Focus | Methodology and Statistics

eISSN 2005-8330
https://doi.org/10.3348/kjr.2024.0049
Korean J Radiol 2024;25(7):656-661

Korean Journal of Radiology
KJR

Statistical Methods for Comparing Predictive Values in Medical Diagnosis

Chanrim Park[1], Seo Young Park[2], Hwa Jung Kim[3,4], Hee Jung Shin[5]

[1]Biomedical Research Institute, Seoul National University Hospital, Seoul, Republic of Korea
[2]Department of Statistics and Data Science, Korea National Open University, Seoul, Republic of Korea
[3]Department of Preventive Medicine, University of Ulsan College of Medicine, Seoul, Republic of Korea
[4]Department of Clinical Epidemiology and Biostatistics, Asan Medical Center, Seoul, Republic of Korea
[5]Department of Radiology and Research Institute of Radiology, Asan Medical Center, Seoul, Republic of Korea

Evaluating the performance of a binary diagnostic test, including artificial intelligence classification algorithms, involves measuring sensitivity, specificity, positive predictive value, and negative predictive value. Particularly when comparing the performance of two diagnostic tests applied on the same set of patients, these metrics are crucial for identifying the more accurate test. However, comparing predictive values presents statistical challenges because their denominators depend on the test outcomes, unlike the comparison of sensitivities and specificities. This paper reviews existing methods for comparing predictive values and proposes using the permutation test. The permutation test is an intuitive, non-parametric method suitable for datasets with small sample sizes. We demonstrate each method using a dataset from MRI and combined modality of mammography and ultrasound in diagnosing breast cancer.

Keywords: PPV; NPV; Diagnostic tests; Comparing predictive values; Permutation test

INTRODUCTION

In studies evaluating and comparing the performance of two different diagnostic tests, including artificial intelligence classification algorithms, the tests are often administered to the same patient cohort, and their outcomes are compared against the true disease status. There are several metrics to quantify the agreement between test results and the true disease status; however, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) are the most commonly utilized measures of test performance [1]. It is important to note that PPV and NPV can only be estimated from data that reflects the prevalence of the population; these measures cannot be estimated in case-control studies due to the altered prevalence inherent in the study design [2]. Calculating point estimates and confidence intervals for these measures is typically straightforward since they are fundamentally proportions. Often, hypothesis testing to determine superior test performance involves assessing differences in these measures between the two tests. The differences in sensitivity or specificity are usually examined using McNemar's test, which assesses differences in proportions of two paired binary variables and can be applied to data summarized in a 2 x 2 contingency table with mutually exclusive cells, tabulating the outcomes of two tests on a set of subjects.

Nevertheless, PPV and NPV cannot be compared using McNemar's test because data for estimating these values cannot be summarized in a standard 2 x 2 contingency table with mutually exclusive cells. As a result, many published studies do not perform statistical inference on the differences in PPV or NPV between two tests, instead reporting separate confidence intervals for the predictive values of each test [3].

Various methodologies have been developed to address the challenge of comparing predictive values between two diagnostic tests performed on the same set of patients [4-7].

Received: January 16, 2024 **Revised:** May 2, 2024
Accepted: May 4, 2024
Corresponding author: Seo Young Park, PhD, Department of Statistics and Data Science, Korea National Open University, 86 Daehak-ro, Jongno-gu, Seoul 03087, Republic of Korea
• E-mail: biostat81@knou.ac.kr

These methods are based on the asymptotic distribution of specific statistics. This review introduces both parametric tests and non-parametric tests, such as permutation tests, for comparing predictive values. The discussion focuses on four statistical methods that are supported by existing R packages, promoting practical application of these methods in research settings. We begin with a motivating example using a breast cancer diagnosis dataset and review the four statistical methods designed to compare predictive values.

Motivating Example

The dataset described here is derived from radiological diagnosis of breast cancer involving 300 high-risk patients, 233 of whom were diagnosed with breast cancer, representing a disease prevalence of approximately 77.7%. Three distinct imaging techniques were evaluated: diffusion-weighted imaging (DWI), dynamic contrast enhanced imaging (DCE), and the combined modality of mammography and ultrasound (MG/US). The PPVs for these imaging techniques were 0.974 for DWI, 0.955 for DCE, and 0.954 for MG/US, demonstrating high PPVs with only marginal differences between them. In the following sections, we will compare performance metrics using DWI and MG/US as an illustrative example. We will then detail the methodologies for comparing PPVs and NPVs, applying these methods to analyze the comparative PPVs of the three techniques.

Performance Measures of Diagnostic Tests

Accurate medical diagnostic tests with binary outcomes are essential for effective patient treatment. Sensitivity, specificity, PPV, and NPV are critical measures used to evaluate diagnostic tests. Sensitivity and specificity assess the probability of correct test results given the disease status: sensitivity is the probability of a positive test result given diseased status, while specificity is the probability of a negative test result given non-diseased status. PPVs and NPVs determine the probability of disease status given the test results; PPV measures the probability of diseased status given a positive test result, and NPV is defined as the probability of non-diseased status given a negative test result. These predictive values are particularly crucial in radiology, where determining patient conditions based on test results is necessary without a definitive ‘true value’ for reference. This paper aims to compare the performance of two diagnostic tests, referred to as Test 1 and Test 2.

McNemar’s test, widely used to compare sensitivities and specificities between two paired tests (e.g., DWI vs. MG/US), assesses differences in proportions of two paired binary variables. The data are summarized in a 2 x 2 contingency table that tabulates the outcomes of the two tests on a patient set. To compare sensitivities of two diagnostic tests, we create a 2 x 2 contingency table using data from patients whose true disease status is positive (Fig. 1), while specificity comparisons utilize data from patients confirmed as disease-negative (Fig. 1). For instance, comparing the sensitivities of DWI and MG/US involves examining the section of the table that specifically records results for the cancer group. Since the denominators for DWI’s sensitivity and MG/US’s sensitivity are identical—totaling the number of patients whose true disease status is positive (i.e., 176 + 32 + 11 + 13 = 232)—the comparison becomes an evaluation of the two marginal sums of the test-positive category (i.e., ‘176 + 11 = 187’ and ‘176 + 32 = 208’) in Figure 1. Specificity comparisons rely on patients whose true disease status is negative, comparing the two marginal sums of the test-negative category.

However, McNemar’s test is not suitable for comparing predictive values due to inherent challenges related to the different denominators in each test’s predictive values. Specifically, the denominator for Test 1’s PPV is the number of subjects with a positive result in Test 1, and similarly, the denominator for Test 2’s PPV is the number of Test 2 positives. For example, in the comparison between DWI and MG/US, the denominator for DWI’s PPV is the total of DWI positive results (i.e., 176 + 11 + 4 + 1 = 192), and the denominator for MG/US’s PPV is determined by its positive results (i.e., 176 + 32 + 4 + 6 = 218). Although there is often overlap between the subjects with positive results for Test 1 and those for Test 2, these two groups are not completely identical. Therefore, the data required to estimate PPV or NPV cannot be neatly summarized into a standard 2 x 2 table with mutually exclusive cells, rendering McNemar’s test inappropriate for this purpose.

Furthermore, the two independent proportions test (z-test) is not suitable either, as it is intended for comparing two proportions from two independent populations. Since the subjects testing positive in Test 1 often substantially overlap with those testing positive in Test 2, we cannot consider the PPV of Test 1 and the PPV of Test 2 as independent proportions. This overlap also applies to the NPV. Consequently, comparing the predictive values of two diagnostic tests within the same patient group to determine which test performs better poses a significant challenge.

To address this issue, several statistical methods have been developed to compare the predictive values of two diagnostic tests [4-7]. We review these methods and propose a feasible alternative: the permutation test. The characteristics of the existing methods and the permutation test are summarized in Table 1.

Existing Methods

The estimated PPVs of Test 1 and Test 2, based on Figure 1, can be expressed as follows:

$$\widehat{PPV_1} = \frac{n_1 + n_2}{n_1 + n_2 + n_5 + n_6}, \quad \widehat{PPV_2} = \frac{n_1 + n_3}{n_1 + n_3 + n_5 + n_7}$$

Similarly, the estimated NPVs for Test 1 and Test 2 are:

$$\widehat{NPV_1} = \frac{n_7 + n_8}{n_3 + n_4 + n_7 + n_8}, \quad \widehat{NPV_2} = \frac{n_6 + n_8}{n_2 + n_4 + n_6 + n_8}$$

In the subsequent sections, we review methodologies for comparing predictive values. The null hypothesis for comparing the PPVs of Test 1 and Test 2 is H_0: $PPV_1 = PPV_2$, and for NPVs, it is H_0: $NPV_1 = NPV_2$.

Leisenring et al. (2000)

Leisenring et al. [4] proposed comparing predictive values using generalized linear models, specifically through the generalized estimating equation method. For the PPV comparison, consider a logistic regression model with the true disease status (1 if positive, 0 if negative) as the response variable, and an indicator variable for the test (1 if Test 1, 0 if Test 2) as the sole independent variable [4]. This model is fitted to the subset of data that has a positive test result and utilizes a robust sandwich variance estimator to accommodate the correlation among multiple observations from the same patient. The beta coefficient of the independent variable in this model indicates how much more indicative a positive result from Test 1 is of the disease than a positive result from Test 2. Hence, a significance test for the beta coefficient serves as a test for the difference in PPVs. Similarly, if we fit the same model to the subset of data with a negative test result, we can use the beta coefficient to compare NPVs. Although the Wald test is commonly used for significance testing of the beta coefficient, Leisenring et al. [4] found that the score test performs better than the Wald test in their simulation study.

Moskowitz and Pepe (2006)

Moskowitz and Pepe [5] discussed two methods for comparing the predictive values of two different diagnostic tests: The first based on relative predictive value, and the second using a regression framework that considers discordant pairs. We will focus on the first method as follows: The metrics they utilized for comparison are the relative PPV (rPPV) defined as rPPV $= \frac{PPV_1}{PPV_2}$, and the relative NPV (rNPV) defined as rNPV $= \frac{NPV_1}{NPV_2}$. The estimated relative positive and NPVs derived from the observed data (Fig. 1) are:

$$\widehat{rPPV} = \frac{\widehat{PPV_1}}{\widehat{PPV_2}} = \frac{(n_1 + n_2) \ / \ (n_1 + n_2 + n_5 + n_6)}{(n_1 + n_3) \ / \ (n_1 + n_3 + n_5 + n_7)}$$

$$\widehat{rNPV} = \frac{\widehat{NPV_1}}{\widehat{NPV_2}} = \frac{(n_7 + n_8) \ / \ (n_3 + n_4 + n_7 + n_8)}{(n_6 + n_8) \ / \ (n_2 + n_4 + n_6 + n_8)}$$

Hypothesis testing to compare predictive values can then be performed using the fact that $\frac{\log \widehat{rPPV}}{\sqrt{\hat{\sigma}_P^2/N}}$ and $\frac{\log \widehat{rNPV}}{\sqrt{\hat{\sigma}_N^2/N}}$ asymptotically follow a normal distribution.

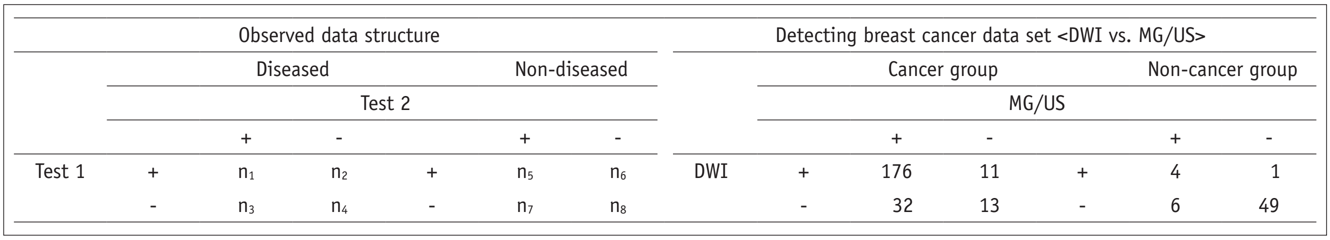
A similar approach was proposed by Wang et al. [6].

Kosinski (2013)

Kosinski [7] proposed using a generalized linear model

Table 1. Characteristics of the existing methods for comparing positive predictive values

Author/method	Statistic	R function	Advantages	Pitfalls
Moskowitz and Pepe [5]	Relative predictive values	pv.rpv	Can derive required sample size for study design	May not be ideal for covariate adjustment
Leisenring et al. [4]	Generalised score statistic	pv.gs	Allow adjusting covariate. Shows better power than Wald statistic	Considerably complex, less intuitive
Kosinski [7]	Weighted generalised score statistic	pv.wgs	Better type I error control than Leisenring's method	Less intuitive
Permutation test	Exact *P*-value	ppv.permutation.test	Intuitive; simple non-parametric test	Different seeds yield different *P*-value



Observed data structure						
		Diseased			Non-diseased	
		Test 2				
		+	-		+	-
Test 1	+	n_1	n_2	+	n_5	n_6
	-	n_3	n_4	-	n_7	n_8

Detecting breast cancer data set <DWI vs. MG/US>						
		Cancer group			Non-cancer group	
		MG/US				
		+	-		+	-
DWI	+	176	11	+	4	1
	-	32	13	-	6	49

Fig. 1. Data structure used to compute positive predictive values and negative predictive values. DWI = diffusion-weighted imaging, MG/US = mammography and ultrasound

similar to the one used in Leisenring's method, but introduced a different statistic for testing the equality of predictive values. They suggested the weighted generalized score (WGS) test statistic, which is an enhancement of the score statistic used in generalized linear models. They demonstrated that hypothesis testing using WGS exhibits superior type I error performance compared to other methods.

Suggested Method: Permutation Test

The permutation test is an intuitive and straightforward non-parametric method based on random sampling [8]. In this test, rather than mathematically deriving the distribution of a statistic under the null hypothesis, the null distribution of a statistic is obtained by permuting the group indicator [9]. To test the difference in predictive values of Test 1 and Test 2, the empirical distribution of the statistic $\widehat{PPV}_1$ - $\widehat{PPV}_2$ under the null hypothesis can be obtained by shuffling the labels of each test. The *P*-value is then calculated by comparing the null distribution with the observed difference $\widehat{PPV}_1$ - $\widehat{PPV}_2$ in the original data.

Takahashi and Yamamoto's [10] exact test, which also uses permutations, involves calculating all possible shuffling patterns to derive the null distribution. Although this method provides high precision, it can become computationally intensive, especially with large datasets, making it more suitable for small clinical trials. In contrast, our permutation method relies on random sampling instead of exhaustive computation. Our approach involves randomly shuffling the labels or outcomes to estimate the null distribution, which allows for greater flexibility and scalability. By sampling from a broader set of possible outcomes, our method can efficiently handle large datasets.

R Implementation

The execution of existing methods for comparing two predictive values (PPVs) in a paired study design can be achieved using the R software, utilizing the DTComPair Package (version 1.2.2) [11]. This package supports the implementation of three previously discussed methods: the generalised score statistic by Leisenring et al. [4], the relative predictive values by Moskowitz and Pepe [5], and the weighted generalised score statistic by Kosinski [7]. Additionally, permutation tests can be conducted using the R syntax provided in the Supplementary Materials.

To determine whether the PPVs or NPVs of the two tests are significantly different, the dataset should be structured with three columns indicating the true disease status, the results of Test 1, and the results of Test 2. The structure of the example dataset is illustrated in Supplementary Table 1.

With the DTComPair package, the tab.paired() function allows for the creation of two contingency tables for diseased and non-diseased groups. The acc.paired() function can be used to extract sensitivities, specificities, PPVs, NPVs, and their confidence intervals for each group. The *P*-value based on the generalized score statistics can be computed using the pv.gs() function, while the *P*-value for relative predictive values can be obtained with the pv.rpv() function. The *P*-value based on the WGS can be calculated using the pv.wgs() function. For conducting a permutation test, the syntax and an example are provided in the Supplementary Materials.

Applications

The practical application of the aforementioned comparative techniques was explored using a breast cancer diagnosis dataset, which included three imaging techniques: DWI, DCE, and MG/US. Three specific contrasts were defined for this study: <DWI vs. DCE>, <DWI vs. MG/US>, and <DCE vs. MG/US>. The DTComPair package (R software v.4.2.2) [11] was employed to compute the generalized score statistic, relative predictive values, and WGS. Permutation tests were also conducted using specially devised code. The null hypothesis for the PPV comparisons posited $PPV_1 = PPV_2$, and for NPV comparisons $NPV_1 = NPV_2$. A difference was deemed statistically significant if the *P*-value was less than 0.05. The individual observations for each comparison are summarized in Supplementary Tables 2-4.

A comprehensive tabulation of performance measures, including PPVs, NPVs, sensitivities, and specificities, was calculated for each technique within these groups (Table 2). Notably, DWI presented the highest PPV and specificity but also demonstrated the lowest NPV and sensitivity. Conversely, DCE displayed the highest NPV and sensitivity. MG/US exhibited the lowest PPV and specificity.

In relation to the PPV comparison within the three identified groups, only the DWI vs. DCE comparison showed a statistically significant difference, as detected by the permutation test among the four methods. Other group comparisons did not yield significant differences (Table 3). Intriguingly, the *P*-value for the PPV difference between DWI and DCE was less than 0.05, but this was only the case for the permutation test.

Regarding the comparison of NPVs across the techniques, all three group comparisons highlighted significant differences in NPVs with all four methods (Table 3). In this example study, *P*-values derived from permutation tests for both PPVs and NPVs were observed to be smaller than those generated by the other methods. However, it should be emphasized that a lower *P*-value does not necessarily indicate that one statistical method is superior to others. For comparing PPVs, a significant difference was found only in the DWI vs. DCE comparison, while all three groups showed differences in NPVs across the techniques.

Table 2. Values of the three techniques' performances for breast cancer diagnosis

	PPV	NPV	Sensitivity	Specificity
DWI	0.974	0.579	0.807	0.925
DCE	0.955	0.966	0.991	0.836
MG/US	0.954	0.676	0.893	0.746

PPV = positive predictive value, NPV = negative predictive value, DWI = diffusion-weighted imaging, DCE = dynamic contrast enhanced imaging, MG/US = mammography and ultrasound

CONCLUSION

We reviewed various methods for comparing predictive values in diagnostic tests. Leisenring et al. [4] proposed utilizing the generalized score statistic from a generalized linear model. Secondly, Moskowitz and Pepe [5] suggested relative predictive values, derived from the ratio of the estimated predictive value of Test 1 to that of Test 2. Similar to the method suggested by Leisenring et al. [4], Kosinski [7] recommended using a generalized linear model to test the difference in predictive values with a WGS. Unlike comparing sensitivities and specificities, the methods for comparing two predictive values suggested in previous research are complex.

We propose the permutation test as an alternative for comparing predictive values. While McNemar's statistic is applicable only for comparing sensitivities and specificities, the permutation test can be applied to compare sensitivities, specificities, and predictive values concurrently. Naturally, there are some limitations to the permutation test. For instance, with different random seeds for permutation, *P*-values will differ, potentially leading to differing conclusions. Despite these limitations, we suggest the permutation test for comparing predictive values because its intuitiveness and simplicity make it plausible for use by

Table 3. PPV and NPV differences and *P*-values of the four methods per technique set

	Difference*	*P*-value			
		Moskowitz and Pepe [5]	Leisenring [4]	Kosinski [7]	Permutation test
PPV					
DWI vs. DCE	0.019	0.054	0.052	0.075	0.032†
DWI vs. MG/US	0.020	0.102	0.100	0.113	0.113
DCE vs. MG/US	0.001	0.958	0.958	0.957	1
NPV					
DWI vs. DCE	0.387	< 0.001†	< 0.001†	< 0.001†	0†
DWI vs. MG/US	0.097	0.008†	0.007†	0.008†	0.006†
DCE vs. MG/US	0.290	< 0.001†	< 0.001†	< 0.001†	0†

*Differences are expressed as absolute values, $^\dagger P < 0.05$.
PPV = positive predictive value, NPV = negative predictive value, DWI = diffusion-weighted imaging, DCE = dynamic contrast enhanced imaging, MG/US = mammography and ultrasound

medical practitioners.

Supplement

The Supplement is available with this article at https://doi.org/10.3348/kjr.2024.0049.

Conflicts of Interest

Seo Young Park, who holds the respective position of Statistical Consultant of the *Korean Journal of Radiology*, was not involved in the editorial evaluation or decision to publish this article. The remaining authors have declared no conflicts of interest.

Author Contributions

Conceptualization: Chanrim Park, Seo Young Park, Hwa Jung Kim. Resources: Hee Jung Shin. Software: Chanrim Park, Seo Young Park. Writing—original draft: Chanrim Park. Writing—review & editing: Seo Young Park.

ORCID IDs

Chanrim Park
https://orcid.org/0009-0000-2661-7111
Seo Young Park
https://orcid.org/0000-0002-2702-1536
Hwa Jung Kim
https://orcid.org/0000-0003-1916-7014
Hee Jung Shin
https://orcid.org/0000-0002-3398-1074

Funding Statement

None

REFERENCES

1. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ* 2015;351:h5527
2. Celentano DD, Szklo M. *Gordis epidemiology*. 6th ed. Philadelphia: Elsevier, 2019
3. Mercaldo ND, Lau KF, Zhou XH. Confidence intervals for predictive values with an emphasis to case-control studies. *Stat Med* 2007;26:2170-2183
4. Leisenring W, Alonzo T, Pepe MS. Comparisons of predictive values of binary medical diagnostic tests for paired designs. *Biometrics* 2000;56:345-351
5. Moskowitz CS, Pepe MS. Comparing the predictive values of diagnostic tests: sample size and analysis for paired study designs. *Clin Trials* 2006;3:272-279
6. Wang W, Davis CS, Soong SJ. Comparison of predictive values of two diagnostic tests from the same sample of subjects using weighted least squares. *Stat Med* 2006;25:2215-2229
7. Kosinski AS. A weighted generalized score statistic for comparison of predictive values of diagnostic tests. *Stat Med* 2013;32:964-977
8. Bakeman R, Robinson BF, Quera V. Testing sequential association: estimating exact p values using sampled permutations. *Psychol Methods* 1996;1:4-15
9. Ludbrook J. Advantages of permutation (randomization) tests in clinical and experimental pharmacology and physiology. *Clin Exp Pharmacol Physiol* 1994;21:673-686
10. Takahashi K, Yamamoto K. An exact test for comparing two predictive values in small-size clinical trials. *Pharm Stat* 2020;19:31-43
11. Stock C, Hielscher T, Discacciati A. DTComPair: comparison of binary diagnostic tests in a paired study design [accessed on January 3, 2024]. Available at: http://CRAN.R-project.org/package=DTComPair