Austin Biometrics and Biostatistics

Open Access

Austin Publishing Group

Research Article

Follow-up Design for Comparing Two Binary Diagnostic Tests

Kenta Murotani[1]*, Akihiro Hirakawa[2], Yoshiko Aoyama[3] and Takashi Yanagawa[3]

[1]Center for Clinical Research, Aichi Medical University, Japan

[2]Center for Advanced Medicine and Clinical Research, Nagoya University Hospital, Japan

[3]The Biostatistics Center, Kurume University, Japan

***Corresponding author:** Kenta Murotani, Center for Clinical Research, Aichi Medical University, 1-1 Yazakokarimata, Nagakute, Aichi, Japan

Received: January 06, 2015; **Accepted:** May 11, 2015; **Published:** May 26, 2015

Abstract

Most conventional methods of comparing two diagnostic tests require patients whose true disease statuses are known. We deal with in this paper a problem of comparing two binary diagnostic tests (referred to as new and standard tests) in a follow-up design, where there are no gold standards. Assume that each patient is examined twice by new and standard tests, respectively. We employed a comparison measure Ψ, which is compared on the basis of the odds ratio of the new and standard test. It is not possible to estimate Ψ from the full likelihood function based on the design, even if two independent multinomial distributions are assumed to the data. Therefore, we focus only on data from discordant pairs between new and standard tests. We construct conditional likelihood conditioned on those pairs and estimate parameters involved in the conditional likelihood. An estimate of Ψ is obtained by plugging those estimates in Ψ. The asymptotic normality of the estimator of Ψ is shown based on delta method and a confidence interval of Ψ is developed. A method of sample size determination for this design is also proposed. Simulation is conducted to study the behavior of the proposed method by considering several scenarios.

Keywords: Follow-up design; Diagnostic test; Comparison; No gold standard

Introduction

Accurate diagnosis of the patient is crucial when planning the treatment of a disease. After determining the accurate diagnosis has been determined, the patient can begin receiving adequate treatment. An accurate evaluation and selection of the diagnostic method plays an important role in the patients' health. A medical method that aims at determining whether a patient is affected by a disease is called a ''diagnostic test''. Particularly, diagnostic tests that evaluate the strength of suspicion of certain diseases on binary (''positive'' and ''negative'') are called ''binary diagnostic tests''. To determine which of the two binary diagnostic tests is statistically better, the sensitivity and specificity must be closely examined [1,2]. Sensitivity and specificity are defined by the following equation: Sensitivity = Pr (T=1|D=1), Specificity = Pr (T=0|D=0), Where T indicates the diagnostic results according to the binary diagnostic test, and D indicates the actual condition of the disease. The T (D)=1 indicates positivity (disease), and 0 indicates negativity (not disease). Sensitivity is the conditional probability for patients who are actually disease to be diagnosed as positive, and specificity is the conditional probability for patients who are not actually disease to be diagnosed as negative. In both cases, values closer to 1 mean that the diagnostic test is accurate. If an observation of each patient's actual disease condition (D) is conducted, the sensitivity and specificity can be estimated simply by calculating the proportion.

However, an accurate observation of the value of D involves methods that are often invasive for the patient. In the case of cancers, for example, the value of D can be assessed only by collecting cell samples through biopsy or surgery, and by determining the diagnosis in a comprehensive manner by using pathological and histological methods.

An example of an actual test is that of Berg et al. [3], who performed biopsy in patients with elevated risks of breast cancer for the determination of a definitive diagnosis, to examine whether ''mammography alone'' and ''mammography combined with ultrasound'' was effective as diagnostic tests for breast cancer detection. Similarly, in Japan, a large-scale randomized controlled trial of breast cancer screening methods (mammography alone vs. mammography combined with ultrasound) is being conducted on 100,000 women in their 40s [4]. In this study, the definitive diagnosis was determined on the basis of biopsy or surgery for patients whose overall screening results indicated a need for thorough examination. These examples involve two important issues. In other words, when sensitivity and specificity are evaluated directly for comparison, then information related to the definitive diagnosis is required, and the problem is that this imposes a huge burden both on the patient and on the health care workers.

Therefore, in this paper, we propose a methodology for the comparison of two binary diagnostic tests (referred to hereinafter as ''new test'' and ''standard test'') in the absence of a definitive diagnosis, and discuss the follow-up design by using the said methodology. The characteristic of this method is that each patient was twice subjected to the new test and the standard test, respectively, both for a short period and focus was given to findings in which discordant results were obtained from the new and standard tests. This research paper comprises the following: Section 2 summarizes the criteria considered while comparing the two diagnostic tests; Section 3, we propose the methodology; Section 4, numerical simulations are performed using several scenarios; Section 5, a discussion is provided.

Comparison measure

$T_N, T_S \in \{0,1\}$ are random variables representing the results of the

Citation: Murotani K, Hirakawa A, Aoyama Y and Yanagawa T. Follow-up Design for Comparing Two Binary Diagnostic Tests. Austin Biom and Biostat. 2015;2(2): 1016.

diagnosis according to the two binary diagnostic tests, namely the new test and the standard test. Murotani et al. [5] previously summarized the criteria for comparing the standard test and the new test as (C1), (C2), (C3), (C4) as follows:

(C1) $Pr(T_N=1 \mid D=1) > Pr(T_S=1 \mid D=1)$ and $Pr(T_N=0 \mid D=0) > Pr(T_S=0 \mid D=0)$,

(C2) $Pr(D=1 \mid T_N=1) > Pr(D=1 \mid T_S=1)$ and $Pr(D=0 \mid T_N=1) > Pr(D=0 \mid T_S=0)$,

(C3) $Pr(T_N=1 \mid D=1) + Pr(T_N=1 \mid D=1) > Pr(T_S=1 \mid D=1) + Pr(T_S=0 \mid D=0)$, and

$$(C4)\ \frac{Pr(T_N=1 \mid D=1)Pr(T_N=0 \mid D=0)}{Pr(T_N=0 \mid D=1)Pr(T_N=1 \mid D=0)} > \frac{Pr(T_S=1 \mid D=1)Pr(T_S=0 \mid D=0)}{Pr(T_S=0 \mid D=1)Pr(T_S=1 \mid D=0)}.$$

(C1) is compared on the basis of the sensitivity and specificity. In (C1) and (C2), the conditions are reversed. In other words, comparison was made on the basis of the probability for the patients actual condition to be "presence of disease" ("absence of disease") when (C2) was diagnosed as positive (negative). Therefore, the diagnostic tests were compared on the basis of their capability to predict the diagnosis. (C3) was compared on the basis of the size of the sum of sensitivity and specificity. This is the equivalent to selecting a diagnostic test with a large Area under the Curve (AUC). (C4) was compared on the basis of the odds ratio of the new test and standard test.

The meanings of the (C4) criteria were as follows: When T_N was $T_N=1$, the predictive capacity was expressed as follows:

$O_1=Pr(D=1 \mid T_N=1)/Pr(D=0 \mid T_N=1)$.

When T_N was $T_N=0$, the predictive capacity was expressed as follows:

$O_2=Pr(D=0 \mid T_N=0)/Pr(D=1 \mid T_N=0)$.

The larger the predictive value of $T_N=1$, the greater the value of O_1. The larger the predictive value of $T_N=0$, the greater the value of O_2; in other words, the lower the value of O_2^{-1}. Therefore, the ratio of the two (O_1/O_2) expresses the strength of the relationship between the new test and D. Higher values of the ratio would indicate that the new test is a good diagnostic test. Similarly, the standard test was also defined by the odds ratio, and the (C4) of the new test was compared with that of the standard test on the basis of the meaning of the odds ratio. In this paper, the diagnostic tests were compared on the basis of the meaning of (C4).

The parameters summarizing the (C4) criteria are defined by the following equation:

$$\psi = \frac{Pr(T_N=1 \mid D=1)Pr(T_N=0 \mid D=0)}{Pr(T_N=0 \mid D=1)Pr(T_N=1 \mid D=0)} \Big/ \frac{Pr(T_S=1 \mid D=1)Pr(T_S=0 \mid D=0)}{Pr(T_S=0 \mid D=1)Pr(T_S=1 \mid D=0)}.$$

According to the (C4) criteria, the following interpretations can be made, depending on the value of Ψ: {Ψ >1 if T_N is superior to T_S; Ψ =1 if T_N and T_S are equal; Ψ <1 is inferior to T_S.

Thus, Ψ is a criterion for the comparison of the two diagnostic tests.

If Ψ can be estimated on the basis of the data, then the two binary diagnostic tests can be compared on the basis of the estimated value. In addition, if the distribution associated with the estimator of Ψ can be calculated, then a hypothesis testing pertaining to Ψ as well as the estimation of the confidence interval can also be conducted, and a follow-up design for the comparison of two binary diagnostic tests, including the planning of the number of cases, can be proposed. In the absence of definitive diagnosis (in the absence of observation of D), and on the basis of the data obtained by application of the new test and the standard test twice on each patient, the estimate of Ψ and its asymptotic distribution were calculated under several assumptions. From the next section, we discuss the methodology in concrete terms.

Methodology

Notation and definition

$\{T_{Nij}(T_{Sij}),\ j=1,2,\ldots,n\}$was a random variable representing the diagnostic results of the new test (standard test) that the *i* patient underwent for the j^{th} time; $\{D_i, i=1,2,\ldots,n\}$ was a random variable representing the actual status of the i^{th} individual's disease. This implies that D_i does not depend on *j*, but the actual status of the disease remained unchanged at the time of the first and second application of the new test and the standard test. This can be ensured by applying the two diagnostic tests in a relatively short period, during which the actual condition of the disease remains unchanged. D_i is a non-observed random variable. T_{Nij},T_{Sij}, and D_i are binary random variables in which 1 means positive (disease) and 0 means negative (not disease). In addition, it was assumed that $p=Pr(D_i=1)$ for all i.

The value p represents the prevalence rate. If $\{\varepsilon_{Nij},\varepsilon_{Sij}\}$ are considered as instances of T_{Nij},T_{Sij}, the data obtained from the application of the new test and standard test twice to *n* patients without definitive diagnosis are expressed as $(\varepsilon_{Ni1},\varepsilon_{Si1},\varepsilon_{Ni2},\varepsilon_{Si2}), i=1,2,\ldots,n$.

The cell probability p_{ikl} was $p_{ikl}=Pr(T_{Nij}=k,T_{Sij}=\ell), k,\ell\in\{0,1\}$. In addition, regarding p_{ikl}, if the actual condition of the disease is known, then $q_{Dkli}=Pr(T_{Nij}=k,T_{Sij}=\ell \mid D=1)$,

and $q_{\bar{D}kli}=Pr(T_{Nij}=k,T_{Sij}=l \mid D=0)$ for *i,j,k* and *l*. Here, $p_{ikl}, q_{Dkli}, q_{\bar{D}kli}$ were independent of *j*, but this meant that the cell probability remained unchanged in both the first and the second diagnostic results.

Design based approach

In this section, we consider the probability distribution on the basis of the method of extraction of individuals and to construct the likelihood. The new and standard tests, respectively, were applied twice on the i^{th} patient, and therefore, the j^{th} j=1,2 diagnostic results can be summarized in 2×2 contingency tables. When the two-dimensional random variable representing the diagnostic results obtained at the time when the new and standard tests were applied on the i^{th} patient (T_{Ni1},T_{Si1}), and the second diagnostic results of the new and standard tests (T_{Ni2},T_{Si2}) follow a mutually independent multinomial distribution, the likelihood for the i^{th} patient can be expressed in the following equation:

$$P_{i00}^{(1-\varepsilon_{Ni1})(1-\varepsilon_{Si1})}P_{i01}^{(1-\varepsilon_{Ni1})\varepsilon_{Si1}}P_{i10}^{\varepsilon_{Ni1}(1-\varepsilon_{Si1})}P_{i11}^{\varepsilon_{Ni1}\varepsilon_{Si1}}$$
$$\times P_{i00}^{(1-\varepsilon_{Ni2})(1-\varepsilon_{Si2})}P_{i01}^{(1-\varepsilon_{Ni2})\varepsilon_{Si2}}P_{i10}^{\varepsilon_{Ni2}(1-\varepsilon_{Si2})}P_{i11}^{\varepsilon_{Ni2}\varepsilon_{Si2}}.$$

In addition, because the actual status of the disease is unknown, the cell probability p_{ikl} will be the mixture probability of the mixing ratio p, as represented by $p_{ikl}=pq_{Dkli}+(1-p)q_{\bar{D}kli}$. In summary, the overall likelihood function (L) of n patients is provided as follows:

$$L=\prod_{i=1}^{n}\left\{(pq_{D00i}+(1-p)q_{\overline{D}00i})^{\sum_{j=1}^{2}(1-\varepsilon_{Nij})(1-\varepsilon_{Sij})}\right\}\left\{(pq_{D01i}+(1-p)q_{\overline{D}01i})^{\sum_{j=1}^{2}(1-\varepsilon_{Nij})(1-\varepsilon_{Sij})}\right\}$$

$$\times\left\{(pq_{D10i}+(1-p)q_{\overline{D}10i})^{\sum_{j=1}^{2}\varepsilon_{Nij}(1-\varepsilon_{Sij})}\right\}\left\{(pq_{D11i}+(1-p)q_{\overline{D}11i})^{\sum_{j=1}^{2}\varepsilon_{Nij}\varepsilon_{Sij})}\right\}.$$

Here q_{D10i} q_{D01i}/q_{D01i} q_{D10i}, does not depend on {i,j} and the results of the new and standard test are mutually independent when conditioned with the actual disease status, Ψ can be expressed by the following equation.

$$\psi=\frac{q_{D10i}q_{\overline{D}01i}}{q_{D01i}q_{\overline{D}10i}}.$$

When Ψ is estimated based on the overall likelihood L, it is important to know whether L is an exponential family. If L is an exponential family, then it is sufficient estimated on the basis of the conditional likelihood of Ψ when sufficient sample statistics on nuisance parameters other than Ψ are conditioned? However, unfortunately, L is not an exponential family. Therefore, it is difficult to estimate the Ψ.

Conditional approach

When the overall likelihood is constructed by assuming the multinomial distribution estimated on the basis of the design, the cell probability will be the mixture probabilities of the not diseased group and that of the diseased group where the prevalence is a mixing ratio. Thus, the overall likelihood was not an exponential family, and it was not possible to estimate Ψ based on sufficient statistics. In this section, we limit the data to those used in the analysis, and propose a new approach composed of conditional likelihood functions.

First, we assume the following (E1):

(E1) The data, in which the results of the new test and standard test were consistent with each other, are not related to the comparison of diagnostic tests.

If (E1) is expressed in other words, it insists on the fact that at the time of the analysis, there is no need to take into consideration the data in which the new test and standard test produced the same results. Based on an assumption (E1), considerations are only given to the pairs of data in which the diagnostic results differed from each other (discordant pairs) in the new test and standard test. Therefore, the following sets of A, B_1 and B_2 are defined depending on the number of times the new test and standard test.

$A=\{i: (T_{Ni1},T_{Si1}, T_{Ni2}, T_{Si2})= (0,1,0,1),(0,1,1,0),(1,0,0,1),(1,0,1,0)\}$,

$B_1=\{i: (T_{Ni1},T_{Si1}, T_{Ni2}, T_{Si2})= (0,1,1,0),(0,1,0,0),(1,0,1,1),(1,0,0,0)\}$,

$B_2=\{i: (T_{Ni1},T_{Si1}, T_{Ni2}, T_{Si2})= (1,1,0,1),(0,0,0,1),(1,1,1,0),(0,0,1,0)\}$.

"A" represented a set of individuals in whom the results of the new test and standard test differed from each other, both the first time and the second time they were conducted. B_1 (B_2) represents a set of individuals in whom the results of the 'new test' and 'standard test' differed from each other the first time (the second time) they were conducted.

For $A\cup B_1\cup B_2$, $T_{ij'}^{*}$ is defined by the following equation.

$$T_{ij'}^{*}=\begin{cases}1 & \text{if}(T_{Nij'},T_{Sij'})=(1,0)\\ 0 & \text{if}(T_{Nij'},T_{Sij'})=(0,1)\end{cases},\ j'=1,2.,$$

where

$$\Pr\left(T_{ij'}^{*}=1\right)=\frac{\Pr(T_{Nij'}=1,T_{Sij'}=0)}{\Pr(T_{Nij'}=1,T_{Sij'}=0)+\Pr(T_{Nij'}=0,T_{Sij'}=1)},$$

$$\Pr\left(T_{ij'}^{*}=1\right)=1-\Pr(T_{ij'}^{*}=0). \qquad (1)$$

For $i\in A$, the observed values of T_{i1}^{*},T_{i2}^{*} are $(\varepsilon_{i1}^{*},\varepsilon_{i2}^{*})$. In the same manner, for $i\in B_1$, the observed value of T_{i1}^{*} is $\varepsilon_{iB_1}^{*}$, for $i\in B_2$, the observed value of T_{i2}^{*} is $\varepsilon_{iB_2}^{*}$. In addition, for the i^{th} individual, M_i is defined as $M_i=2$ for $i\in A$, and as $M_i=1$ for $i\in B_1\in B_2$. In addition, (A1), (A2), (A3) are assumed as follows:

(A1) $i\in A, \Pr(T_{i1}^{*}=\varepsilon_{i1}^{*},T_{i2}^{*}=\varepsilon_{i2}^{*}\,|\,D_i=\varepsilon_i)=\prod_{j'=1}^{2}\Pr(T_{ij'}^{*}=\varepsilon_{ij'}^{*}\,|\,D_i=\varepsilon_i)$,

(A2) $\alpha=\Pr(T_{ij'}^{*}=1\,|\,D_i=1),\beta=\Pr(T_{ij'}^{*}=0\,|\,D_i=0)$, $j'=1,2, i\in A\cup B_1\cup B_2$, and

(A3) $(T_{i1}^{*},T_{i2}^{*}), i\in A, T_{iB_1}^{*}, i\in B_1, T_{iB_2}^{*}, i\in B_2$ are mutually independent.

(A1) assumes that for the i^{th} individual, T_{i1}^{*},T_{i2}^{*} are mutually independent under the actual status of the disease. Assumptions similar to this have previously been used by Hui and Walter [6] and Yanagawa and Kasagi [7], and are commonly known as conditional independence. Because this assumption is somewhat strong, Vacek [8] and Torrance-Rynard and Walter [9] have examined the effect of the divergence from the assumption on the estimation of the sensitivity and specificity.

(A2) assumes that from the perspective of T_{ij}^{*} the sensitivity and specificity is constant, and does not depend on i or j. (A3) assumes that each individual is independent of the other individuals. The following important relationship exists between and the two parameters α and β.

$$\psi=\frac{\alpha\beta}{(1-\alpha)(1-\beta)}. \qquad (2)$$

This relational equation shows that the conditional maximum likelihood estimator of Ψ can be obtained if α and β, which maximize L_c are plugged in into the right side of (2). Under (A1), (A2) and (A3), the conditional likelihood function L_c is provided by the following equation (Appendix 1):

$$L_c(p,\alpha,\beta)=\prod_{i\in A}\left\{(1-p)(1-\beta)^{\varepsilon_{i1}^{*}+\varepsilon_{i2}^{*}}\beta^{M_i-\varepsilon_{i1}^{*}-\varepsilon_{i2}^{*}}+p\alpha^{\varepsilon_{i1}^{*}+\varepsilon_{i2}^{*}}(1-\alpha)^{M_i-\varepsilon_{i1}^{*}-\varepsilon_{i2}^{*}}\right\}$$
$$\times\prod_{i\in B_1}\left\{(1-p)(1-\beta)^{\varepsilon_{iB_1}^{*}}\beta^{M_i-\varepsilon_{iB_1}^{*}}+p\alpha^{\varepsilon_{iB_1}^{*}}(1-\alpha)^{M_i-\varepsilon_{iB_1}^{*}}\right\}$$
$$\times\prod_{i\in B_2}\left\{(1-p)(1-\beta)^{\varepsilon_{iB_2}^{*}}\beta^{M_i-\varepsilon_{iB_2}^{*}}+p\alpha^{\varepsilon_{iB_2}^{*}}(1-\alpha)^{M_i-\varepsilon_{iB_2}^{*}}\right\}. \qquad (3)$$

Asymptotic distribution

The α and β, which maximize the L_c are termed $\hat{\alpha}$ and $\hat{\beta}$. Under such circumstances, the plug-in estimator of Ψ is provided by the following equation:

$$\hat{\psi}=\frac{\hat{\alpha}\hat{\beta}}{(1-\hat{\alpha})(1-\hat{\beta})}.$$

$\mathrm{Var}\left(\log\hat{\psi}\right)$ is referred to as V_Ψ. When actually calculated, the V_Ψ is a asymptotically given by the following equation.

$$V_\psi\approx\frac{1}{n}\left\{\left(\frac{1}{\alpha}+\frac{1}{1-\alpha}\right)^2\mathrm{Var}\left(\sqrt{n}(\hat{\alpha}-\alpha)\right)+\left(\frac{1}{\beta}+\frac{1}{1-\beta}\right)^2\mathrm{Var}\left(\sqrt{n}(\hat{\beta}-\beta)\right)\right.$$
$$\left.+2\left(\frac{1}{\alpha}+\frac{1}{1-\alpha}\right)\left(\frac{1}{\beta}+\frac{1}{1-\beta}\right)\mathrm{Cov}\left(\sqrt{n}(\hat{\alpha}-\alpha),\sqrt{n}(\hat{\beta}-\beta)\right)\right\}.$$

When the asymptotic normality of $\hat{\alpha},\hat{\beta}$ and the delta method are used $\log\hat{\psi}\to_L N(\log\psi,V_\psi)$, as $n\to\infty$ can be derived (Appendix 2), where $\to_L$ shows a convergence in law.

Table 1: Combination of the true probability of occurrence and true prevalence.

		Pr((TN,Ts)ID = 1)				Pr((TN,Ts)ID = 0)			
Pattern	p	(0,0)	(0,1)	(1,0)	(1,1)	(0,0)	(0,1)	(1,0)	(1,1)
1	0.05	0.1	0.15	0.2	0.55	0.6	0.1	0.2	0.1
2	0.2	0.1	0.15	0.2	0.55	0.6	0.1	0.2	0.1
3	0.05	0.1	0.05	0.15	0.7	0.5	0.2	0.1	0.2
4	0.2	0.1	0.05	0.15	0.7	0.5	0.2	0.1	0.2

Using an asymptotic distribution, the 95% confidence interval of Ψ is given by the following equation:

$$\exp\left(\log\widehat{\psi}-1.96\sqrt{\widehat{\mathrm{V}}_{\psi}}\right)\le\psi\le\exp\left(\log\widehat{\psi}+1.96\sqrt{\widehat{\mathrm{V}}_{\psi}}\right).$$

Follow-up design

In the previous section, the estimator and asymptotic distribution of Ψ, which was used as an index for the comparison of two binary diagnostic tests, were calculated by focusing on the discordant pairs in the data obtained by applying diagnostic tests twice on patients without definite diagnosis. Here, we would like to describe the design of follow-up trial for the comparison of diagnostic tests using Ψ as a primary endpoint. To design a trial, a known distribution of the primary endpoint is required.

The $\log\widehat{\psi}$ follows $\log\widehat{\psi}\sim\mathrm{N}\left(\log\psi,\mathrm{V}_{\psi}\right)$ asymptotically, and the tested hypothesis is the following: H_0: logΨ=0 vs. H_0: logΨ≠0. This is the framework of a standard single-arm trial. If the values of logΨ and $V_Ψ$, and the level of significance and power are fixed, then the sample size needed for the detection of differences will be determined. However, because $V_Ψ$ is a quantity, which is difficult to understand intuitively, it can be predicted that $V_Ψ$ may be difficult to estimate during the design phase. To prevent this, we propose that the trial be started without determining $V_Ψ$, and that $V_Ψ$ is estimated at a time when an n_0 number of individuals have been accumulated after the beginning of the trial, and that the sample size needed for the detection of the differences be designed by using the estimate of variance. The order of the $V_Ψ$ can be evaluated according to the following equation:

$$\mathrm{V}_{\psi}=\frac{A}{n}+\mathrm{o}_{\mathrm{p}}\left(\frac{1}{n}\right),\ \text{as}\ \mathrm{n}\to\infty,$$

Where, A is a constant. After the beginning of the trial, an estimation of the variance is performed at a time when an n_0 number of individuals have accumulated, and the resulting value is termed $V_{Ψ0}$. In such cases, the variance can be estimated according to the below equation, at a time when an n_1 number of cases have been accumulated for an arbitrary $n_1>n_0$.

$$\mathrm{V}_{\psi_1}\approx\frac{n_0}{n_1}\mathrm{V}_{\psi_0}$$

Based on the above, when considering $\log\Psi_1$ as the difference to detect, Z_k as the upper-tail percentage points for the standard normal distribution, a as the level of significance, and 1 - b as the power, the sample size (n_1) needed for the detection of the difference with a probability higher than 1 - b can be designed according to the following equation,

$$\mathrm{n}_1=\frac{\left(Z_{a/2}+Z_b\right)^2\mathrm{V}_{\psi_1}}{\left(\log\psi_1\right)^2}.$$

Using the approximation of $\mathrm{V}_{\psi_1}\approx(n_0/n_1)\mathrm{V}_{\psi_0}$, we obtain the following equation,

$$\mathrm{n}_1=\frac{\left(Z_{a/2}+Z_b\right)\sqrt{\mathrm{n}_0\mathrm{V}_{\psi_0}}}{\left|\log\psi_1\right|}.$$

Table 2: The true values of α, β, Ψ and Log Ψ.

Pattern	α	β	Ψ	Log Ψ
1, 2	0.57	0.33	0.67	-0.41
3, 4	0.75	0.67	6	1.79

Simulation

Several concrete situations are designed, and the behavior of the $\log\widehat{\psi}$ according to the proposed method was examined numerically. $\Pr(T_N,T_S|D=1)$ and $\Pr(T_N,T_S|D=0)$ as well as the prevalence p=Pr(D=1) were put. Here, pattern 1 to pattern 4 was taken into account (Table1).

The differences between the patterns depended on 4 combinations involving whether the prevalence was high (low), and whether the new test was better (worse) than the standard test. In pattern 1, the prevalence was low (p=0.05), and the new test inferior to the standard test (logΨ < 0). In pattern 2, the prevalence was high (p=0.2), and the new test inferior to the standard test (logΨ < 0). In pattern 3, the prevalence was low (p=0.05), and the new test superior to the standard test (logΨ > 0). In pattern 4, the prevalence was high (p=0.2), and the new test superior to the standard test (logΨ > 0). The true values of α,β and Ψ were calculated based on (1), (2), and the true conditional probability established in Table 1. In pattern 1, for example, α=0.2/(0.2+0.15)=0.57, β=0.1/(0.1+0.2)=0.67, Ψ=(0.57×0.333)/(1-0.57)×(1-0.33)=0.67, logΨ=log(0.67)=-0.41. The true values of α,β,Ψ and logΨ in other patterns are summarized in Table 2.

For each pattern, data composed of random numbers $\left\{\left(\varepsilon_{Ni1},\varepsilon_{Si1},\varepsilon_{Ni2},\varepsilon_{Si2}\right);i=1,2,...,n\right\}$ were generated, a set consisting of A, B_1 and B_2 was formed, and data sets consisting exclusively of discordant pairs were generated. Next, $\hat{p}$, $\hat{\alpha}$ and $\hat{\beta}$ maximizing the likelihood (3) were calculated; the estimate Ψ was calculated on the basis of $\hat{\psi}=\hat{\alpha}\hat{\beta}/\left(1-\hat{\alpha}\right)\left(1-\hat{\beta}\right)$; and logΨ was calculated. The calculation was repeated 1,000 times, and the sample mean of the estimates of logΨ, Standard Error (SE), bias and Mean Squared Error (MSE) were calculated. A bias was defined as a subtraction of the true value from the sample mean. In other words, if the bias had a positive value, it showed an overestimate, and if it had a negative value, then it showed an underestimate. The sample size extracted at the beginning was set to n= 500, 1000, 2000, 5000, and 10,000 (Note that this is not the number of discordant pairs). All calculations were performed using the statistical software R (Ver. 3.1.1). The results were as follows.

In patterns 1 and 2, the new test was bad (true log= -0.41) the prevalence p was p=0.05 in pattern 1 and p = 0.2 in pattern 2. The prevalence was the only parameter that showed a difference between both the patterns. The results of the estimations are summarized in Table 3. Even when n is increased, the bias is not stable in pattern 1. Except for n=2000, a slight tendency to overestimate was found. On the other hand, the bias in pattern 2 is more unstable than that of pattern 1. MSE was lower in pattern 2 than in pattern 1, and estimations showing better accuracy at high prevalence were conducted. Next, we show the results of pattern 3 and pattern 4.

For pattern 3 and pattern 4, the new test was superior to the standard test (true log=1.79); and the prevalence p was p=0.05 in pattern 3 and p = 0.2 in pattern 4. The accuracy was higher with pattern 4 than with pattern 3 (i.e. high prevalence leads to the reduction of S.E.). In addition, for both patterns 3 and 4, an increase in sample size

Table 3: Results of the simulation of pattern I and pattern 2.

	Pattern 1 (p= 0.05)				Pattern 2 (p= 0.20)			
n	mean	s. e.	bias	MSE	mean	s.e.	bias	MSE
500	-0.493	0.022	-0.088	1.094	-0.462	0.019	-0.056	11808
1000	-509	0.021	-0.103	0.997	-0.392	0.018	1014	0.7
2000	-0.384	0.021	0.021	0.949	-0.421	0.016	-0.016	0.586
5000	-11409	0.019	-0.003	0.78	-0.342	0.015	0.063	0.48
10000	-0.427	0.017	-0.021	0.67	-0.348	0.013	0.057	395

Table 4: Results of the simulation of pattern 3 and pattern 4.

n	Pattern 3 (p = 0.05)				Pattern 4 (p=0.20)			
	mean	s. e.	bias	MSE	mean	s.e.	bias	MSE
500	2.312	0.078	0.52	0.437	1.824	0.036	0.033	0.037
1000	1.993	0.042	0.201	0.088	1.68	0.024	-0.112	0.028
2000	1.659	0.026	-0.133	0.037	1.571	0.017	-0.22	0.056
5000	1.567	0.02	-0.225	0.062	1.515	0.011	-0.277	0.08
10000	1320	0.016	-0.272	0.081	131.4	11008	-0.278	0.079

was accompanied by a tendency to averagely underestimate logΨ. The numerical results from Tables 3 and 4 are summarized in Figure 1. The error bars in the figure show the 95% confidence interval for the mean, and the dotted line represents the true value of logΨ. The lower half corresponds to patterns 1 and 2, and the upper half corresponds to patterns 3 and 4. Triangles show values in case of p=0.05; circles show values in case of p=0.2 (Figure 1).

Discussion

In this paper, we propose a parameter Ψ for the comparison of diagnostic tests on the basis of data obtained from the application of each binary diagnostic test twice in patients with no definitive diagnosis. The asymptotic distribution of logΨ was also calculated on the basis of conditions in which the data were limited to discordant pairs; further, the method for designing the sample size was also discussed. The influence of the restricted focus on discordant pairs on the estimation results is probably an issue that will need to be evaluated in the future. Comparisons with the estimation of logΨ from the overall likelihood can be conducted, but when the estimation is based on the overall likelihood, then the number of parameters increases and application of the diagnostic tests twice in each individual does not allow for a sufficient degree of freedom and makes it impossible to conduct simultaneous estimation of all parameters. In this way, for estimating all parameters, the necessary for application of the diagnostic tests for estimating all parameters between the proposed method and the overall likelihood based method is different. Therefore, a comparison between two approaches is complicated.

The results of the numerical simulation showed an average tendency to underestimate when the true value of logΨ was positive. The fact that logΨ was positive implied that the conditions were more excellent with the new test than with the standard test. From a researchers' perspective, trials can be carried out with certitude that the new test is a better diagnostic test than the standard test. This is believed to pose no particularly major problem because even if the new test is actually good, it can be interpreted as comparing conditions in a conservative manner. However, the theoretical reasons for underestimating need to be further evaluated. When the simulation results were discussed on the basis of the relationship with prevalence, the estimations were more highly accurate when the prevalence was high than when it was low. When the prevalence was high, individuals with D =1 were potentially included in large numbers. For such individuals, the accuracy of the estimation of parameters (α) conditioned at D =1 was higher, and as a result, the accuracy of logΨ was considered to improve.

Our methodology allows designing the necessary number of cases at a time when n_0 individuals have been accumulated after the start of the trial. In such cases, the problematic issue comprises "what the value of n_0 should be in order to be considered sufficient;" but the results of numerical simulations have shown that even in the worst case (pattern 3 and n=500), the SE of logΨ was about 0.078. Therefore, the evaluation of dispersion might be good if performed at n_0=500.

This study was conceived exclusively for patients without a definitive diagnosis; however, after the start of the trial, we expected that while the trial was underway, the definitive diagnosis of some individuals might be determined. With the current methodology, there is no other choice but to conduct analyses by treating such

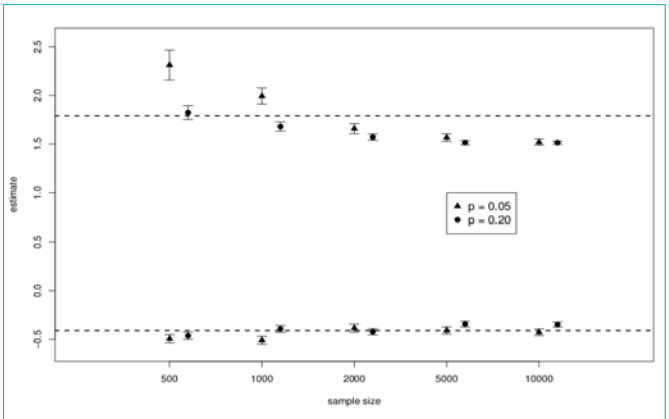

Figure 1: Sample mean and 95%confidence interval of the estimated values of log in each Pattern.

individuals in the same manner as those whose definitive diagnosis has not yet been determined are treated. However, it is also beneficial to estimate information pertaining to the definitive diagnosis in mid-course of the trial and to develop a methodology allowing for estimation that is more accurate. This issue will be the topic of another paper.

References

1. Pepe MS. The Statistical Evaluation of Medical Tests for Classification and Prediction. New York: Oxford University Press. 2003.
2. Jin H, Lu Y. A non-inferiority test of areas under two parametric ROC curves. Contemp Clin Trials. 2009; 30: 375-379.
3. Berg WA, Blume JD, Cormack JB, Mendelson EB, Lehrer D, Bohm-Velez M, et al. Combined screening with ultrasound and mammography vs mammography alone in women at elevated risk of breast cancer. JAMA. 2008; 299: 2151-2163.
4. Ohuchi N, Ishida T, Kawai M, Narikawa Y, Yamamoto S, Sobue T. Randomized controlled trial on effectiveness of ultrasonography screening for breast cancer in women aged 40-49 (J-START): research design. Jpn J Clin Oncol. 2011; 41: 275-277.
5. Murotani K, Aoyama Y, Nagata S, Yanagawa T. Exact method for comparing two diagnostic tests with multiple readers based on categorical measurements. J Biometrics. 2009; 30: 69-79.
6. Hui SL, Walter SD. Estimating the error rates of diagnostic tests. Biometrics. 1980; 36: 167-171.
7. Yanagawa T, Kasagi F. Estimating prevalence and incidence of disease from a diagnostic test. Statistical Theory and Data Analysis, Amsterdam: Elsevier. 1985.
8. Vacek PM. The effect of conditional dependence on the evaluation of diagnostic tests. Biometrics. 1985; 41: 959-968.
9. Torrance-Rynard VL, Walter SD. Effects of dependent errors in the assessment of diagnostic test performance. Stat Med. 1997; 16: 2157-2175.

Austin Biom and Biostat - Volume 2 Issue 2 - 2015
Submit your Manuscript | www.austinpublishinggroup.com

Citation: Murotani K, Hirakawa A, Aoyama Y and Yanagawa T. Follow-up Design for Comparing Two Binary Diagnostic Tests. Austin Biom and Biostat. 2015;2(2): 1016.