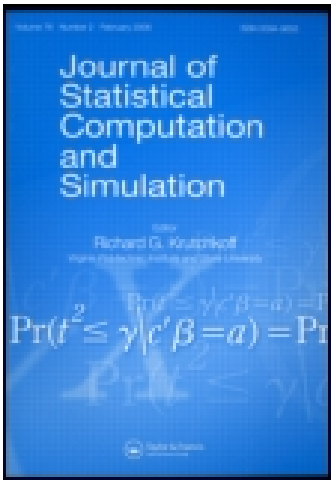

Journal of Statistical Computation and Simulation

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/gscs20

Comparing diagnostic tests: test of hypothesis for likelihood ratios

Nimet Anil Dolgun [a], Harika Gozukara [a] & Ergun Karaagaoglu [a]

[a] Department of Biostatistics, Hacettepe University, Ankara, Turkey

Published online: 31 May 2011.

To cite this article: Nimet Anil Dolgun, Harika Gozukara & Ergun Karaagaoglu (2012) Comparing diagnostic tests: test of hypothesis for likelihood ratios, Journal of Statistical Computation and Simulation, 82:3, 369-381, DOI: 10.1080/00949655.2010.531480

To link to this article: http://dx.doi.org/10.1080/00949655.2010.531480

Journal of Statistical Computation and Simulation
Vol. 82, No. 3, March 2012, 369–381

Taylor & Francis
Taylor & Francis Group

Comparing diagnostic tests: test of hypothesis for likelihood ratios

Nimet Anil Dolgun*, Harika Gozukara and Ergun Karaagaoglu

Department of Biostatistics, Hacettepe University, Ankara, Turkey

(*Received 23 May 2009; final version received 9 October 2010*)

Likelihood ratios (LRs) are used to characterize the efficiency of diagnostic tests. In this paper, we use the classical weighted least squares (CWLS) test procedure, which was originally used for testing the homogeneity of relative risks, for comparing the LRs of two or more binary diagnostic tests. We compare the performance of this method with the relative diagnostic likelihood ratio (rDLR) method and the diagnostic likelihood ratio regression (DLRReg) approach in terms of size and power, and we observe that the performances of CWLS and rDLR are the same when used to compare two diagnostic tests, while DLRReg method has higher type I error rates and powers. We also examine the performances of the CWLS and DLRReg methods for comparing three diagnostic tests in various sample size and prevalence combinations. On the basis of Monte Carlo simulations, we conclude that all of the tests are generally conservative and have low power, especially in settings of small sample size and low prevalence.

Keywords: binary diagnostic tests; comparing likelihood ratios; classical weighted least squares; relative diagnostic likelihood ratio; diagnostic likelihood ratio regression, type I error; statistical power; statistical simulation

AMS Subject Classification: 62P10; 62F03; 92B15; 92C50; 37M05

1. Introduction

In many clinical studies, researchers compare two or more diagnostic procedures or tests that are used to indicate the presence or absence of a particular disease, in a situation where the true disease status is known. A traditional comparison of the two competing binary diagnostic tests considers the comparison of sensitivities and specificities of each test, whether separately or in summary [1]. Several methods have been proposed for comparing the sensitivities, specificities, predictive values, the area under the receiver operator characteristic (ROC) curves, and Youden's indices of the tests, both in paired and unpaired study designs [2,3]. The use of positive and negative likelihood ratios (LRs) rather than sensitivity and specificity as measures of diagnostic ability has some advantages [4,5]. Biggerstaff [6] emphasized the usefulness of the LRs for assessing the tests' relative diagnostic abilities and recommended the use of positive and negative LRs over the use of sensitivity and specificity for binary test comparison. His paper also presented

*Corresponding author. Email: anilbarak@yahoo.com

ISSN 0094-9655 print/ISSN 1563-5163 online

http://dx.doi.org/10.1080/00949655.2010.531480
http://www.tandfonline.com

a simple graphical approach for comparing the LRs of two or more diagnostic tests. However, this approach was somewhat non-inferential and lacked some important aspects of significance testing. Simel *et al.* [7] presented confidence intervals for the LRs of binary tests and demonstrated a sample size estimation procedure for diagnostic test studies based on the desired LR confidence interval. Leisenring and Pepe [8] proposed the diagnostic likelihood ratio regression (DLRReg) method for comparing the LRs of binary diagnostic tests in paired/unpaired designs and for clustered data. Nofuentes and Castillo [9] recently proposed a method for comparing the LRs of two or more binary diagnostic tests in paired designs. However, the comparison of the LRs in unpaired data has not been widely studied in the statistical literature. Unpaired designs are common in clinical studies, especially when the patients are unlikely to have more than one diagnostic test for some reason (e.g. time limitations, ethical considerations, etc.). Pepe [2] derived empirical estimates of the relative diagnostic likelihood ratios (rDLRs) using asymptotic distribution theory, in the specific case when the data are unpaired and independent.

In this paper, we present the use of the classical weighted least squares (CWLS) test procedure as a tool for comparing LRs of two or more binary diagnostic tests. Throughout this study, we consider only binary diagnostic tests that are applied to independent groups of patients, each of which undergoes a different test (unpaired design). Also, we only consider methods which provide an omnibus comparison of the LRs when the number of diagnostic tests is more than two. Therefore, available multiple comparison methods are not included in the study. In Section 2, after giving some basic definitions of LRs and the previous works that are used to compare the LRs of the diagnostic tests, we describe how the CWLS test procedure can be applied to compare the LRs. In Section 4, we present the results of a simulation study to compare the size and power of the CWLS, rDLR, and DLRReg methods when they are used to compare two diagnostic tests. We also examine performances of the CWLS and DLRReg methods for comparing three diagnostic tests. In Section 5, we present a brief example to illustrate the test procedures for comparing LRs, and in Section 6 we discuss our conclusions.

2. General information

2.1. *LRs of binary tests*

The diagnostic ability of a binary diagnostic test is usually measured in terms of its sensitivity (Sen), which is the probability of a positive test result when the patient is diseased, and its specificity (Spe), which is the probability of a negative test result when the patient is non-diseased. Another way of describing the diagnostic ability of a test is the use of LRs. The positive LR (LR^+) is defined as the ratio of the probability of correctly classifying a diseased patient to the probability of incorrectly classifying a non-diseased patient, while the negative LR (LR^-) is defined as the ratio of the probability of incorrectly classifying a diseased patient to the probability of correctly classifying a non-diseased patient (Table 1). For a given diagnostic test, larger values of LR^+ and smaller values of LR^- indicate greater diagnostic ability.

The positive and negative LRs algebraically combine sensitivity and specificity, and they provide measures of test accuracy that are directly related to the predictive values of a test. Unlike predictive values, they are functionally independent of the disease prevalence in the population. These properties of LRs make them a more appropriate means of comparing binary diagnostic tests.

Biggerstaff [6] presented an ROC curve-based graph in which two or more LRs are compared according to the four regions defined. His method was somewhat non-inferential and lacked some important aspects of significance testing, since his comparison was based only on the magnitudes of LRs. He also recommended the computation of confidence intervals for the LRs and incorporation of these into the graphic to provide a more formal inference.

Table 1. 2×2 Contingency tables of disease status and test result for k competing binary diagnostic tests.[a]

First test	Disease status			kth test	Disease status	
Test result	D^+	D^-	...	Test result	D^+	D^-
T^+	a_1	b_1	...	T^+	a_k	b_k
T^-	c_1	d_1	...	T^-	c_k	d_k
Total	$n_1(D^+)$	$n_1(D^-)$	...	Total	$n_k(D^+)$	$n_k(D^-)$
$\widehat{\text{Sen}}_1 = a_1/n_1(D^+)$			...	$\widehat{\text{Sen}}_k = a_k/n_k(D^+)$		
$\widehat{\text{Spe}}_1 = d_1/n_1(D^-)$			...	$\widehat{\text{Spe}}_k = d_k/n_k(D^-)$		
$\widehat{\text{LR}}_1^+ = \widehat{\text{Sen}}_1/(1-\widehat{\text{Spe}}_1)$			...	$\widehat{\text{LR}}_1^+ = \widehat{\text{Sen}}_k/(1-\widehat{\text{Spe}}_k)$		
$\widehat{\text{LR}}_1^- = (1-\widehat{\text{Sen}}_1)/\widehat{\text{Spe}}_1$			...	$\widehat{\text{LR}}_k^- = (1-\widehat{\text{Sen}}_k)/\widehat{\text{Spe}}_k$		

[a] D^+, diseased population; D^-, non-diseased population; T^+, positive test result; T^-, negative test result.

2.2. *The rDLR*

Let us consider two binary diagnostic tests applied to two different groups of patients, where a_k is the number of true positives, d_k is the number of true negatives, b_k is the number of false positives, and c_k is the number of false negatives for $k = 1, 2$ tests. The sensitivity of the kth test is defined as $\widehat{\text{Sen}}_k = a_k/n_k(D^+)$ where $n_k(D^+) = a_k + c_k$ is the number of patients with disease. The specificity of the kth test is defined as $\widehat{\text{Spe}}_k = d_k/n_k(D^-)$ where $n_k(D^-) = b_k + d_k$ is the number of patients without disease. Also, the positive and negative LRs are $\widehat{\text{LR}}_k^+ = \widehat{\text{Sen}}_k/(1 - \widehat{\text{Spe}}_k)$ and $\widehat{\text{LR}}_k^- = (1 - \widehat{\text{Sen}}_k)/\widehat{\text{Spe}}_k$. Pepe [2] defined the 'rDLR' as the ratio of the LRs of two competing diagnostic tests. Taking the first test as a reference, the positive rDLR ($\widehat{\text{rDLR}}^+$) was defined as $\widehat{\text{rDLR}}^+ = \widehat{\text{LR}}_2^+/\widehat{\text{LR}}_1^+$, and the negative rDLR was analogously defined as $\widehat{\text{rDLR}}^- = \widehat{\text{LR}}_2^-/\widehat{\text{LR}}_1^-$. Pepe [2] showed that under the null hypothesis ($H_0 : \text{rDLR}^+ = 1$ for positive rDLR and $H_0 : \text{rDLR}^- = 1$ for negative rDLR) and in large samples, $\log(\widehat{\text{rDLR}}^+)$ and $\log(\widehat{\text{rDLR}}^-)$ are normally distributed with mean 0 and variances

$$\widehat{\text{var}}(\log(\widehat{\text{rDLR}}^+)) = \frac{1-\widehat{\text{Sen}}_1}{n_1(D^+)\widehat{\text{Sen}}_1} + \frac{\widehat{\text{Spe}}_1}{n_1(D^-)(1-\widehat{\text{Spe}}_1)} + \frac{1-\widehat{\text{Sen}}_2}{n_2(D^+)\widehat{\text{Sen}}_2} + \frac{\widehat{\text{Spe}}_2}{n_2(D^-)(1-\widehat{\text{Spe}}_2)}$$

and

$$\widehat{\text{var}}(\log(\widehat{\text{rDLR}}^-)) = \frac{\widehat{\text{Sen}}_1}{n_1(D^+)(1-\widehat{\text{Sen}}_1)} + \frac{1-\widehat{\text{Spe}}_1}{n_1(D^-)(\widehat{\text{Spe}}_1)} + \frac{\widehat{\text{Sen}}_2}{n_2(D^+)(1-\widehat{\text{Sen}}_2)} + \frac{1-\widehat{\text{Spe}}_2}{n_2(D^-)(\widehat{\text{Spe}}_2)},$$

respectively. Thus, under H_0, the test statistic $z = \log(\widehat{\text{rDLR}})/\sqrt{\widehat{\text{var}}(\log(\widehat{\text{rDLR}}))}$ follows a standard normal distribution. In this method, LRs of two diagnostic tests are compared using the z test statistic such as, if $\log(\widehat{\text{rDLR}}^+)$ is significantly different from '0' then $\widehat{\text{LR}}_2^+$ is statistically different from $\widehat{\text{LR}}_1^+$. However, in order to apply this approach to compare more than two diagnostic procedures, one requires an adjustment due to multiple testing. In such situations, the use of omnibus hypothesis testing should be considered.

2.3. *The DLRReg*

Leisenring and Pepe [8] proposed a regression method which allows for direct assessment of covariate effects on LRs for binary diagnostic tests. With this approach one can easily compare different diagnostic tests and determine how the diagnostic accuracy (in terms of LRs) varies with different patient or environmental characteristics. The DLRReg models for DLR^+ with p covariates and for DLR^- with q covariates are

$$\ln(DLR^+(X_1)) = \alpha_0 + \alpha_1 X_{11} + \cdots + \alpha_p X_{1p},$$
$$\ln(DLR^-(X_2)) = \beta_0 + \beta_1 X_{21} + \cdots + \beta_q X_{2q},$$

where αs and $X_1 = (X_{11}, \ldots, X_{1p})$ are parameters and matrix of covariates associated with DLR^+ and βs and $X_2 = (X_{21}, \ldots, X_{2q})$ are associated with DLR^-. In order to compare the LRs of two tests, the model can be rewritten as

$$\ln(DLR^+(X)) = \alpha_0 + \alpha_1 X,$$
$$\ln(DLR^-(X)) = \beta_0 + \beta_1 X,$$

where X is the dummy variable for comparing one test to another. The model also can be extended to compare more than two tests by simply adding extra dummy variables to the model. Also, the quantities $\exp(\alpha_1)$ and $\exp(\beta_1)$ give the estimates of $rDLR^+$ and $rDLR^-$, respectively.

The standard LR test statistic $-2(L_{\text{null}} - L_{\text{full}})$, where L_{full} is the log-likelihood for the full model (which includes all dummy variables for $k \geq 2$ tests comparison) and L_{null} is the log-likelihood of the null model (which includes only a constant term), can be used as an omnibus test of equality of the LRs. The resulting test statistic is compared with the corresponding critical chi-square value with degrees of freedom (df) $df = df_{\text{full}} - df_{\text{null}}$. Further details of estimation and applications can be found in Leisenring and Pepe [8].

It is also important to note that in DLRReg method, there is only one test statistic for negative and positive LR comparison. To be more specific, since the DLRReg model simultaneously models the effects on the positive and negative LRs, we get only one LR test statistic which corresponds to both positive and negative LR comparison. As a result of this, the omnibus DLRReg method does not clearly indicate whether the positive or negative LRs of the diagnostic tests differ from each other.

2.4. *The CWLS test procedure*

Let us consider k binary diagnostic tests applied to different groups of patients. Table 1 displays the frequencies of k test results when applied to k different diseased and non-diseased groups. We use the subscript i $(i = 1, \ldots, k)$ to denote the observed frequencies of the corresponding diagnostic test results.

When comparing the LRs of k binary diagnostic tests, we consider the null hypothesis $H_0 : \text{LR}_1 = \text{LR}_2 = \cdots = \text{LR}_k$, where LR_i is either LR_i^+ or LR_i^-. Since LR_i^+ and LR_i^- are the ratios of two independent binomial probabilities (sensitivity and specificity), they are algebraically identical to relative risk ratios [7]. Simel *et al.* [7] presented an approximate confidence interval formula for positive and negative LRs using a Taylor series approximation and the logarithms of the LRs. Using these results, the CWLS test procedure [10,11], which was originally used for testing the homogeneity of relative risks, can be easily applied for testing the null hypothesis $H_0 : \text{LR}_1 = \text{LR}_2 = \cdots = \text{LR}_k$.

The CWLS test statistic uses a weighted average of the logarithms of the $\widehat{\text{LR}}$s, say $\hat{L}_i$ ($\hat{L}_i^+ = \log(\widehat{\text{LR}}_i^+)$ for the log-positive LR and $\hat{L}_i^- = \log(\widehat{\text{LR}}_i^-)$ for the log-negative LR), where the

weights (W_i) are the reciprocals of the variances of $\hat{L}_i$ ($W_i^+ = 1/\widehat{\text{var}}(\hat{L}_i^+)$ and $W_i^- = 1/\widehat{\text{var}}(\hat{L}_i^-)$). By applying the delta method [12], it is easily seen that the estimated asymptotic variances of $\hat{L}_i^+$ and $\hat{L}_i^-$ are $\widehat{\text{var}}(\hat{L}_i^+) = (1 - \widehat{\text{Sen}}_i)/a_i + (\widehat{\text{Spe}}_i)/b_i$ and $\widehat{\text{var}}(\hat{L}_i^-) = (\widehat{\text{Sen}}_i)/c_i + (1 - \widehat{\text{Spe}}_i)/d_i$, respectively.

Under $H_0 : \text{LR}_1 = \text{LR}_2 = \cdots = \text{LR}_k$, $\hat{L}_i$s are normally distributed with mean 0 and the variances given above. Therefore, the quantity $\chi^2_{\text{Total}} = \sum_{i=1}^k W_i \hat{L}_i^2$ gives the total variance (in terms of chi-square) in independent $k > 2$ tables. χ^2_{Total} is partitioned into two components, namely between tables variance (which is $\chi^2_{\text{Homogeneity}}$ and is equal to CWLS estimator in our context) and within tables variance (which is $\chi^2_{\text{Association}}$), where $\chi^2_{\text{Total}} = \sum_{i=1}^k W_i \hat{L}_i^2$ and $\chi^2_{\text{Association}} = (\sum_{i=1}^k W_i \hat{L}_i)^2 / \sum_{i=1}^k W_i$. Thus, the CWLS test is equal to the $\chi^2_{\text{Homogeneity}} = \chi^2_{\text{Total}} - \chi^2_{\text{Association}}$ which is

$$\text{CWLS} = \sum_{i=1}^{k} W_i \hat{L}_i^2 - \frac{(\sum_{i=1}^k W_i \hat{L}_i)^2}{\sum_{i=1}^k W_i} \tag{1}$$

and is tested against the critical chi-square value with degrees of freedom $df_{\text{Total}} - df_{\text{Association}} = k - 1$. Under the null hypothesis, the test statistic (1) asymptotically follows the χ^2 distribution

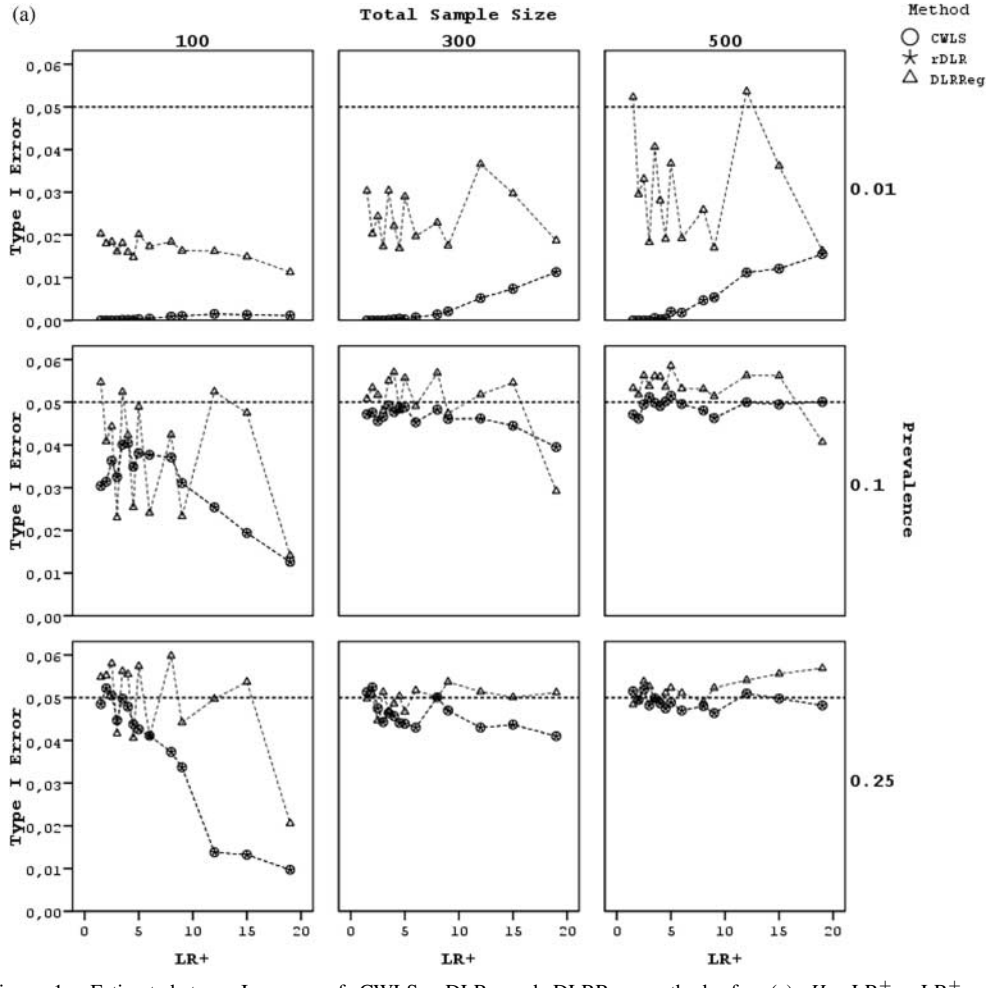

Figure 1. Estimated type I errors of CWLS, rDLR, and DLRReg methods for (a) $H_0 : \text{LR}_1^+ = \text{LR}_2^+$ and (b) $H_0 : \text{LR}_1^- = \text{LR}_2^-$.

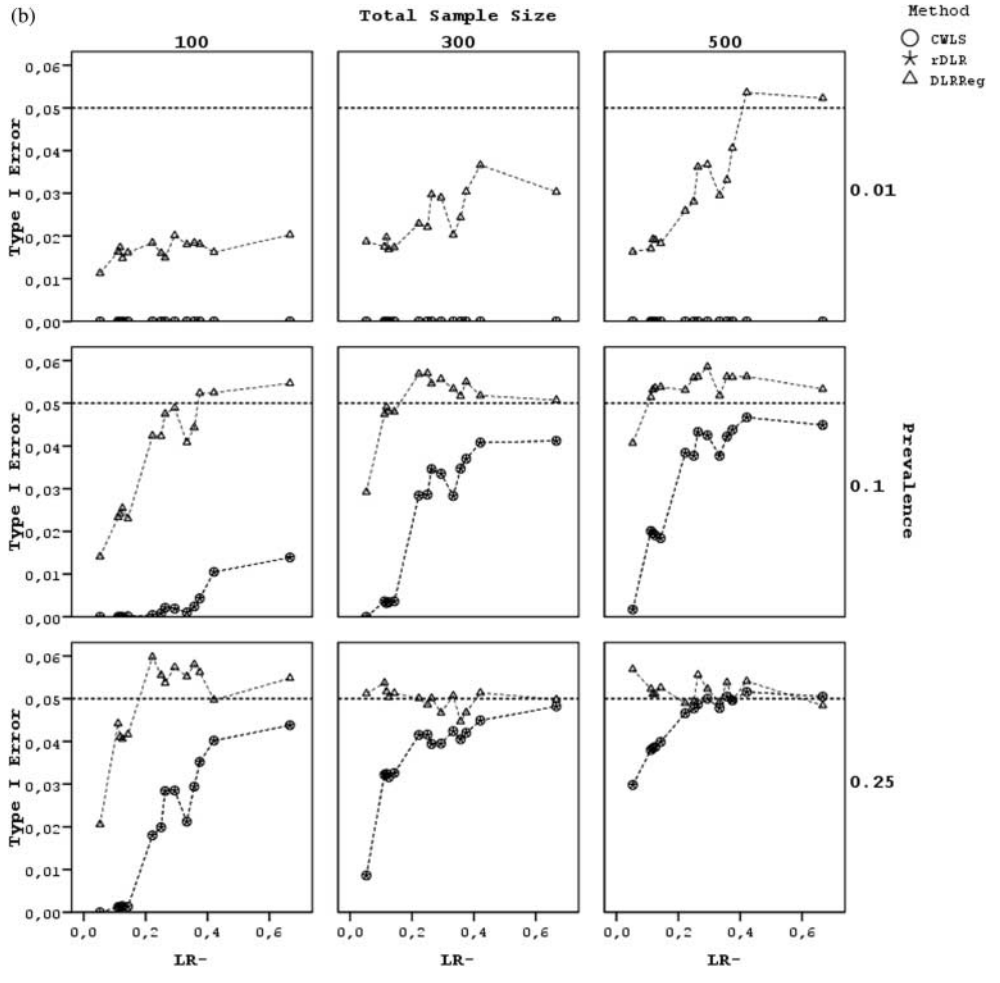

Figure 1. Continued.

with $k-1$ degrees of freedom [10,11] when all the observed frequencies in Table 1 are sufficiently large. Thus, we reject H_0 at the significance level α if CWLS $> \chi^2_{\alpha,k-1}$, where $\chi^2_{\alpha,k-1}$ is the upper $100(\alpha)$th percentile of the chi-square distribution with $k-1$ degrees of freedom. Also, note that the χ^2 test statistic of the CWLS is equal to z^2 test statistic of the rDLR when $k=2$.

3. Description of the simulation study

In order to evaluate and compare the size (type I error) and power of the CWLS, rDLR and DLRReg methods, we performed a Monte Carlo simulation. Three factors were considered that could affect the size and power of the CWLS, rDLR and DLRReg methods, namely, the total sample size n_i^T (where $n_i^T = n_i(D^+) + n_i(D^-)$), the estimated disease prevalence pr_i (where $\text{pr}_i = n_i(D^+)/n_i^T$), and the LRs, LR_i, under the null or the alternative hypothesis. We considered situations in which the estimated disease prevalence is 0.01, 0.10, or 0.25 and the total sample size is 100, 300, or 500. Also, since the values of $\text{LR}_i^+ < 1$ and $\text{LR}_i^- > 1$ suggest that the diagnostic tests under consideration are swapping the meanings of 'positive' with 'negative', we specified

the values of sensitivity and specificity in the range [0.60–0.95], so that the resulting LR_i have practical meaning. For each combination of n_i^T, pr_i and LR_i, $i = 1, \ldots, k$, we generated 10,000 independent sets of cell counts for k 2×2 tables. Upon obtaining the LR_i from Sen_i and Spe_i, we generated cell counts a_i and d_i from independent binomial distributions with parameters $(Sen_i, n_i(D^+))$ and $(Spe_i, n_i(D^-))$, respectively. Note that, when one of the cell counts in the 2×2 tables equals 0, the resulting estimate of LR_i is either 0 or infinity. In such cases, we applied the commonly used *ad hoc* adjustment procedure of adding 0.50 to each cell frequency in that particular table when estimating LR_i.

We obtained the estimated size and power for a given test as the percentage of times the test rejects the null hypothesis when the LR_i are held constant or varied, respectively. We simulated a total of 126 configurations for the size and 261 configurations for the power of the test statistics, when the number of diagnostic tests is two ($k = 2$). Furthermore, we examined the performance of the CWLS and DLRReg approaches when $k = 3$ with a total of 126 and 315 configurations for the size and power, respectively. The simulation program is developed in R version 2.9.1 software by the authors.

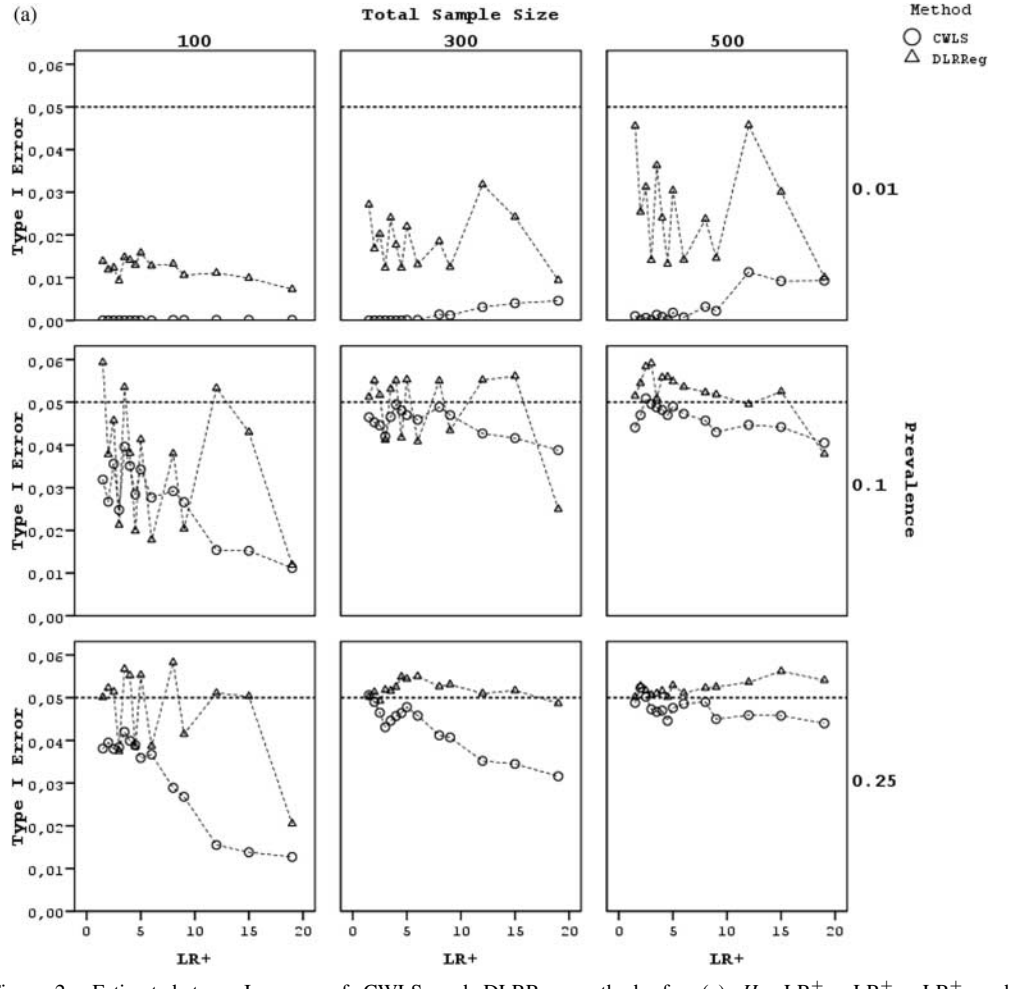

Figure 2. Estimated type I errors of CWLS and DLRReg methods for (a) $H_0 : LR_1^+ = LR_2^+ = LR_3^+$ and $H_0 : LR_1^- = LR_2^- = LR_3^-$.

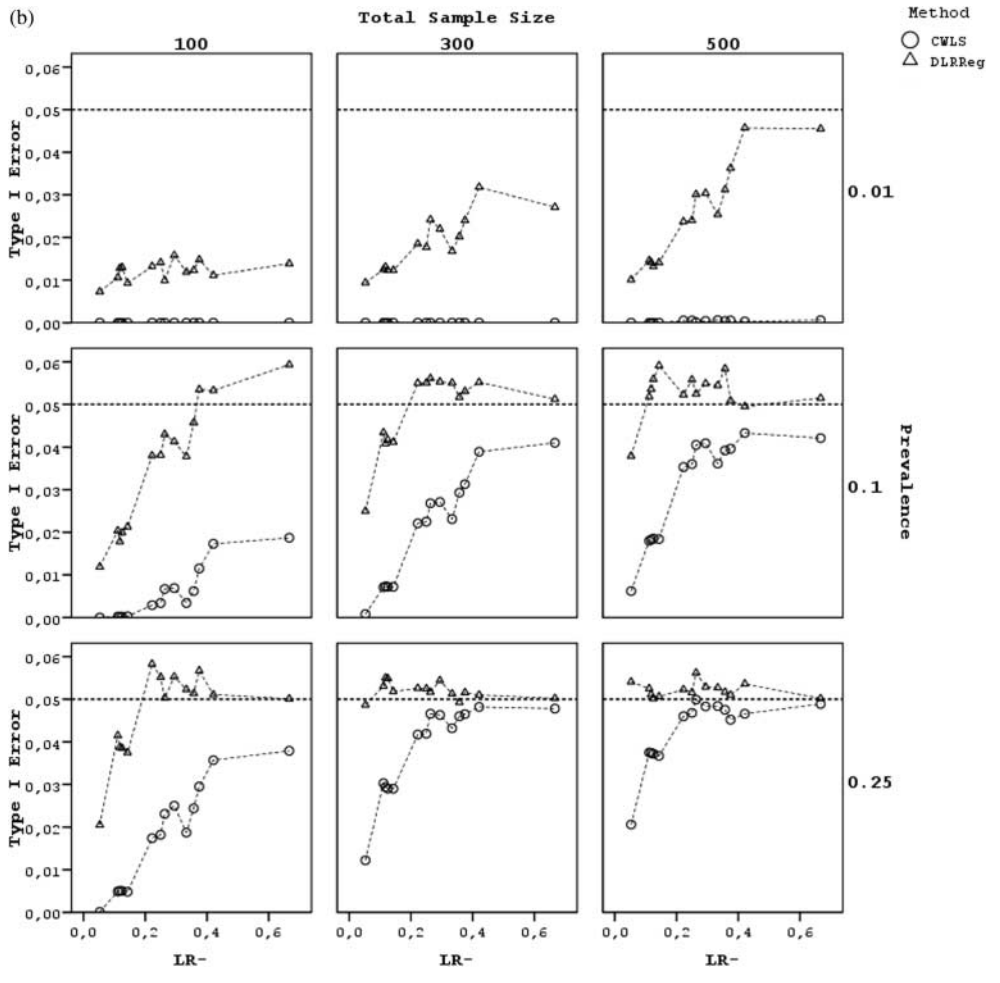

Figure 2. Continued.

4. Results

4.1. *The estimated type I error*

Figure 1(a) and (b) summarizes the estimated type I errors of the CWLS, rDLR, and DLRReg methods at the significance level 0.05 for the given combinations of total sample size, disease prevalence, and LRs (positive and negative LR, respectively) when the number of diagnostic tests is two ($k = 2$). According to Figure 1, when testing $H_0 : \mathrm{LR}_1^+ = \mathrm{LR}_2^+$ and $H_0 : \mathrm{LR}_1^- = \mathrm{LR}_2^-$, we observe that the CWLS and rDLR methods give exactly the same type I error results for all simulation settings. The DLRReg method, on the other hand, has higher type I error rates than CWLS and rDLR methods in most situations. All of the test procedures are conservative, in settings of low prevalence ($\mathrm{pr}_i = 0.01$) and small sample size ($n_i^T = 100$), and they reach desired 0.05 level only in settings of higher prevalence and larger sample size.

Figure 2(a) and (b) summarizes the estimated type I errors of the CWLS and DLRReg procedures at the significance level 0.05 for the given combinations of total sample size, disease prevalence, and LRs (positive and negative LR, respectively) when the number of diagnostic tests is three ($k = 3$). The results that can be derived from Figure 2 are similar to those of Figure 1.

Table 2. The estimated power of CWLS, rDLR, and DLRReg methods, for testing hypotheses $H_0 : \mathrm{LR}_1^+ = \mathrm{LR}_2^+$ and $H_0 : \mathrm{LR}_1^- = \mathrm{LR}_2^-$ at 0.05 significance level, for given combinations for LR_i, pr_i, and n_i^T (abbreviated version).

						$H_0 : \mathrm{LR}_1^+ = \mathrm{LR}_2^+$		$H_0 : \mathrm{LR}_1^- = \mathrm{LR}_2^-$		Omnibus
LR_1^+	LR_2^+	LR_1^-	LR_2^-	pr_i	n_i^T	CWLS	rDLR	CWLS	rDLR	DLRReg
1.5	2	0.667	0.333	0.01	100	0.0000	0.0000	0.0000	0.0000	0.0195
					300	0.0000	0.0000	0.0000	0.0000	0.0332
					500	0.0000	0.0000	0.0000	0.0000	0.0683
				0.1	100	0.0857	0.0856	0.0279	0.0279	0.1297
					300	0.2832	0.2832	0.3389	0.3389	0.3189
					500	0.4426	0.4426	0.5597	0.5597	0.4946
				0.25	100	0.1701	0.1701	0.2556	0.2555	0.2692
					300	0.4395	0.4395	0.7221	0.7221	0.6818
					500	0.6484	0.6484	0.9108	0.9108	0.8858
2	3	0.333	0.143	0.01	100	0.0000	0.0000	0.0000	0.0000	0.1887
					300	0.0033	0.0033	0.0000	0.0000	0.5813
					500	0.0988	0.0988	0.0000	0.0000	0.8305
				0.1	100	0.2616	0.2616	0.0033	0.0033	0.2255
					300	0.7183	0.7183	0.1594	0.1594	0.6669
					500	0.9037	0.9037	0.3544	0.3544	0.8844
				0.25	100	0.3511	0.3511	0.0995	0.0992	0.2876
					300	0.7910	0.7910	0.5334	0.5334	0.7181
					500	0.9488	0.9488	0.7649	0.7649	0.9160
2	4	0.571	0.25	0.01	100	0.0010	0.0010	0.0000	0.0000	0.2406
					300	0.0251	0.0251	0.0000	0.0000	0.7157
					500	0.2568	0.2567	0.0000	0.0000	0.9197
				0.1	100	0.3895	0.3895	0.0622	0.0622	0.3495
					300	0.8605	0.8605	0.4843	0.4843	0.8224
					500	0.9795	0.9795	0.7305	0.7305	0.9693
				0.25	100	0.5129	0.5129	0.3940	0.3940	0.4665
					300	0.9449	0.9449	0.8818	0.8818	0.9207
					500	0.9948	0.9948	0.9819	0.9819	0.9915
1.5	4.5	0.667	0.125	0.01	100	0.0082	0.0082	0.0000	0.0000	0.7718
					300	0.4496	0.4496	0.0000	0.0000	0.9986
					500	0.8617	0.8617	0.0072	0.0072	1.0000
				0.1	100	0.8822	0.8822	0.1774	0.1774	0.8588
					300	0.9998	0.9998	0.9097	0.9097	0.9999
					500	1.0000	1.0000	0.9925	0.9925	0.9995
				0.25	100	0.9489	0.9489	0.8226	0.8226	0.9237
					300	0.9999	0.9999	1.0000	1.0000	0.9997
					500	1.0000	1.0000	1.0000	1.0000	1.0000
2	6	0.333	0.444	0.01	100	0.3382	0.3382	0.0000	0.0000	0.9977
					300	0.5886	0.5886	0.0000	0.0000	0.9999
					500	0.6670	0.6670	0.0011	0.0011	1.0000
				0.1	100	0.6847	0.6847	0.0034	0.0034	0.9959
					300	0.9815	0.9815	0.0572	0.0572	1.0000
					500	0.9995	0.9995	0.1053	0.1052	1.0000
				0.25	100	0.7815	0.7815	0.0445	0.0445	0.9939
					300	0.9977	0.9977	0.1525	0.1525	1.0000
					500	1.0000	1.0000	0.2529	0.2529	1.0000

Table 3. The estimated power of CWLS and DLRReg methods for testing hypotheses $H_0 : \mathrm{LR}_1^+ = \mathrm{LR}_2^+ = \mathrm{LR}_3^+$ and $H_0 : \mathrm{LR}_1^- = \mathrm{LR}_2^- = \mathrm{LR}_3^-$ at 0.05 significance level for given combinations of LR_i, pr_i, and n_i^T (abbreviated version).

								Power of CWLS		Power of DLRReg
LR_1^+	LR_2^+	LR_3^+	LR_1^-	LR_2^-	LR_3^-	pr_i	n_i^T	$H_0 : \mathrm{LR}_1^+ = \mathrm{LR}_2^+ = \mathrm{LR}_3^+$	$H_0 : \mathrm{LR}_1^- = \mathrm{LR}_2^- = \mathrm{LR}_3^-$	Omnibus
1.5	2	2.5	0.667	0.333	0.357	0.01	100	0.0000	0.0000	0.1548
							300	0.0012	0.0000	0.6196
							500	0.0711	0.0054	0.8778
						0.1	100	0.1660	0.0583	0.2910
							300	0.5439	0.3466	0.7668
							500	0.7916	0.5838	0.9474
						0.25	100	0.2926	0.2621	0.3844
							300	0.7618	0.7614	0.8833
							500	0.9410	0.9397	0.9862
2	4	6	0.333	0.25	0.118	0.01	100	0.0169	0.0000	0.9207
							300	0.6550	0.0002	0.9995
							500	0.9416	0.0060	0.9998
						0.1	100	0.8956	0.0161	0.9193
							300	1.0000	0.1596	1.0000
							500	1.0000	0.3953	1.0000
						0.25	100	0.9127	0.1089	0.8942
							300	1.0000	0.6164	1.0000
							500	1.0000	0.8654	1.0000
2	3	3	0.333	0.143	0.143	0.01	100	0.0000	0.0000	0.1476
							300	0.0002	0.0000	0.5648
							500	0.0181	0.0026	0.8547
						0.1	100	0.2243	0.0089	0.2017
							300	0.7519	0.1724	0.7048
							500	0.9376	0.3656	0.9163
						0.25	100	0.3388	0.1156	0.2805
							300	0.8288	0.5620	0.7625
							500	0.9647	0.7900	0.9365
6	6	8	0.118	0.118	0.222	0.01	100	0.0008	0.0000	0.0561
							300	0.0294	0.0000	0.2679
							500	0.0971	0.0002	0.5095
						0.1	100	0.0700	0.0026	0.1151
							300	0.1696	0.0722	0.4498
							500	0.2760	0.1854	0.6964
						0.25	100	0.0601	0.0490	0.2053
							300	0.1629	0.3285	0.5755
							500	0.2532	0.5506	0.8199
3	6	6	0.143	0.118	0.118	0.01	100	0.0046	0.0000	0.5462
							300	0.3774	0.0000	0.9877
							500	0.8555	0.0002	0.9999
						0.1	100	0.5923	0.0003	0.5494
							300	0.9824	0.0117	0.9865
							500	0.9996	0.0271	1.0000
						0.25	100	0.5857	0.0095	0.5192
							300	0.9792	0.0464	0.9662
							500	0.9994	0.0602	0.9988

When testing $H_0 : \mathrm{LR}_1^+ = \mathrm{LR}_2^+ = \mathrm{LR}_3^+$ and $H_0 : \mathrm{LR}_1^- = \mathrm{LR}_2^- = \mathrm{LR}_3^-$, the CWLS and DLRReg test procedures are generally conservative, especially in settings of low prevalence ($\mathrm{pr}_i = 0.01$) and small sample size ($n_i^T = 100$), whereas the DLRReg method has higher type I error rates in most settings. Both tests reach the desired 0.05 level in settings where the prevalence is high and the sample size is large.

4.2. *The estimated power*

Table 2 summarizes a part of the results obtained for the power of the CWLS, rDLR, and DLRReg methods when testing hypotheses $H_0 : \mathrm{LR}_1^+ = \mathrm{LR}_2^+$ and $H_0 : \mathrm{LR}_1^- = \mathrm{LR}_2^-$ at the significance level 0.05 for given combinations of LR_i, pr_i, and n_i^T. We observe that the power of CWLS and rDLR methods are the same for all simulation settings when $k = 2$. On the other hand, the DLRReg method is more powerful in most of the situations. Moreover, all test procedures have low power, especially in settings of small sample size ($n_i^T = 100$) and low prevalence ($\mathrm{pr}_i = 0.01$), and the powers increase as the total sample size and prevalence increases.

Table 3 summarizes a part of the results obtained for the power of the CWLS and DLRReg methods when testing hypotheses $H_0 : \mathrm{LR}_1^+ = \mathrm{LR}_2^+ = \mathrm{LR}_3^+$ and $H_0 : \mathrm{LR}_1^- = \mathrm{LR}_2^- = \mathrm{LR}_3^-$ at the significance level 0.05 for given combinations of LR_i, pr_i, and n_i^T. The results that can be derived from Table 3 are similar to those of Table 2. The CWLS and DLRReg methods have low power, especially in settings of small sample size ($n_i^T = 100$) and low prevalence ($\mathrm{pr}_i = 0.01$), and the powers increase as the total sample size and prevalence increases. Again the DLRReg method has higher power values in most of the situations.

5. An example

We present a brief example to illustrate the CWLS, rDLR, and DLRReg test procedures for comparing LRs. Two different studies of enzyme-linked immunosorbent assay (ELISA) in the diagnosis of Lyme disease are held, of which the sensitivity and specificity are reported as 40% (23/57), 94% (130/139) [13] for the first study, and 78% (45/58), 89% (101/113) [14] for the second study, respectively. The LRs for the first study are $\widehat{\mathrm{LR}}_1^+ = 6.23$ and $\widehat{\mathrm{LR}}_1^- = 0.64$, and, for the second study, they are $\widehat{\mathrm{LR}}_2^+ = 7.31$ and $\widehat{\mathrm{LR}}_2^- = 0.25$. The CWLS test statistics for $H_0 : \mathrm{LR}_1^+ = \mathrm{LR}_2^+$ and $H_0 : \mathrm{LR}_1^- = \mathrm{LR}_2^-$ are 0.121 and 11.92, respectively. Also, according to the rDLR method, $\log(\widehat{\mathrm{rDLR}^+}) = -0.16$ and $\log(\widehat{\mathrm{rDLR}^-}) = 0.93$ with the estimated variances $\widehat{\mathrm{var}}(\log(\widehat{\mathrm{rDLR}^+})) = 0.21$ and $\widehat{\mathrm{var}}(\log(\widehat{\mathrm{rDLR}^-})) = 0.07$, yielding the z test statistics as -0.35 and 3.45. Thus, for both tests, we reject $H_0 : \mathrm{LR}_1^- = \mathrm{LR}_2^-$ at the level of $\alpha = 0.05$, but we cannot reject $H_0 : \mathrm{LR}_1^+ = \mathrm{LR}_2^+$. In order to get the estimates in the DLRReg method, we used a STATA do-file named lrreg.ado which is written by Leisenring and Longton [15]. According to the DLRReg model, the estimates are $\widehat{\alpha}_0 = 1.99$, $\widehat{\alpha}_1 = -0.16$, $\widehat{\beta}_0 = -1.38$, and $\widehat{\beta}_1 = 0.93$, where the log-likelihood of the null model is $L_{\mathrm{null}} = -150.066$ and the log-likelihood for the full model is $L_{\mathrm{full}} = -140.889$. The LR test statistic yields $-2(L_{\mathrm{null}} - L_{\mathrm{full}}) = 18.35$. When the resulting test statistic is compared with the corresponding critical chi-square value with $df = 2$, we conclude that at least one coefficient ($\widehat{\alpha}_1$ or $\widehat{\beta}_1$) is statistically significant.

6. Discussion

In this study, by using the fact that the LRs are identical to relative risks, we presented the use of the CWLS test procedure, which was originally used for testing the homogeneity of relative

risks, for comparing the LRs of two or more binary diagnostic tests. We examined the size and power of the CWLS, rDLR, and DLRReg methods and observed that, CWLS and rDLR methods' performances were the same in all settings of the simulation study when $k = 2$. All of the tests were generally conservative and had low power especially in settings of small sample size and low prevalence. Also, the DLRReg method had higher type I errors and higher powers in most of the situations.

We also examined the performances of the CWLS and DLRReg methods for comparing three diagnostic tests in various sample size and prevalence combinations, and the results were similar to those for $k = 2$. They were generally conservative and had low power especially in settings of small sample size and low prevalence. This was not surprising for the CWLS test procedure because it does not perform well in small sample size and sparse data settings [16,17] since it uses the asymptotic distribution theory. The DLRReg method had higher type I errors and higher powers than the CWLS method in most of the situations.

The simulations were based on a cohort design, and it is clear that no test procedure provided satisfactory power, especially in settings of small sample size and low prevalence. In such situations, one would typically follow a case-control design. In a case control study, note that $n_i(D^+)$ and $n_i(D^-)$ are fixed by design and the relative frequency of diseased patients is usually much higher than in the population from which the cases and controls are drawn, at least if the disease prevalence is relatively low. The inferences for the LRs are exactly the same in a case control study, and the performance of the CWLS test procedure depends only on the magnitude of $n_i(D^+)$ and $n_i(D^-)$.

LRs tell us two different characteristics of the tests' performance, namely the ability of the tests to rule in and rule out a disease. For that reason, the LR^+ and LR^- are tested separately with the CWLS test procedure throughout the paper. However, in some cases, one can think of testing the LR^+ and LR^- overall instead of testing them separately. In such cases, the DLRReg method can be used to test LR^+ and LR^- in an omnibus manner and this method will be advantageous than the CWLS and rDLR methods (which test the hypotheses separately and allow the type I error rate to inflate), since it controls the type I error rate at the desired level, like other omnibus hypothesis testing procedures such as ANOVA. Also, one should be aware that the large sample covariance of $\log(\widehat{LR}^+)$ and $\log(\widehat{LR}^-)$ depends on sample sizes; it is simply $-(1/n_i(D^+) + 1/n_i(D^-))$ [2]. Another way of overall testing of LRs could be by using the Odds Ratio ($OR_i = LR_i^+/LR_i^-$), which algebraically combines the positive and negative LRs [18]. The CWLS test procedure is also valid for comparing the odds ratios of two or more binary diagnostic tests.

Another use of the CWLS test procedure is in meta-analytic studies to compare the performance of a diagnostic test that is applied on different populations, as given in the example, or subgroups of patients, for evaluating the spectrum effect [19]. The CWLS test procedure also has some advantages that make the procedure very flexible. First, it allows us to compare more than two diagnostic tests. Second, if the test statistic is significant, in the next step, one could partition the chi-square into appropriate components in order to identify which diagnostic tests have significantly different LRs.

In summary, we conclude that the CWLS test procedure can be used as a generalization of the rDLR method when $k > 2$ and that it merits consideration as a method for comparing more than two LRs of diagnostic tests in settings of large sample size and high prevalence. Also, the DLRReg method is a very useful tool to determine how the diagnostic accuracy (in terms of LRs) varies with different patient or environmental characteristics and to compare different diagnostic tests. Moreover, when one wants to test both positive and negative LRs simultaneously, this method is more advantageous than the CWLS and rDLR methods, since it controls the type I error rate at the desired level.

Acknowledgements

We would like to thank the referee whose suggestions and questions improved the clarity and quality of the paper. We are also grateful to Mr. A. Kerem Uludag for his help in programming the Monte Carlo simulation.

References

[1] M.H. Zweig and G. Campbell, *Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine*, Clin. Chem. 39 (1993), pp. 561–577.
[2] M.S. Pepe, *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford University Press, New York, 2003.
[3] X.H. Zhou, N.A. Obuchowski, and D.K. McClish, *Statistical Methods in Diagnostic Medicine*, Wiley, New York, 2002.
[4] R.H. Fletcher, S.W. Fletcher, and E.H. Wagner, *Clinical Epidemiology. The Essentials*, Williams & Wilkins, Baltimore, 1996.
[5] R. Jaeschke, G.H. Guyatt, and D.L.Sackett, *How to use an article about a diagnostic test*, J. Am. Med. Assoc. 271 (1994), pp. 703–707.
[6] B.J. Biggerstaff, *Comparing diagnostic tests: A simple graphic using likelihood ratios*, Stat. Med. 19 (2000), pp. 649–663.
[7] D.L. Simel, G.P. Samsa, and D.B. Matchar, *Likelihood ratios with confidence: Sample size estimation for diagnostic test studies*, J. Clin. Epidemiol. 44 (1991), pp. 763–770.
[8] W. Leisenring and M.S. Pepe, *Regression modelling of diagnostic likelihood ratios for the evaluation of medical diagnostic tests*, Biometrics 54 (1998), pp. 444–452.
[9] J.A.R. Nofuentes and J.D.L. Castillo, *Comparison of the likelihood ratios of two binary diagnostic tests in paired designs*, Stat. Med. 26 (2007), pp. 4179–4201.
[10] J.L. Fleiss, *Statistical Methods for Rates and Proportions*, Wiley, New York, 1981.
[11] J.M. Lachin, *Biostatistical Methods: The Assessment of Relative Risks*, Wiley, New York, 2000.
[12] A. Agresti, *Categorical Data Analysis*, Wiley, New York, 1990.
[13] F. Dressler, J.A. Whalen, B.N. Reinhardt, and A.C. Steere, *Western blotting in the serodiagnosis of Lyme disease*, J. Infect. Dis. 167 (1993), pp. 392–400.
[14] B.J.B. Johnson, K.E. Robbins, R.E. Bailey, B.L. Cao, S.L. Sviat, R.B. Craven, L.W. Mayer, and D.T. Dennis, *Serodiagnosis of Lyme disease: Accuracy of a two-step approach using a flagella-based ELISA and immunoblotting*, J. Infect. Dis. 174 (1996), pp. 346–353.
[15] W. Leisenring and G. Langton, Downloadable stata programs and help files for diagnostic likelihood ratio regression (lrreg.ado, lrreg_ll.ado, lrreg.hlp) in the webpage for *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford University Press, New York, 2003. Available at http://labs.fhcrc.org/pepe/book/index.html.
[16] K.-J. Lui and C. Kelly, *Tests for homogeneity of the risk ratio in a series of* 2×2 *tables*, Stat. Med. 19 (2000), pp. 2919–2932.
[17] K.-J. Lui, *A Monte Carlo evaluation of five interval estimators for the relative risk in sparse data*, Biom. J. 48 (2006), pp. 131–143.
[18] A.S. Glas, J.G. Lijmer, M.H. Prins, G.J. Bonsel, and P.M. Bossuyt, *The diagnostic odds ratio: A single indicator of test performance*, J. Clin. Epidemiol. 56 (2003), pp. 1129–1135.
[19] S.A. Mulherin and W.C. Miller, *Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation*, Ann. Intern. Med. 137 (2002), pp. 598–602.