts in order to avoid estimation problems of likelihood ratios and their (co)variances. For instance, zero cell counts for one of the J samples would also result in zero cell counts for all corresponding H bootstrap data sets.

From the previously constructed global hypothesis test it is possible to obtain the marginal hypothesis tests i.e. $H_0 : \log(LR_i^+) = \log(LR_j^+)$ and $H_0 : \log(LR_i^-) = \log(LR_j^-)$. Thus, the statistic to test $H_0 : \log(LR_i) = \log(LR_j)$ versus $H_1 :$

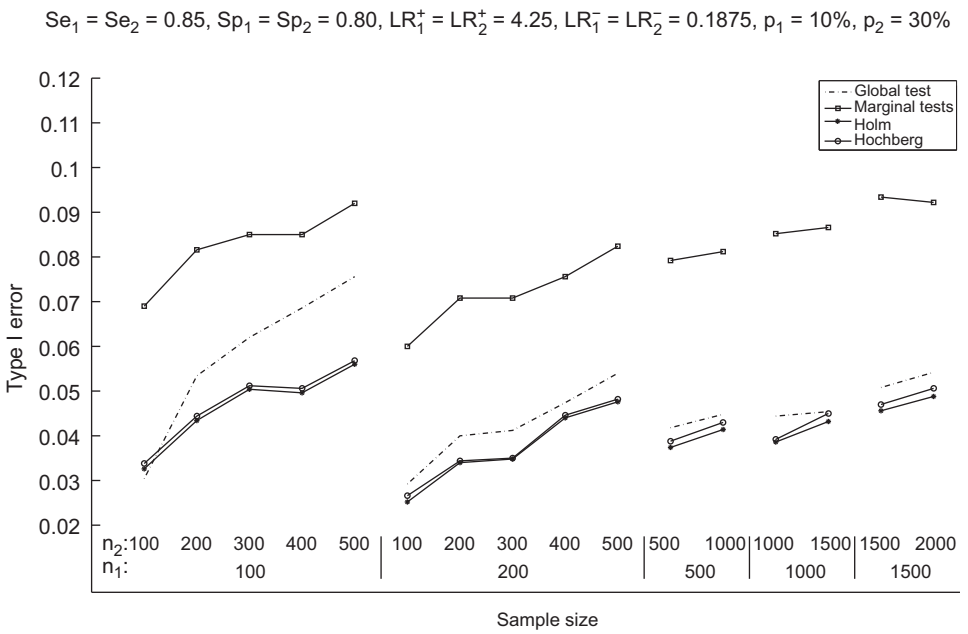

Fig. 1. Type I errors of hypothesis tests for two binary diagnostic tests.

$\log(LR_i) \neq \log(LR_j)$ is

$$\frac{\log(\widehat{LR}_i) - \log(\widehat{LR}_j)}{\sqrt{\widehat{Var}(\log(\widehat{LR}_i)) + \widehat{Var}(\log(\widehat{LR}_j))}} \to N(0,1), \tag{4}$$

where $\widehat{LR}$ is $\widehat{LR}^+$ or $\widehat{LR}^-$. This marginal hypothesis test is similar to that proposed by Altman and Bland (2003) to compare two relative risks with independent samples and the method proposed by Pepe (2003).

In this section, we have constructed a global hypothesis test based on the chi-square distribution to simultaneously compare the logarithms of the positive likelihood ratios and the logarithms of negative likelihood ratios of multiple binary tests with independent samples. An alternative method to this global hypothesis test consists of solving the marginal hypothesis tests, $H_0 : \log(LR_i^+) = \log(LR_j^+)$ and $H_0 : \log(LR_i^-) = \log(LR_j^-)$, and applying a multiple comparison procedure such as the method proposed by Holm (1979) or that proposed by Hochberg (1988), which are less conservative methods than the traditional Bonferroni (1936) approach.

3. Simulation studies

We carried out simulation experiments to study the type I error and the power of the global hypothesis test (Eq. (2)) when comparing the likelihood ratios of two and of three binary diagnostic tests, respectively. In both cases we have taken as the nominal error $\alpha = 5\%$, and we have also studied the type I error and the power of three alternative methods to solve the joint hypothesis test: (1) solve the marginal hypothesis tests to an error rate of $\alpha = 5\%$, (2) solve the marginal hypothesis tests and apply Holm's (1979) method to an error rate of $\alpha = 5\%$, (3) solve the marginal hypothesis tests and apply Hochberg's (1988) method to an error rate of $\alpha = 5\%$.

3.1. Two binary diagnostic tests

The simulation experiments consisted of generating 5000 pairs of random samples with multinomial distributions with different sizes (n_1,n_2={100, 200, 300, 400, 500, 1000, 1500, 2000}) whose probabilities have been calculated from the probabilities given in Table 1. As values of the disease prevalence in each population we took p_1,p_2={10%, 30%, 50%}, and as sensitivities and specificities we took {0.75, 0.80, 0.85, 0.90, 0.95}, which are values that appear quite frequently in clinical practice. The simulation experiments were designed in such a way that in all of the samples generated it is possible to estimate the logarithms of the likelihood ratios and their variances–covariances (i.e. there were no zero cell counts).

3.1.1. Type I error

Figs. 1 and 2 present part of the results (additional results are available upon request) obtained for the type I error of the different methods to solve the joint hypothesis test, thereby estimating the variances–covariances through the delta

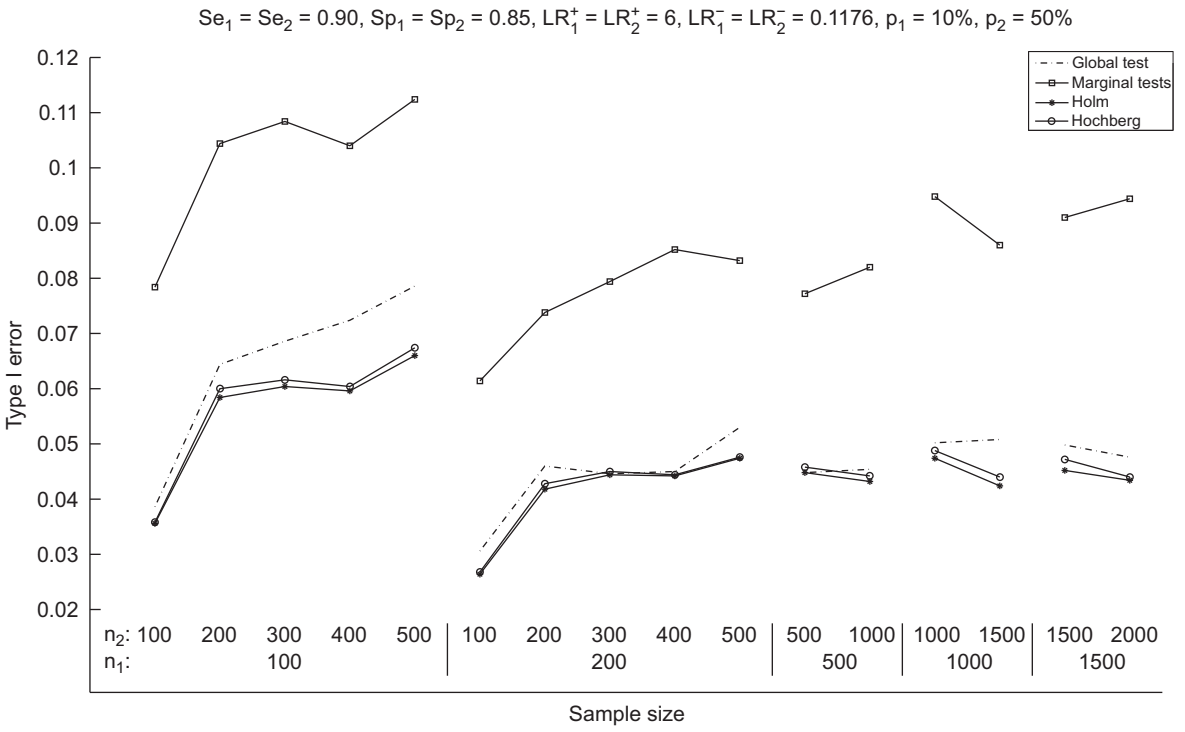

Fig. 2. Type I errors of hypothesis tests for two binary diagnostic tests.

method. A numerical representation of the results is available in Appendix A (Tables A.1 and A.2). From the results the following conclusions are obtained:

(1) Global test: The sample sizes and the prevalences have an important effect on the type I error of the global hypothesis test. In general terms, when the two samples are relatively large ($n_i \geq 200$) the type I error fluctuates around the nominal error regardless of the two prevalences, so that the global hypothesis test has the classic performance of an asymptotic hypothesis test. When one of the samples is relatively small ($n_i = 100$) and the corresponding prevalence is also small ($p_i = 10\%$) and the other sample is larger ($n_j \geq 300$ or 400), the type I error usually overwhelms the nominal error. Consequently, when a sample and its prevalence are relatively small and the two sample sizes are very unbalanced (for example $n_1 = 100$, $p_1 = 10\%$ and $n_2 \geq 300$), the type I error is slightly greater than the nominal error of 5%. This behavior may be due to the fact that with these sample sizes the multinomial distribution does not have a good convergence to the normal distribution, since with larger sample sizes the type I error does not overwhelm the nominal error.
(2) Marginal test with $\alpha = 5\%$: When the joint hypothesis test is solved applying the two marginal hypothesis tests to an error rate of $\alpha = 5\%$, the type I error overwhelms the nominal error, so that this method can lead to erroneous results.
(3) Holm's method and Hochberg's method: The type I error of Holm's (Hochberg's) method is almost always lower than the type I error of the global hypothesis test. In general, Holm's (Hochberg's) method is a conservative one and its type I error is almost always lower than the nominal error. Regarding both methods, although the type I error of Holm's method is slightly lower compared to that of Hochberg's method, there are no important differences between both type I errors.

Similar conclusions are obtained for the other values of sensitivities, specificities and prevalences, and when the variances–covariances are estimated using bootstrap methods.

3.1.2. Power

Figs. 3 and 4 summarize part of the results regarding the power of each one of the four methods to solve the joint hypothesis test. Appendix A (Tables A.3 and A.4) includes a numerical representation of the results. All of the methods require both samples to be large so that the power is large (greater than 80% or 90%), between 500 and 1500 individuals depending on the values of the likelihood ratios. The method based on the marginal hypothesis tests to an error rate of $\alpha = 5\%$ has a power greater than the other methods due to the fact that its type I error is also greater than the type I error of the other methods. Regarding Holm's method and Hochberg's method, there are no important differences between the

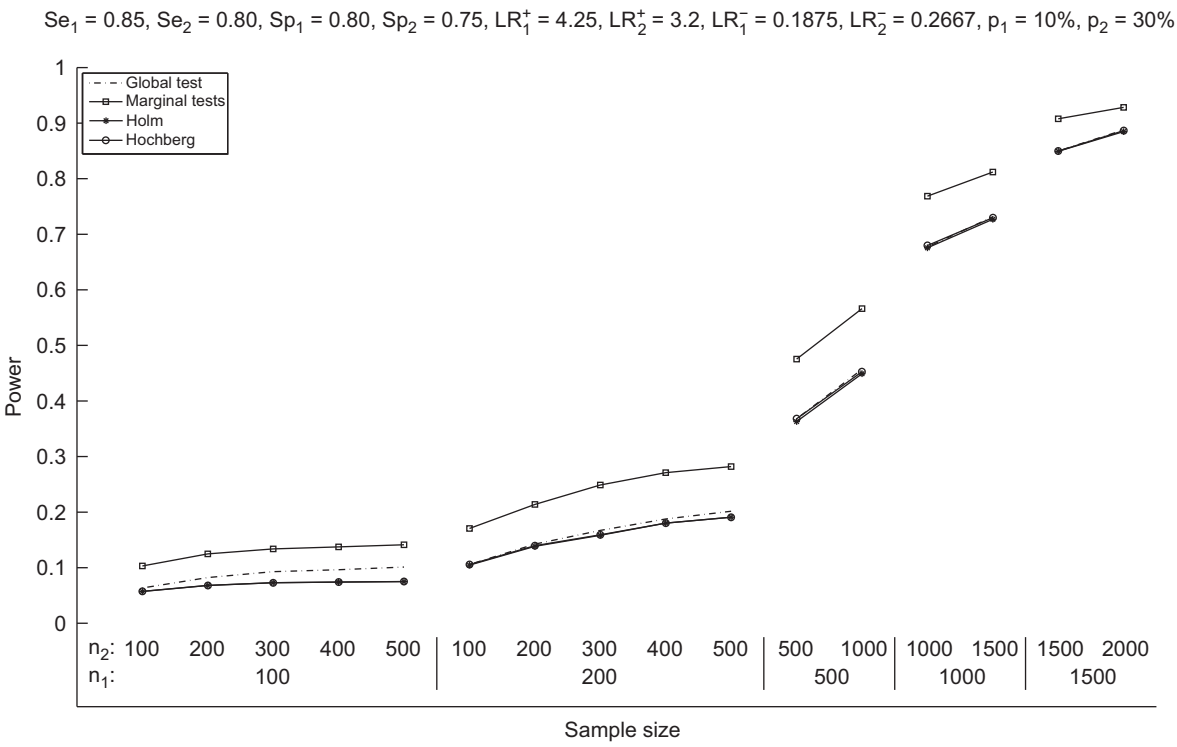

Fig. 3. Powers of hypothesis tests for two binary diagnostic tests.

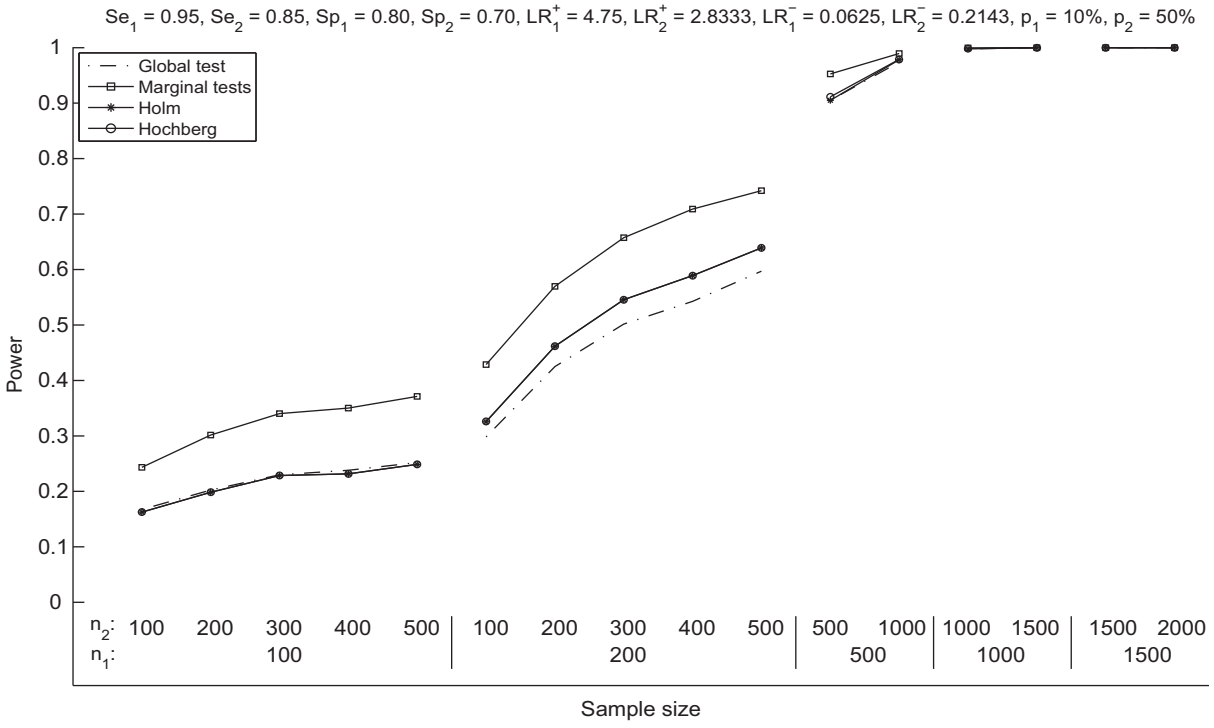

Fig. 4. Powers of hypothesis tests for two binary diagnostic tests.

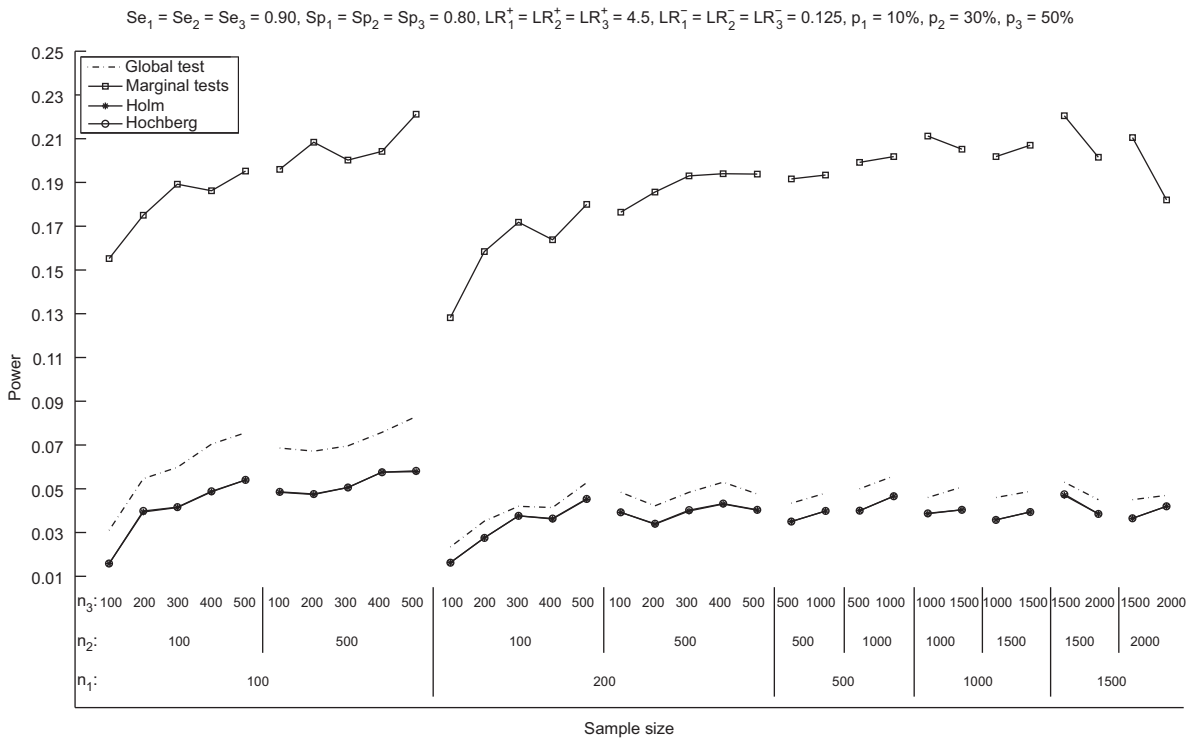

Fig. 5. Type I errors of hypothesis tests for three binary diagnostic tests.

powers of both methods. Similar conclusions are obtained for the rest of the values of sensitivities, specificities and prevalences, and when the variances–covariances are estimated using bootstrap methods.

3.2. Three binary diagnostic tests

In order to compare the likelihood ratios of three binary diagnostic tests, the simulation experiments consisted of generating 5000 sets of random samples with multinomial distributions of different sizes (n_1,n_2,n_3={100, 200, 300, 400, 500, 1000, 1500, 2000}), where the probabilities of each multinomial distribution were calculated from the probabilities given in Table 1. As values of the disease prevalence in each population and as sensitivities and specificities we took the same values as in the case of two diagnostic tests. As in the previous case, the simulation experiments were designed so that in all of the random samples it is possible to estimate the logarithms of the likelihood ratios and their variances–covariances.

3.2.1. Type I error

In Fig. 5 we show part of the results obtained for the type I error of the different methods to compare the likelihood ratios of the three diagnostic tests (cf. Table A.5). The conclusions that are obtained for three binary tests are very similar to those obtained when comparing two diagnostic tests:

(1) Global test: The sample sizes and the prevalences have an important effect on the type I error. In general terms, when the samples are relatively large ($n_i \geq 200$) the type I error fluctuates around the nominal error and when a sample and its prevalence are small ($n_i = 100$ and $p_i = 10\%$) and the other samples are larger ($n_j \geq 300$ or 400) (regardless of the corresponding prevalences), the type I error usually overwhelms the nominal error.
(2) Marginal test with $\alpha = 5\%$: when the joint hypothesis test is solved applying the two marginal hypothesis tests to an error rate of $\alpha = 5\%$, the type I error clearly overwhelms the nominal error, and therefore this method leads to erroneous results.
(3) Holm's method and Hochberg's method: In general terms, both Holm's method and Hochberg's method are conservative methods, and their type I errors are almost always lower than the nominal error.

Similar conclusions are obtained for the other values of sensitivities, specificities and prevalences, and when the variances–covariances are estimated using bootstrap methods.

3.2.2. Power

Fig. 6 presents the results regarding the power of each one of the four methods to solve the joint hypothesis test (cf. Table A.6). In general terms, it is necessary to have large samples, of between 500 and 1500 individuals (depending on the values of the likelihood ratios), so that the power of the global hypothesis test is high (greater than 80% or 90%). The method based on the marginal hypothesis tests to an error rate of $\alpha=5\%$ has a greater power than the other methods due to the fact that its type I error is greater than the type I error of the other three methods. Regarding Holm's method and Hochberg's method, there are no important differences between the power of both methods. Similar conclusions are obtained for the rest of the values of sensitivities, specificities and prevalences, and when the variances–covariances are estimated using bootstrap methods.

4. Hypothesis tests for multi-level diagnostic tests

The likelihood ratios can be estimated for diagnostic tests with more than two levels (i.e. multi-level diagnostic tests). Thus, for example a diagnostic test can lead to three results: negative, non-negative non-positive, and positive. We propose a global test to simultaneously compare the likelihood ratios of multiple multi-level tests.

Let us consider J multi-level diagnostic tests in such a way that each one of them can lead to K results. Let $T_1, \ldots, T_J$ be the random variables that model the results of the diagnostic tests, such that $T_j=k$ when the jth test indicates a result k, with $k=1, \ldots, K$, and let D be the binary random variable that models the result of the gold standard. The jth multi-level diagnostic test is applied to a random sample sized n_j, and all of the samples are independent of each other. Consequently, each random sample is extracted from a population, and all of the J populations are independent of each other. In Table 2

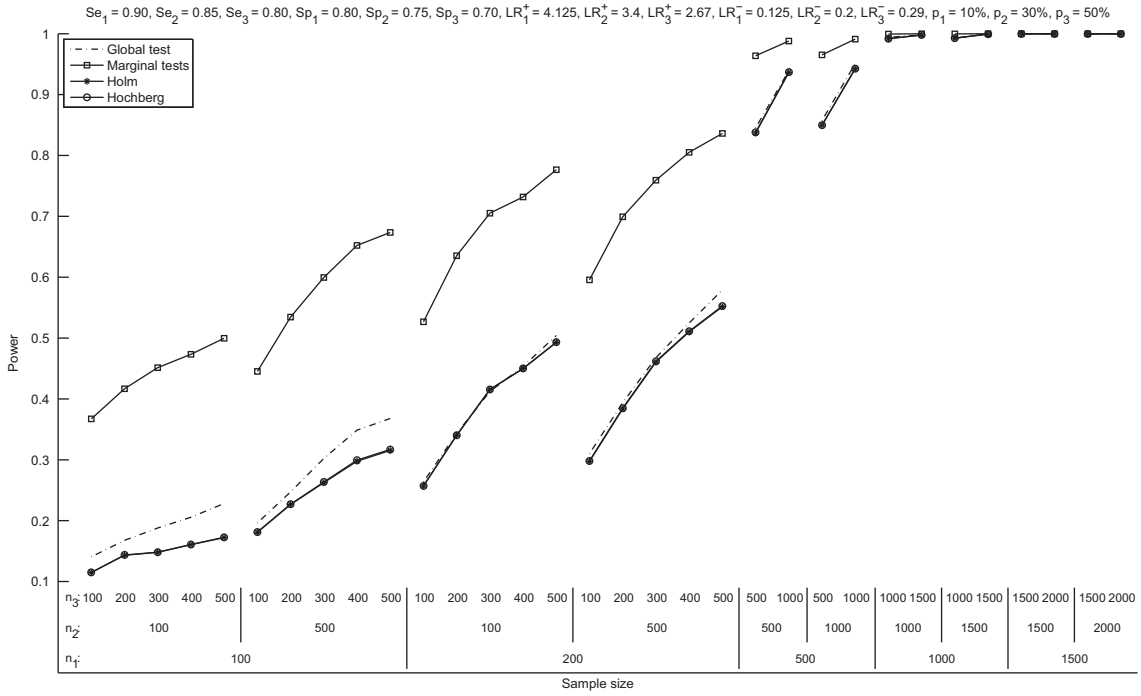

Fig. 6. Powers of hypothesis tests for three binary diagnostic tests.

Table 2
Observed frequencies and probabilities for jth multi-level test.

	$T_j=1$	…	$T_j=k$	…	$T_j=K$
Observed frequencies					
$D=1$	s_{1j1}	…	s_{1jk}	…	s_{1jK}
$D=0$	s_{0j1}	…	s_{0jk}	…	s_{0jK}
Probabilities					
$D=1$	$p_j p_{j1\|1}$	…	$p_j p_{jk\|1}$	…	$p_j p_{jK\|1}$
$D=0$	$q_j p_{j1\|0}$	…	$q_j p_{jk\|0}$	…	$q_j p_{jK\|0}$

$p_{jk|i}=P(T_j=k|D=i)$, $q_j=1-p_j$, $j=1, \ldots, J$, $k=1, \ldots, K$, $i=0, 1$.

we show the frequencies obtained by applying the jth multi-level test to a random sample sized n_j, along with the probabilities of each cell in the multinomial distribution. For the jth multi-level test, the likelihood ratio in each level is defined (Simel et al., 1991) as

$$LR_j(k) = \frac{P(T_j = k|D = 1)}{P(T_j = k|D = 0)}, j = 1,\ldots,J,\ k = 1,\ldots,K. \tag{5}$$

Let the probabilities $p_{ijk} = P(D = i, T_j = k)$ with $i = 0,1$, $j = 1, \ldots, J$ and $k = 1, \ldots, K$. For the jth multi-level diagnostic test, the logarithm of the likelihood ratio in the kth level is written in terms of the previous probabilities as

$$\log(LR_j(k)) = \log\left\{\frac{1-\sum_{h=1}^{K} p_{1jh}}{\sum_{h=1}^{K} p_{1jh}}\frac{p_{1jk}}{p_{0jk}}\right\}. \tag{6}$$

Following a similar procedure to that of Section 2, let $\boldsymbol{\omega}_j = (p_{1j1}, \ldots, p_{1jK}, p_{0j1}, \ldots, p_{0jK})^T$ be a vector sized $2K$ whose components are the probabilities of a multinomial distribution. The maximum likelihood estimators of the probabilities p_{ijk} are $\hat{p}_{ijk} = s_{ijk}/n_j$ and the variances–covariances matrix of the vector $\hat{\boldsymbol{\omega}}_j$ is $\sum_{\hat{\boldsymbol{\omega}}_j} = \{diag(\boldsymbol{\omega}_j) - \boldsymbol{\omega}_j\boldsymbol{\omega}_j^T\}/n_j$.

For the jth multi-level test assume the vector $\log(\mathbf{LR}_j) = (\log(LR_j(1)), \ldots, \log(LR_j(K)))^T$ and applying the delta method, the variance-covariance matrix of $\log(\widehat{\mathbf{LR}}_j)$ is

$$\sum_{\log(\widehat{\mathbf{LR}}_j)} = \left(\frac{\partial \log(\mathbf{LR}_j)}{\partial \boldsymbol{\omega}_j}\right) \sum_{\hat{\boldsymbol{\omega}}_j} \left(\frac{\partial \log(\mathbf{LR}_j)}{\partial \boldsymbol{\omega}_j}\right)^T. \tag{7}$$

Let $\log(\mathbf{LR}) = (\log(LR_1(1)), \ldots, \log(LR_1(K)), \ldots, \log(LR_J(1)), \ldots, \log(LR_J(K)))^T$ be a vector of size JK. As each diagnostic test is applied to a random sample and all of the samples are independent of each other, the likelihood ratios of the diagnostic tests are also independent and the estimated variance-covariance matrix of the vector $\log(\widehat{\mathbf{LR}})$ is

$$\hat{\sum}_{\log(\widehat{\mathbf{LR}})} = \begin{pmatrix} \hat{\sum}_{\log(\widehat{\mathbf{LR}}_1)} & & & 0 \\ & \ddots & & \\ & & \hat{\sum}_{\log(\widehat{\mathbf{LR}}_j)} & \\ & & & \ddots \\ 0 & & & \hat{\sum}_{\log(\widehat{\mathbf{LR}}_J)} \end{pmatrix}.$$

The joint hypothesis test to simultaneously compare the likelihood ratios of the J diagnostic tests in all of the levels is

$$H_0 : \begin{cases} \log(LR_1(1)) = \ldots = \log(LR_j(1)) = \ldots = \log(LR_J(1)) \\ \vdots \\ \log(LR_1(k)) = \ldots = \log(LR_j(k)) = \ldots = \log(LR_J(k)) \\ \vdots \\ \log(LR_1(K)) = \ldots = \log(LR_j(K)) = \ldots = \log(LR_J(K)) \end{cases}$$

H_1 : at least one equality is not true,

which is equivalent to the hypothesis test

$$\begin{aligned} &H_0 : \boldsymbol{\varphi}\log(\mathbf{LR}) = \mathbf{0} \\ &H_1 : \boldsymbol{\varphi}\log(\mathbf{LR}) \neq \mathbf{0}, \end{aligned} \tag{8}$$

where $\boldsymbol{\varphi}$ is a full rank matrix whose dimensions are $K(J-1) \times JK$ and whose values are known constants. For example, for two diagnostic tests with three levels ($J = 2$ and $K = 3$)

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

and for three diagnostic tests with three levels ($J = 3$ and $K = 3$)

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

Finally, the statistic for the joint hypothesis test in Eq. (8) is

$$Q^2 = \log(\widehat{\mathbf{LR}})^T \boldsymbol{\varphi}^T \left(\boldsymbol{\varphi} \hat{\sum_{\log(\widehat{\mathbf{LR}})}} \boldsymbol{\varphi}^T \right)^{-1} \boldsymbol{\varphi} \log(\widehat{\mathbf{LR}}) \xrightarrow[n\to\infty]{} \chi^2_{K(J-1)}. \tag{9}$$

If the global hypothesis test is significant to the error rate α, research into the causes of the significance is carried out through the paired comparisons of diagnostic tests in each level, and the contrast statistic is

$$\frac{\log\left(\widehat{LR}_i(k)\right) - \log\left(\widehat{LR}_j(k)\right)}{\sqrt{\widehat{Var}\left(\log(\widehat{LR}_i(k))\right) + \widehat{Var}\left(\log(\widehat{LR}_j(k))\right)}} \to N(0,1),$$

and applying a multiple comparison method (Holm or Hochberg).

5. Application

The results obtained in Section 2 were applied to two examples, one with two binary diagnostic tests and another with three binary diagnostic tests.

5.1. Two binary tests

Weiner et al. (1979) studied the diagnosis of arterial disease in men and women through three diagnostic tests (exercise stress test, clinical history and resting EKG) using as the gold standard a coronary arteriography. Table 3 summarizes the results obtained for the clinical history (modeled through the variable T_1) and the gold standard (modeled through the variable D) in a sample of 1465 men, and the results obtained for the resting EKG (modeled through the variable T_2) and the gold standard in a sample of 580 women. For the men it holds that $\widehat{LR}_1^+ = 1.71$ ($\log(\widehat{LR}_1^+) = 0.54$) and $\widehat{LR}_1^- = 0.12$ ($\log(\widehat{LR}_1^-) = -2.13$), and for the women $\widehat{LR}_2^+ = 1.38$ ($\log(\widehat{LR}_2^+) = 0.32$) and $\widehat{LR}_2^- = 0.85$ ($\log(\widehat{LR}_1^-) = -0.16$). The estimated variance-covariance matrix of $\log(\widehat{\mathbf{LR}}) = (\log(\widehat{LR}_1^+), \log(\widehat{LR}_1^-), \log(\widehat{LR}_2^+), \log(\widehat{LR}_2^-))^T$ is

$$\hat{\sum_{\log(\widehat{\mathbf{LR}})}} \approx \begin{pmatrix} 0.00187 & -0.00324 & 0 & 0 \\ -0.00324 & 0.02035 & 0 & 0 \\ 0 & 0 & 0.01505 & -0.00835 \\ 0 & 0 & -0.00835 & 0.00487 \end{pmatrix},$$

and applying Eq. (3), the contrast statistic for the joint hypothesis test

$H_0 : \log(LR_1^+) = \log(LR_2^+)$ and $\log(LR_1^-) = \log(LR_2^-)$

$H_1 : \log(LR_1^+) \neq \log(LR_2^+)$ and/or $\log(LR_1^-) \neq \log(LR_2^-)$,

Table 3
Diagnosis of arterial disease.

	Clinical history (men)			Resting EKG (women)		
	$T_1=1$	$T_1=0$	Total	$T_2=1$	$T_2=0$	Total
$D=1$	969	54	1023	67	102	169
$D=0$	245	197	442	118	293	411
Total	1214	251	1465	185	395	580

Table 4
Prenatal diagnosis of genetic disorders.

	Transcervical CVS			Mid-trimester amniocentesis			Early amniocentesis		
	$T_1=1$	$T_1=0$	Total	$T_2=1$	$T_2=0$	Total	$T_3=1$	$T_3=0$	Total
$D=1$	21	3	24	20	1	21	53	2	55
$D=0$	19	820	839	2	944	946	3	1997	2000
Total	40	823	863	22	945	967	56	1999	2055

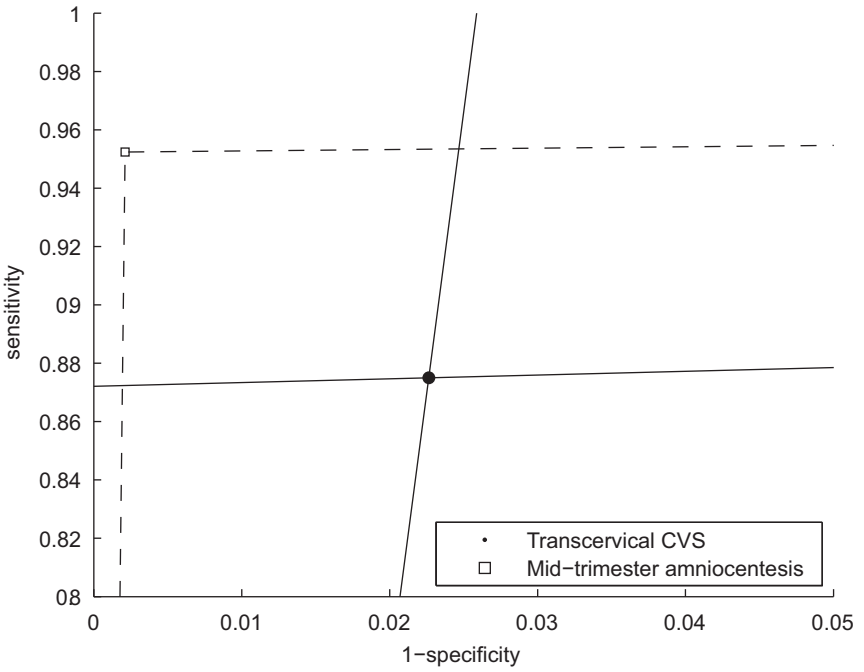

Fig. 7. Likelihood ratios graph of transcervical CVS and mid-trimester amniocentesis.

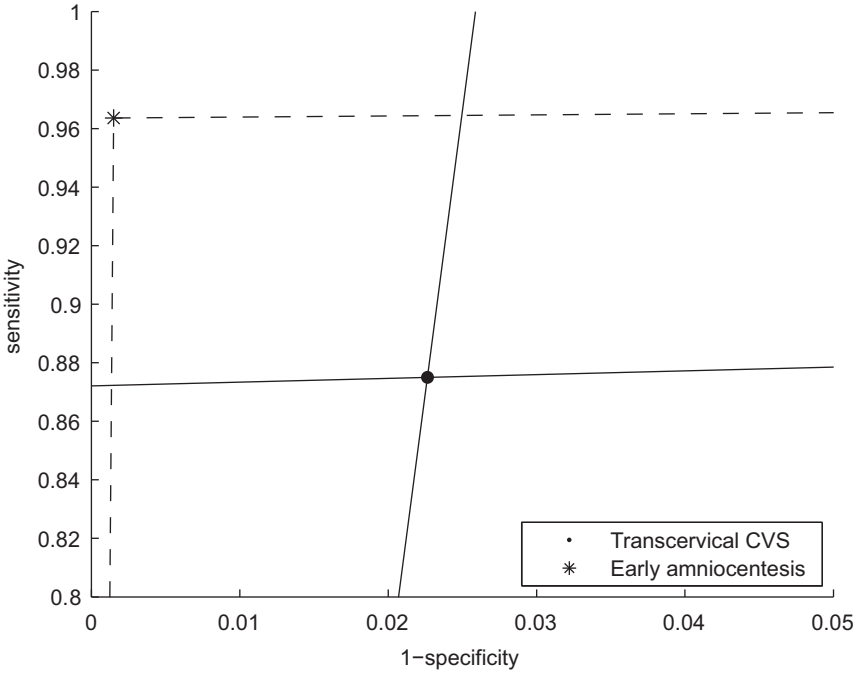

Fig. 8. Likelihood ratios graph of transcervical CVS and early amniocentesis.

is $Q^2=194.532$ (2 degrees of freedom, p-value is nearly 0), and therefore the null hypothesis of equality of the positive likelihood ratios and of the equality of the negative likelihood ratios is rejected. In order to study the causes of the significance the two marginal hypothesis tests are solved and Holm's method or Hochberg's method is applied. It is found that we cannot reject the null hypothesis $H_0 : \log(LR_1^+) = \log(LR_2^+)$ ($z=1.64$, p-value$=0.1013$), but the null hypothesis $H_0 : \log(LR_1^-) = \log(LR_2^-)$ can be rejected ($z=12.39$, p-value is nearly 0). Therefore, we cannot conclude that a positive result in the clinical history in men is more indicative of the presence of coronary arterial disease than a positive result in the

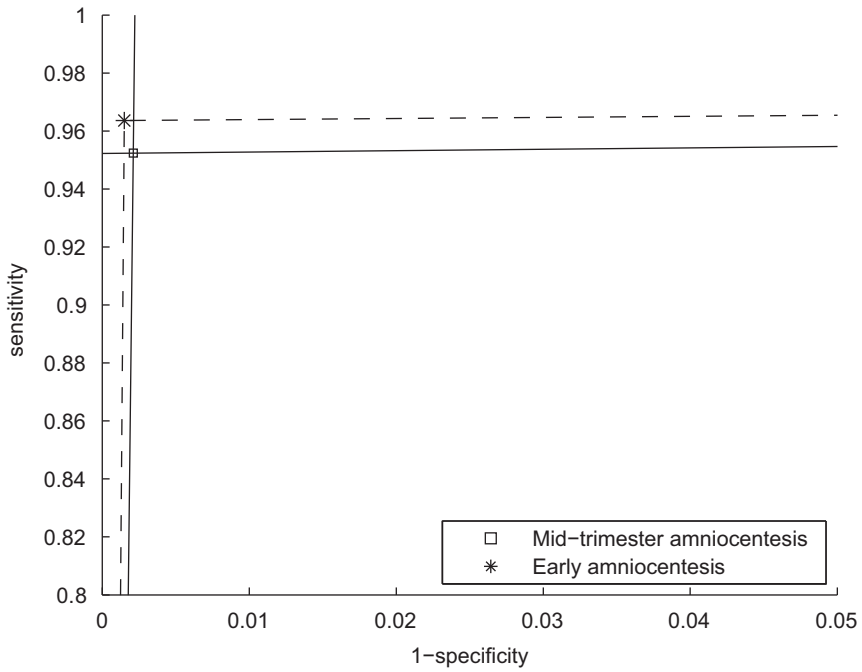

Fig. 9. Likelihood ratios graph of mid-trimester amniocentesis and early amniocentesis.

resting EKG in women; but a negative result in the clinical history in men is more indicative of the absence of the disease than a negative result in the resting EKG in women.

5.2. Three binary tests

Transcervical chorionic villus sampling (CVS), mid-trimester amniocentesis and early amniocentesis are three invasive approaches for prenatal diagnosis of genetic disorders (Alfirevic et al., 2003). Table 4 presents data for each of these diagnostic methods, where each of the methods was used for a separate sample of women (i.e. 863 women had transcervical CVS, 967 women had mid-trimester amniocentesis and 2055 women had early amniocentesis), and the golden standard to obtain the karyotype of the fetus or liveborn child (i.e. abnormal versus normal). For transcervical CVS $\widehat{LR}_1^+ = 38.64$ and $\widehat{LR}_1^- = 0.13$, mid-trimester amniocentesis resulted in $\widehat{LR}_2^+ = 450.48$ and $\widehat{LR}_2^- = 0.05$, while $\widehat{LR}_3^+ = 642.42$ and $\widehat{LR}_3^- = 0.04$ for early amniocentesis. Likelihood ratios graphs for these diagnostic tests are presented in Figs. 7–9 (Biggerstaff, 2000). These figures suggest an increased diagnostic ability of early and mid-trimester amniocentesis as compared to transcervical CVS. In the following, a formal statistical test is performed to compare the diagnostic methods. The estimated variance-covariance matrix of $\log(\widehat{\mathbf{LR}}) = \left(\log(\widehat{LR}_1^+), \log(\widehat{LR}_1^-), \log(\widehat{LR}_2^+), \log(\widehat{LR}_2^-), \log(\widehat{LR}_3^+), \log(\widehat{LR}_3^-)\right)^T$ equals

$$\hat{\sum}_{\log(\widehat{\mathbf{LR}})} \approx \begin{pmatrix} 0.05739 & -0.04286 & 0 & 0 & 0 & 0 \\ -0.04286 & 0.29169 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.50132 & -0.04868 & 0 & 0 \\ 0 & 0 & -0.04868 & 0.95238 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.33352 & -0.01868 \\ 0 & 0 & 0 & 0 & -0.01868 & 0.48182 \end{pmatrix}.$$

The contrast statistic (Eq. (3)) for the joint hypothesis test

$H_0 : \log(LR_1^+) = \log(LR_2^+) = \log(LR_3^+)$ and $\log(LR_1^-) = \log(LR_2^-) = \log(LR_3^-)$

H_1 : at least one equality is not true

is $Q^2 = 28.433$ (4 degrees of freedom, p-value is 0.00001), and therefore the null hypothesis of equality of the positive likelihood ratios and of the equality of the negative likelihood ratios is rejected. Solving the marginal hypothesis tests and applying Hochberg's or Holm's method reveals that the null hypotheses $H_0 : \log(LR_1^+) = \log(LR_2^+)$ ($z = 3.29$, p-value $= 0.001$) and $H_0 : \log(LR_1^+) = \log(LR_3^+)$ can be rejected ($z = 4.50$, p-value is 0.000007). Based on these results we can conclude that

diagnosis of abnormality using mid-trimester amniocentesis or early amniocentesis is more indicative of the presence of genetic disorders than the diagnosis of abnormality via transcervical CVS.

6. Discussion

The positive and negative likelihood ratio are, along with sensitivity and specificity, some of the most important measures to assess and compare the performance of binary diagnostic tests. The likelihood ratios only depend on the sensitivity and specificity of the diagnostic test, so that they are measures that only depend on the intrinsic ability of the diagnostic test to distinguish between diseased individuals and non-diseased individuals and, therefore, positive and negative likelihood ratios are dependent on each other. In this article, we first propose a global hypothesis test to simultaneously compare the likelihood ratios of multiple binary diagnostic tests in unpaired designs. Therefore, due to the fact that these parameters of interest have an asymmetrical distribution, the logarithmic transformation of the likelihood ratios has been considered, making the distribution more symmetric. Simulation experiments were carried out to study the type I error and the power of several methods to solve the hypothesis test for simultaneous comparison of the likelihood ratios of two and three binary diagnostic tests. The results of the simulation experiments have shown that the method based on the chi-square distribution is the approach with the best performance, both in terms of type I error and power, and in general it is necessary to have large samples so that the power is high. Based on the results from the simulation experiments, we propose the following steps to compare the likelihood ratios of multiple binary tests in unpaired designs. Apply the global hypothesis test based on the chi-square distribution (Eq. (3)) to an error rate of α. If the global hypothesis test is not significant to an error rate of α, the homogeneity hypothesis of the likelihood ratios is not rejected, otherwise, employ the marginal hypothesis tests (cf. Eq. (4)) and Holm's method or Hochberg's method in order to investigate the causes of the significance. This strategy has been used to analyze the data from Section 5.1 and Section 5.2. In case of low sample sizes and if they are unbalanced and some of the prevalences are small, one can directly apply the marginal hypothesis tests along with Holm's method or Hochberg's method.

Furthermore, the likelihood ratios are also valid parameters to assess and compare the performance of diagnostic tests with more than two levels. In this study, we have constructed a global hypothesis test to simultaneously compare the likelihood ratios of multiple multi-level diagnostic tests with independent samples. The hypothesis test is an extension of that obtained for binary diagnostic tests.

Acknowledgments

Jan Luts is a Postdoctoral Fellow of the Research Foundation—Flanders (FWO-Vlaanderen); the Spanish Ministry of Science, Grant number MTM2009-08886; the Department for Innovation, Science and Business of the Autonomous Government of Andalusia, Spain, Grant number FQM-01459; GOA Ambiorics; GOA MaNet; CoE EF/05/006; FWO G.0341.07 (Data fusion); IWT: TBM070706-IOTA3; Belgian Federal Science Policy Office IUAP P6/04 (DYSCO).

Appendix A

See Table A1–A6 for more details.

Table A.1
Type I errors of hypothesis tests for two binary diagnostic tests.

$Se_1 = Se_2 = 0.85$, $Sp_1 = Sp_2 = 0.80$, $LR_1^+ = LR_2^+ = 4.25$, $LR_1^- = LR_2^- = 0.1875$, $p_1 = 10\%$, $p_2 = 30\%$

n_1	n_2	Global test	Marginal tests	Holm	Hochberg
100	100	0.0304	0.0690	0.0326	0.0338
	200	0.0534	0.0816	0.0434	0.0444
	300	0.0620	0.0850	0.0504	0.0512
	400	0.0686	0.0850	0.0496	0.0506
	500	0.0756	0.0920	0.0560	0.0568
200	100	0.0292	0.0600	0.0252	0.0266
	200	0.0400	0.0708	0.0340	0.0344
	300	0.0412	0.0708	0.0348	0.0350
	400	0.0474	0.0756	0.0440	0.0446
	500	0.0540	0.0824	0.0476	0.0482
500	500	0.0418	0.0792	0.0374	0.0388
	1000	0.0448	0.0812	0.0414	0.0430
1000	1000	0.0444	0.0852	0.0386	0.0392
	1500	0.0454	0.0866	0.0432	0.0450
1500	1500	0.0508	0.0934	0.0456	0.0470
	2000	0.0542	0.0922	0.0488	0.0506

Table A.2
Type I errors of hypothesis tests for two binary diagnostic tests.

$Se_1 = Se_2 = 0.90$, $Sp_1 = Sp_2 = 0.85$, $LR_1^+ = LR_2^+ = 6$, $LR_1^- = LR_2^- = 0.1176$, $p_1 = 10\%$, $p_2 = 50\%$

n_1	n_2	Global test	Marginal tests	Holm	Hochberg
100	100	0.0386	0.0784	0.0356	0.0358
	200	0.0644	0.1044	0.0584	0.0600
	300	0.0686	0.1084	0.0604	0.0616
	400	0.0724	0.1040	0.0596	0.0604
	500	0.0786	0.1124	0.0660	0.0674
200	100	0.0306	0.0614	0.0264	0.0268
	200	0.0460	0.0738	0.0418	0.0428
	300	0.0446	0.0794	0.0444	0.0450
	400	0.0450	0.0852	0.0442	0.0444
	500	0.0530	0.0832	0.0474	0.0476
500	500	0.0448	0.0772	0.0448	0.0458
	1000	0.0454	0.0820	0.0432	0.0442
1000	1000	0.0502	0.0948	0.0474	0.0488
	1500	0.0508	0.0860	0.0424	0.0440
1500	1500	0.0498	0.0910	0.0452	0.0472
	2000	0.0476	0.0944	0.0434	0.0440

Table A.3
Powers of hypothesis tests for two binary diagnostic tests.

$Se_1 = 0.85$, $Se_2 = 0.80$, $Sp_1 = 0.80$, $Sp_2 = 0.75$, $LR_1^+ = 4.25$, $LR_2^+ = 3.2$, $LR_1^- = 0.1875$, $LR_2^- = 0.2667$, $p_1 = 10\%$, $p_2 = 30\%$

n_1	n_2	Global test	Marginal tests	Holm	Hochberg
100	100	0.0636	0.1032	0.0572	0.0574
	200	0.0822	0.1248	0.0680	0.0682
	300	0.0928	0.1338	0.0730	0.0730
	400	0.0964	0.1374	0.0741	0.0742
	500	0.1012	0.1412	0.0750	0.0752
200	100	0.1058	0.1706	0.1046	0.1056
	200	0.1424	0.2138	0.1384	0.1396
	300	0.1672	0.2488	0.1584	0.1592
	400	0.1876	0.2710	0.1802	0.1802
	500	0.2016	0.2820	0.1908	0.1908
500	500	0.3680	0.4754	0.3630	0.3680
	1000	0.4570	0.5662	0.4490	0.4526
1000	1000	0.6770	0.7686	0.6756	0.6797
	1500	0.7315	0.8120	0.7270	0.7298
1500	1500	0.8498	0.9078	0.8492	0.8496
	2000	0.8878	0.9284	0.8848	0.8864

Table A.4
Powers of hypothesis tests for two binary diagnostic tests.

$Se_1 = 0.95$, $Se_2 = 0.85$, $Sp_1 = 0.80$, $Sp_2 = 0.70$, $LR_1^+ = 4.75$, $LR_2^+ = 2.8333$, $LR_1^- = 0.0625$, $LR_2^- = 0.2143$, $p_1 = 10\%$, $p_2 = 50\%$

n_1	n_2	Global test	Marginal tests	Holm	Hochberg
100	100	0.1674	0.2434	0.1626	0.1626
	200	0.2024	0.3016	0.1984	0.1984
	300	0.2300	0.3402	0.2286	0.2286
	400	0.2380	0.3502	0.2316	0.2316
	500	0.2520	0.3712	0.2486	0.2486
200	100	0.2982	0.4286	0.3254	0.3260
	200	0.4250	0.5696	0.4616	0.4616
	300	0.5016	0.6574	0.5456	0.5456
	400	0.5426	0.7090	0.5890	0.5890
	500	0.5970	0.7422	0.6390	0.6390

Table A.4 (*continued*)

$Se_1 = 0.95, Se_2 = 0.85, Sp_1 = 0.80, Sp_2 = 0.70, \quad LR_1^+ = 4.75, LR_2^+ = 2.8333, LR_1^- = 0.0625, LR_2^- = 0.2143, \quad p_1 = 10\%, p_2 = 50\%$

n_1	n_2	Global test	Marginal tests	Holm	Hochberg
500	500	0.9052	0.9526	0.9050	0.9114
	1000	0.9740	0.9894	0.9782	0.9788
1000	1000	0.9980	0.9996	0.9980	0.9984
	1500	0.9998	1	1	1
1500	1500	1	1	1	1
	2000	1	1	1	1

Table A.5
Type I errors of hypothesis tests for three binary diagnostic tests.

$Se_1 = Se_2 = Se_3 = 0.90, Sp_1 = Sp_2 = Sp_3 = 0.80, \quad LR_1^+ = LR_2^+ = LR_3^+ = 4.5, LR_1^- = LR_2^- = LR_3^- = 0.125, \quad p_1 = 10\%, p_2 = 30\%, p_3 = 50\%$

n_1	n_2	n_3	Global test	Marginal tests	Holm	Hochberg
100	100	100	0.0310	0.1552	0.0158	0.0158
		200	0.0546	0.1750	0.0394	0.0398
		300	0.0598	0.1892	0.0414	0.0416
		400	0.0704	0.1862	0.0486	0.0488
		500	0.0756	0.1952	0.0540	0.0540
	500	100	0.0686	0.1960	0.0484	0.0486
		200	0.0672	0.2084	0.0474	0.0476
		300	0.0696	0.2002	0.0506	0.0506
		400	0.0758	0.2042	0.0576	0.0576
		500	0.0830	0.2212	0.0578	0.0582
200	100	100	0.0234	0.1282	0.0162	0.0162
		200	0.0352	0.1584	0.0276	0.0276
		300	0.0420	0.1718	0.0376	0.0376
		400	0.0414	0.1638	0.0362	0.0364
		500	0.0528	0.1800	0.0452	0.0454
	500	100	0.0484	0.1764	0.0392	0.0392
		200	0.0422	0.1856	0.0338	0.0340
		300	0.0484	0.1930	0.0398	0.0402
		400	0.0530	0.1940	0.0430	0.0432
		500	0.0476	0.1938	0.0402	0.0404
500	500	500	0.0434	0.1916	0.0350	0.0350
		1000	0.0480	0.1934	0.0398	0.0398
	1000	500	0.0500	0.1992	0.0400	0.0400
		1000	0.0558	0.2018	0.0466	0.0466
1000	1000	1000	0.0460	0.2112	0.0386	0.0388
		1500	0.0508	0.2052	0.0404	0.0404
	1500	1000	0.0460	0.2018	0.0358	0.0358
		1500	0.0488	0.2070	0.0394	0.0394
1500	1500	1500	0.0530	0.2205	0.0470	0.0475
		2000	0.0450	0.2015	0.0385	0.0385
	2000	1500	0.0450	0.2105	0.0365	0.0365
		2000	0.0470	0.1820	0.0420	0.0420

Table A.6
Powers of hypothesis tests for three binary diagnostic tests.

$Se_1 = 0.90, Se_2 = 0.85, Se_3 = 0.80, Sp_1 = 0.80, Sp_2 = 0.75, Sp_3 = 0.70, \quad LR_1^+ = 4.125, LR_2^+ = 3.4, LR_3^+ = 2.67, LR_1^- = 0.125, LR_2^- = 0.2, LR_3^- = 0.29,$
$p_1 = 10\%, p_2 = 30\%, p_3 = 50\%$

n_1	n_2	n_3	Global test	Marginal tests	Holm	Hochberg
100	100	100	0.1406	0.3672	0.1148	0.1148
		200	0.1676	0.4168	0.1432	0.1436

Table A.6 (*continued*)

$Se_1 = 0.90, Se_2 = 0.85, Se_3 = 0.80, Sp_1 = 0.80, Sp_2 = 0.75, Sp_3 = 0.70, \quad LR_1^+ = 4.125, LR_2^+ = 3.4, LR_3^+ = 2.67, LR_1^- = 0.125, LR_2^- = 0.2, LR_3^- = 0.29,$
$p_1 = 10\%, p_2 = 30\%, p_3 = 50\%$

n_1	n_2	n_3	Global test	Marginal tests	Holm	Hochberg
		300	0.1880	0.4514	0.1480	0.1482
		400	0.2056	0.4734	0.1606	0.1608
		500	0.2282	0.4996	0.1722	0.1726
	500	100	0.1962	0.4452	0.1808	0.1814
		200	0.2472	0.5344	0.2266	0.2270
		300	0.3022	0.5996	0.2624	0.2634
		400	0.3486	0.6522	0.2976	0.2994
		500	0.3678	0.6736	0.3154	0.3170
200	100	100	0.2638	0.5268	0.2570	0.2572
		200	0.3432	0.6352	0.3402	0.3404
		300	0.4122	0.7052	0.4156	0.4156
		400	0.4534	0.7318	0.4498	0.4500
		500	0.5042	0.7766	0.4930	0.4932
	500	100	0.3102	0.5956	0.2974	0.2982
		200	0.3938	0.6992	0.3838	0.3850
		300	0.4676	0.7594	0.4606	0.4618
		400	0.5248	0.8052	0.5102	0.5112
		500	0.5790	0.8362	0.5516	0.5528
500	500	500	0.8452	0.9638	0.8370	0.8378
		1000	0.9392	0.9880	0.9364	0.9370
	1000	500	0.8574	0.9652	0.8496	0.8498
		1000	0.9510	0.9912	0.9428	0.9428
1000	1000	1000	0.9942	0.9994	0.9918	0.9920
		1500	0.9980	1.0000	0.9980	0.9980
	1500	1000	0.9926	0.9998	0.9926	0.9928
		1500	0.9998	1.0000	0.9994	0.9994
1500	1500	1500	1	1	1	1
		2000	1	1	1	1
	2000	1500	1	1	1	1
		2000	1	1	1	1

References

Alfirevic, Z., Mujezinovic, F., Sundberg, K., 2003. Amniocentesis and chorionic villus sampling for prenatal diagnosis. Cochrane Database Syst. Rev. (3), CD003252.

Altman, D.G., Bland, J.M., 2003. Interaction revisited: the difference between two estimates. Brit. Med. J. 326, 219.

Biggerstaff, B.J., 2000. Comparing diagnostic tests: a simple graphic using likelihood ratios. Stat. Med. 19 (5), 649–663.

Bonferroni, C.E., 1936. Teoria statistica delle classi e calcolo delle probabilità. Pubbl. R Ist. Super. Sci. Economiche Commerciali Firenze 8, 3–62.

Hochberg, Y., 1988. A sharper Bonferroni procedure for multiple tests of significance. Biometrika 75 (4), 800–802.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scand. J. Stat. 6 (2), 65–70.

Leisenring, W., Pepe, M.S., 1998. Regression modelling of diagnostic likelihood ratios for the evaluation of medical diagnostic tests. Biometrics 54 (2), 444–452.

Macaskill, P., Walter, S.D., Irwig, L., Franco, E.L., 2002. Assessing the gain in diagnostic performance when combining two diagnostic tests. Stat. Med. 21 (17), 2527–2546.

Pepe, M.S., 2003. The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford University Press, New York.

Roldán Nofuentes, J.A., Luna del Castillo, J.D., 2007. Comparison of the likelihood ratios of two binary diagnostic tests in paired designs. Stat. Med. 26 (22), 4179–4201.

Sackett, D.L., Haynes, R.B., Tugwell, P., 1985. Clinical Epidemiology: A Basic Science for Clinical Medicine. Little & Brown, Boston.

Simel, D.L., Samsa, G.P., Matchar, D.B., 1991. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. J. Clin. Epidemiol. 44 (8), 763–770.

Weiner, D.A., Ryan, T.J., McCabe, C.H., Kennedy, J.W., Schloss, M., Tristani, F., Chaitman, B.R., Fisher, L.D., 1979. Exercise stress testing. Correlations among history of angina, ST-segment response and prevalence of coronary-artery disease in the coronary artery surgery study (CASS). N. Engl. J. Med. 301 (5), 230–235.