Pak. J. Statist.
2017 Vol. 33(4), 301-314

THE PERFORMANCES OF TWO DIAGNOSTICS TESTS: MCNEMAR AND NEWCOMBE GRAPHICAL APPROACH

Mustafa Agah Tekindal[1], Can Ateş[2], Özlem Güllü Kaymaz[3] and **Yasemin Yavuz[4]**

[1] (PhD) (Biostatistics), Izmir, Turkey. Email: matekindal@gmail.com
[2] Faculty of Medicine Department of Biostatistics, Van Yüzüncü Yıl University Van, Turkey. Email: can.ates@gmail.com
[3] Faculty of Sciences, Department of Statistics, Ankara University Ankara, Turkey. Email: ozlem.gullu@gmail.com
[4] Faculty of Medicine, Department of Biostatistics, Ankara University Ankara, Turkey. Email: genc@medicine.ankara.edu.tr

ABSTRACT

This study aims to address a simultaneous comparison of sensitivity and specificity of two different tests by taking into consideration of prevalence and clinical significance level of the tests under the condition that the result of the gold standard test is known. The proposed method involves the point and interval estimation of the weighted mean (for *f*) concerning the specificity and sensitivity differences of two tests. Getting a value between 0 and 1, the parameter λ represents the clinical cost of false positive and false negative and the occasion posing serious problems for the prevalence value belonging to the population to which tests are administered. Data is acquired from a study carried out at the Department of Cytopathology of Ankara University Medical School. In that study, the impact of the use of various antibodies in determining thyroid nodules as benign or malign via Preoperative Fine-Needle Aspiration Biopsy in order to achieve diagnostic accuracy was investigated. Considering the prevalence of Newcombe graphic approach that compares both the sensitivity and specificity, it can be concluded that Newcombe graphical approach gives better results than the classical McNemar test for the same sample group used in the study Simultaneously comparing the sensitivity and specificity of tests by taking into consideration the clinical costs of false positive/false negative values and prevalence, the Newcombe graphic approach method is also quite advantageous in terms of clinical interpretation.

KEYWORDS

Sensitivity and specificity, simultaneous designs, comparison of paired ratios, Newcombe graphical approach.

1. INTRODUCTION

A frequent goal of medical research is to compare the effectiveness of different diagnostic tests. The assessment is based on a patient sample classified as "diseased" or "disease free" using a "gold standard" rule. Each patient classified by diagnostic tests for the disease, "positive" or "negative" as it refers [1].

Several methods have been proposed to compare the performance of two diagnostic tests for the same data. Greenhouse and Mantel [2] compared sensitivities within a certain common specificity, and Linnett [3] examined the need for strength and sample size. Several methods are based on comparing the areas under the ROC (Receiver Operating Characteristic) curve generated by modifying the positive definition. Hanley and McNeil [4] compared all areas under ROC curves. Delong, Delong and Clarke-Pearson [5] also use the U-statistic theory to bring a more non-parametric solution to the Hanley and McNeil procedures. Wieand et al. [6] identify a test procedure family that includes the above approaches as special cases, comparing sensitivities in the common property interval. Campbell [7] compares two ROC curves using a supermodel and uses the boot loader to test the hypothesis that two theoretical ROC curves are equal. Recent studies have been included below.

Bloch [1] presented methods for comparing two diagnostic tests for the same positive threshold value for the same sample. The first method is to compare the predicted risks of diagnostic tests and the other method is to compare kappa coefficients. These methods are explained on the example.

In his study, Newcombe [8] compared sensitivity and specificity with simulated design and graphical approach in simultaneous designs. This approach makes point and range estimation by weighting the difference between sensitivity and specificity of two diagnostic tests.

Wang et al. [9], proposed a weighted least squares method to compare diagnostic tests in the same sample.

Cairns et al. [10] used Newcombe's graphical approach to compare the sensitivity and specificity of two diagnostic tests with a single measure.

Kim and Lee [11] provided a theoretical framework for comparing sensitivity and specificity values separately using the Mc Nemar test under paired conditions and simulated these values under certain paired conditions by conducting a simulation study.

The fundamental criteria used in evaluating the performance of diagnostic and screen tests are sensitivity and specificity. While sensitivity refers to the ratio of diagnostic/screen test yielding a positive result in actually ill cases, specificity is the ratio of diagnostic/screen test yielding a negative result in actually healthy cases [8].

New methods with the claim of being better are proposed every passing day in parallel with the advancements in technology. This brings about the need for new statistical methods for comparing the effectiveness of the proposed tests. There are various methods used for comparing two diagnostic/screen tests developed for the same purpose. The first of these is the comparison of sensitivity and specificity in paired designs separately one via McNemar test.

The second approach involves the comparison of general accuracy rates of two tests. This approach contains inconveniences because it ignores the prevalence and the relative weights of false positive and false negative results.

Another method developed by Newcombe makes a simultaneous comparison of sensitivity and specificity in paired designs by taking into account the prevalence and the

clinical costs of false positive and false negative values. The results provided by this method are also quite advantageous in terms of clinical interpretation [8].

The performance of the two diagnostic tests is compared according to the results obtained from patients and healthy subjects. The McNemar test and similar range estimation methods can be used to compare the sensitivities of the two tests, but they do not take into account variations that may occur in the sample and differences in specificity. The aim of this study is to compare the sensitivity and specificity of two different tests with each other by taking into consideration of the prevalence and clinical significance under the condition that the result of the gold standard test is known.

2. MATERIAL AND METHOD

It is possible to compare T1 (Diagnostic test 1) and T2 (Diagnostic test 2) developed for the same purpose through two designs i.e. unpaired and paired.

Assuming that G represents gold standard test, T1 and G are administered to the subject N_1 in unpaired designs while T2 and G are administered to the subject N_2 in paired designs. The fact that T1 and T2 are administered to subjects simultaneously in paired designs provides an obvious advantage by making it possible to work with less number of subjects. The unpaired design is suggested to be used when T1 and T2 cannot be administered to the same subjects for any reason.

The present study focuses on paired designs. The method that is most frequently used in comparing sensitivity and specificity separately in paired designs is McNemar test. Relevant notations are demonstrated in Table 1 and Table 2.

When the sensitivity difference between two tests is $\theta_1 = \eta_{1.} - \eta_{.1} = \eta_{12} - \eta_{21}$, McNemar method can be used for testing the hypothesis, $H_0 : \theta_1 = 0$ (both tests have the same sensitivity) [12,13].

$$\chi^2 = \left(a_{12} - a_{21}\right)^2 \Big/ \left(a_{12} - a_{21}\right) \qquad (1)$$

In this test, a_{12} and a_{21} display discordant cell frequencies. *p value* can be calculated exactly, or z value can be calculated for large samples, or continuity correction can be made [14, 8, 15].

Table 1
Notations for the Simultaneous Comparison of Sensitivity and Specificity of Two Tests in Paired Designs for Actual Rates

GOLD STANDARD POSITIVE					GOLD STANDARD NEGATIVE				
	Test 2					Test 2			
Test 1		+	−	Total	Test 1		+	−	Total
	+	η_{11}	η_{12}	$\eta_{1.}$		+	ξ_{11}	ξ_{12}	$\xi_{1.}$
	−	η_{21}	η_{22}	$\eta_{2.}$		−	ξ_{21}	ξ_{22}	$\xi_{2.}$
	Total	$\eta_{.1}$	$\eta_{.2}$	1		Total	$\xi_{.1}$	$\xi_{.2}$	1
Sensitivity Difference $\theta_1 = \eta_{1.} - \eta_{.1} = \eta_{12} - \eta_{21}$					Specificity Difference $\theta_2 = \xi_{2.} - \xi_{.2} = \xi_{21} - \xi_{12}$				

Table 2
Notations for the Simultaneous Comparison of Sensitivity and Specificity of Two Tests in Paired Designs for the Frequencies Observed

GOLD STANDARD POSITIVE					GOLD STANDARD NEGATIVE				
	Test 2					Test 2			
Test 1		+	−	Total	Test 1		+	−	Total
	+	a_{11}	a_{12}	$a_{1.}$		+	b_{11}	b_{12}	$b_{1.}$
	−	a_{21}	a_{22}	$a_{2.}$		−	b_{21}	b_{22}	$b_{2.}$
	Total	$a_{.1}$	$a_{.2}$	M		Total	$b_{.1}$	$b_{.2}$	$N-M$

In 1995, Lu and Bean [16] stated that the relative importance of sensitivity and specificity could change, thus both sensitivity and specificity should be taken into consideration when comparing tests.

Newcombe developed a representation by weighing the clinical costs of false negatives and false positives. The costs of both test procedures may vary substantially in the course of time. Such varying costs are not included in the process simply. Let's assume that c_1 is the clinical cost of false negative result, c_2 is the clinical cost of false positive result, and 0 is the clinical cost of correct classification. To evaluate any screening program properly, the following must be the case: c_1 (the cost of false negative) $> c_2$ (the cost of false positive) > 0.

π is the estimated prevalence belonging to the population to which Test 1 and Test 2 are administered. Usually, it is not equal to M/N (M: The number of ill people $/N$: The number of all [healthy and ill] people), but lower than M/N.
Let's assume that E_1 (expected value) shows the performance loss of T_1 test in comparison to the gold standard test.

If

$$E_1 = \pi(1-\eta_{1.})c_1 + (1-\pi)(1-\xi_2)c_2 \quad (2)$$

and

$$E_2 = \pi(1-\eta_{.1})c_1 + (1-\pi)(1-\xi_{.2})c_2 \quad (3)$$

refers to performance loss for T2 test, the following is obtained [8]:

$$\begin{aligned} &E_2 - E_1 = \pi c_1\theta_1 + (1-\pi)c_2\theta_2 \\ &\theta_1 \text{ and } \theta_2 \\ &f = \lambda\theta_1 + (1-\lambda)\theta_2 \\ &\frac{\lambda}{1-\lambda} = \frac{\pi c_1}{(1-\pi)c_2} \rightarrow \lambda = \frac{1}{1+\frac{(1-\pi)c_2}{\pi c_1}} \end{aligned} \quad (4)$$

Getting a value between 0 and 1, λ is a parameter that is to be calculated by using the clinical cost proposed for false negative and false positive values and the prevalence value belonging to the population to which tests are administered [8].

λ depends on the balance between c_1/c_2 and π. It allows the researcher to evaluate results in accordance with the purposes set for the study. If the cost of false negative is more important, the case will be as follows: $c_1 \le c_2$. In such a case, $c_1/c_2 \rightarrow 0$ will be true and π will be different from 0. In addition, $\lambda \rightarrow 1$ will be true, and f will replace θ_1. On the contrary, if the prevalence of disease is low enough and the cost of false positive is more important, $\pi \rightarrow 0$ will be true, and c_1/c_2 value will be different from 0. $\lambda \rightarrow 0$ will be true, and f will replace θ_2.

To create an appropriate confidence interval method for the difference of two ratios in unpaired designs, two Wilson score confidence interval methods [17] can be combined with squaring and addition operations. This process can be regulated as in the equation 5 for independent chance X_1 and X_2 variables along with $w_1 = +1$ and $w_2 = -1$ provided here.

$$\text{var}(w_1X_1 + w_2X_2) = {w_1}^2 \text{ var } X_1 + {w_2}^2 \text{ var } X_2 \quad (5)$$

The means of independent random samples whose variance is known as σ^2 and whose mean is selected from the normal distribution with μ_1 and μ_2 are defined as $\bar{x}_1$ and $\bar{x}_2$. Let's assume that lower limit is (l_i) and upper limit is (u_i). If the confidence interval for $\mu_i\,(i=1,2)$ is $100\%(1-\alpha)$, the confidence interval $100\%(1-\alpha)$ for $\mu_1 - \mu_2$ [18, 19, 20, 21] is to be as follows:

$$\bar{x}_1 - \bar{x}_2 - \sqrt{\left((\bar{x}_1 - l_1)^2 + (u_2 - \bar{x}_2)^2\right)} \text{ and } \bar{x}_1 - \bar{x}_2 + \sqrt{\left((\bar{x}_2 - l_2)^2 + (u_1 - \bar{x}_1)^2\right)} \quad (6)$$

The methods developed determine an interval for the difference between two ratios in paired and unpaired designs. This interval will be independent of deviations [8].

For θ_i, the confidence interval $100\%(1-\alpha)$ is re-calculated through the representation of $i=1,2 \rightarrow (L_i, U_i)$. The process in unpaired designs is re-applied for any $\lambda \in [0,1]$ value. However, this time, the confidence interval $100\%(1-\alpha)$ of $w_1 = \lambda$ and $w_2 = 1-\lambda$ and f is calculated as in the equation 7[22, 23, 24, 25].

$$\hat{f} - \sqrt{\left(\lambda^2\left(\hat{\theta}_1 - L_1\right)^2 + (1-\lambda)^2\left(\hat{\theta}_2 - L_2\right)^2\right)}$$
$$\hat{f} + \sqrt{\left(\lambda^2\left(U_1 - \hat{\theta}_1\right)^2 + (1-\lambda)^2\left(U_2 - \hat{\theta}_2\right)^2\right)} \qquad (7)$$

Here, the positive values of square root results are taken. This is reduced to $(L_1 U_1)$ in $\lambda = 1$, and to $(L_2 U_2)$ in $\lambda = 0$ [23, 24, 225].

The graph of f under the limits determined above for the values of λ in the 0-1 interval allows preferring one test instead of another by taking into consideration any assumed prevalence value and loss ratios. Researchers can determine the λ value depending on both prevalence (known or assumed) and the significance level of clinical cost [26, 27]. In addition, the order of tests can be changed to determine which test is to be used as well as its degree of influence. As it is simply as follows: $\frac{\partial^2 L}{\partial^2 \lambda} < 0 < \frac{\partial^2 U}{\partial^2 \lambda}$, the graph to be drawn will be concave for closed form all the time [28]. Since it is generally as follows: $M < N/2$ (The number of ill people / [The total number of people /2]), the interval of sensitivity difference (θ_1) (showed with $\lambda = 1$ in graph) will be wider than the interval of specificity difference (θ_2) (showed with $\lambda = 0$ in graph).

A study carried out at the Department of Cytopathology of Ankara University Medical School investigated the impact of the use of various antibodies in determining thyroid nodules as benign or malign via Preoperative Fine-Needle Aspiration Biopsy (FNAB) in order to achieve diagnostic accuracy. To this end, the antibodies of HBME-1, CD56 and CITED-1 were studied immunehistochemically. Histopathological examination was accepted as gold standard. At the end of the evaluation made through gold standard test, 44 people were found to have benign thyroid nodules while 46 people were found to have malign thyroid nodules. In this study, the performance of the diagnostic test has been assessed by comparing the two methods: the classical methods (McNemar) and Newcombe graphical approach.

The performances of the three tests in comparison to one another were evaluated via Newcombe method. Analyses were performed through MINITAB 16.0 and the Ms Office Excel Macro developed by Newcombe.

3. FINDINGS

Table 3
The Performances of PHME1 and PCD56 Tests when the Gold Standard was Positive and Negative

GOLD STANDARD POSITIVE					GOLD STANDARD NEGATIVE				
		PCD56					PCD56		
		+	–	**Total**			+	–	**Total**
PHME1	+	37	3	40	**PHME1**	+	5	1	6
	–	0	4	4		–	10	30	40
	Total	37	7	44		**Total**	15	31	46
Sensitivity Difference PHME1 STAINING – PCD56 STAINING = 0.0682 95% confidence limits -0.0267 and 0.1776					Specificity Difference PHME1 STAINING – PCD56 STAINING = 0.1957 95% confidence limits 0.0566 and 0.3320				

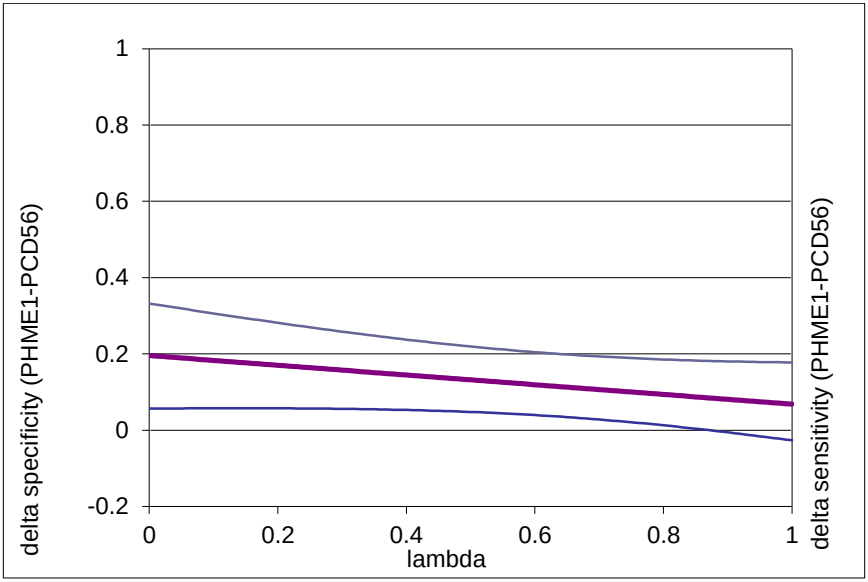

Figure 1: The Simultaneous Representation of Sensitivity and Specificity Based on the Prevalence (lambda) between PHME1 Staining and PCD56 Staining

The performance of the PHME1 and PCD56 tests in case where the gold standard is positive and negative; and the simultaneous evaluation of sensitivity and specificity by the Newcombe test according to the prevalence (lambda) between the PHME1 staining and the PCD56 staining are displayed in Table 3 and Figure 1, respectively.

Table 4
The Performances of PHME1 and PCITED Tests when the Gold Standard was Positive and Negative

GOLD STANDARD POSITIVE					GOLD STANDARD NEGATIVE				
		PCITED					PCITED		
		+	–	Total			+	–	Total
PHME1	+	38	2	40	PHME1	+	5	1	6
	–	4	0	4		–	8	32	40
	Total	42	2	44		Total	13	33	46
Sensitivity Difference PHME1 STAINING – PCITED STAINING = –0.0455 95% confidence limits 0.1727 and 0.0772					Specificity Difference PHME1 STAINING – PCITED STAINING = 0.1522 95% confidence limits 0.0226 and 0.2835				

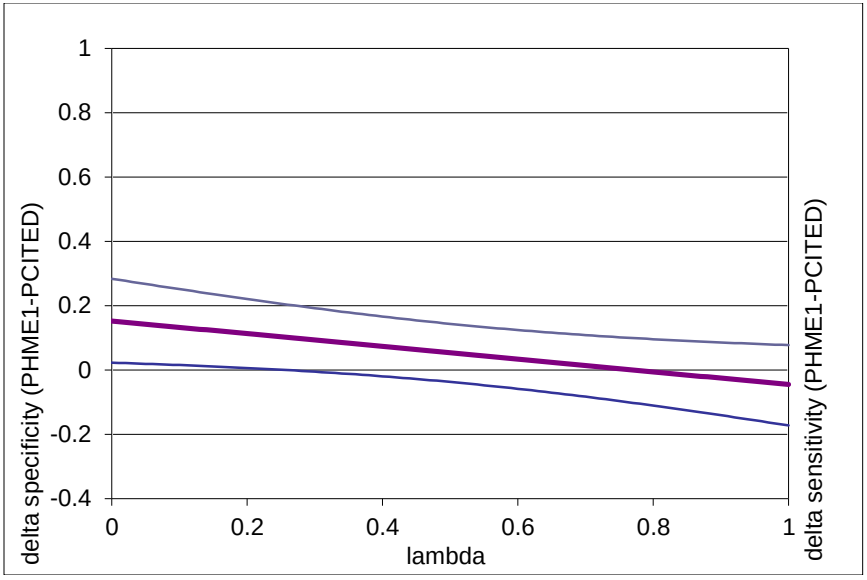

Figure 2: The Simultaneous Representation of Sensitivity and Specificity based on the Prevalence (lambda) between PHME1 Staining and PCITED Staining

The performance of the PHME1 and PCITED tests in case where the gold standard is positive and negative; and the simultaneous evaluation of sensitivity and specificity by the Newcombe test according to the prevalence (lambda) between the PHME1 staining and the PCITED staining are displayed in Table 4 and Figure 2, respectively.

Table 5
The Performances of PCITED and PCD56 Tests when the Gold Standard was Positive and Negative

GOLD STANDARD POSITIVE					GOLD STANDARD NEGATIVE				
	PCD56					PCD56			
PCITED		+	–	Total	PCITED		+	–	Total
	+	35	7	42		+	8	5	13
	–	2	0	2		–	7	26	33
	Total	37	7	44		Total	15	31	46
Sensitivity Difference PCITED STAINING – PCD56 STAINING = 0.1136 95% confidence limits -0.0249 and 0.2552					Specificity Difference PCITED STAINING – PCD56 STAINING = 0.0435 95% confidence limits -0.1080 and 0.1924				

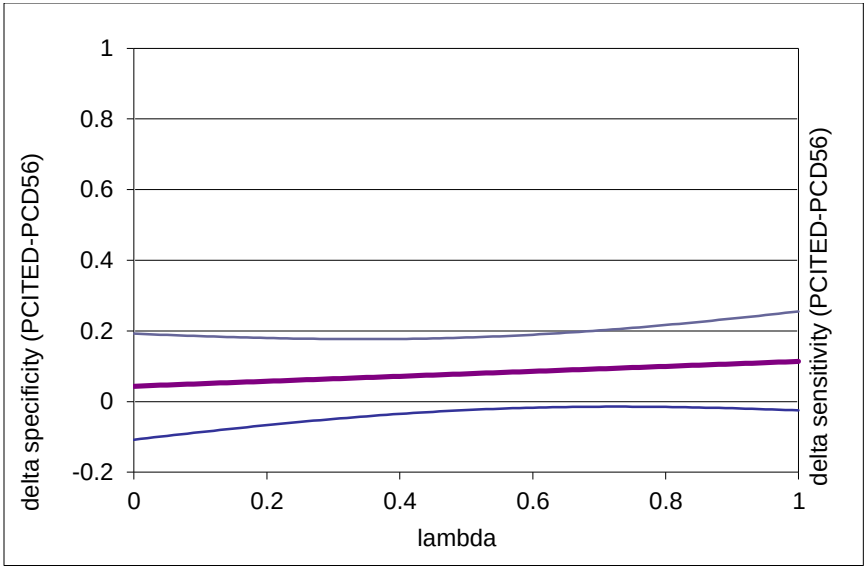

Figure 3: The Simultaneous Representation of Sensitivity and Specificity based on the Prevalence (lambda) between PCITED Staining and PCD56 Staining

The performance of the PCITED and PCD56 tests in case where the gold standard is positive and negative; and the simultaneous evaluation of sensitivity and specificity by the Newcombe test according to the prevalence (lambda) between the PCITED staining and the PCD56 staining are displayed in Table 5 and Figure 3, respectively.

McNemar Test Performance:

If there was no relationship between the disease and the risk factor, one would anticipate that the number of pairs in which the case was subjected to the risk factor while

the control was not subjected to the risk factor to be equal to the number of pairs in which this time the control was subjected to the risk factor while the case was not. The present study involved three discordant pairs, with the case and the control having different exposure to the risk factor. There were three (100.00%) pairs in which the control was subjected to the risk factor while the case was not; and 0 (0.00%) pairs in which the case was subjected to the risk factor while the control was not. The answer to this question lies in the p value: If there is not any relationship between risk factor and disease, how probable is it to determine a discrepancy that is not less large than this figure between the numbers of the two types of discordant pairs? A small p value is a proof that indicates the presence of a relationship between the risk factor and the disease.

The performances of PHME1 and PCD56 tests with the gold standard positive and the gold standard negative are below:

Gold Standard Positive; the two-tailed p value equals 0.2482. According to conventional criteria, this difference is regarded to be "not statistically significant". McNemar's test with the continuity correction was applied to calculate the p value. Chi squared equals 1.333 with 1 degree of freedom.

Gold Standard Negative; the two-tailed p value equals 0.0159. According to conventional criteria, this difference is regarded to be "statistically significant". McNemar's test with the continuity correction was applied to calculate the p value. Chi squared equals 5.818 with 1 degree of freedom.

The performances of PHME1 and PCITED tests with the gold standard positive and the gold standard negative are below:

Gold Standard Positive; the two-tailed p value is equal to 0.6831. According to conventional criteria, this difference is regarded to be "not statistically significant". McNemar's test with the continuity correction was applied to calculate the p value. Chi squared is equal to 0.167 with 1 degree of freedom.

Gold Standard Negative; the two-tailed p value is equal to 0.0455. According to conventional criteria, this difference is regarded to be "statistically significant". McNemar's test with the continuity correction was applied to calculate the p value. Chi squared is equal to 4.000 with 1 degree of freedom.

The performances of PCITED and PCD56 tests with the gold standard positive and the gold standard negative are below:

Gold Standard Positive; the two-tailed p value is equals to 0.1824. According to conventional criteria, this difference is regarded to be "not statistically significant". McNemar's test with the continuity correction was applied to calculate the p value. Chi squared is equal to 1.778 with 1 degree of freedom.

Gold Standard Negative; The two-tailed p value is equal to 0.7728. According to conventional criteria, this difference is regarded to be "not statistically significant". McNemar's test with the continuity correction was applied to calculate the p value. Chi squared is equal to 0.083 with 1 degree of freedom.

4. DISCUSSION

Wang et al. [9], compared the predictive values of two diagnostic tests for the same sample using the weighted least squares method (WLS). 608 coronary artery disease (CAD) and 263 CAD free subjects have 0.89 and 0.88 positive predictive value (PPV) and 0.78 and 0.65 negative predictive value (NPV) of clinical history and exercise stress testing (EST), respectively. Both WLS and generalized estimating equation (GEE) methods were applied to the data. In a way to compare PPV and NPV, the two statistics from WLS were found to be very close to those from GEE. For this reason, the PPV values of EST and clinical history (0.89 and 0.88) were not different in order to predict CAD. However, the clinical history has a preferred NPV value (0.78 vs. 0.65) compared to EST. As a result, it can be concluded that WLS and GEE methods yielded similar results. The authors noted that the WLS method using the difference between two PPVs or NPVs was based on the simulation study and the results they reached from the actual data had similar test size or power in the GEE score test. The WLS method is suggested, since both the difference-based approach is easier to use and the WLS is more understandable by the researchers.

In the Cairns et al. [10] rapid diagnostic tests (RDTs) were used to distinguish between those who are truly having malaria and those not. Nevertheless, they used Newcombe's graphical approach, with the understanding that combining correct management of positives and negatives as a single summary measure could be misleading. According to the results obtained, if the PMR (Positive Management Rate)is prioritized $(R > 1)$, i.e. false negatives (missedmalaria cases) are highlighted, the lambda is raised. It is sharper than the low prevalence and therefore has more effect on the weighted difference of the PMR. Conversely, if a false positive is given priority $(R < 1)$, then the NMR (Negative Management Rate) has more effect. When the confidence interval of the weighted average exceeds 0, the test group is no longer significantly better than the control group. In their study, graphical representations were given according to five different values of the prevalence.

Kim and Lee [11] have studied data, used by Luong et al., on the diagnosis of invasive pulmonary aspergillosis (IPA) in lung transplant recipients. Using the cases including 97 negative controls and 51 infected patients, they reanalyzed the data to compare the PCR assays and the GM assay. They performed McNemar's test separately for infected patients and non-infected patients and compared the results with the result of another McNemar's test performed on all patients (infected and uninfected). While the results for separate tests were meaningful, the result for the latter test with combined data was meaningless. This indicates that the improper use of McNemar's test may lead to very different results from the results those need to be obtained via separately performed analysis. The authors mentioned that even though the McNemar test is widely used in comparing two tests, in order to address the situations as in their study there are alternative methods within the literature such as graphical approach.

In this study, when the performances of the tests were evaluated by Newcombe method, the following results are obtained.

When the antibodies of HBME-1 and CD56 were compared, when $\lambda = 0$, the difference between the specificity of tests was found to be 0.20 (0.06-0.33); and when $\lambda = 1$, the difference between the sensitivity of tests was found to be 0.07 (-0.027-0.18) (Table 3, Figure 1). Since the interval value within the λ = 0.87 instantly included the value of 0, it is not appropriate to evaluate the circumstances where $\lambda = 1$ in terms of sensitivity. However, when λ = 0.50 (i.e. the cost of false positive and the cost of false negative had the same level of significance and the prevalence was close to 50%), the antibody of HBME-1 was found to be superior to the antibody of CD56 in terms of both sensitivity and specificity. In all circumstances where $\lambda \leq 0.87$, the antibody of HBME-1 was found to be superior to the antibody of CD56 in terms of specificity (Table 3, Figure 1).

When the antibodies of HBME-1 and CITED-1 were compared, when $\lambda = 0$, the difference between the specificity of tests was found to be 0.15 (0.02-0.28); and when $\lambda = 1$ the difference between the sensitivity of tests was found to be -0.05 (-0.17-0.07) (Table 4, Figure 2). Since the interval value within the λ = 0.19 instantly included the value of 0, it is not appropriate to evaluate both the circumstances where $\lambda = 1$ in terms of sensitivity and those where λ = 0.50 in terms of sensitivity and specificity (Table 4, Figure 2).

When the antibodies of CITED-1 and CD56 were compared, since for all λ values, the interval instantly included the value of 0, it was determined that there was no difference between the performances of tests in comparison to one another (Table 5, Figure 3).

5. CONCLUSION

The relative benefits of two tests can be identified by determining their distance to the gold standard test used. This yields definite results, but is very expensive or usually invasive. In some instances, these tests can be used only after death (Alzheimer) [29, 30, 31].

In such disciplines as histopathology where there is a particular test procedure, the repetition of gold standard test may not yield perfect results, or this test may not be repeatable. The classical methods proposed so far allow comparing the sensitivity and specificity of two tests in paired designs separately.

Presented as an alternative to these methods, the Newcombe graphical approach is a graphical approach that permits simultaneous comparison of sensitivity and specificity between two tests of the two predictors of the two false positives. When applied to the relevant sample, the Newcombe graphical approach allows the simultaneous comparison of the positive and negative cases between two different management strategies by drawing the difference in the correct probability. This approach therefore provides a simple visual way to summarize the two aspects of case management performance, and extend results from one setting to areas with different prevalence.

Considering the prevalence of Newcomb graphic approach that compares both the sensitivity and specificity, as a result of this study it can be concluded that Newcombe graphical approach gives better results than the classical Mc Nemar test for the same sample group used in the study. Simultaneously comparing the sensitivity and specificity

of tests by taking into consideration the clinical costs of false positive/false negative values and prevalence, Newcombe graphic approach method is also quite advantageous in terms of clinical interpretation.

ACKNOWLEDGEMENT

Special thanks to Prof. Newcombe for his invaluable assistance and motivation/ support for this the study. Thank you who Tuğba Taşkın Türkmenoğlu, [Dr. (MD), Ankara University School of Medicine Department of Pathology, Cytopathology Department] and Koray CEYHAN, **[**Prof. Dr. Ankara University School of Medicine, Department of Pathology, Cytopathology Department] for their help in the study.

REFERENCES

1. Bloch D.A. (1997). Comparing Two Diagnostic Tests Against the Same "Gold Standard" in the Same Sample, *Biometrics*, 53, 73-85
2. Greenhouse, S.S. and Mantel, N. (1950). The evaluation of diagnostic tests. *Biometrics*, 6, 399-412.
3. Linnet, K. (1987). Comparison of quantitative diagnostic tests: Type 1 error, power, and sample size. *Statistics in Medicine*, 6, 147-158.
4. Hanley, J.A. and McNeil, B.J. (1983). A method of comparing the areas under receiver operating characteristics curves derived from the same cases. *Radiology*, 148, 839-843.
5. Delong, E.R., Delong, D.M. and Clarke-Pearson, D.L. (1988). Comparing the areas under two or more correlated receiver operating characteristics curves: A nonparametric approach. *Biometrics*, 44, 837-845.
6. Wieand, S., Gail, M.H., James, B.R., James, K.L. (1989). A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika*, 76, 585-592.
7. Campbell, G. (1994). General methodology I advances in statistical methodology for the evaluation of diagnostic and laboratory tests. *Statistics in Medicine*, 13, 499-508.
8. Newcombe, R.G. (2001). Simultaneous comparison of sensitivity and specificity of two tests in the paired design: a straightforward graphical approach. *Statistics in Medicine*, 20(6), 907-915.
9. 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. *Statist. Med*. 25, 2215-2229.
10. Cairns, M.E., Leurant, B. and Milligan, J.M. (2014). Composite endpoints for malaria case management: not simplifying the picture? *Malaria Journal*, 13, 494.
11. Kim, S. and Lee, W. (2017). Does McNemar's test compare the sensitivities and specificities of two diagnostic tests? *Statistical Methods in Medical Research*, 26(1), 142-154.
12. McNemar Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 17(2), 153-157.
13. Scott, W.A. (1955). Reliability of content analysis: the case of nominal scale coding. *Public Opinion Quarterly* 19(38), 321-325.

14. Lancaster, H.O. (1949). The combination of probabilities arising from data in discrete distributions. *Biometrika*, 36(3/4), 370-382.
15. Stone M. (1969). The role of significance testing. Some data with a message. *Biometrika*, 56(3), 485-493.
16. Lu, Y. and Bean, J.A. (1995). On the sample size for one-sided equivalence of sensitivities based upon McNemar's test. *Statistics in Medicine*, 14(16), 1831-1839.
17. Wilson, E.B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158), 209-212.
18. Agresti, A. and Coull, B.A. (1998). Approximate is better than 'exact' for interval estimation of binomial proportions. *American Statistician*, 52(2), 119-126.
19. Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(3), 37-46.
20. Cox, D.R. and Hinkley, D.V. (1974). *Chapter 4 Significance Test: Simple Null Hypotheses, Theoretical Statistics*, Chapman and Hall, 1st Edition, London, 89-91.
21. Tango, T. (1998). Equivalence test and confidence interval for the difference in proportions for the paired-sample design. *Statistics in Medicine*, 17(8), 891-908.
22. Mee, R.W. (1990). Confidence intervals for probabilities and tolerance regions based on a generalization of the Mann-Whitney statistic. *Journal of the American Statistical Association*, 85(411), 793-800.
23. Newcombe, R.G. (1998). Improved confidence intervals for the difference between binomial proportions based on paired data. *Statistics in Medicine*, 17(22), 2635-2650.
24. Newcombe, R.G. (1998). Interval estimation for the difference between independent proportions. A comparative evaluation of eleven methods. *Statistics in Medicine*, 17(8), 873-890.
25. Newcombe, R.G. (1998). Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine*, 17(8), 857-872.
26. Hanley, J.A. and McNeil, B.J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143:1, 29-36.
27. Hanley, J.A. and McNeil, B.J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148(3), 839-843.
28. Hope, R.L., Chu, G., Hope, A.H., Newcombe, R.G., Gillespie, P.E. and Williams, S.J. (1996). A comparison of three faecal occult blood tests in the detection of colorectal neoplasia. *Gut*, 39(5), 722-725.
29. Kyle, P.M., Campbell, S., Buckley, D., Kissane, J., Swiet, M., Albano, J., Millar, J.G. and Redman, C.W. (1996). A comparison of the inactive urinary kallikrein: creatinine ratio and the angiotensin sensitivity test for the prediction of pre-eclampsia. *BJOG: An International Journal of Obstetrics & Gynaecology*, 103(10), 981-987.
30. Kyle, P.M., Redman, C.W.G., De Swiet, M., Millar, J.G.A. (1997). Comparison of the inactive urinary kallikrein: creatinine ratio and the angiotensin sensitivity test for the prediction of pre-eclampsia. *British Journal of Obstetrics and Gynaecology*, 104(8); 971.
31. Temel, G.O. and Kanik, E.A. (2011). Çok Testli Çok Değerlendiricili ROC Çalışmalarında Tanı Testleri Arasındaki İlişkinin Diagnostik Doğruluk Sonuçlarına Etkisi. *Turkiye Klinikleri Journal of Biostatistics*, 3(2), 63-73.