MedCrave
Step into the World of Research

Biometrics & Biostatistics International Journal

Review Article

Open Access

CrossMark

An extended McNemar test for comparing correlated proportion of positive responses

Abstract

The area under a ROC curve (AUC) is an important summary measures useful in assessing the accuracy of a diagnostic test in discriminating true disease status when the data for measurement is paired. This assessment is most important when the AUCs of different diagnostic test procedures are compared. These comparisons are not without some problem associated with it such as the inability of some test such as the McNemar test to adjust for the possible presence of ties in the data, thereby leading to erroneous conclusions in data analysis occasioned by committing Type II error more often than not. This is evident when the use of the traditional McNemar test in data analysis yielded high value of variance and low chi-square value thereby making one to accept a false null hypothesis more often than expected. To be able to tackle this challenge, we extend the usual McNemar test adopted by adjusting for the possible presence of ties in the data when measurements of data may be on any scale. The extended McNemar test can enable one to easily estimate the probability that randomly selected pair of subjects from two diagnostic test procedures respond positive or both respond negative and it can be used to test the null hypothesis of equality of proportion of positive responses in two diagnostic test procedures. An extensive simulation study was carried out to determine the Type I error and statistical power of the existing and extended tests and the application of the tests to standard and real data, was carried out and result showed that in all the McNemar test demonstrates superior statistical power and less conservative type I error compared to DeLong area test, Bandos et al area test and the usual McNemar area test and so compares favorably.

Keywords: extended mcnemar test, positive response, correlated data, nonparametric test, diagnostic tests, type ii error

Volume 8 Issue 4 - 2019

Okeh Uchechukwu Marius,[1] Obiora-Ilouno Happiness[2]

[1]Department of Mathematics and Computer Science, Ebonyi State University Abakaliki, Nigeria

[2]Department of Statistics, Nnamdi Azikiwe University Awka, Nigeria

Correspondence: Okeh Uchechukwu, Department of Mathematics and Computer Science, Ebonyi State University Abakaliki, Nigeria, Tel +2348034304278, Email uzomaokey@ymail.com

Received: May 09, 2019 | **Published:** July 08, 2019

Introduction

The receiver operating characteristic (ROC) curve is a standard tool used to evaluate the performance of a diagnostic test when measurement of test results are either continuous or ordinal.[1] In 1950s the methodology of ROC was first developed by electrical and radar engineers during World War II for signal detection theory in battle fields.[2] In an ROC curve, the true positive rate (TPR) is plotted against the false positive rate (FPR) across all possible cut-off values in other to make meaningful decision. The area under the ROC curve (AUC) is a summary index for measuring the diagnostic accuracy. AUC ranges from 0 to 1 inclusive and the greater the value of AUC close to 1, the better the discriminatory power of the diagnostic procedure. Often times, the aim of many diagnostic studies is to compare the accuracy of diagnostic tests to determine the superiority of one test over another test for a certain condition or disease when data measurement may be on any scale. Statistical inference may be based on parametric, nonparametric or semi-parametric statistics. If the statistical inference is nonparametric, the difference between correlated AUCs for paired data was first proposed by DeLong et al.,[3] and it is based upon asymptotic theory for *U*-statistics.[4] But the validity of this or any other method relays on large sample size and when the sample size is small, the validity of the test for the difference between two or more AUCs may not be achieved. Two permutation tests for paired receiver operating characteristic (ROC) studies currently exist: one proposed by Venkatraman & Begg[5] and the more recent test of Bandos et al.,[6] The test of Bandos et al.,[6] directly tests for an equality of AUCs, while the test of Venkatraman & Begg[5] is more general and tests for equality of the underlying ROC curves. As a result, the test of Venkatraman & Begg[5] is less powerful for testing equality of AUCs. Both permutation tests are executed by permuting the labels of the two tests within each diseased and non-diseased subject. Such an approach implicitly assumes that both tests are exchangeable within subject and requires an appropriate transformation, such as ranks, for tests differing in scale. Bandos et al.,[6] compared the performance of their test to that of DeLong et al.,[3] using simulation and found that the permutation test had greater power than the nonparametric test developed by DeLong et al.,[3] when there was moderate correlation between two tests, large AUCs, and small sample sizes.

When comparing two diagnostic procedures, the difference between AUCs is often used and to control for the sources of changes arising from changes due to subjects which represents a reasonable size of the overall changes of the AUC, a paired data is recommended. This is because paired data usually induces positive correlation between the test results of the same subjects. Based on the use of paired data, Sumi et al.,[7] adopted the usual McNemar[8] test for comparing two correlated marginal probability of positive responses in diagnostic test procedures. This paper is an extension of this work for evaluating the performance of two diagnostic tests in terms of the proportion of positive responses and the comparison of this method with the existing tests by DeLong et al.,[3] Bandos et al.,[6] Sumi et al.[7]

Estimation of AUC

In estimating the AUC, two main factors have to be considered namely, the design of the study and the distribution of test result.[9] Under the study design, test results or dataset can be classified into three types namely: (i) paired data (ii) unpaired data and (iii) partially paired data. For the paired and partially-paired set of data, correlation between AUCs is considered. Under the distribution type of test

result, three approaches for estimating the AUCs are considered namely: (i) A parametric approach (ii) A semi-parametric approach (iii) A non-parametric approach, in this paper, our focus will be on the non-parametric method. All the approaches to estimating the AUC differ in the way the distribution functions of both populations are estimated based on their sample values. Basically the nonparametric (empirical) method of estimating AUC is stated as follows.

Given that there are two diagnostic tests, let n be the total number of subjects without disease and m as the total number of subjects with disease. Suppose X^b *and* $Y^b(b=1,2)$ represent the subjects without disease and with disease respectively. Therefore the corresponding bivariate outcomes for the two diagnostic procedures on the same N non-diseased and M diseased subjects should be $x_i^b(i=1,2,...,n)$ *and* $y_j^b(j=1,2,...,m)$. Bivariate cumulative distribution functions are denoted by $F(x^1,x^2)$ *and* $G(y^1,y^2)$ and their corresponding margina $F_b(x^b), G_b(y^b)(b=1,2)$. Bamber[10] noted that the AUC is equal to $P(Y>X)$. Let $AUC_b(b=1,2)$ be the AUCs of diagnostic procedures. The formula suggested by Hanley and McNeil (1982) for computing the AUC is given as

$$AUC=\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}g(X_i,Y_j) \qquad 1$$

Where m=number of diseased subjects, n=number of non-diseased subjects. Also X_i *and* Y_j are respectively the test result of the i[th] and j[th] subject without and with disease and g is the indicator function comparing X_i *with* Y_j such that

$$g(X_i,Y_j)=\begin{cases}1, if\ Y_j>X_i\\0.5, if\ Y_j=X_i\\0, otherwise\end{cases} \qquad 2$$

Therefore for the b^{th} diagnostic test procedure the AUC can be computed as

$$AUC_b=\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}g(X_i^b,Y_j^b) \qquad 3$$

To carry out significant test for the differences between two or more correlated AUCs, it is necessary to consider the distribution of the test result which also determines the procedure to be adopted in estimating the AUCs and its variance-covariance matrix. By the comparison of areas under the two ROC curves, we can estimate which one of two diagnostic tests is more suitable for discriminating non-diseased subjects from diseased subjects or any other two conditions of interest.[7] Braun & Alonzo[11] proposed a modified rank test that does not require such a transformation and showed that the modified test has the same power as Bandos et al.,[6] Bantis & Feng.,[12] focused on comparing two correlated ROC curves at a given specificity level. They proposed parametric approaches, transformations to normality, and nonparametric kernel-based approaches. Extensions of their methods also involved inference for the AUC and accommodating covariates. They evaluated the robustness of their techniques through simulations, compared to other known approaches and presented a real data application involving prostate cancer screening. They approaches perform satisfactorily in terms of size and power. The limitation of Bantis & Feng[12] method is that their Box-Cox version does not take into account the variability of the transformation parameter. Finally, to increase the ability to detect the crossing alternative, Yu et al.,[13] suggested a two-stage test, where the first stage uses the test derived by DeLong et al.,[3] to test the equality of the two AUCs and the second stage uses a modified area test to test two partial AUCs.

Existing nonparametric tests for comparing correlated AUC$_s$ in paired sample design

A number of tests exist for comparing two or more AUCs or proportion of positive responses for the matched sample case.

DeLong et al's conventional nonparametric method for comparing AUCs

DeLong et al.,[3] developed a totally nonparametric approach to compare two correlated AUCs of two diagnostic tests for paired samples of subjects by using the theory of generalized U statistics. In other words, they developed a conventional fully nonparametric approach leading to an asymptotically normal test statistic. This method is important as it helps to study the behavior of the type I error and the statistical power of the conventional nonparametric test for comparing two AUCs over a wide range of relevant parameters and against various alternatives. The test by Delong et al is limited by the fact that the AUC has an unbiased non-parametric estimator called the indicator variable that requires the comparison of all the number of subjects responding positive and negative, thus working with very large number of observations, so that computational time could be long. In estimating AUC, sigmoid function is sometimes used instead of indicator function or variable.[14] However, DeLong et al.,[3] method is based only on a continuous scale of measurement. The method of structural components is used to generate an estimated covariance matrix and the resulting test statistic has asymptotically a chi-square distribution.

Suppose $X_i, i=1,2,....,n$ denote test results for a sample of n non-diseased subjects, and $Y_j, j=1,2,....,m$ denote the test results for m diseased subjects. For each (X_i,Y_j) pair, an indicator function $I(X_i,Y_j)$ is defined as follows:

$$I(X_i,Y_j)=\begin{cases}1\ if\ Y_j>X_i\\0.5\ if\ Y_j=X_i\\0 if\ Y_j<X_i.\end{cases} \qquad 4$$

The average of these values for I over all nm comparisons is the Wilcoxon or Mann-Whitney U statistic:

$$U=\frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}I(X_i,Y_j). \qquad 5$$

Where U is equivalent to the AUC under the trapezoidal ROC curve Wieand et al.,[15] obtained by connecting the ROC data points by straight lines, and the expected value of U, $E(U)$, according to Hajian-Tilaki & Hanley.,[16] is the area under the theoretical (population) ROC curve $(\theta): E(U)=\theta=prob(Y>X)$.

An alternative representation, used by DeLong et al.,[3] is to define the components of the U statistic for each of the n non-diseased subjects and for each of the m diseased subjects:

$$Var_N(X_i)=\frac{1}{m}\sum_{j=1}^{m}I(X_i,Y_j)\ and\ Var_D(Y_j)=\frac{1}{n}\sum_{i=1}^{n}I(X_i,Y_j). \qquad 6$$

Where $Var_N(X_i)$ *and* $Var_D(Y_j)$ are called "pseudo-values" or "pseudo-accuracies." The pseudo-value $Var_N(X_i)$ for the i^{th} subject

Citation: Marius OU, Happiness OI. An extended McNemar test for comparing correlated proportion of positive responses. *Biom Biostat Int J.* 2019;8(4):125–137. DOI: 10.15406/bbij.2019.08.00281

in the non-diseased group is defined as the proportion of *Y*'s in the sample of diseased subjects where *Y* is greater than X_i. While $Var_D(Y_j)$ for the j^{th} subject in the diseased group is defined as the proportion of *X*'s in the sample of non-diseased subjects whose *X* is less than Y_j. $\{Var_N\}$ **a***nd* $\{Var_D\}$ can be used in place of the original diagnostic test results $\{X\}$ and $\{Y\}$ to construct the empirical ROC curve. The average of the $\{\bar{V}ar_N\}$ *and* $\{\bar{V}ar_D\}$ sample are respectively given as

$$\bar{V}ar_N = \frac{1}{n}\sum_{i=1}^{n} Var_N(X_i) = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} I(X_i, Y_j) = U. \qquad 7$$

and

$$\bar{V}ar_D = \frac{1}{m}\sum_{j=1}^{m} Var_D(Y_j) = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} I(X_i, Y_j) = U. \qquad 8$$

Therefore

$$A\hat{U}C = U = \frac{1}{n}\sum_{i=1}^{n} Var_N(X_i) = \frac{1}{m}\sum_{j=1}^{m} Var_D(Y_j) \qquad 9$$

Thus, the average of the values for $n\{Var_N\}$ and the average of those for $m\{Var_D\}$ are both equivalent to the *U* statistic, which is why there are called pseudo-accuracy measures. As was shown by Hettmansperger.,[17] the estimate of variance of the *U* statistic (which he called *W* instead of *U*) can be expressed as the sum of variances of $\bar{V}_N, \bar{V}_D$, and a third component, $U(1-U)/nm$. DeLong et al.,[3] omitted the third component, since it is negligible when *n* and *m* are large. They explained that for a single diagnostic test, the variance of AUC is given as

$$Var[A\hat{U}C] = Var[\hat{U}] = \frac{S_{T_D}^2}{m} + \frac{S_{T_N}^2}{n}. \qquad 10$$

Where $S_{T_D}^2$ *and* $S_{T_N}^2$ are respectively the sample variances for the diseased and non-diseased components and are defined as

$$S_{T_N}^2 = \frac{\sum_{i=1}^{n}\left[Var_N(X_i) - AUC\right]^2}{n-1} \; and \; S_{T_D}^2 = \frac{\sum_{j=1}^{m}\left[Var_D(Y_j) - AUC\right]^2}{m-1} \qquad 11$$

The null hypothesis of interest is to compare the equality of AUCs from two diagnostic test procedures when the data is paired and by extension if the period of measurement of test results are the same and the test statistic according to DeLong et al.,[3] is the Z-test given as

$$Z = \frac{AUC_1 - AUC_2}{\sqrt{Var(AUC_1 - AUC_2)}} \qquad 12$$

Where

$$Var(AUC_1 - AUC_2) = Var(AUC_1) + Var(AUC_2) - 2Cov(AUC_1, AUC_2)$$

If the two diagnostic tests are not matched to the same subjects, the two AUCs are independent and the covariance term would be zero. In other to estimates the AUCs for the two diagnostic test procedures, Delong et al.,[3] considered that each variance of AUC be defined as

$$Var(AUC_b) = \frac{S_{T_{Db}}}{m_b} + \frac{S_{T_{Nb}}}{n_b} \qquad 13$$

Where

$$S_{T_{Nb}} = \frac{\sum_{i=1}^{n_b}\left[Var_N(X_{ib}) - AUC_b\right]^2}{n_b - 1}, b = 1,2.$$

and

$$S_{T_{Db}} = \frac{\sum_{j=1}^{m_b}\left[Var_D(Y_{jb}) - AUC_b\right]^2}{m_b - 1}, b = 1,2.$$

The variance of the components $Var_N(X_{ib})$ *and* $Var_D(Y_{jb})$ are respectively defined as

$$Var_N(X_{ib}) = \frac{\sum_{j=1}^{m_b} I\left(X_{bi}, Y_{bj}\right)}{m_b - 1} and \; Var_D(X_{ib}) = \frac{\sum_{i=1}^{n_b} I\left(X_{bi}, Y_{bj}\right)}{n_b - 1}, b = 1,2. \qquad 14$$

Where

$$I\left(X_{ib}, Y_{jb}\right) = \begin{cases} 1 \; if \; Y > X \\ 0.5 \; if \; Y = X \\ 0 \; if \; Y < X \end{cases}$$

$$AUC_b = \frac{1}{m_b}\sum_{j=1}^{m_b} Var_D(Y_{bj}) = \frac{1}{n_b}\sum_{i=1}^{n_b} Var_N(X_{bi}), b = 1,2. \qquad 15$$

Note here that Y_{bj} *and* X_{bi} are the observed diagnostic test results for the *jth and ith* subjects in group b diagnostic test procedures that are diseased and non-diseased respectively.

Also

$$Cov\left(AUC_1, AUC_2\right) = \frac{S_{T_{D_1}T_{D_2}}}{m} + \frac{S_{T_{N_1}T_{N2}}}{n} \qquad 16$$

Where

$$S_{T_{D_1}T_{D_2}} = \frac{1}{m-1}\sum_{j=1}^{m}\left[Var_D(Y_{1j}) - AUC_1\right]\left[Var_D(Y_{2j}) - AUC_2\right]$$

And

$$S_{T_{N_1}T_{N_2}} = \frac{1}{n-1}\sum_{i=1}^{n}\left[Var_N(X_{1i}) - AUC_1\right]\left[Var_N(X_{2i}) - AUC_2\right]$$

Here $S_{T_{D_1}T_{D_2}}$ is the pooled variances of diseased test result for the first and second diagnostic test procedure or process, $S_{T_{N_1}T_{N_2}}$ is the pooled variances of the non-diseased test result for the first and second diagnostic test process or procedure, $Var_D(Y_{1j})$ is the variance of the positive diagnostic test result for the j^{th} subject in the first diagnostic test process, $Var_D(Y_{2j})$ is the variance of the positive diagnostic test result for the j^{th} subject in the second diagnostic test process, $Var_N(X_{1i})$ is the variance of the negative diagnostic test result for the i^{th} subject in the first diagnostic test process and $Var_N(X_{2i})$ is the variance of the negative diagnostic test result for the i^{th} subject in the second diagnostic test process. When the variances are estimated, one can calculate the AUC for the two diagnostic tests and then make comparison.

Bandos et al permutation nonparametric test for comparing AUCs

Bandos et al.,[6] derived exact and asymptotic permutation test methods to test the equality of two correlated ROC curves which are designed to have increased power to detect difference in the AUC. The test of Bandos et al.,[6] directly tests for an equality of AUCs. This approach implicitly assumes that both diagnostic test procedures

are exchangeable within subject and requires an appropriate transformation, such as ranks, for diagnostic test procedures differing in scale. Bandos et al.,[6] compared the performance of their test to that of DeLong et al.,[3] via simulation and found that the permutation test had greater power than the nonparametric test developed by DeLong et al.,[3] when there was moderate correlation between diagnostic tests, large AUCs, and small sample sizes. Bandos et al.,[6] test is limited by the fact that it requires the exchangeability of the diagnostic test procedures and do requires also the transformations of the original data. It also requires diagnostic tests that are measured on identical scales and so may prove to be less powerful in settings in which the diagnostic test results are skewed Braun & Alonzo.[11] If $\left\{X_i^b\right\}_{i=1}^n, \left\{Y_j^b\right\}_{j=1}^m$ be the test results of the diagnostic procedure b for n actually non-diseased and m actually diseased subjects and $\left\{x_i^b\right\}_{i=1}^n, \left\{y_j^b\right\}_{j=1}^m$ be approximately transformed test results, an unbiased nonparametric estimator for the AUC for diagnostic procedure or test b can be written as $A\hat{U}C_b$. For a paired sample design, the difference in two AUCs can be estimated as,

$$A\hat{U}C_1 - AU\hat{C}_2 = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\psi\left(X_i^1, Y_j^1\right)}{nm} - \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\psi\left(X_i^2, Y_j^2\right)}{nm} \quad 17$$

Where

$$\psi\left(X_i^1, Y_j^1\right) - \psi\left(X_i^2, Y_j^2\right) = \begin{cases} 1, if\ x_i^1 < y_j^1, x_i^2 > y_j^2 \\ 0.5, if\ x_i^1 < y_j^1, x_i^2 = y_j^2\ or\ x_i^1 = y_j^1, x_i^2 > y_j^2 \\ 0, if\ x_i^1 < y_j^1, x_i^2 < y_j^2\ or\ x_i^1 > y_j^1, x_i^2 > y_j^2 or\ x_i^1 = y_j^1, x_i^2 = y_j^2 \\ -0.5, if\ x_i^1 > y_j^1, x_i^2 = y_j^2\ or\ x_i^1 = y_j^1, x_i^2 < y_j^2 \\ -1, if\ x_i^1 > y_j^1, x_i^2 < y_j^2 \end{cases}$$

Being a member of U statistics, the non-parametric estimator of the AUC difference is known to be asymptotically normally distributed under quite general condition Hoeffding.[4] Based on this property and the additional assumption of exchangeability, they constructed a simple asymptotic test procedure with test statistic

$$\frac{A\hat{U}C_1 - A\hat{U}C_2}{\sqrt{Var_\Omega\left(A\hat{U}C_1 - A\hat{U}C_2\right)}} \xrightarrow{d} N(0,1) \quad 18$$

Where Ω is the parameter space.

Sumi et al (McNemar Test) nonparametric method for comparing AUCs

Sumi et al.,[7] proposed a method for comparing two proportion of positive responses. This test is based on McNemar.,[8] for the comparison of two diagnostic tests for continuous and discrete binary scale data that are matched. Their McNemar[8] test is based on the comparison of the equality of the proportion of positive responses in two diagnostic tests. Here each subject's test result is either positive coded 1 or negative coded 0 on each of two diagnostic processes and interest is in testing whether the proportion of 'positive' responses are the same on the first and second diagnostic procedure taken into account the correlation of the two diagnostic test results. This test is limited by the fact that it does not provide evidence of inferiority or superiority of one diagnostic test over another. Any test capable of this should have one sided alternative hypothesis Zhou et al.,[18] The test assumes the use of summarized data which leads to loss of information and reliability in decisions about the data analyzed. Such summarized data could have many ties and if not adjusted for will reduce the power of any test statistic employed for the analysis. It is worthy of mentioning that McNemar[8] test is concerned with matched pairs of dichotomous test results. Here the result of each diagnostic test are all into two categories, positive coded 1 and negative coded 0.The resulting data is presented in a 2x2 contingency table where row represents the result of one diagnostic test while the column represent the result of another diagnostic test. Here each cell represents the number of observed cases with the particular combination of test results. Depending on the scale of measurement of test results whether continuous or binary, one can compare the two test procedures by constructing a 2x2 contingency table after which McNemar[8] test can be applied and the result compared with the result obtained using the conventional non-parametric test suggested by DeLong et al.,[3] and the permutation test by Bandos et al.,[6] For two diagnostic tests producing the continuous test results as $\left\{X_i^b\right\}_{i=1}^n$ *and* $\left\{Y_j^b\right\}_{j=1}^m$ in the bth diagnostic test, the subjects are ordered so that $\left\{X_i^b\right\}_{i=1}^n$ *and* $\left\{Y_j^b\right\}_{j=1}^m$ becomes the transferred results in the bth diagnostic test for n real negative and m real positive subjects. Suppose we have an optimal cut-off value of c_b for bth diagnostic test, then we classify all results above c_b as positive and results less than or equal to c_b as negative so that the 2x2 contingency table can be constructed for each diagnostic procedure. The resulting table 1 is From Table 1, A_b =number of subjects who are diseased and who actually tested positive $(y_j^b > c_b)$, B_b =number of subjects who do not have disease and actually tested positive $(x_i^b > c_b)$, C_b =number of subjects with disease and actually tested negative $(y_j^b \le c_b)$, D_b c= number of subjects without disease who actually tested negative $(x_i^b \le c_b)$. Now each diagnostic test result is used to obtain a 2x2 contingency table based on the optimal cut-off value, so that one can verify if the diagnostic test procedure has any effect on the true observed (True) status. To test for the significance of any observed change using the McNemar8 test, one sets up a fourfold table of frequencies representing the first and the second sets of responses (test results) from the same subjects. If both diagnostic test procedures have significant effects, in other words, there are correlated, we can combine the two diagnostic test procedures thus obtaining a matched pair data from the combination of these two diagnostic tests and we obtain a contingency Table 2.

Table 1 A 2x2 contingency table for bth (b=1, 2) diagnostic test procedure

Test result for diagnostic procedure	Observed (True) status		Total
	Nondiseased(-)	*Nondiseased(-)*	
positive(+*ve*)	A_b	B_b	$A_b + B_b$
negative(−*ve*)	C_b	D_b	$C_b + D_b$
Total	$A_b + C_b$	$B_b + D_b$	n_b

In Table 2, P_A represents probability of positive test results on both test procedures, P_B is the probability of positive test result in diagnostic test procedure 1 but negative test result in diagnostic test procedure 2, P_C *and* P_D are similarly defined. A, B, C, and D are the corresponding frequencies representing test results on both diagnostic tests.For instance, A represents the frequency that diagnostic test 1 and diagnostic test 2 subjects both respond positive while D represents the frequency that diagnostic test 1 and diagnostic test 2 subjects both respond negative and n represents the pairs of diagnostic test 1- diagnostic test 2 subjects studied. From Table 2, the proportion of diagnostic test 1 subjects studied who respond positive is

Citation: Marius OU, Happiness OI. An extended McNemar test for comparing correlated proportion of positive responses. *Biom Biostat Int J.* 2019;8(4):125–137. DOI: 10.15406/bbij.2019.08.00281

Table 2 A 2x2 contingency table for two diagnostic test procedures

Diagnostic test 2			Total
Diag test I	**Positive(+ve)**	**Negative(-ve)**	
Positive($+ve$)	A (P_A)	B (P_B)	$A+B$ (P_A+P_B)
Negative($-ve$)	C (P_C)	D (P_D)	$C+D$ (P_C+P_D)
Total	$A+C$ (P_A+P_C)	$B+D$ (P_B+P_D)	N (1)

$$p_1 = \frac{A+C}{N} \qquad 19$$

while the proportion of diagnostic test 2 studied who respond positive is

$$p_2 = \frac{A+B}{N} \qquad 20$$

The difference between the proportions of diagnostic test 1 and diagnostic test 2 subjects who respond positive is

$$p_2 - p_1 = \frac{B+C}{N} \qquad 21$$

which is independent of A and D, the number of test results in which the diagnostic test 1 and diagnostic test 2 subjects both respond positive or both respond negative respectively.

The standard error of the difference between the two proportions of positive responses is

$$Se(p_2 - p_1) = \frac{\sqrt{B+C}}{N} \qquad 22$$

which is also unaffected by A and D

If π_1 *and* π_2 are respectively the proportions of diagnostic test 1 and diagnostic test 2 in the sampled populations who respond positive then a null hypothesis that may be of interest is whether the two diagnostic test procedures are equal in their performances as

$$H_0 : \pi_2 - \pi_1 = 0 \; versus \; H_1 : \pi_2 - \pi_1 \neq 0 \qquad 23$$

Its equivalent is to test whether the marginal probabilities of positive result on the diagnostic test 1 and diagnostic test 2 Sumi et al.,[7] based on Table 2 are equal

$$H_0 : P_A + P_B = P_A + P_C \; versus \; H_1 : P_A + P_B \neq P_A + P_C \qquad 24$$

The McNemar test statistic (1947) follows a chi-square distribution with 1 degree of freedom for testing the null hypothesis of Equ.23/24 is

$$\chi^2 = \left(\frac{(p_2 - p_1)}{Se(p_2 - p_1)}\right)^2 = \frac{(B-C)^2}{B+C} (without \; continuity \; correction) \qquad 25$$

$$\chi^2 = \frac{\left(|B-C|-1\right)^2}{B+C} (with \; continuity \; correction) \qquad 26$$

which has a chi-square distribution with 1 degree of freedom. The null hypothesis of equal population proportions is rejected at the α level of significance in favour of the alternative hypothesis if

$$\chi^2 \geq \chi^2_{1-\alpha;1} \qquad 27$$

McNemar test used here employs a continuous distribution to approximate a discrete probability distribution by recommending for continuity for correction in calculating the test statistic. When the sample size is small in the interest of accuracy, the exact binomial probability for the data should be used Sumi et al.,[7] McNemar test unlike the DeLong et al.,[3] and Bandos et al.,[6] methods is applicable both for continuous and discrete binary scale data irrespective of having knowledge of true disease status (gold standard).

The identified problem statement associated with this study is that the usual McNemar[8] test cannot adjust for the possible presence of ties in data, thereby making the variance value high while the chi-square value remained low such that Type II error is often times committed. To be able to solve this problem, this study is aimed at comparing correlated proportion of positive responses in two diagnostic test procedures by extending the usual McNemar test statistic to accommodate for ties in the data.

Extended McNemar test

This extension is based on the previous work by Sumi et al.,[7] who applied the usual McNemar[8] in comparing correlated marginal probability of positive responses from two diagnostic test procedures. The usual McNemar[8] test assumes that the data to be used are presented in a summarized form rather than being in a raw form that needs to be processed. Most times, these data may be quantitative in nature and as such may be continuous also meaning that the chances of getting any tied data is at least zero in theory but practically, there exist ties in the data. This is one of the limitations of the usual McNemar test that needs attention. It these ties are adjusted for, the power of the test statistic used for data analysis is increased. To extend the usual McNemar test adopted by Sumi et al.,[7] to allow for the possible presence of ties in the data, let (t_{v2}, t_{v1}) be the test results of subjects from diagnostic test 2 case 1 respectively for the vth pair of subjects who are undergoing diagnostic test 2 and 1 say respectively where v=1,2,..,N pairs of subjects in diagnostic test 2 and 1.Assuming that the data is measured on at least interval scale.

$$Let \; T_v = \begin{cases} 1, if \; t_{v2} \; and \; t_{v1} \; are \; test \; results \; of \; subjects \; for \; diagnostic \; test \; 2 \; and \; 1 \; responding \; positive \; and \; negative \; to \; the \; condition \; respectively \\ 0, if \; t_{v2} \; and \; t_{v1} \; are \; test \; results \; of \; subjects \; for \; diagnostic \; test \; 2 \; and \; 1 \; responding \; both \; positive \; or \; responding \; both \; negative \\ -1, if \; t_{v2} \; and \; t_{v1} \; are \; test \; results \; of \; subjects \; for \; diagnostic \; test \; 2 \; and \; 1 \; responding \; negative \; and \; positive \; to \; the \; condition \; respectively \end{cases} \qquad 28$$

For the v[th] pair of subjects in diagnostic test 2 and 1, where v=1,2,..,N,where N is the total number of pairs.

Let

$$\pi^+ = P(T_v = 1) : \pi^0 = P(T_v = 0); and \; \pi^- = P(T_v = -1) \qquad 29$$

Where

$$\pi^+ + \pi^0 + \pi^- = 1 \qquad 30$$

Therefore let

Citation: Marius OU, Happiness OI. An extended McNemar test for comparing correlated proportion of positive responses. *Biom Biostat Int J.* 2019;8(4):125–137. DOI: 10.15406/bbij.2019.08.00281

$$W = \sum_{v=1}^{N} T_v \qquad 31$$

Where W is the total number of subjects in the matched pairs of subjects who test or respond positive. Based on the above specifications, the expected value of T_v is

$$E(T_v) = \pi^+ - \pi^- \qquad 32$$

While

$$Var\,(T_v) = \pi^+ + \pi^- - (\pi^+ - \pi^-)^2 \qquad 33$$

From equations 6 and 7, expected value of W is

$$E(W) = n(\pi^+ - \pi^-) \qquad 34$$

Adding from equation 8

$$Var(W) = N(\pi^+ + \pi^- - (\pi^+ - \pi^-)^2) \qquad 35$$

Note that $\pi^+, \pi^0 and\ \pi^-$ are respectively the probabilities that for a randomly selected pair of subjects from diagnostic tests 2 and 1, the subjects from diagnostic test 2 on the average responds positive and the subjects from diagnostic test 1 responds negative or the subjects from diagnostic test 2 and 1 both respond positive or the subjects from both diagnostic tests respond negative, or the subjects from diagnostic test 2 responds negative and subjects from diagnostic test 1 responds positive. The sample estimates of these probabilities are respectively defined as

$$\hat{\pi}^+ = \frac{p^+}{N}; \hat{\pi}^0 = \frac{p^0}{N} and\ \hat{\pi}^- = \frac{p^-}{N} \qquad 36$$

where $p^+, p^0\ and\ p^-$ represents respectively the frequencies 1's,0's and -1's in the distribution given in $T_v, v = 1,2,...,N$. That is, $p^+, p^0\ and\ p^-$ are respectively the number of diagnostic test 2 and 1 subject pairs in which the diagnostic test 2 respond positive and the diagnostic test 1 respond negative or the diagnostic test 2 and 1 subjects both respond positive or both respond negative or the diagnostic test 2 responds negative and the diagnostic test 1 subject responds positive. These frequencies are expressed in terms of diagnostic tests 2 and 1 in Table 3.

Table 3 Fourfold Table for presenting Data on paired samples

Diagnostic test 2			**Total**
Diag test 1	**Positive Response (+ve)**	**Negative Response(-ve)**	
Positive Response (+ve)	$n_{11} = p^{0+} = A$	$n_{12} = p^+ = B$	$n_{11} + n_{12} = A + B$
Negative Response (−ve)	$n_{21} = p^- = C$	$n_{22} = p^{0-} = D$	$n_{21} + n_{22} = C + D$
Total	$n_{11} + n_{21} = A + C$	$n_{12} + n_{22} = B + D$	$n_{..}(= N)$

There are respectively represented from Table 3 as

$$p^+ = n_{12}; p^0 = n_{11} + n_{22} = p^{0+} + p^{0-}; p^- = n_{21} \qquad 37$$

Where

$$p^{0+} = n_{11}; p^{0-} = n_{22} \qquad 38$$

are respectively the number of diagnostic test 2 and 1 subject pairs where diagnostic test 2 and 1 subjects both respond positive or both respond negative and $\hat{\pi}^{0+}\ and\ \hat{\pi}^{0-}$ are the corresponding relative frequencies.

But $\pi^+ - \pi^-$ measures the difference in rate of positive responses by subjects in the diagnostic test 2 and diagnostic test 1 procedure and its estimate of the sample is

$$\pi^+ - \pi^- = \frac{W}{N} = \frac{p^+ - p^-}{N} \qquad 39$$

And the variance is estimated from Equ 35 as

$$Var(\hat{\pi}^+ - \hat{\pi}^-) = \frac{Var(W)}{N^2} = \frac{\hat{\pi}^+ + \hat{\pi}^- - (\hat{\pi}^+ - \hat{\pi}^-)^2}{N} \qquad 40$$

But the McNemar test statistic is $\chi^2 = \left(\frac{(p_2 - p_1)}{Se(p_2 - p_1)}\right)^2 = \frac{(B-C)^2}{B+C}$ with the numerator given as

$$W^2 = \left(N(\hat{\pi}^+ - \hat{\pi}^-)\right)^2 = (p^+ - p^-)^2 \qquad 41$$

Now a test statistic explaining the difference between positive response rates for diagnostic test 2 and 1 subjects can be developed by noting that π^+ represents the proportion of pairs of subjects out of a total of N pairs in which the subject from diagnostic test 2 procedure and was given say T_2 treatment in a given pair responds positive and the subject from diagnostic test 1 in the pair and given treatment T_1 say, responds negative; π^0 represents the proportion of the total number of N pairs of subjects with the members of the pair both responding positive or both responding negative and π^- is the proportion of pairs out of a total of 'N' pairs in which the subject from diagnostic test 2 procedure and was given say T_2 in a given pair responds negative and the subject from diagnostic test 1 in the pair and given treatment T_1 responds positive. The diagnostic test 2 and 1 differential positive response rate is given as $\pi^+ - \pi^-$ with their sample estimate and variance given respectively by Eqns 39 and 40. If the sampled proportion is given respectively as $p_1 = \frac{A+C}{N}\ and\ p_2 = \frac{A+B}{N}$ based on Table 1, we obtain more important and detailed information given as

$$P_1 = \frac{n_{11} + n_{21}}{N} = \frac{p^{0+} + p^-}{N} = \hat{\pi}^{0+} + \hat{\pi}^- \qquad 42$$

And

$$P_2 = \frac{n_{11} + n_{12}}{N} = \frac{p^{0+} + p^+}{N} = \hat{\pi}^{0+} + \hat{\pi}^+ \qquad 43$$

$$where\ \hat{\pi}^{0+} = \frac{p^{0+}}{N}\ and\ \hat{\pi}^{0-} = \frac{p^{0-}}{N} \qquad 44$$

$$such\ that\ \hat{\pi}^{0} = \hat{\pi}^{0+} + \hat{\pi}^{0-} \qquad 45$$

Now the null hypothesis H_0 of interest is to test that the proportions of subjects responding positive in the diagnostic test 2 and 1 procedures or treatment conditions T_2 and T_1 differ by some value β_0 .This is equivalent to testing the null hypothesis given as

$$H_0 : \pi^{+} - \pi^{-} = \beta_0 \ versus\ H_1 : \pi^{+} - \pi^{-} \neq \beta_0 \ (-1 \leq \beta_0 \leq 1) \qquad 46$$

While the test statistic is given by

$$\chi^2 = \frac{(W - n\beta_0)^2}{N\left(\hat{\pi}^{+} + \hat{\pi}^{-} - (\hat{\pi}^{+} - \hat{\pi}^{-})^2\right)} \qquad 47$$

Or equivalently

$$\chi^2 = \frac{n((\hat{\pi}^{+} - \hat{\pi}^{-}) - \beta_0)^2}{\hat{\pi}^{+} + \hat{\pi}^{-} - (\hat{\pi}^{+} - \hat{\pi}^{-})^2} \qquad 48$$

which with 1 degree of freedom is approximately chi-square distributed for sufficiently large 'n'. The null hypothesis of equal population proportion of positive responses is rejected at the α level of significance in favour of the alternative hypothesis if

$$\chi^2 \geq \chi^2_{1-\alpha;1} \qquad 49$$

Note therefore that under null hypothesis H_0, the numerators of the extended test statistic of Equs 47 and 48 are as in the usual McNemar[8] test statistic independent of $n_{11} = p^{0+} and\ n_{22} = p^{0-}$ the number of pairs in which diagnostic test 2 and 1 subjects in each pair both respond positive or both respond negative to the conditions of interest while for equations 47 and 48, the denominator is also independent of n_{11} and n_{22}.Hence both the extended test statistic and the usual McNemar[8] test statistic are not affected by those pairs in which the subjects in each pair both respond positive or both respond negative to the disease or treatments condition. Unlike the usual McNemar test statistic, the extended McNemar[8] test has by specifications been adjusted and corrected for the possible presence of ties in the data. In addition, the variance of the extended McNemar test statistic in Eqn 48 is smaller than the variance of the usual McNemar test statistic stated in between eqns 40 and 41.This is because of the fact that $Var(\hat{\pi}^{+} - \hat{\pi}^{-}) = \frac{Var(W)}{N^2} = \frac{\hat{\pi}^{+} + \hat{\pi}^{-} - (\hat{\pi}^{+} - \hat{\pi}^{-})^2}{N}$ and $Se(p_2 - p_1) = \frac{\sqrt{B + C}}{N}$ so that

$$Var(\hat{\pi}^{+} - \hat{\pi}^{-}) = \frac{\hat{\pi}^{+} + \hat{\pi}^{-} - (\hat{\pi}^{+} - \hat{\pi}^{-})^2}{N} = \frac{n_{12} + n_{21}}{N^2} - \frac{(n_{12} + n_{21})^2}{N^3} \leq \frac{n_{12} + n_{21}}{N^2} = Var(P_2 - P_1),$$

$$since\ \frac{(\hat{\pi}^{+} - \hat{\pi}^{-})^2}{N} = \frac{(n_{12} + n_{21})^2}{N^3} \geq 0, for\ all\ \pi^{+} \neq \pi^{-}, or\ n_{12} \neq n_{21}$$

In conclusion, the extended McNemar test statistic is relatively more efficient and so is most likely to be more powerful than the usual McNemar test statistic whenever the diagnostic test 2 and 1 test results of subjects have differences in positive response rates $(\hat{\pi}^{+} \neq \hat{\pi}^{-}; or\ P_1 \neq P_2)$ to the conditions of interest. It is note worthy that $N(\hat{\pi}^{+} - \hat{\pi}^{-})^2$ is the reduced value in the variance of W since by specifications of equation 28 it has been adjusted for the possible presence of ties between the responses of diagnostic test 2 and 1 procedures. The major difference between the usual McNemar[8] test and the extended McNema[8] test is that there is adjustment of possible presence of tied observations in the later test, the extended McNemar[8] test statistic will likely have smaller variance and larger calculated chi-square value than the usual McNemar[8] test statistic, thus leading to the more chances of committing Type II error in the usual McNemar[8] test more often than in the extended McNemar[8] test.

Application to simulation study

We carried out computer simulations here to evaluate the performance of the extended McNemar test. We performed extensive simulations to evaluate and compare Type I errors (empirical test sizes) and statistical power of the extended McNemar[8] test, usual (traditional) McNemar test, conventional nonparametric test of DeLong et al.,[3] and asymptotic test of Bandos et al.,[6] Here we assumed equal correlation coefficient across the two diagnostic test procedures for diseased and non-diseased test results of subjects measured on continuous and discrete binary scales and the sample sizes are 20,60,100 and 180. These test results of subject were generated from a standard bi-variate normal distribution having mean and variance respectively for the two diagnostic tests as $\mu_1, \sigma_1^2 \ and\ \mu_2, \sigma_2^2$ when measurement of data is on continuous scale. The AUC for diagnostic test 1 and 2 procedures are respectively given as $AUC_1 = \Phi\left(\frac{\mu_1}{\sqrt{(1 + \sigma_1^2)}}\right) and\ AUC_2 = \Phi\left(\frac{\mu_2}{\sqrt{(1 + \sigma_2^2)}}\right)$ where Φ is the standard normal cumulative distribution function. Under binary random variable X for one diagnostic test procedure, if the test result of subject is positive, it is coded 1 and if the test result is negative, it is coded 0. If binary variables (X,Y) is assumed for correlated diagnostic test procedures, the joint distribution of X and Y is determined. The correlation coefficient (r) of X and Y is determined and having the range $-1 \leq r \leq +1$. For data on binary scale of measurement, correlated binary test results were generated with required probabilities of positive responses (P_1, P_2) to obtain specific difference $\left(\pi^{+} - \pi^{-} = \beta_0\ or\ \pi_b - \pi_t \geq \beta_0\right)$ between the probability of positive responses for the two diagnostic test procedures for the extended McNemar test and the proposed chi-square test respectively. The binary test results for the non-diseased subjects, are generated by fixing the probability of positive responses as 0.30 and 0.35. This procedure of simulating binary data is in line with the previous works of Leisch et al.,[19] and Islam et al.,[20] who discussed the algorithm for simulating correlated binary test results. The SAS version 9 is the statistical software used to perform the simulation study.

The range of values of the correlation coefficient r for the extensive simulation for continuous test results and values of parameters (a and b) for estimating mean (μ_1, μ_2) and variance (σ_1^2, σ_2^2) parametrically as drawn to obtain the difference between two AUCs ranges from 0 to 0.3. For binary test results, the correlation coefficient r is also taken to range from 0.25 to 0.75 and the probability of positive responses were drawn so as to obtain the difference between probability of positive responses of subjects for the two diagnostic test procedures and it ranged from 0 to 0.2. For either binary or continuous scenario considered, we used 2000 replications in running the simulations. Table 5 compares the empirical test size (Type I error) and the statistical power of the extended McNemar[8] test to the usual McNemar[8] test proposed by Sumi et al.,[7] to the conventional nonparametric test developed by DeLong et al.,[3] and to asymptotic permutation test developed by Bandos et al.,[6] for comparing two diagnostic test procedures for continuous test results. This comparison was similarly carried out for binary test results. The estimates of Type I error as well as estimates of the statistical power are obtained when

the proportion of positive responses or the true AUCs for the two diagnostic test procedures are the same and different respectively as can be seen in Table 5&6. The rejection regions for the two tests are determined using 5% as level of significance.

For smaller AUCs, the extended McNemar test indicates a more conserved empirical test size (type I error) and thereafter an increased statistical power when compared to the traditional McNemar[8] test by Sumi et al.,[7] conventional nonparametric method by DeLong et al.,[3] and asymptotic permutation test by Bandos et al.,[6] when the test results is continuous. But when the correlation coefficient is moderate and for increased sample size for the two diagnostic test procedure, stability appears to be more in the scenario considered (continuous case) and the five tests mentioned above tends to be very close in terms of their empirical test size and statistical power. The extended McNemar[8] test shows more false positive rate (FPR) when the correlation coefficient r is smaller. This is because the McNemar[8] test are most suitably used when the data is correlated. However, when the correlation coefficient r is increased, the estimate of FPR reduces drastically. In the same way, when the AUCs is increased, the estimates of the empirical test size (type I error) for every sample sizes and all values of correlation coefficients can be compared. The extended McNemar test discriminates better than the traditional McNemar test by Sumi et al.,[7] conventional nonparametric test by DeLong et al.,[3] and the permutation test Bandos et al.,[6] when the AUCs are getting higher and for lower values of correlation coefficients.

When the AUCs values are high and for moderate values of the correlation coefficient, the other three tests namely the usual McNemar[8] test by Sumi et al.,[7] test by DeLong et al.,[3] and test by Bandos et al.,[6] gives better statistical power than the extended McNemar test but when the sample sizes increases, the extended McNemar[8] test provides very close statistical power to the others. In considering the binary test results in all aspects of parameter settings and for either big or small sample sizes, the extended McNemar[8] test shows lower conservative empirical test size (Type I error) and shows higher statistical power when compared to tests by Sumi et al.,[7] DeLong et al.,[3] and Bandos et al.,[6] Finally, in the continuous case situation, the results of the simulation shows that the proposed chi-square test and the extended McNemar[8] test gives very close harmony of Type I error to the significant level α but when the values of AUCs are low this harmony yields or provides among the diagnostic test procedures moderate and very high correlation coefficient. Also having greater or higher sample sizes in the continuous case also makes the extended McNemar[8] test have statistical power that is very comparable to other existing nonparametric methods of comparing correlated AUCs. In addition, for the discrete binary case, the extended McNemar[8] test possesses higher operating characteristics than other existing tests considered in all the settings of parameter. The performance of the extended McNemar[8] test may be impaired in a simulation study when the test result is continuous because of the problem of choosing or finding an optimal cut-off value for classifying the test results of subjects. To make this point clearer, we in the next section will adopt a known standard data set that already has a real cut-off value and we will conduct a bootstrap power analysis so as to compare the statistical power of all the four tests namely, extended McNemar[8] test, usual McNemar[8] test by Sumi et al.,[7] conventional test by DeLong et al.,[3] and permutation test by Bandos et al (Table 4&5).

Table 4 Empirical type I error and statistical power when comparing two diagnostic tests for continuous test results. [- Area of diagnostic test 1; - Area of diagnostic test 2; D, DeLong et al.,[3] Test; B, Bandos et al.,[6] Test; S, Sumi et al.,[7] Test; EM, Extended McNemar[8] Test]

AUC	Mean		Variance	Sample size		$\tilde{n} = 0.25$				$\tilde{n} = 0.50$				$\tilde{n} = 0.75$			
AUC_1 AUC_2	μ_2	μ_2	$\sigma_1^2 = \sigma_2^2$	N	M	D[a]	B[b]	S[c]	EM[d]	D[a]	B[b]	S[c]	EM[d]	D[a]	B[b]	S[c]	EM[d]
Type I error and statistical power																	
0.6, 0.7 Type I error	.38	.38	1.0	20	20	.049	.040	.065	.069	.048	.044	.059	.061	.051	.052	.050	.049
				60	60	.045	.043	.072	.080	.047	.048	.054	.067	.050	.050	.048	.049
				100	100	.058	.057	.095	.096	.040	.040	.061	.063	.045	.045	.056	.057
				140	140	.047	.047	.087	.091	.043	.042	.083	.086	.043	.042	.076	.083
				180	180	.043	.042	.097	.099	.042	.042	.072	.078	.046	.046	.071	.080
0.6, 0.8 Power	.38	.76	1.0	20	20	.121	.090	.183	.189	.171	.162	.204	.240	.225	.214	.199	.209
				60	60	.188	.177	.334	.357	.297	.287	.387	.398	.397	.386	.453	.462
				100	100	.229	.085	.458	.472	.449	.439	.553	.572	.587	.575	.632	.641
				140	140	.441	.430	.678	.692	.637	.628	.781	.796	.800	.791	.876	.886
				180	180	.608	.604	.841	.855	.808	.801	.914	.935	.936	.932	.962	.978
0.6, 0.9 Power	.38	1.23	1.0	20	20	.404	.364	.468	.472	.570	.523	.558	.576	.723	.626	.603	.589
				60	60	.705	.678	.803	.825	.870	.838	.883	.898	.955	.942	.926	.918
				100	100	.682	.849	.939	.952	.975	.967	.978	.989	.997	.991	.990	.997
				140	140	.978	.976	.995	.898	.998	.998	.998	.998	1.000	1.000	1.000	1.000
				180	180	.996	.996	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Citation: Marius OU, Happiness OI. An extended McNemar test for comparing correlated proportion of positive responses. *Biom Biostat Int J.* 2019;8(4):125–137. DOI: 10.15406/bbij.2019.08.00281

Table Continued

AUC	Mean		Variance	Sample size		$\tilde{n} = 0.25$				$\tilde{n} = 0.50$				$\tilde{n} = 0.75$			
AUC_1 AUC_2	μ_2	μ_2	$\sigma_1^2 = \sigma_2^2$	N	M	D[a]	B[b]	S[c]	EM[d]	D[a]	B[b]	S[c]	EM[d]	D[a]	B[b]	S[c]	EM[d]
Type I error and statistical power																	
0.7 , 0.8 Power	.38	1.84	1.0	20	20	.816	.766	.762	.778	.938	.903	.835	.907	.985	.968	.883	.878
				60	60	.990	.983	.982	.986	.998	.998	.991	.996	1.000	1.000	.998	.998
				100	100	.998	.997	.999	.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
				140	140	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
				180	180	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.7, 0.9 Type I error	.79	.80	1.0	20	20	.047	.041	.048	.049	.046	.044	.039	.049	.049	.051	.029	0.019
				60	60	0.40	.038	.044	.048	.049	.048	.047	.057	.050	.049	.041	.032
				100	100	.061	.057	.047	.058	.036	.035	.046	.048	.046	.048	.039	.028
				140	140	.033	.033	.065	.072	.051	.050	.051	.056	.049	.049	.050	.042
				180	180	.048	.048	.064	.066	.041	.041	.051	.048	.054	.054	.049	.047
0.7, 0.9 Power	.79	1.25	1.0	20	20	.136	.125	.117	.123	.196	.186	.128	.120	.253	.245	.150	.140
				60	60	.231	.220	.228	.236	.350	.339	.271	.243	.470	.459	.324	.309
				100	100	.362	.348	.353	.361	.526	.512	.417	.401	.678	.668	.493	.417
				140	140	.561	.551	.576	.583	.744	.733	.679	.662	.870	.858	.769	.703
				180	180	.729	.723	.755	.763	.903	.898	.669	.619	.969	.966	.911	.879
0.8, 0.9 Power	.79	1.85	1.0	20	20	.531	.497	.356	.462	.696	.656	.414	.389	.824	.780	.467	.412
				60	60	.857	.832	.693	.721	.959	.946	.778	.742	.990	.980	.841	.810
				100	100	.953	.943	.892	.898	.995	.993	.929	.911	1.000	.998	.969	.931
				140	140	.998	.997	.984	..998	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
				180	180	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

D[a], Conventional AUC DeLong et al.,[3]; B[b], Approximation to permutation AUC test Bandos et al.,[6]; M[c], McNemar[8] testSumi et al.,[7]; EM[d], Extended McNemar[8] test (new method)

Table 5 Empirical type I error and statistical power when comparing two diagnostic tests for discrete binary test results. [- Area of diagnostic test 1; - Area of diagnostic test 2; D, DeLong et al.,[3] Test; B, Bandos et al.,[6] Test; S, Sumi et al.,[7] Test; EM, Extended McNemar[8] Test]

Probability of positive response			Sample size		$\rho = 0.25$				$\rho = 0.50$				$\rho = 0.75$			
p_1	p_2	$\pi^+ - \pi^-$	N	M	D[a]	B[b]	S[c]	EM[d]	D[a]	B[b]	S[c]	EM[d]	D[a]	B[b]	S[c]	EM[d]
Type I error and statistical power																
0.60 Type I error	0.60	0.00	20	20	.065	.059	.027	.019	.077	.062	.022	.017	.071	.054	.024	.018
			60	60	.059	.054	.038	.027	.069	.068	.039	.024	.069	.065	.048	.032
			100	100	.068	.066	.048	.034	.080	.074	.056	.037	.093	..092	.059	.037
			140	140	.084	.081	.068	.047	.097	.095	.080	.062	.107	.104	.091	.078
			180	180	.115	.112	.079	.058	.124	.122	.093	.056	.135	.132	.120	.96
0.60 Power	0.70	0.10	20	20	.061	.054	.062	.051	.062	.049	.069	.078	.076	.052	.080	.093
			60	60	.071	.064	.131	.109	.073	.069	.146	.163	.075	.068	.220	.287
			1000	100	.079	.069	.204	.178	.076	.068	.248	.287	.097	.094	.336	.413
			140	140	.089	.087	.298	.242	.092	.087	.380	.421	.117	.109	.559	.624
			180	180	.102	.098	.439	.217	.112	.110	.557	.734	.146	.140	.764	.813

Citation: Marius OU, Happiness OI. An extended McNemar test for comparing correlated proportion of positive responses. *Biom Biostat Int J.* 2019;8(4):125–137. DOI: 10.15406/bbij.2019.08.00281

Table Continued

Probability of positive response			Sample size		$\rho = 0.25$				$\rho = 0.50$				$\rho = 0.75$			
p_1	p_2	$\pi^+ - \pi^-$	N	M	D[a]	B[b]	S[c]	EM[d]	D[a]	B[b]	S[c]	EM[d]	D[a]	B[b]	S[c]	EM[d]
Type I error and statistical power																
0.60 Power	0.80	0.20	20	20	.112	.102	.147	.181	.146	.106	.183	.192	.184	.140	.236	.261
			60	60	.182	.165	.343	.479	.231	.213	.408	.524	.303	.268	.584	.692
			100	100	.243	.222	.510	.611	.320	..293	.609	.741	.445	.422	.794	.847
			140	140	.376	.263	.719	.876	.489	.459	.846	.919	.643	.609	.959	.980
			180	180	.521	.497	.907	.968	.625	.603	.960	.987	.806	.787	.893	.907
0.70 Power	0.7	0.00	20	20	.065	.056	.031	.027	.071	.057	.023	.019	.069	.060	.024	.020
			60	60	.057	.054	.042	.035	.059	.058	.041	.017	.066	.060	.048	.034
			100	100	.076	.071	.055	.036	.084	.079	.058	.041	.095	.094	.061	.043
			140	140	0.85	.084	.060	.041	.094	.090	.075	.035	.136	.130	.086	.063
			180	180	.098	.096	.086	.063	.118	.116	.098	.062	.163	.159	.129	.108
0.70 Type I error	0.80	0.10	20	20	.062	.051	.064	.073	.060	.046	.075	.092	.076	.054	.085	.092
			60	60	.069	.065	.137	.148	.070	.068	.160	.177	.078	.068	.345	.269
			100	100	.076	.072	.214	.265	.086	.082	.267	.281	.098	.095	.381	.420
			140	140	.089	.084	.352	.368	.097	.092	.432	..525	.130	.122	.583	.674
			180	180	.110	.103	.480	.519	.110	.106	.606	.718	.166	.157	.784	.819
0.70 Power	0.90	0.20	20	20	.127	.108	.157	.168	.152	.112	.196	.227	.198	.153	.256	.280
			60	60	.198	.184	.372	.428	.251	.238	.445	.632	.336	.301	.627	.684
			100	100	.278	.259	.564	.687	.357	.325	.665	.728	.473	.444	.839	.872
			140	140	.422	.406	.785	.938	.501	.473	.883	.921	.704	.671	.973	.981
			180	180	.584	.565	.931	.981	.696	.674	.778	.835	.852	.820	.989	.991

Da, Conventional AUC test DeLong et al.,[3]; B[b], Approximation to permutation AUC test Bandos et al.,[6]; M[c], McNemar[8] testSumi et al.,[7]; EM[d], Extended McNemar[8] test (new method)

Application of tests to standard data

In other to demonstrate the workability of the new non-parametric method (extended McNemar test) for comparing correlated proportion of positive responses, we consider a practical data set adopted from Venkatraman & Begg[5] who carried out a distribution free procedure for comparing ROC curves from a paired experiment. This study was aimed at evaluating the performance of two diagnostic test results obtained from the anterior and posterior nodes in the cause of diagnosing Melanoma.

To demonstrate the feasibility of the extended McNemar test, we made use of the data from this study whose objective was to investigate the performance of two diagnostic test results obtained from anterior and posterior nodes for diagnosing Melanoma. The data presented in Table 4 in Venkatraman & Begg[5] provide the results using a clinical scoring system and a dermoscopic scoring scheme. The purpose of the analysis is to determine whether the dermoscope contributes similar diagnostic information. The null hypothesis is that the dermoscope contributes the same information as the clinical scoring system. This is the same as testing the null hypothesis that the sizes of anterior and posterior nodes possess equivalent diagnostic information. Using these data, estimates of proportion of positive responses for the two diagnostic tests 1 and 2 procedures are 0.725 and 0.652 respectively and the estimated correlation coefficient between the two diagnostic tests is 0.157. To test equivalence of the accuracy of these two diagnostic tests, the conventional test by DeLong et al.,[3] asymptotic permutation test by Bandos et al.,[6] the usual McNemar[8] test by Sumi et al.,[7] and the extended McNemar8 test are in agreement of significant different performances yielding two tailed p-values of 0.0048,0.017,0.0028,0.0019 respectively.

Bootstrap power analysis for comparing the statistical power of tests

The bootstrap is a powerful nonparametric approach Efron.[21] In an effort to obtain better and more specific knowledge regarding statistical power of tests, we have conducted a bootstrapping study where for each of considered sample sizes, 2000 random samples were taken from the data and rejection rates are computed.

Table 6 shows that given all sample sizes, the extended McNemar test provides the highest superior rejection rate followed by the McNemar[8] test by Sumi et al.,[7] and so on. At increased sample sizes, tests by DeLong et al.,[3] Bandos et al.,[6] and Sumi et al.,[7] shows rejection rates very closed to the Extended McNemar[8] test.

Citation: Marius OU, Happiness OI. An extended McNemar test for comparing correlated proportion of positive responses. *Biom Biostat Int J.* 2019;8(4):125–137. DOI: 10.15406/bbij.2019.08.00281

Table 6 Bootstrapping Test for obtaining the statistical power of different tests

Sample size		Rejection rate			
N	M	D[a]	B[b]	S[c]	EM[d]
20	20	0.67	0.538	0.679	0.685
60	60	0.769	0.737	0.819	0.827
100	100	0.869	0.857	0.889	0.89
140	140	0.919	0.911	0.929	0.931
180	180	0.946	0.938	0.977	0.994

Application to real life data

The new test for comparing correlated proportion of positive responses can be applied to real life data on gestational diabetes mellitus (GDM). Actually a random sample of 1113 pregnant women who tested positive for 50g Glucose Challenge Test (GCT) indicating that their plasma blood glucose level were at least 140 mg/dl after 1 hour. These same numbers of pregnant women were subsequently recalled and further subjected to two competing diagnostic test procedures, namely, 2-hour 75g OGTT and 3-hours 100g OGTT at various gestation periods according to the standard of World Health Organization[22] and National Diabetes Data Group.[23] These two diagnostic test procedures are paired. Women who were known diabetics, or who were suffering from any chronic illness were excluded from the study. The data is measured on a continuous scale and is dichotomized using at 7.8mmol/l or at least 140 mg/dl as cut-off value which is the recommended cut-off value for diagnosing GDM WHO.[22] Pregnant women whose test result is at least 7.8mmol/l is considered diseased (positive, coded 1) otherwise; they are not diseased (negative, coded 0). The data for the GDM response variables (tests results) for diagnostic test 1 and 2 procedures, namely 75g OGTT and 100g OGTT are paired and hence correlated for the 1113 pregnant women considered for this study. The null hypothesis of interest is testing the equality of the proportion of positive responses for the two diagnostic test procedures. The dichotomized data for the two diagnostic tests are as usual cross classified and presented in a contingency table to demonstrate the feasibility of the new nonparametric methods as well as the existing methods considered. We therefore obtain the sample estimates $\hat{\pi}^+, \hat{\pi}^0$ *and* $\hat{\pi}^-$, variance estimates and the McNemar test statistic and test the null hypothesis. In applying the extended McNemar test to the data, we evaluate the values of T_v of Eqn 29 where t_{v1} *and* t_{v2} are test results respectively by the subjects in the vth pair of diagnostic test 1 and diagnostic test 2 procedures for $v = 1,2,....,1113$. From the values of T_v, we have that

$$p^+ = n_{12} = 270, p^0 = n_{11} + n_{22} = p^{0+} + p^{0-} = 134 + 157 = 291; p^- = n_{21}.$$

From Eqn 36, we have the sample estimates as

$$\hat{\pi}^+ = \frac{270}{1113} = 0.2426; \hat{\pi}^0 = \frac{291}{1113} = 0.2615; \hat{\pi}^- = \frac{556}{1113} = 0.4995;$$

$$But\ \hat{\pi}^0 = \frac{291}{1113} = \frac{134}{1113} + \frac{157}{1113} = 0.1204 + 0.1411 = \hat{\pi}^{0+} + \hat{\pi}^{0-}.$$

$$Also\ W = p^+ - p^- = n_{12} - n_{21} = 270 - 556 = -286.$$

From Eqn 11, we have the estimated variance of W as

$$Var(W) = (1113)(0.2426 + 0.4995 - (0.2426 - 0.4995)^2) = (1113)(0.7421 - 0.0660) = (1113)(0.6761) = 752.4993.$$

Therefore to test the null hypothesis of equation 46 using the extended McNemar test statistic we have from Eqn 47 with $\beta_0 = 0$ that $\chi^2 = \frac{(270-556)^2}{752.4993} = \frac{81796}{752.4993} = 108.69(P - value = 0.0012)$ which with 1 degree of freedom is statistically significant showing that diagnostic test 1 and diagnostic test 2 do have differential effect of GDM on pregnant women. In other words, the probability of positive responses from the two diagnostic test procedures for the pregnant women differs significantly. To differ this result, we make use of the usual McNemar[8] test which was adopted by Sumi et al.,[7]

to analyze the GDM data that the estimated variance of P_2-P_1 is

$$Var(P_2 - P_1) = \frac{p^+ + p^-}{N^2} = \frac{n_{12} + n_{21}}{N^2} = \frac{270 + 556}{(1113)^2} = \frac{825}{1238769} = 0.000667.$$

Its test statistic for the H_0 of Eqn 36 with $\beta_0 = 0$ is $\chi^2 = \frac{(270-556)^2}{270+556} = \frac{81796}{826} = 99.03(P - value = 0.0028)$ which with 1 degree of freedom is also statistically significant. Even though the extended McNemar[8] test statistic and the usual McNemar[8] test statistic had both lead to the rejection null hypothesis, the relative sizes of the calculated chi-square values and the p-values obtained indicates that the usual McNemar[8] test statistic as adopted by Sumi et al.,[7] has greater chances of leading to Type II error more often than the extended McNemar[8] test statistic. Also, we note that the estimated variance of $\hat{\pi}^+ - \hat{\pi}^- is\ \text{var}(\hat{\pi}^+ - \hat{\pi}^-) = \frac{0.2426 + 0.4995 - (0.2426 - 0.4995)^2}{1113} = \frac{0.7421 - 0.0660}{1113} = \frac{0.6761}{1113} = 0.000607$ which is $0.000667 - 0.000607 = 0.00006 = \frac{0.0660}{1113} = \frac{(\hat{\pi}^+ - \hat{\pi}^-)^2}{N}$, smaller as expected than the variance of P_2-P_1 obtained when the usual or unmodified McNemar test is used.

Application of existing tests to the real life data

Applying the tests on the real life data, we obtain the following estimates of AUCs for the two diagnostic tests, the correlation coefficients between the test results of the two diagnostic test procedures and the p-values after testing for the equality of performance of the two diagnostic test procedures as

From Table 7 results indicates that all tests showed significant difference since the p-values are less than the chosen level of significant of 5 percent at increased sample size of 1113 for the data on GDM. Overall result shows that the extended McNemar[8] test are in agreement of significant different in their performances and therefore out performs other tests considered in this work.

Discussion

The extended McNemar[8] test statistic shown in this work apart from being simple to calculate, easy to understand and readily applicable, has proved that it is more powerful than the usual McNemar[8] test based on the fact that it provides for the possible presence of ties in the data used for analysis. From the analysis, it was seen that even though the extended McNemar[8] test statistic and the usual McNemar[8]

test statistic had both lead to the rejection null hypothesis, the relative sizes of the calculated chi-square values and the p-values obtained indicates that the usual McNemar[8] test statistic as adopted by Sumi et al.,[7] has greater chances of leading to Type II error more often than the extended McNemar8 test statistic. The proposed chi-square test does not require the knowledge of the true disease status or the gold standard may not be known. This is not the same with other traditional tests such as Bandos et al.,[6] and Delong et al.,[3] which must require the knowledge of true status (gold standard) in estimating the AUC.

The extended McNemar[8] test as an alternative method of evaluating the accuracy of diagnostic tests can be used in testing the null hypothesis that the proportion of positive responses are equal in two diagnostic test procedures.

Table 7 Comparison of the tests by estimates obtained from the data on GDM

S/n	Tests	p_1	p_2	$\hat{AUC}_1$	$\hat{AUC}_2$	Correlation coefficient (r)	p-value
1	Extended McNemar[8]	0.7214	0.7022	0.91183	0.9012	0.1654	0.0007
2	Sumi et al.,[7]	0.6765	0.6532	0.8675	0.8564	0.1754	0.0012
3	Bandos et al.,[6]	0.6375	0.6253	0.7392	0.7235	0.2732	0.00014
4	DeLong et al.,[3]	0.6453	0.6359	0.6443	0.6248	0.2401	0.0016

It is known that in the study of the statistical methods for diagnosis, one of the most interesting topics is the comparison of the accuracy of two binary diagnostic tests in relation to the same gold standard. The extended McNemar[8] test used in comparing the accuracy of two diagnostic tests does not make any reference to the gold standard in its comparison. This is indeed an innovation in statistical methods for diagnosis.

Summary and conclusions

The extended McNemar[8] test is applied to correlated data so as to compare the discriminatory abilities of two different test procedures. The data analysis using these methods involved computer simulation, standard data and real life data analysis carried out and result showed that the extended McNemar[8] test can be good alternative to the test by Sumi et al.,[7] test by DeLong et al.,[3] and test by Bandos et al.,[6] whose limitations were outlined in this paper. The McNemar test is therefore simple to communicate to the potential users of the procedures and it is easy to be applied in discriminating diagnostic test procedures even by non-statisticians. The summary of the finding are as follows:

1. In comparison to other tests, extended McNemar[8] test statistic is a very suitable alternative having the highest statistical power among the analysis carried out and so has the capacity to discriminate between diseased and non-diseased subjects in a better way.
2. The extended McNemar8 test does not require the knowledge of true status of subjects or any other gold standard in carrying out its analysis.
3. The proposed extended McNemar[8] test offers reliable statistical inferences even in small sample problems and circumvents the long period normally experienced while estimating the test statistics for the Delong et al(1988) and Bandos et al.,[6] which leads to computer memory loss and time.
4. The extended McNemar[8] test adjusts for the possible presence of ties in the data and therefore eliminates erroneous conclusions occasioned by using data without adjustment.
5. The variance of the extended McNemar[8] test statistic is smaller than the variance of the usual McNemar[8] test statistic and is relatively more efficient and is more powerful than the usual McNemar[8] test statistic. The calculated chi-square value of the extended McNemar[8] test is larger than that of the usual McNemar[8] test so that the chances of committing Type II error are reduced.
6. The extended McNemar8 test shows more false positive rate (FPR) when the correlation coefficient r is smaller than other tests considered. This is because the McNemar[8] tests are most suitably used when the data is correlated.
7. Considering all the applications to data, results showed that the extended McNemar test discriminates better than the traditional McNemar test by Sumi et al.,[7] conventional nonparametric test by DeLong et al.,[3] and the permutation test Bandos et al.,[6] when the AUCs are getting higher and for lower values of correlation coefficients.
8. The extended McNemar test enables the researcher to readily estimate not only the chances that among a randomly selected pair of diagnostic test 1 and 2 test results of subjects, the diagnostic test 1 responds positive and the diagnostic test 2 responds negative; or the diagnostic test 1 responds negative and the diagnostic test 2 responds positive, but also even when both the diagnostic test 1 and 2 test results of subjects have similar responses, it enables one easily estimate the probability that both respond positive or both respond negative. We therefore conclude as follows: The extended McNemar test statistic is more powerful than the usual McNemar test and indeed test by DeLong et al.,[3] and the permutation test Bandos et al.[6] Using any test statistic, the presence of ties in a data needs to be adjusted for before carrying out data analysis to avoid committing as much Type II error as possible so that decisions based on data analysis will not be erroneous.

Acknowledgements

I wish to appreciate Dr. Happiness Ilouno and Dr C.H Nwankwo of the Department of Statistics Nnamdi Azikiwe[24] University Awka for their valuable moral support during the period of putting up this work. Their advice and contributions cannot be forgotten in a hurry.

Competing interests

The authors declare that they have no competing interests.

References

1. JA Hanley, BJ McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29–36.
2. Green DM, Swets JA. *Signal Detection Theory and Psychophysics*. Wiley: New York; 1966.
3. DeLong ER, DeLong DM, Clarke–Pearson DL. Comparing the area under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44(3):837–845.
4. Hoeffding W. A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*. 1948;19(3):293–325.
5. Venkatraman ES, Begg CB. A distribution free procedure for comparing receiver operating characteristic curves from a paired experiment. *Biometrika*. 1996;83:835–848.
6. Bandos AI, Rockette HE, Gur D. A permutation test sensitive to differences in areas for comparing ROC curves from a paired design. *Statistics in Medicine*. 2005;24(18):2873–2893.
7. Nahid Sultana Sumi, M. Ataharul Islam, Akhtar Hossain. Evaluation and Computation of Diagnostic tests: A simple Alternative. *2010 mathematics subject classification*. 2010;92(08):62–107.
8. Ismael A Vergara, Tomás Norambuena, Evandro Ferrada, et al. StAR: a simple tool for the statistical comparison of ROC curves. BMC Bioinformatics. 2008;9:265.
9. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. Journal of Mathematical Psychology. 1975;12:387–415.
10. Thomas M Braun, Todd A Alonzo. A modified sign test for comparing paired ROC curves. Biostatistics. 2008;9(2):364–372.
11. Leonidas E Bantis, Ziding Feng. Comparison of two correlated ROC curves at a given specificity or sensitivity level. *Stat Med*. 2016;35(24):4352–4367.
12. Yu W, Park E, Chang YC. Comparison of Paired ROC Curves through a Two–Stage Test. Journal of Biopharmaceutical Statistics. 2015;25(5):881–902.
13. Calders T, Jaroszewicz S. Efficient AUC Optimization for Classification, Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'07). 2007;42–53.
14. Weiand S, MH Gail, BR James, et al. A Family of non–parametric statistics for comparing diagnostic markers with paired or unparied data. *Boimetrika*. 1989;76:585–592.
15. Karim O Hajian–Tilaki, James A Hanley. Comparison of Three Methods for Estimating the Standard Error of the Area under the Curve in ROC Analysis of Quantitative Data. *Acad Radiol*. 2002;9(11):1278–1285.
16. Hettmansperger TP. *Statistical inference based on ranks*. New York: NY, Wiley; 1984.
17. Zhou XH, Obuchowski NA, McClish DK. *Statistical Methods in Diagnostic Medicine*. New York: John Wiley and Sons, Inc; 2011.
18. F Leisch, A Weingessel, K Hornik. *On the generation of correlated artificial binary data. Austria: Working paper series*. Working paper No. 13, Vienna University of Economics and Business Administration; 1998.
19. MA Islam, RI Chowdhury, L Briollais. A bivariate binary model for testing dependence in outcomes. *Bull Malays Math Sci Soc*. 2012;35(4);845–858.
20. Efron B, Tibshirani RJ. *An introduction to the bootstrap*. New York: Chapman & Hall; 1993.
21. *Definition, Diagnosis and Classification of Diabetes Mellitus and its Complications*. USA: WHO. 1999.
22. National Diabetes Data Group. Classification and diagnosis of diabetes mellitus and other categories of glucose intolerance. *Diabetes*. 1979;28(12):1039–1057.
23. Nahid Sultana Sumi, Akhtar Hossain. A study on parametric approaches to compare areas under two correlated ROC curves. *Bangladesh J Sci Res*. 2012;25(1):61–72.
24. Leonidas E Bantis, Ziding Feng. Comparison of two correlated ROC curves at a given specificity or sensitivity level. *Stat Med*. 2016;35(24):4352–4367.

Citation: Marius OU, Happiness OI. An extended McNemar test for comparing correlated proportion of positive responses. *Biom Biostat Int J*. 2019;8(4):125–137. DOI: 10.15406/bbij.2019.08.00281