Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2015, Article ID 128930, 7 pages
http://dx.doi.org/10.1155/2015/128930

Research Article

Efficient Noninferiority Testing Procedures for Simultaneously Assessing Sensitivity and Specificity of Two Diagnostic Tests

Guogen Shan,[1] Amei Amei,[2] and Daniel Young[3]

[1]*Epidemiology and Biostatistics Program, Department of Environmental and Occupational Health, School of Community Health Sciences, University of Nevada Las Vegas, Las Vegas, NV 89154, USA*
[2]*Department of Mathematical Sciences, University of Nevada Las Vegas, Las Vegas, NV 89154, USA*
[3]*Division of Health Sciences, University of Nevada Las Vegas, Las Vegas, NV 89154, USA*

Correspondence should be addressed to Guogen Shan; guogen.shan@unlv.edu

Received 28 May 2015; Revised 3 August 2015; Accepted 6 August 2015

Academic Editor: Qi Dai

Sensitivity and specificity are often used to assess the performance of a diagnostic test with binary outcomes. Wald-type test statistics have been proposed for testing sensitivity and specificity individually. In the presence of a gold standard, simultaneous comparison between two diagnostic tests for noninferiority of sensitivity and specificity based on an asymptotic approach has been studied by Chen et al. (2003). However, the asymptotic approach may suffer from unsatisfactory type I error control as observed from many studies, especially in small to medium sample settings. In this paper, we compare three unconditional approaches for simultaneously testing sensitivity and specificity. They are approaches based on estimation, maximization, and a combination of estimation and maximization. Although the estimation approach does not guarantee type I error, it has satisfactory performance with regard to type I error control. The other two unconditional approaches are exact. The approach based on estimation and maximization is generally more powerful than the approach based on maximization.

1. Introduction

Sensitivity and specificity are often used to summarize the performance of a diagnostic or screening procedure. Sensitivity is the probability of positive diagnostic results given the subject having disease, and specificity is the probability of a negative outcome as the diagnostic result in the nondiseased group. Diagnostic tests with high values of sensitivity and specificity are often preferred and they can be estimated in the presence of a gold standard. For example, two diagnostic tests, the technetium-99m methoxyisobutylisonitrile single photon emission computed tomography (Tc-MIBI SPECT) and the computed tomography (CT), were compared for diagnosing recurrent or residual nasopharyngeal carcinoma (NPC) from benign lesions after radiotherapy in the study by Kao et al. [1]. The gold standard in their study is the biopsy method. The sensitivity and specificity are 73% and 88% for the CT test and 73% and 96% for the Tc-MIBI SPECT test.

Traditionally, noninferiority of sensitivity and specificity between two diagnostic procedures is tested individually using the the McNemar test [2–6]. Recently, Tange et al. [7] developed an approach to simultaneously test sensitivity and specificity in noninferiority studies. Lu and Bean [2] were among the first researchers to propose a Wald-type test statistic for testing a nonzero difference in sensitivity or specificity between two diagnostic tests for paired data. Later, it was pointed out by Nam [3] that the test statistic by Lu and Bean [2] has unsatisfactory type I error control. A new test statistic based on a restricted maximum likelihood method was then proposed by Nam [3] and was shown to have good performance with actual type I error rates closer to the desired rates. This test statistic was used by Chen et al. [8] to compare sensitivity and specificity simultaneously in the presence of a gold standard. Actual type I error rates for a compound asymptotic test were evaluated on some specific points in the sample space. It is well known that the asymptotic method behaves poorly when the sample size is small. Therefore, it is not necessary to comprehensively evaluate type I error rate [9–14].

An alternative to an asymptotic approach is an exact approach conducted by enumerating all the possible tables for given total sample sizes of diseased and nondiseased subjects. The first commonly used unconditional approach is a method based on maximization [15]. In the unconditional approach, only the number of subjects in the diseased and nondiseased group is fixed, not the total number of responses from both groups. The latter is considered as the usual conditional approach by treating both margins of the table as fixed. The p value of the unconditional approach based on maximization is calculated as the maximum of the tail probability over the range of a nuisance parameter [15]. This approach has been studied for many years and it can be conservative due to a smaller actual type I error rate as compared to the test size in small sample settings. One possible reason leading to the conservativeness of this approach is the spikes in the tail probability curve. Storer and Kim [16] proposed another unconditional approach based on estimation which is also known as the parametric bootstrap approach. The maximum likelihood estimate (MLE) is plugged into the null likelihood for the nuisance parameter. Other estimates may be considered if the MLE is not available [7]. Although this estimation based approach is often shown to have type I error rates being closer to the desired size than asymptotic approaches, it still does not respect test size.

A combination of the two approaches based on estimation and maximization has been proposed by Lloyd [4, 17] for the testing of noninferiority with binary matched-pairs data, which can be obtained from a case-control study and a twin study. The p value of the approach based on estimation is used as a test statistic in the following maximization step. It should be noted that there could be multiple estimation steps before the final maximization step. The final step must be a maximization step in order to make the test exact. This approach has been successfully extended for the testing trend with binary endpoints [5, 18]. The rest of this paper is organized as follows. Section 2 presents relevant notation and testing procedures for simultaneously testing sensitivity and specificity. In Section 3, we extensively compare the performance of the competing tests. A real example is illustrated in Section 4 for the application of asymptotic and exact procedures. Section 5 is given to discussion.

2. Testing Approaches

Each subject in a study is evaluated by two dichotomous diagnostic tests, T_1 and T_2, in the presence of a gold standard. Suppose each subject, either diseased or nondiseased, was already determined by the gold standard before performing the two diagnostic tests. Within the diseased group, n_{ij} ($i = 0, 1$; $j = 0, 1$) is the number of subjects with diagnostic results $T_1 = i$ and $T_2 = j$, where $T_k = 0$ and $T_k = 1$ represent negative and positive diagnostic results from the kth test ($k = 1, 2$), respectively, with p_{ij} being the associated probability. The total number of diseased subjects is $n = n_{00} + n_{10} + n_{01} + n_{11}$. Similarly, m_{ij} ($i = 0, 1$; $j = 0, 1$) is the number of subjects with diagnostic results $T_1 = i$ and $T_2 = j$ in the nondiseased group, q_{ij} is the associated probability, and $m = m_{00}+m_{10}+m_{01}+m_{11}$ is the total number of nondiseased patients. Such data can be organized in a $2 \times 2 \times 2$ contingency table (Table 1), where $\mathbf{N} = (n_{00}, n_{10}, n_{01}, n_{11})$ and $\mathbf{M} = (m_{00}, m_{10}, m_{01}, m_{11})$. It is reasonable to assume that the diseased group is independent of the nondiseased group.

Table 1: Test results from two diagnostic tests when a gold standard exists.

Diagnostic result	Diseased group		Nondiseased group	
	$T_2 = 1$	$T_2 = 0$	$T_2 = 1$	$T_2 = 0$
$T_1 = 1$	$n_{11}(p_{11})$	$n_{10}(p_{10})$	$m_{11}(q_{11})$	$m_{10}(q_{10})$
$T_1 = 0$	$n_{01}(p_{01})$	$n_{00}(p_{00})$	$m_{01}(q_{01})$	$m_{00}(q_{00})$

In a study with given total sample sizes n and m in the diseased and the nondiseased groups, respectively, sensitivities of diagnostic tests T_1 and T_2 are estimated as $\widehat{\text{sen}}_1 = (n_{11} + n_{10})/n$ and $\widehat{\text{sen}}_2 = (n_{11} + n_{01})/n$. Similarly, $\widehat{\text{spe}}_1 = (m_{00} + m_{01})/m$ and $\widehat{\text{spe}}_2 = (m_{00} + m_{10})/m$ are specificities for T_1 and T_2, respectively. The estimated difference between their sensitivities is

$$\widehat{\theta}_{\text{sen}} = \widehat{\text{sen}}_1 - \widehat{\text{sen}}_2 = \frac{n_{10} - n_{01}}{n}, \tag{1}$$

and the estimated difference between their specificities is

$$\widehat{\theta}_{\text{spe}} = \widehat{\text{spe}}_1 - \widehat{\text{spe}}_2 = \frac{m_{01} - m_{10}}{m}. \tag{2}$$

The hypotheses for noninferiority of sensitivity and specificity between T_1 and T_2 are given in the format of compound hypotheses as

$$\begin{aligned} H_0&: \theta_{\text{sen}} \le -\delta_{\text{sen}}, \\ &\text{or } \theta_{\text{spe}} \le -\delta_{\text{spe}}, \end{aligned} \tag{3}$$

against

$$\begin{aligned} H_a&: \theta_{\text{sen}} > -\delta_{\text{sen}}, \\ &\theta_{\text{spe}} > -\delta_{\text{spe}}, \end{aligned} \tag{4}$$

where δ_{sen} and δ_{spe} are the clinical meaningful differences between T_1 and T_2 in sensitivity and specificity, $\delta_{\text{sen}} > 0$ and $\delta_{\text{spe}} > 0$. For example, investigators may consider a difference in sensitivity of less than 0.2 not clinically important ($\delta_{\text{sen}} = 0.2$).

A test statistic for the hypotheses H_0: $\theta_{\text{sen}} \le -\delta_{\text{sen}}$ versus H_a: $\theta_{\text{sen}} > -\delta_{\text{sen}}$ is

$$Z_{\text{sen}}(\mathbf{N}) = \frac{\widehat{\theta}_{\text{sen}} + \delta_{\text{sen}}}{\widehat{\sigma}_{\text{sen}}}, \tag{5}$$

where $\widehat{\theta}_{\text{sen}}$ is the estimated difference in sensitivities and $\widehat{\sigma}_{\text{sen}}$ is the estimated standard error of $\widehat{\theta}_{\text{sen}}$. The estimate of $\widehat{\sigma}_{\text{sen}}$ based on a restricted maximum likelihood estimation approach [3, 19, 20] is used, and the associated form is $\widehat{\sigma}_{\text{sen}} = \sqrt{(2\widehat{p}_{01} - \delta_{\text{sen}}(\delta_{\text{sen}} + 1))/n}$, where

$$\begin{aligned} &\widehat{p}_{01} = \frac{\left(\sqrt{B^2 - 8A} - B\right)}{4}, \\ &\text{with } A = \frac{\delta_{\text{sen}}(\delta_{\text{sen}} + 1)\, n_{01}}{n},\ B = -\widehat{\theta}_{\text{sen}}(1 - \delta_{\text{sen}}) - 2\left(\frac{n_{01}}{n} + \delta_{\text{sen}}\right). \end{aligned} \tag{6}$$

There are two reasons for using this estimate instead of some other estimates [2]. First, it has been shown to perform well [8, 20]. Second, it is applicable to a 2×2 contingency table with off-diagonal zero cells. We are going to consider the exact approaches by enumerating all possible tables with some of them having zero cells in off-diagonals. The traditional estimate for σ_{sen} does not provide a reasonable estimate of variance for such tables.

The test statistic for sensitivity in (5) follows a normal distribution asymptotically. The null hypothesis H_0: $\theta_{\text{sen}} \leq -\delta_{\text{sen}}$ would be rejected if the test statistic Z_{sen} in (5) is greater than or equal to z_α, where z_α is the upper α percentile of the standard normal distribution.

As mentioned by many researchers, the asymptotic approach has unsatisfactory type I error control especially in small or medium sample settings. An alternative is an exact approach by enumerating all possible tables for a given total of sample sizes. The first exact unconditional approach considered here is a method based on maximization (referred to as the M approach) [15]. The p value of this approach is calculated as the maximum of the tail probability. In this approach, the worst possible value for the nuisance parameter is found in order to calculate the p value, where $\mathbf{N}_{\text{obs}}$ is the observed data of $\mathbf{N}$. The tail set based on the test statistic Z_{sen} for this approach is

$$R_{Z_{\text{sen}}}\left(\mathbf{N}_{\text{obs}}\right) = \left\{\mathbf{N}; Z_{\text{sen}}\left(\mathbf{N}\right) \geq Z_{\text{sen}}\left(\mathbf{N}_{\text{obs}}\right)\right\}. \tag{7}$$

It is easy to show that $(n_{10}, n_{01} \mid n)$ follows a trinomial distribution with parameters $(n; p_{10}, p_{01})$. Then, the M p value is expressed as

$$P_M\left(\mathbf{N}_{\text{obs}}\right) = \max_{p_{01}\in\Theta} \sum_{\mathbf{N}\in R_{Z_{\text{sen}}}(\mathbf{N}_{\text{obs}})} \Pr\left(n_{10}, n_{01}; p_{01}\right), \tag{8}$$

where $\Theta = (\delta_{\text{sen}}, \min(1, (1+\delta_{\text{sen}})/2))$ is the search range for the nuisance parameter p_{01} and $\Pr(n_{10}, n_{01}; p_{01}) = (n!/n_{10}!n_{01}!(n-n_{10}-n_{01})!)(p_{01}-\delta_{\text{sen}})^{n_{10}} p_{01}^{n_{01}} (1-2p_{01}+\delta_{\text{sen}})^{n-n_{10}-n_{01}}$ is the probability density function for a trinomial distribution.

The M approach could be conservative when the actual type I error is much less than the test size [5, 9]. To overcome this disadvantage of exact unconditional approaches, Lloyd [21] proposed a new exact unconditional approach based on estimation and maximization (referred to as the $E + M$ approach). The first step in this approach is to compute the p value for each table based on the estimation approach [16], also known as parametric bootstrap. We refer to this approach as the E approach. The nuisance parameter in the null likelihood is replaced by the maximum likelihood estimate and the E p value is calculated as

$$P_E\left(\mathbf{N}_{\text{obs}}\right) = \sum_{\mathbf{N}\in R_{Z_{\text{sen}}}(\mathbf{N}_{\text{obs}})} \Pr\left(n_{10}, n_{01}; \widehat{p}_{01}\right). \tag{9}$$

It should be noted that the E approach does not guarantee type I error rate. Once the E p values are calculated for each table, they will be used as a test statistic in the next M step for the p value calculation. The $E + M$ p value is then given by

$$P_{E+M}\left(\mathbf{N}_{\text{obs}}\right) = \max_{p_{01}\in\Theta} \sum_{\mathbf{N}\in R_E(\mathbf{N}_{\text{obs}})} \Pr\left(n_{10}, n_{01}; p_{01}\right), \tag{10}$$

where $R_E(\mathbf{N}_{\text{obs}}) = \{\mathbf{N}; P_E(\mathbf{N}) \leq P_E(\mathbf{N}_{\text{obs}})\}$ is the tail set. The refinement from the E step in the $E + M$ approach could possibly increase the actual type I error rate of the testing procedure which may lead to power increase for exact tests.

Monotonicity is an important property in exact testing procedures to reduce the computation time and guarantee that the maximum of the tail probability is attained at the boundary for noninferiority hypotheses. Berger and Sidik [22] showed that monotonicity is satisfied for paired data for testing one-sided hypothesis based on the NcNemar test. Most importantly, the dimension of nuisance parameters is reduced from two to one [17]. We provide the following theorem to show the monotonicity of the test statistic Z_{sen}.

Theorem 1. *Monotonicity property is satisfied for Z_{sen} under the null hypothesis: $Z_{sen}(n_{10}, n_{01}+1) \leq Z_{sen}(n_{10}, n_{01})$ and $Z_{sen}(n_{10}, n_{01}) \leq Z_{sen}(n_{10}+1, n_{01})$.*

Proof. Let $x_1 = n_{10}$ and $x_2 = n_{10} + 1$. For a given n_{01},

$$\begin{aligned} Z_{\text{sen}}\left(x_1\right) - Z_{\text{sen}}\left(x_2\right) &= \frac{\widehat{\theta}_{\text{sen}}\left(x_1\right)\widehat{\sigma}_{\text{sen}}\left(x_2\right) - \widehat{\theta}_{\text{sen}}\left(x_2\right)\widehat{\sigma}_{\text{sen}}\left(x_1\right) + \delta_{\text{sen}}\widehat{\sigma}_{\text{sen}}\left(x_2\right) - \delta_{\text{sen}}\widehat{\sigma}_{\text{sen}}\left(x_1\right)}{\widehat{\sigma}_{\text{sen}}\left(x_1\right)\widehat{\sigma}_{\text{sen}}\left(x_2\right)} \\ &= \frac{\left(\widehat{\sigma}_{\text{sen}}\left(x_2\right) - \widehat{\sigma}_{\text{sen}}\left(x_1\right)\right)\left(\widehat{\theta}_{\text{sen}}\left(x_1\right) + \delta_{\text{sen}}\right) + \widehat{\sigma}_{\text{sen}}\left(x_1\right)\left(\widehat{\theta}_{\text{sen}}\left(x_1\right) - \widehat{\theta}_{\text{sen}}\left(x_2\right)\right)}{\widehat{\sigma}_{\text{sen}}\left(x_1\right)\widehat{\sigma}_{\text{sen}}\left(x_2\right)} \\ &= \frac{\left[\widehat{\sigma}_{\text{sen}}\left(x_2\right) - \widehat{\sigma}_{\text{sen}}\left(x_1\right)\right]\left(\widehat{\theta}_{\text{sen}}\left(x_1\right) + \delta_{\text{sen}}\right) - \widehat{\sigma}_{\text{sen}}\left(x_1\right)/n}{\widehat{\sigma}_{\text{sen}}\left(x_1\right)\widehat{\sigma}_{\text{sen}}\left(x_2\right)}. \end{aligned} \tag{11}$$

Under the null hypothesis, $\widehat{\theta}_{\text{sen}}(x_1) + \delta_{\text{sen}} \leq 0$. In order to show $Z_{\text{sen}}(x_2) \geq Z_{\text{sen}}(x_1)$, we only need to prove that $\widehat{\sigma}_{\text{sen}}(x_2) - \widehat{\sigma}_{\text{sen}}(x_1) \geq 0$. From (6), we know that

$$\widehat{p}_{01} = \frac{\sqrt{B^2 - 8A} - B}{4} = \frac{-8A}{4\left(\sqrt{B^2 - 8A} + B\right)}, \tag{12}$$

where A and B are given in (6). It is obvious that B is a decreasing function of n_{10} and A is a positive constant number when n_{01} is fixed and $\widehat{p}_{01}$ is an increasing function of n_{10}, which leads to $\widehat{\sigma}_{\text{sen}}(x_2) - \widehat{\sigma}_{\text{sen}}(x_1) \geq 0$. It follows that $Z_{\text{sen}}(x_2) \geq Z_{\text{sen}}(x_1)$.

For a given n_{10}, similar proof will lead to a result of $Z_{\text{sen}}(n_{10}, n_{01}+1) \leq Z_{\text{sen}}(n_{10}, n_{01})$. □

The probability of the tail set for either the M approach or the $E + M$ approach has two nuisance parameters, p_{01} and p_{10}. Applying the theorem for the monotonicity property, type I error of the test occurs on the boundary of the two-dimensional nuisance parameter space, $p_{01} = p_{10}$. Therefore, there is only one nuisance parameter, p_{01}, in the definition of the two exact p values.

For testing the specificity, the asymptotic approach, the M approach, the E approach, and the $E + M$ approach can be similarly applied to test the hypotheses H_0: $\theta_{\text{spe}} \leq -\delta_{\text{spe}}$ against H_a: $\theta_{\text{spe}} > -\delta_{\text{spe}}$. The test statistic [3, 19, 20] would be

$$Z_{\text{spe}} = \frac{\widehat{\theta}_{\text{spe}} + \delta_{\text{spe}}}{\widehat{\sigma}_{\text{spe}}}, \tag{13}$$

where $\widehat{\sigma}_{\text{spe}} = (2\widehat{q}_{10} - \delta_{\text{spe}}(\delta_{\text{spe}}+1))/n$ is the estimated standard error of $\widehat{\theta}_{\text{spe}}$, $\widehat{q}_{10} = (\sqrt{D^2 - 8C} - C)/4$, $C = \delta_{\text{spe}}(\delta_{\text{spe}}+1)m_{10}/m$, and $D = -\widehat{\theta}_{\text{spe}}(1 - \delta_{\text{spe}}) - 2(m_{10}/m + \delta_{\text{spe}})$. Under the null hypothesis, one can show that the monotonicity of Z_{spe} is in a similar way to Z_{sen}.

When there are two diagnostic tests available, we may want to simultaneously confirm the noninferiority of sensitivity and specificity for the two tests. The population from the diseased group and the nondiseased group can be reasonably assumed to be independent of each other. Then, the joint probability is a product of two probabilities:

$$\begin{aligned} &\Pr\left(\mathbf{N}, \mathbf{M} \mid \mathbf{N} \in R\left(\mathbf{N}_{\text{obs}}\right), \mathbf{M} \in R\left(\mathbf{M}_{\text{obs}}\right)\right) \\ &\quad = \Pr\left(\mathbf{N} \mid \mathbf{N} \in R\left(\mathbf{N}_{\text{obs}}\right)\right) \Pr\left(\mathbf{M} \mid \mathbf{M} \in R\left(\mathbf{M}_{\text{obs}}\right)\right), \end{aligned} \tag{14}$$

where R is the rejection region. Let α_{sen} and α_{spe} be the significance levels for testing sensitivity and specificity separately. We can reject the compound null hypothesis H_0: $\theta_{\text{sen}} \leq -\delta_{\text{sen}}$ or $\theta_{\text{spe}} \leq -\delta_{\text{spe}}$ at the significance level of α when the sensitivity null hypothesis is rejected at the level of α_{sen} and the specificity null is rejected at the level of α_{spe}, where $\alpha_{\text{sen}}\alpha_{\text{spe}} = \alpha$. For simplicity, we assume $\alpha_{\text{sen}} = \alpha_{\text{spe}} = \sqrt{\alpha}$.

3. Numerical Study

We already know that both the asymptotic approach and the E approach do not guarantee type I error rate; however, it is still interesting to compare type I error control for the following four approaches: (1) the asymptotic approach, (2) the E approach, (3) the M approach, and (4) the $E + M$ approach. We select three commonly used values of δ_{sen} and δ_{spe}, 0.05, 0.1, and 0.2. For each configuration of δ_{sen} and δ_{spe}, actual type I error rates are presented in Table 2 for sample size $n = m = 20$ and in Table 3 for sample size $n = m = 50$ at the significance level of $\alpha = 0.05$. It can be seen from both tables that the asymptotic approach generally has inflated type I error rates. Both the M approach and the $E+M$ approach are exact tests and respect the test size as expected. Although the E approach does not guarantee type I error rate, the performance of the E approach is much better than the asymptotic approach regarding the type I error control. Even for large sample size, the M approach is still conservative. The $E + M$ approach has an actual type I error rate which is very close to the nominal level when $n = m = 50$.

Table 2: Actual type I error rates $n = m = 20$.

δ_{sen}	δ_{spe}	A approach	M approach	E approach	$E + M$ approach
0.05	0.05	0.1285	0.0343	0.0499	0.0489
	0.1	0.0894	0.0380	0.0489	0.0490
	0.2	0.0877	0.0401	0.0479	0.0480
0.1	0.05	0.0894	0.0380	0.0489	0.0490
	0.1	0.0621	0.0421	0.0481	0.0492
	0.2	0.0610	0.0444	0.0470	0.0481
0.2	0.05	0.0877	0.0401	0.0479	0.0480
	0.1	0.0610	0.0444	0.0470	0.0481
	0.2	0.0599	0.0468	0.0460	0.0471

Table 3: Actual type I error rates $n = m = 50$.

δ_{sen}	δ_{spe}	A approach	M approach	E approach	$E + M$ approach
0.05	0.05	0.0821	0.0300	0.0492	0.0498
	0.1	0.0731	0.0341	0.0489	0.0493
	0.2	0.0677	0.0356	0.0486	0.0498
0.1	0.05	0.0731	0.0341	0.0489	0.0493
	0.1	0.0650	0.0387	0.0486	0.0489
	0.2	0.0603	0.0404	0.0482	0.0494
0.2	0.05	0.0677	0.0356	0.0486	0.0498
	0.1	0.0603	0.0404	0.0482	0.0494
	0.2	0.0559	0.0422	0.0479	0.0499

The asymptotic approach will not be included in the power comparison due to inflated type I error rates. We include the E approach in the power comparison with the M approach and the $E + M$ approach due to the good performance of type I error control in the E approach. The power is a function of four parameters: p_{01}, θ_{sen}, q_{10}, and θ_{spe}

$$\begin{aligned} \text{Power}_{\phi} = &\sum_{\mathbf{N} \in R_{\text{sen}}} \Pr\left(n_{10}, n_{01}; p_{01}, \theta_{\text{sen}}\right) \\ &\cdot \sum_{\mathbf{M} \in R_{\text{spe}}} \Pr\left(m_{10}, m_{01}; q_{10}, \theta_{\text{spe}}\right), \end{aligned} \tag{15}$$

where $\phi = E, M$ and $E + M$ approaches and R_{sen} and R_{spe} are the rejection region for the diseased group and the nondiseased group at a significance level of $\sqrt{\alpha}$ based on the ϕ approach. Given the two parameters q_{10} and θ_{spe} in the nondiseased group, the power is a function of θ_{sen} for a given

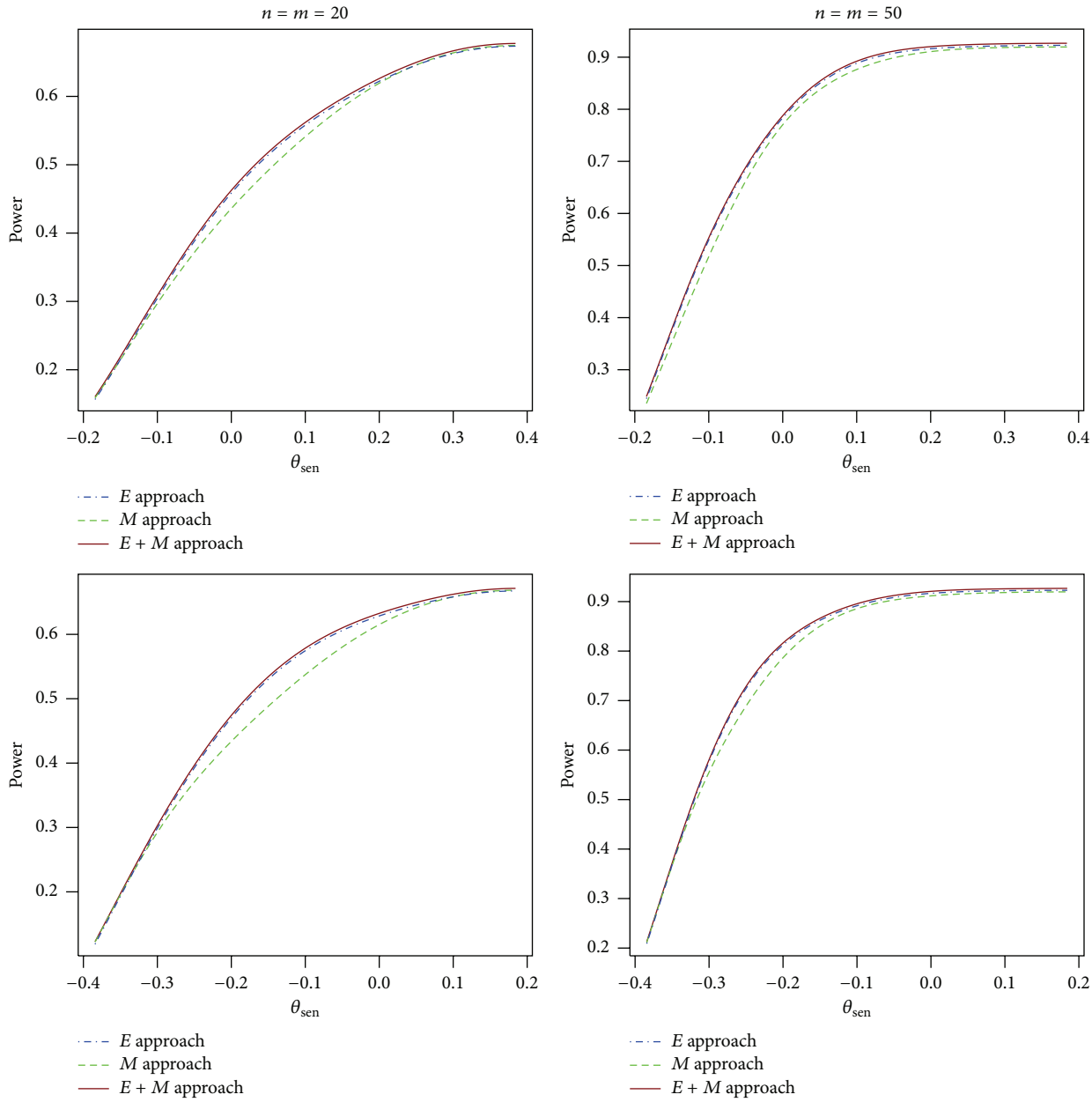

Figure 1: Power curves for the E approach, the M approach, and the $E + M$ approach for balanced data with $\theta_{spe} = 0$, $q_{10} = 0.2$, $p_{01} = 0.3$, $\delta_{sen} = 0.2$, and $\delta_{spe} = 0.2$ for the first row and $\theta_{spe} = 0$, $q_{10} = 0.2$, $p_{01} = 0.4$, $\delta_{sen} = 0.4$, and $\delta_{spe} = 0.2$ for the second row.

p_{01}. We compared multiple configurations of the parameters. Typical comparison results for balanced data are presented in Figure 1. The power difference between the E approach and the $E + M$ approach is often negligible and both are generally more powerful than the M approach. We also compared the power for unbalanced data with the ratio of sample size 1/2, 1/3, 2, and 3. Similar results are observed as compared to the balanced data; see Figure 2. We also observe similar results in comparing the power as a function of θ_{spe} for the given θ_{sen}, p_{01}, and q_{10}.

4. An Example

Kao et al. [1] compared diagnostic tests to detect recurrent or residual NPC in the presence of a gold standard test. Simultaneous comparison of sensitivity and specificity is conducted between the CT test (T_1) and the Tc-MIBI SPECT test (T_2), with $n = 11$ and $m = 25$. The diagnostic results using these two tests are presented in Table 4. The sensitivity and specificity are 73% and 88% for the CT test and 73% and 96% for the Tc-MIBI SPECT test. The clinical meaningful

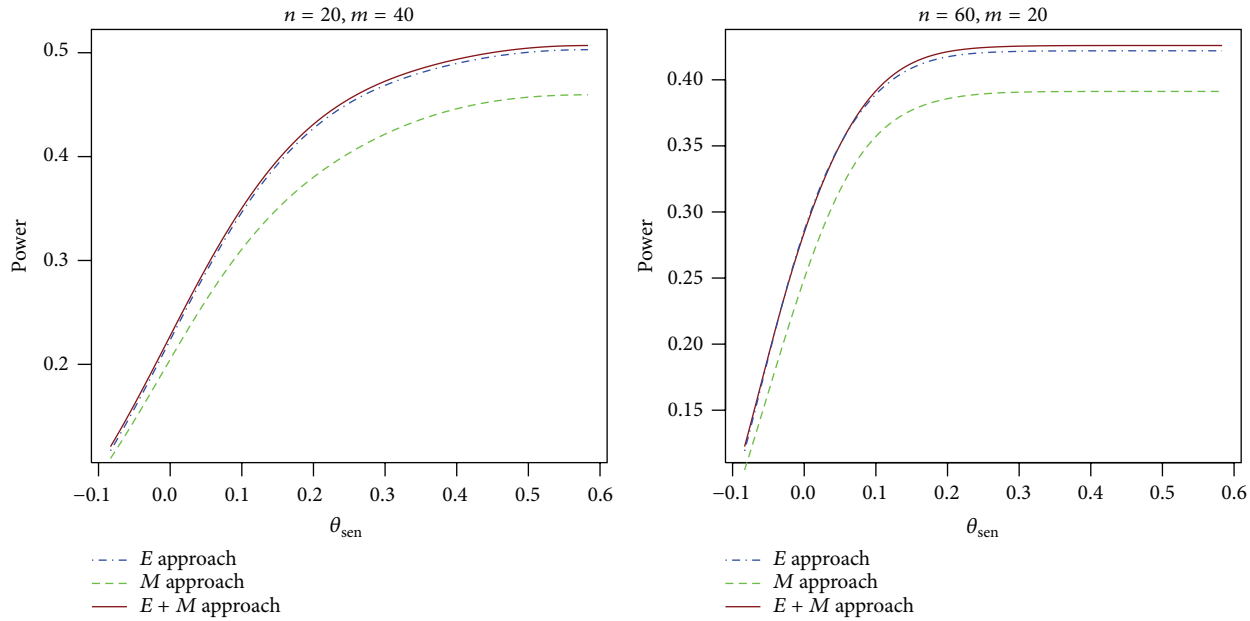

Figure 2: Power curves for the E approach, the M approach, and the $E + M$ approach for unbalanced data with $\theta_{spe} = 0$, $q_{10} = 0.3$, $p_{01} = 0.2$, $\delta_{sen} = 0.1$, and $\delta_{spe} = 0.1$.

Table 4: Results of CT and Tc-MIBI SPECT diagnoses of NPC in the presence of a gold standard.

Diagnostic result	Diseased group (NPC: +)		Nondiseased group (NPC: −)	
	CT: +	CT: −	CT: +	CT: −
Tc-MIBI SPECT: +	5	3	1	0
Tc-MIBI SPECT: −	3	0	2	22

difference in sensitivity and specificity is assumed to be $\delta_{sen} = 0.01$ and $\delta_{spe} = 0.01$, respectively. Four testing procedures are used to calculate the p value: (1) the asymptotic approach; (2) the E approach; (3) the M approach; and (4) the $E + M$ approach. The p values based on the asymptotic, E, M, and $E + M$ approaches are 0.0677, 0.0317, 0.0764, and 0.0418, respectively. Both the E approach and the $E + M$ approach reject the null hypothesis at a 5% significance level, while the asymptotic approach and the M approach do not. It should be noted that the two tests have the same sensitivities which may contribute to the significant result even with a small difference between the two tests.

5. Discussion

In this paper, the asymptotic approach, the E approach, the M approach, and the $E + M$ approach are considered for testing sensitivity and specificity simultaneously in the presence of a gold standard. Although the E approach does not guarantee type I error rate, it has good performance regarding type I error rate control and the difference between the E approach and the $E+M$ approach is negligible. Since the computational time is not an issue for this problem and the $E + M$ approach is an exact method, the $E + M$ approach is recommended for use in practice due to the power gain as compared to the M approach.

Tang [9] has studied the E approach and the M approach for comparing sensitivity and specificity when combining two diagnostic tests. The E approach has been shown to be a reliable testing procedure. We would consider comparing the $E+M$ approach with the E approach in this context as a future work. The intersection-union method may be considered for testing sensitivity and specificity [8].

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors would like to thank the Associate Editor and the three reviewers for their valuable comments and suggestions. The authors also thank Professor Michelle Chino for her valuable comments. Shan's research is partially supported by the NIH Grant 5U54GM104944.

References

[1] C.-H. Kao, Y.-C. Shiau, Y.-Y. Shen, and R.-F. Yen, "Detection of recurrent or persistent nasopharyngeal carcinomas after radiotherapy with technetium-99m methoxyisobutylisonitrile single photon emission computed tomography and computed tomography: comparison with 18-fluoro-2-deoxyglucose positron emission tomography," *Cancer*, vol. 94, no. 7, pp. 1981–1986, 2002.

[2] Y. Lu and J. A. Bean, "On the sample size for one-sided equivalence of sensitivities based upon McNemar's test," *Statistics in Medicine*, vol. 14, no. 16, pp. 1831–1839, 1995.

[3] J.-M. Nam, "Establishing equivalence of two treatments and sample size requirements in matched-pairs design," *Biometrics*, vol. 53, no. 4, pp. 1422–1430, 1997.

[4] C. J. Lloyd and M. V. Moldovan, "A more powerful exact test of noninferiority from binary matched-pairs data," *Statistics in Medicine*, vol. 27, no. 18, pp. 3540–3549, 2008.

[5] G. Shan, C. Ma, A. D. Hutson, and G. E. Wilding, "An efficient and exact approach for detecting trends with binary endpoints," *Statistics in Medicine*, vol. 31, no. 2, pp. 155–164, 2012.

[6] G. Shan and W. Wang, "Exact one-sided confidence limits for Cohen's kappa as a measurement of agreement," *Statistical Methods in Medical Research*, 2014.

[7] N.-S. Tang, M.-L. Tang, and S.-F. Wang, "Sample size determination for matched-pair equivalence trials using rate ratio," *Biostatistics*, vol. 8, no. 3, pp. 625–631, 2007.

[8] J. J. Chen, H.-M. Hsueh, and J.-P. Liu, "Simultaneous noninferiority test of sensitivity and specificity for two diagnostic procedures in the presence of a gold standard," *Biometrical Journal*, vol. 45, no. 1, pp. 47–60, 2003.

[9] M.-L. Tang, "On simultaneous assessment of sensitivity and specificity when combining two diagnostic tests," *Statistics in Medicine*, vol. 23, no. 23, pp. 3593–3605, 2004.

[10] I. S. F. Chan, N.-S. Tang, M.-L. Tang, and P.-S. Chan, "Statistical analysis of noninferiority trials with a rate ratio in small-sample matched-pair designs," *Biometrics*, vol. 59, no. 4, pp. 1170–1177, 2003.

[11] G. Shan, "A better confidence interval for the sensitivity at a fixed level of specificity for diagnostic tests with continuous endpoints," *Statistical Methods in Medical Research*, 2014.

[12] W. Wang and G. Shan, "Exact confidence intervals for the relative risk and the odds ratio," *Biometrics*, 2015.

[13] G. Shan and W. Wang, "ExactCIdiff: an R package for computing exact confidence intervals for the difference of two proportions," *The R Journal*, vol. 5, no. 2, pp. 62–71, 2013.

[14] G. Shan, "Exact unconditional testing procedures for comparing two independent Poisson rates," *Journal of Statistical Computation and Simulation*, vol. 85, no. 5, pp. 947–955, 2015.

[15] D. Basu, "On the elimination of nuisance parameters," *Journal of the American Statistical Association*, vol. 72, no. 358, pp. 355–366, 1977.

[16] B. E. Storer and C. Kim, "Exact properties of some exact test statistics for comparing two binomial proportions," *The Journal of the American Statistical Association*, vol. 85, no. 409, pp. 146–155, 1990.

[17] C. J. Lloyd, "A new exact and more powerful unconditional test of no treatment effect from binary matched pairs," *Biometrics*, vol. 64, no. 3, pp. 716–723, 2008.

[18] G. Shan, C. Ma, A. D. Hutson, and G. E. Wilding, "Some tests for detecting trends based on the modified Baumgartner-Weiß-Schindler statistics," *Computational Statistics & Data Analysis*, vol. 57, no. 1, pp. 246–261, 2013.

[19] T. Tango, "Equivalence test and confidence interval for the difference in proportions for the paired-sample design," *Statistics in Medicine*, vol. 17, no. 8, pp. 891–908, 1998.

[20] J.-P. Liu, H.-M. Hsueh, E. Hsieh, and J. J. Chen, "Tests for equivalence or non-inferiority for paired binary data," *Statistics in Medicine*, vol. 21, no. 2, pp. 231–245, 2002.

[21] C. J. Lloyd, "Exact p-values for discrete models obtained by estimation and maximization," *Australian and New Zealand Journal of Statistics*, vol. 50, no. 4, pp. 329–345, 2008.

[22] R. L. Berger and K. Sidik, "Exact unconditional tests for a 2×2 matched-pairs design," *Statistical Methods in Medical Research*, vol. 12, no. 2, pp. 91–108, 2003.