Published in final edited form as:

Stat Med. 2015 July 10; 34(15): 2325–2333. doi:10.1002/sim.6485.

Comparison of operational characteristics for binary tests with clustered data

Minjung Kwak^a, Sang-Won Um^b, and Sin-Ho Jung^{b,c,*}

^aDepartment of Statistics, Yeungnam University Gyeongsan, Gyeongbuk, 712-749, South Korea

^bSamsung Medical Center, Sungkyunkwan University School of Medicine, Seoul 135-710, South Korea

^cDepartment of Biostatistics and Bioinformatics, Duke University, Durham, NC 27710, USA

Summary

Although statistical methodology is well developed for comparing diagnostic tests in terms of their sensitivity and specificity, comparative inference about predictive values is not. In this paper we consider the analysis of studies comparing operating characteristics of two diagnostic tests that are measured on all subjects and have test outcomes from multiple sites with varying number of sites among subjects. We have developed a new approach for comparing sensitivity, specificity, positive predictive value, and negative predictive value with simple variance calculation and in particular focus on comparing tests using difference of positive and negative predictive values. Simulation studies are conducted to show the performance of our approach. We analyze real data on patients with lung cancer diagnostic tests to illustrate the methodology.

Keywords

Sensitivity; Specificity; Positive predictive value; Negative predictive value; Clustered binary outcome

1 Introduction

Many new medical tests have been developed with recent advances in biotechnology and such tests can be used for various purposes including diagnosis, prognosis, risk prediction, and disease screening. Among frequently used measures for binary tests are accuracy, sensitivity, specificity, positive predictive value, and negative predictive value. Sensitivity is the probability of a positive test among disease subjects (cases), and specificity is the probability of a negative test among healthy subjects (controls). On the other hand, positive predictive value (PPV) is the probability of having the disease, given that the test is positive, and negative predictive value (NPV) is the probability of not having the disease, given that the test is negative. McNemar's test can be used to test equality of the sensitivity or specificity for two tests evaluated on the same subjects. However, relatively little attention has been paid to the comparison between paired PPVs or NPVs. Wang et al. [1] studied the

*sinho.jung@duke.edu.

size and power of the equality test of two PPVs (and of two NPVs in a similar way) using the delta method to derive the asymptotic normality of the log ratio or the difference of two PPVs. Leisenring et al. [2] implemented the generalized estimating equations to solve this problem. They proposed a score statistic and a Wald statistic derived from a marginal regression model. More recently Kosinski [3] proposed a weighted generalized score statistic for simpler computation. His statistic reduces to the score statistic in the independent sample situation. Moskowitz et al. [4] derived sample size formula for the log ratio of two PPVs and they used the multinomial Poisson transformation which transforms the likelihood of the data into a Poisson likelihood with additional parameters to avoid a complicated variance form obtained by the delta method.

Although well established, sensitivity and specificity have some deficiencies in clinical use. This arises mainly from the fact that sensitivity and specificity are population measures, i.e., they summarize the characteristics of the test over a population. However when we consider the results of a diagnostic test from a patient's perspective, if the diagnostic test is positive, then the patient wants to know the chance that he or she actually has the disease. And if the test is negative, the patient may ask about the probability that he or she does not actually have the disease if his/her test comes back to negative. These questions refer to the positive and negative predictive values of a diagnostic test. So PPV and NPV are much more important than sensitivity and specificity from the patient's perspective.

Our study is motivated by a real project to compare the performance of two diagnostic tests to determine the metastatic status of lymph nodes for lung cancer patients. Medi-astinoscopy has been a diagnostic standard, but its sensitivity and negative predictive value are not very high. Furthermore, the technique is invasive and the morbidity and mortality rates are known to be about 2% and 0.08%, respectively (Porte et al.[6], Silvestri et al.[7]). Endobronchial ultrasound-guided transbronchial needle aspiration (EBUS) is a less invasive intervention with broader visual scope so that the investigators would like to compare the new method, EBUS, to the existing diagnostic tool, mediastinoscopy. Both diagnostic tests are conducted for some chosen lymph nodes of each patient, so that the resulting data are clustered paired binary outcomes, where pairs are two diagnostic tests from each lymph node site and clusters are patients. The study was designed to show the non-inferiority of the new test compared to the standard test, but it was shown to be even superior in some operating characteristics. Rao et al. [5] proposed a simple method for the analysis of clustered binary data assuming no specific model for the intracluster correlation. They obtained an estimator of the variance inflation factor due to clustering and adjusted the variance formula derived for independent data by multiplying by the inflation factor. Jung et al. [8] proposed a sample size calculation method in clustered binary data using an optimal weighting scheme that minimizes the variance of the estimator. Hu et al. [9] derived a sample size formula for sensitivity and specificity of diagnostic tests using the sign test for dependent multiple observations per subject under common correlation model. These methods mainly dealt with single sample proportion estimator and the focus has been given on sensitivity or specificity. In addition, Katsis[10] proposed a Bayesian approach to the sample size for two dependent binomial populations using a Dirichlet prior on the proportions.

This paper is novel in that 1) we provide simpler test statistics for the comparison of sensitivity, specificity, positive predictive value, and negative predictive value in quite a general setting 2) our method can be easily applied to clustered data where we allow multiple dependent observations per subject with varying cluster sizes. In Section 2, we describe the data structure and provide statistical testing method for the sensitivity, specificity, accuracy, PPV, and NPV for two clustered test outcomes. We numerically study the performance of our method under various scenarios in Section 3. Section 4 presents an example of the proposed method. Conclusions are made in Section 5.

2 Data and Proposed Method

In clinical practice, clinicians are interested in early detection of a specific disease or decision of medical status for their patients. Usually there exist good treatments available for the disease, but subjects will have some harmful effect if the disease status is erroneously diagnosed. We assume that there are two available diagnostic methods and wish to identify the better one. For example, we prefer the method with a higher PPV. For subject $i = 1, \dots, n$, we observe paired binary outcomes for two diagnostic tests from each of m_i sites, such as lymph node sites. We assume that $\max_i m_i < c (< \infty)$ and a positive diagnostic outcome provides evidence of disease. From site j of subject i , we observe binary random variables d_{ij} denoting the disease status, 0 for non-disease and 1 for disease, and x_{kij} denoting the outcome of diagnostic test $k (= 1, 2)$, 0 for negative and 1 for positive. Thus, the resulting data are summarized as $\{(d_{ij}, x_{1ij}, x_{2ij}), 1 \leq j \leq m_i, 1 \leq i \leq n\}$. Let p_k denote the sensitivity, specificity, accuracy, PPV or NPV for test method k . Now, for each operational characteristic of diagnostic tests, we propose a consistent estimator $\hat{p}_k$ for p_k and derive the asymptotic normality of $\sqrt{n}(\hat{p}_1 - \hat{p}_2)$ using clustered binary data. Hence, we reject $H_0: p_1 = p_2$ in favor of $H_1: p_1 \neq p_2$ if the absolute value of $Z = \sqrt{n}(\hat{p}_1 - \hat{p}_2)/\hat{\sigma}$ is larger than $z_{1-\alpha/2}$, where $\hat{\sigma}^2$ is a variance estimator of $\sqrt{n}(\hat{p}_1 - \hat{p}_2)$ and $z_{1-\alpha/2}$ is the 100(1- α)/2 percentile of the standard normal distribution.

2.1 Sensitivity and Specificity

An unbiased estimator of the sensitivity for diagnostic test $k (= 1, 2)$, p_k , is given by

$$\hat{p}_k = \frac{\sum_{i=1}^n \sum_{j=1}^{m_i} x_{kij} d_{ij}}{\sum_{i=1}^n \sum_{j=1}^{m_i} d_{ij}}.$$

Let $D = \sum_{i=1}^n \sum_{j=1}^{m_i} d_{ij}$ denote the total number of diseased sites across all n subjects. We have

$$\sqrt{n}(\hat{p}_k - p_k) = \frac{\sqrt{n}}{D} \sum_{i=1}^n \sum_{j=1}^{m_i} (x_{kij} - p_k) d_{ij}.$$

Since the n subjects are independent, given the disease status of the patients,

$\varepsilon_{ki} = \sum_{j=1}^{m_i} (x_{kij} - p_k) d_{ij}$ are independent 0-mean random variables. By the central limit theorem for large n , $\sqrt{n}(\hat{p}_k - p_k)$ is asymptotically normal with mean 0 and variance σ_k^2 that can be consistently estimated by

$$\hat{\sigma}_k^2 = \frac{n}{D^2} \sum_{i=1}^n \left\{ \sum_{j=1}^{m_i} (x_{kij} - \hat{p}_k) d_{ij} \right\}^2.$$

Furthermore, for the difference in sensitivity between two diagnostic tests, we have

$$\sqrt{n}(\hat{p}_1 - \hat{p}_2) = \frac{\sqrt{n}}{D} \sum_{i=1}^n \sum_{j=1}^{m_i} (x_{1ij} - x_{2ij}) d_{ij}.$$

Under H_0 : $p_1 = p_2$, $\varepsilon_i = \sum_{j=1}^{m_i} (x_{1ij} - x_{2ij}) d_{ij}$ are independent 0-mean random variables.

Hence, under H_0 , $\sqrt{n}(\hat{p}_1 - \hat{p}_2)$ is approximately normal with mean 0 and variance σ^2 that can be consistently estimated by

$$\hat{\sigma}^2 = \frac{n}{D^2} \sum_{i=1}^n \left\{ \sum_{j=1}^{m_i} (x_{1ij} - x_{2ij}) d_{ij} \right\}^2.$$

The inference on specificities can be derived from that on sensitivities by replacing d_{ij} and x_{kij} with $1 - d_{ij}$ and $1 - x_{kij}$, respectively.

2.2 Accuracy

For site j of patient i , the testing result by diagnostic method k is accurate if $y_{kij} \equiv I(x_{kij} = d_{ij}) = x_{kij}d_{ij} + (1 - x_{kij})(1 - d_{ij})$ equals 1, and inaccurate if $y_{kij} = 0$. So, an unbiased estimator of the accuracy for method k , $p_k = P(x_{kij} = d_{ij})$, is given as

$$\hat{p}_k = \frac{1}{N} \sum_{i=1}^n \sum_{j=1}^{m_i} y_{kij},$$

where $N = \sum_{i=1}^n m_i$ denotes the total number of sites across all n subjects. From

$$\sqrt{n}(\hat{p}_k - p_k) = \frac{\sqrt{n}}{N} \sum_{i=1}^n \sum_{j=1}^{m_i} (y_{kij} - p_k),$$

$\varepsilon_{ki} = \sum_{j=1}^{m_i} (y_{kij} - p_k)$ are independent 0-mean random variables. Hence, $\sqrt{n}(\hat{p}_k - p_k)$ is approximately normal with mean 0 and variance σ_k^2 that can be consistently estimated by

$$\hat{\sigma}_k^2 = \frac{n}{N^2} \sum_{i=1}^n \left\{ \sum_{j=1}^{m_i} (y_{kij} - \hat{p}_k) \right\}^2.$$

For the difference in accuracy between two diagnostic tests, we have

$$\sqrt{n}(\hat{p}_1 - \hat{p}_2) = \frac{\sqrt{n}}{N} \sum_{i=1}^n \varepsilon_i,$$

where $\varepsilon_i = \sum_{j=1}^{m_i} (y_{1ij} - y_{2ij})$, $i = 1, \dots, n$ are independent 0-mean random variables under $H_0: p_1 = p_2$. Hence, under H_0 , $\sqrt{n}(\hat{p}_1 - \hat{p}_2)$ is approximately normal with mean 0 and variance σ^2 that can be consistently estimated by

$$\hat{\sigma}^2 = \frac{n}{N^2} \sum_{i=1}^n \left\{ \sum_{j=1}^{m_i} (y_{1ij} - y_{2ij}) \right\}^2.$$

2.3 PPV and NPV

Noting that the PPV of diagnostic test k is defined by $p_k = P(d_{ij} = 1 | x_{kij} = 1) = P(x_{kij} = 1, d_{ij} = 1) / P(x_{kij} = 1) \equiv a_k / b_k$, a consistent estimator of p_k is obtained by

$$\hat{p}_k = \frac{\sum_{i=1}^n \sum_{j=1}^{m_i} x_{kij} d_{ij}}{\sum_{i=1}^n \sum_{j=1}^{m_i} x_{kij}} \equiv \frac{\hat{a}_k}{\hat{b}_k}$$

where $\hat{a}_k = N^{-1} \sum_{i=1}^n \sum_{j=1}^{m_i} x_{kij} d_{ij}$ and $\hat{b}_k = N^{-1} \sum_{i=1}^n \sum_{j=1}^{m_i} x_{kij}$.

By letting $a_{kij} = x_{kij} d_{ij}$ and $b_{kij} = x_{kij}$, we have

$$\begin{aligned} \hat{p}_k - p_k &= \frac{\hat{a}_k b_k - a_k \hat{b}_k}{\hat{b}_k b_k} \\ &= \frac{b_k(\hat{a}_k - a_k) - a_k(\hat{b}_k - b_k)}{b_k^2} + o_p(n^{-1/2}), \end{aligned}$$

where the ignorable error term $o_p(n^{-1/2})$ is added by replacing the consistent estimator $\hat{b}_k$ with b_k in the denominator. Then, we have

$$\begin{aligned}\sqrt{n}(\hat{p}_k - p_k) &= \frac{\sqrt{n}}{N} \sum_{i=1}^n \sum_{j=1}^{m_i} \left\{ \frac{1}{b_k} (a_{kij} - a_k) - \frac{a_k}{b_k^2} (b_{kij} - b_k) \right\} + o_p(1) \\ &= \frac{\sqrt{n}}{N} \sum_{i=1}^n \varepsilon_{ki} + o_p(1),\end{aligned}$$

where, for $i = 1, \dots, n$, $\varepsilon_{ki} = b_k^{-1} \sum_{j=1}^{m_i} \{a_{kij} - a_k - p_k(b_{kij} - b_k)\}$ are independent random variables with mean 0. Hence, $\sqrt{n}(\hat{p}_k - p_k)$ is approximately normal with mean 0 and variance σ_k^2 that can be consistently estimated by

$$\hat{\sigma}_k^2 = \frac{n}{N^2} \sum_{i=1}^n \hat{\varepsilon}_{ki}^2,$$

where $\hat{\varepsilon}_{ki}$ is obtained from ε_{ki} by replacing a_k , b_k and p_k with their consistent estimators $\hat{a}_k$, $\hat{b}_k$ and $\hat{p}_k$, respectively, i.e. $\hat{\varepsilon}_{ki} = \hat{b}_k^{-1} \sum_{j=1}^{m_i} \{a_{kij} - \hat{a}_k - \hat{p}_k(\hat{b}_{kij} - \hat{b}_k)\}$.

Similarly, under H_0 : $p_1 = p_2$, we have

$$\begin{aligned}\hat{p}_1 - \hat{p}_2 &= (\hat{p}_1 - p_1) - (\hat{p}_2 - p_2) \\ &= N^{-1} \sum_{i=1}^n (\varepsilon_{1i} - \varepsilon_{2i}) + o_p(n^{-1/2}).\end{aligned}$$

For $i = 1, \dots, n$, we observe that $\varepsilon_i = \varepsilon_{1i} - \varepsilon_{2i}$ are independent random variables with mean 0. Hence, under H_0 , $\sqrt{n}(\hat{p}_1 - \hat{p}_2)$ is approximately normal with mean 0 and variance σ^2 that can be consistently estimated by

$$\hat{\sigma}^2 = \frac{n}{N^2} \sum_{i=1}^n (\hat{\varepsilon}_{1i} - \hat{\varepsilon}_{2i})^2.$$

The inference on NPV can be derived from that on PPV by replacing d_{ij} and x_{kij} with $1 - d_{ij}$ and $1 - x_{kij}$, respectively.

3 Simulations

We carried out a simulation study to evaluate the properties of the proposed method under a variety of conditions. First, for sensitivity and specificity with single binary test outcome per subject, we confirmed that our simple variance formulas yield the same numerical results as the lengthy and complicated formulas of Wang et al. [1] that were derived by using the multivariate central limit theorem together with the delta method. So, we focused on testing two PPVs and NPVs with clustered binary test outcomes. We specified the disease prevalence, the number of patients tested n , the positive predictive value, and the negative predictive value for both test $k (= 1, 2)$, PPV_k and NPV_k . In addition, we specify the variance

(σ_1^2) for a subject-specific random effect that is used to induce the correlation among the multiple sites on the same individual and the variance (σ_2^2) for a site-specific random effect that is used to induce the correlation between the two test outcomes performed at each site.

Under each setting, we simulated 5000 data sets and evaluated the size or power of the proposed method. Each simulated data set was generated as follows. First we generate the number of sites m_i for i th patient by generating uniform random number from {2, 4, 6, 8, 10}. Given m_i , we randomly generate disease indicators $d_{ij}, j = 1, \dots, m_i$ for each individual by first drawing a multivariate normal random vector and then turning each random variable into a binary indicator using a percentile corresponding to a specified disease prevalence. For each individual, independent random effects r_i and u_{ij} were generated from zero-mean normal distributions with variance σ_1^2 and σ_2^2 respectively, and binary test outcomes were generated using probit models with the parametrization $P(x_{kij} = 1 | d_{ij}, r_i, u_{ij}) = \Phi\{\alpha_k(1 - d_{ij}) + \beta_k d_{ij} + r_i + u_{ij}\}$ for test $k (= 1, 2)$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. The values of the parameters ($\alpha_1, \beta_1, \alpha_2, \beta_2$) used in these probit models were chosen so that the true and false positive rates are equal to those calculated and they in turn give the desired predictive values. By taking the expectation with respect to the random effects, we obtain the marginal rate for test k as a function of coefficients (α_k, β_k) and the variance of the random effects σ_1^2 and σ_2^2 by

$$P(x_k=1|d)=\Phi\left\{\frac{\alpha_k(1-d)+\beta_k d}{\sqrt{1+\sigma_1^2+\sigma_2^2}}\right\}.$$

We want to examine the empirical type I error rate and power for testing the equality of two PPVs or NPVs. We assume a disease prevalence of 0.25, 0.5, or 0.7, and a sample size n of 100, 200, or 500. The number of sites for each subject m_i takes 2, 4, 6, 8 or 10 with an equal chance of 20%. We present the Kappa statistics as a measurement of correlation between the two test outcomes for each patient in Table 1 under the simulation settings. These are estimated from simulation data. Table 2 reports the empirical size and power for testing the equality of two PPVs with $\alpha = 0.05$. In Table 2 we consider the null hypothesis that the two diagnostic tests have equal PPVs of 0.7 or 0.85. Under the alternative hypothesis, we set $PPV_1 = 0.7$ or 0.85 and $PPV_2 = PPV_1 + 0.05$. In these simulations, NPVs are fixed at 0.9. We observe that the empirical sizes are maintained closely to the nominal level of 0.05 under the simulation settings considered. And the empirical power increases in sample size and prevalence. Also, we have a larger power with $PPV_1 = 0.85$ than with 0.7 since the former has a larger odds ratio between the two prognostic tests and higher Kappa statistics as shown in Table 1.

We compare empirical test power of our proposed method to two previously proposed methods for single site per subject based on 5000 simulation samples. In Table 3 we present the empirical test powers obtained by using weighted least square method by Wang et al.[1], generalized estimating equation method by Leisenring et al. [2], and our proposed method. In Table 3 we consider the null hypothesis that the two diagnostic tests have equal PPVs of

0.75. Under the alternative hypothesis, we set $PPV_1 = 0.85$ and $PPV_2 = 0.75$. In these simulations, NPVs are fixed at 0.85. We consider various scenarios with the disease prevalence of 0.25, 0.5, or 0.75 and the sample sizes of 100, 200, or 500. The proposed method seems to show similar or slightly larger empirical power compared to the previously proposed methods. We observe that the empirical size of the test appears to be close to the nominal level $\alpha = 0.05$ for all three methods and omit the presentation.

We also conducted similar simulations on NPVs and observed almost the same results, so that we decided not to report them here.

4 Example

For illustrative purpose, the methodology developed in this paper is applied to real data from lung cancer diagnostic test. Mediastinoscopy is a technique often used for staging of lymph nodes of lung cancer and involves making an incision above breast bone. Investigators are interested in comparing the performance of a new diagnostic test called EBUS, which is less invasive and has a broader visual scope than mediastinoscopy [11]. The patients with non-small cell lung cancer have come from a single-center prospective trial that is conducted to compare the diagnostic measures such as sensitivity, specificity, PPV and NPV of EBUS and those of mediastinoscopy to detect lymph node metastasis. Some lymph nodes are located in deeper place than others and it is important to confirm the presence of malignancy in differently located lymph nodes. Each examination was performed to evaluate multiple lymph nodes in various stations.

For each subject both diagnostic techniques look into several lymph nodes and determine positivity of cancer malignancy. Also a gold standard involving lymph node dissection is available for each lymph node tested. Each test result is binary outcome. The number of lymph nodes tested per subject is widely varied from three to eighteen and the investigators agree to focus on five lymph nodes which are commonly used to make cancer staging. It is noted that not all five lymph nodes have gold standard test result and that even for the lymph node with gold standard result, some data were coded as missing in either diagnostic test outcome. In the latter case the investigators agree to replace missing diagnostic test outcome with negativity since not being able to see a lymph node is deemed to imply that the unrecognizable lymph node is fine. This result in the number of lymph nodes to vary between two and five per patient. A total of 127 patients, on whom both diagnostic tests were performed, were analyzed with a total of 441 lymph nodes.

Measuring the predictive accuracy of EBUS, we estimate $PPV = 1$ and $NPV = 0.916$. For mediastinoscopy, we estimate $PPV = 1$ and $NPV = 0.901$. Neither test of the equality between the two PPVs or NPVs appears to be statistically significant. The 95% confidence interval for the difference is not available for PPVs and $(-0.048, 0.017)$ for NPVs. Therefore, the two tests EBUS and mediastinoscopy do not seem to have a different PPV or NPV in diagnosing lung cancer. The summary of the equality test of each operating characteristic -sensitivity, specificity, accuracy, PPV, or NPV - for EBUS and mediastinoscopy is presented in Table 4.

In this project, the investigators are actually interested in showing the noninferiority of EBUS to mediastinoscopy in PPV and NPV. We formulate the null hypothesis stating that the PPV (or NPV) for EBUS is at least 10 % worse than the PPV (or NPV) for mediastinoscopy, whereas the alternative hypothesis that we want to prove is stating that the PPV (or NPV) for EBUS is not inferior. Following Silva et al. [12], we obtain the p-values 0 and 1.27×10^{-7} for the noninferiority of PPVs and NPVs, respectively. One-sided lower confidence bound for the difference of PPVs and NPVs are 0 and -0.043 , respectively. Since both lower confidence bounds are larger than -0.1 , the negative of the noninferiority margin, we conclude that EBUS is not inferior to mediastinoscopy in PPV and NPV.

5 Discussion

Recent advances in biotechnology have developed many new medical tests for various purposes including diagnosis, prognosis, and disease screening. Various measures can be used to quantify the test performance. Among frequently used measures for binary tests are sensitivity, specificity, positive predictive value, and negative predictive value. They determine the extent to which the test accurately reflects presence or absence of disease and often used in the various stages of cancer treatment. Several authors have pointed out that the predictive values are more directly applicable in patient care and thus have greater clinical relevance than sensitivity or specificity.

In this article we have developed new methods of comparing operating characteristics - sensitivity, specificity, accuracy, positive predictive value, and negative predictive value - of two diagnostic tests that have clustered binary test outcomes with varying cluster sizes. We have provided consistent variance estimates which can be quite simply calculated for testing the equality of diagnostic performances between two tests. A major advantage of our method is that our method does not need lengthy delta method or variable transformation proposed by Wang et al. [1] and Moskowitz and Pepe [4]. Using simulation studies we have shown that our test statistics comparing the equality of two predictive values perform well in small and moderate sized samples under a variety of conditions. In particular we investigate the diagnostic performance of a newly developed test EBUS to mediastinoscopy by testing the equality of their positive predictive value and negative predictive value. Our test statistic looks familiar and intuitive resembling the one used in testing of the two population proportions. Our approach is useful in that the calculation is quite simple and does not need any modeling.

In case-control study, PPV is a function of sensitivity, specificity, and disease prevalence using Bayes' rule and given by

$$PPV = \frac{\text{sensitivity} \times \text{prevalence}}{\text{sensitivity} \times \text{prevalence} + (1 - \text{specificity}) \times (1 - \text{prevalence})}.$$

So PPV and NPV are dependent on the population chosen and the prevalence of disease which can not be estimated in case-control designs. However, the patients of our motivating example have come from a single-center prospective study conducted to compare the

performance of two diagnostic tests, EBUS and mediastinoscopy, to evaluate lymph node metastasis. Our setting is rather similar to a paired cross-sectional study considered in Wang et al. [1] and Moskowitz and Pepe [4]. In a cross-sectional study on the relevant test population, both PPV and NPV can be estimated directly from the study results.

Predictive values vary with disease prevalence. It has been pointed out that, when the prevalence of disease is very low, the negative predictive value is high even for poorly accurate tests [13]. However, when testing the equality of two predictive values, the prevalence of disease does not seem to affect the performance of the test much in terms of test size and power.

References

1. Wang W, Davis CS, Soong SJ. Comparison of predictive values of two diagnostic tests from the same sample of subjects using weighted least squares. *Statistics in Medicine*. 2006; 25:2215–2229. [PubMed: 16220470]
2. Leisenring W, Alonzo TA, Pepe MS. Comparisons of predictive values of binary medical diagnostic tests for paired designs. *Biometrics*. 2000; 56:345–351. [PubMed: 10877288]
3. Kosinski AS. A weighted generalized score statistic for comparison of predictive values of diagnostic tests. *Statistics in Medicine*. 2012; 31:1002/sim.5587
4. Moskowitz CS, Pepe MS. Comparing the predictive values of diagnostic tests: sample size and analysis for paired study designs. *Clinical Trials*. 2006; 3:272–279. [PubMed: 16895044]
5. Rao JNK, Scott AJ. A simple method for the analysis of clustered binary data. *Biometrics*. 1992; 48:577–585. [PubMed: 1637980]
6. Porte H, Roumilhac D, Eraldi L, Cordonnier C, Puech P, Wurtz A. The role of mediastinoscopy in the diagnosis of mediastinal lymphadenopathy. *European Journal of Cardio-Thoracic Surgery*. 1998; 13:196–199. [PubMed: 9583827]
7. Silvestri GA, Gonzalez AV, Jantz MA, et al. Methods for staging non-small cell lung cancer: Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. *Chest*. 2013; 143:e211S–250S. [PubMed: 23649440]
8. Jung SH, Kang SH, Ahn C. Sample size calculations for clustered binary data. *Statistics in Medicine*. 2001; 20:1971–1982. [PubMed: 11427953]
9. Hu F, Schucany WR, Ahn C. Nonparametric sample size estimation for sensitivity and specificity with multiple observations per subject. *Drug Information Journal*. 2010; 44(5):609–616. [PubMed: 22114363]
10. Katsis A. Sample size for testing homogeneity of two a priori dependent binomial populations using the Bayesian approach. *Journal of Applied Mathematics and Decision Sciences*. 2004; 8(1): 33–42.
11. Um SW, Kim HK, Jung SH, Han J, Lee KJ, Park HY, Choi YS, Shim YM, Ahn MJ, Park K, Ahn YC, Choi JY, Lee KS, Suh JY, Chung MP, Kwon OJ, Kim J, Kim H. Endobronchial Ultrasound versus Mediastinoscopy for Mediastinal Nodal Staging of Non-small Cell Lung Cancer. *Journal of Thoracic Oncology*. To appear.
12. Silva GT, Logan BR, Klein JP. Methods for equivalence and noninferiority testing. *Biology of Blood and Marrow Transplantation*. 2008; 15(1):120–127.
13. Altman DG, Bland JM. Diagnostic tests 2: predictive values. *British Medical Journal*. 1994; 309:102. [PubMed: 8038641]

Table 1

Estimated Kappa statistic under various PPVs and prevalences

Prevalence	PPV_1	PPV_2	Kappa
0.25	0.70	0.70	0.163
	0.70	0.75	0.169
	0.85	0.85	0.467
	0.85	0.90	0.515
0.50	0.70	0.70	0.284
	0.70	0.75	0.310
	0.85	0.85	0.506
	0.85	0.90	0.604
0.75	0.70	0.70	0.160
	0.70	0.75	0.188
	0.85	0.85	0.328
	0.85	0.90	0.426

Table 2

Empirical size under H_0 : $PPV_1 = PPV_2$ and power under H_1 : $PPV_2 = PPV_1 + 0.05$ with $PPV_1 = 0.7$ or 0.85 . n denotes the number of subjects.

Prevalence	n	$PPV_1 = 0.7$		$PPV_1 = 0.85$	
		Size	Power	Size	Power
0.25	50	0.059	0.095	0.065	0.152
	100	0.056	0.107	0.056	0.240
	200	0.053	0.156	0.052	0.412
	500	0.049	0.303	0.047	0.807
0.50	50	0.055	0.326	0.051	0.333
	100	0.056	0.531	0.055	0.554
	200	0.051	0.826	0.047	0.875
	500	0.054	0.996	0.050	0.997
0.75	50	0.057	0.652	0.061	0.674
	100	0.053	0.865	0.054	0.895
	200	0.052	0.937	0.055	0.995
	500	0.046	1.000	0.052	1.000

Table 3

Comparison of empirical powers for equality testing of two PPVs. $PPV_1 = 0.85$, $PPV_2 = 0.75$ and $NPV_1 = NPV_2 = 0.85$. n denotes the number of subjects.

Prevalence	n	WLS ¹	GEE ²	New ³
0.25	100	0.040	0.063	0.073
	200	0.164	0.191	0.208
	500	0.386	0.399	0.404
0.50	100	0.346	0.367	0.390
	200	0.614	0.625	0.652
	500	0.954	0.954	0.966
0.70	100	0.781	0.791	0.821
	200	0.978	0.979	0.982
	500	1.000	1.000	1.000

¹Weighted least square method by Wang et al. [1]

²Generalized estimating equation method by Leisenring et al. [2]

³We set cluster size as $m_i = 1$ for our proposed method.

Table 4

Analysis of the lung cancer data for comparing the diagnostic performance between Mediastinoscopy (M) and EBUS (E)

Parameter	M	E	p-value*	95% CI**
Sens.	0.715	0.764	0.343	(−0.150, 0.052)
Spec.	1.000	1.000	1.000	NA
Accu.	0.921	0.934	0.343	(−0.042, 0.015)
PPV	1.000	1.000	1.000	NA
NPV	0.901	0.916	0.342	(−0.048, 0.017)

* P-value from two-sided tests for testing the equality of two diagnostic tests

** 95% confidence interval for $p_M - p_E$