Method Comparison Study for Diagnostic Devices with Dichotomous Output

Bipasa Biswas
CDRH, FDA, 10903 New Hampshire Avenue, Silver Spring, MD 20993

Abstract
Often *in-vivo* or *in-vitro* diagnostic devices (or tests) are cleared through 510(k) pathway where the subject device is compared to a predicate device which acts as a comparator device. The study comparing the subject device to a predicate device may or may not involve a clinical reference standard (also known as 'gold standard'). Issues related to commonly used but not necessarily appropriate methods to evaluate agreement, between the subject device and a comparator device (predicate or clinical reference standard) with dichotomous output are discussed to show why they are not recommended. Further, measures of agreement to evaluate a subject device compared against a comparator device (not a clinical reference standard) are provided with discussion.

Key Words: Sensitivity, specificity, positive percent agreement, negative percent agreement.

1. Introduction

A diagnostic device with a dichotomous output has two values. The test output indicates the presence or absence of the target condition (condition of interest), where a target condition can "refer to a particular disease, a disease stage, health status, or any other identifiable condition within a patient, such as staging a disease already known to be present, or a health condition that should prompt clinical action, such as the initiation, modification or termination of treatment" (STARD, 2003). A qualitative test can provide a dichotomous result indicating the presence or absence or that the test is positive or negative for the target condition. And a quantitative and/or continuous or an ordinal valued test can be dichotomized using a cut-off or a clinical decision point. Diagnostic devices are henceforth referred to as diagnostic tests in this paper.

The focus of this paper is evaluation of diagnostic tests with a dichotomous result and inappropriate use of some of statistics to assess agreement. This paper discusses subject level assessments where the subject is the unit of measurement and does not discuss repeated or clustered data that may arise due to multiple measurements on a subject. Examples of diagnostic tests with dichotomous output include a qualitative test for Human Papilloma Virus (HPV), a test for detecting Acute Myocardial Infraction (AMI)

in patients presenting with signs and symptoms to the emergency room (ER), an imaging device which classifies lesions as melanoma versus benign.

The goal of a diagnostic clinical performance study (Design Considerations for Pivotal Clinical Investigations for Medical Devices; issued on November 7, 2013) is to establish the performance of an investigational device in the intended use population of the device. The diagnostic clinical performance is assessed based on either sensitivity and specificity pair or the predictive value positive and negative pair or the likelihood ratio positive and negative pair. To compare the diagnostic clinical performance of an investigational device with the diagnostic clinical performance of an established device or method is only possible when a clinical reference standard is used.

When a clinical reference standard is unavailable, the investigational device is sometimes compared with another device in an agreement study. A very high level of agreement may indicate that the investigational device is non-inferior to the established device. However, a high level of agreement is only meaningful if the established device is already known to have an acceptable level of performance.

1.1 Performance Measures

In general, diagnostic devices with dichotomous output e.g. presence or absence of the condition of interest, is evaluated against a clinical reference standard used to establish the true condition. A test with a binary output is represented by a 2x2 table:

Table 1: 2x2 table

Study Population

		Clinical Reference Standard	
		R=1	R=0
Test	T=1	TP	FP
	T=0	FN	TN

The variable R represents the target condition where R=1 means the condition is present and R=0 means the condition is absent and the test is represented by the variable T where T=1 means the test is positive and T=0 means the test is negative. In the above table, “TP” denotes True Positive; “FP” denotes False Positive; “FN” denotes False Negative; “TN” denotes True Negative.

The accuracy of the test is evaluated by either the sensitivity-specificity pair, or the pair of predictive values or the pair of likelihood ratios which are defined as follows.

Sensitivity (TPF) = P(T=1|R=1) estimated by TP/(TP+FN)
Specificity (1-FPF) =P(T=0|R=0) estimated by TN/(TN+FP)

Likelihood ratio positive = sensitivity/(1-specificity)
Likelihood ratio negative= (1-sensitivity)/specificity

Positive predictive value (PPV) = P(R=1|T=1) estimated by TP/(TP+FP)
Negative predictive value (NPV) = P(R=0|T=0) estimated by TN/(FN+TN)
The performance measures PPV and NPV depend on the prevalence of the true target condition.

2. Inappropriate Statistics or Tests to Evaluate Equivalence of Diagnostic Tests with Dichotomous Output

Often overall accuracy/agreement and/or kappa statistics is used to evaluate agreement between two tests and McNemar's test is often used to compare two tests with dichotomous output. Following section elaborates on why these statistics or tests are not recommended for evaluating performance or to assess agreement between two devices/tests.

2.1 Why not Overall accuracy/agreement

If T denotes the test and R denotes the clinical reference standard, $T+(R+)$ and $T-(R-)$ denotes test T (Reference R) positive and test (Reference R) negative respectively. If p (=Pr (R+)) denotes the prevalence and π_{se} (= Pr $(T+|R+)$) and π_{sp} (=Pr $(T-|R-)$) denote the sensitivity and specificity of a test respectively, then the overall percent agreement is

$$OA = Pr(T = R) = Pr(T+|R+)\,Pr(R+) + Pr(T-|R-)\,Pr(R-)$$
$$= p * \pi_{se} + (1-p) * \pi_{sp}$$

Thus, OA is sensitive to p. We see that if $\pi_{se} < \pi_{sp}$ then OA decreases as p increases and if $\pi_{se} > \pi_{sp}$ then OA increases as p increases and only when $\pi_{se} = \pi_{sp} = \pi$ ($OA = p*\pi + (1-p)*\pi = (p+1-p)*\pi = \pi$), OA is independent of p. In addition, for p<<1, a test with poor sensitivity and high specificity will have a high OA, while the test could as well be no better than a random test. For example, if the prevalence is 5%, a test that classifies everything as negative would still have an overall accuracy of 0.95.

As an example we check the effect of prevalence for two tests – one with sensitivity 0.9 and specificity 0.8 and the other with sensitivity 0.8 and specificity 0.9 for prevalence p=0.1 - 0.9 in increments of 0.1.

Table2: Effect of prevalence on OA

P (Prevalence)	OA (Test with SE=0.9 SP=0.8)	OA (Test with SE=0.8 SP=0.9)
0.1	0.81	0.89
0.2	0.82	0.88
0.3	0.83	0.87
0.4	0.84	0.86
0.5	0.85	0.85
0.6	0.86	0.84
0.7	0.87	0.83
0.8	0.88	0.82
0.9	0.89	0.81

We see from the above table that overall accuracy as an evaluation of performance of a diagnostic test is influenced by the prevalence of the target condition.

Note that two tests with different sensitivities and specificities may have the same overall accuracy/agreement. Say T_1 and T_2 are two tests with sensitivities π_{Se_1} and π_{Se_2} and specificities π_{Sp_1} and π_{Sp_2} respectively then OA for T_1 and T_2 are equal if

$$\mathrm{p} * \pi_{Se_1} + (1 - \mathrm{p}) * \pi_{Sp_1} = \mathrm{p} * \pi_{Se_2} + (1 - \mathrm{p}) * \pi_{Sp_2}$$

which holds whenever

$$\frac{p}{1-p} = \frac{\pi_{Sp_2} - \pi_{Sp_1}}{\pi_{Se_1} - \pi_{Se_2}}.$$

For example, if the study population has a prevalence p=0.20 and two tests T_1 and T_2 have sensitivities 0.05 and 0.85 and specificities 0.95 and 0.75 respectively, then both tests have the same OA (=0.77) and yet the two tests have very different sensitivities and specificities and in fact test T_1 is an useless test as it gives a positive result to subjects with and without the target condition with equal probability (0.05). Thus, if performance evaluation was solely based on OA, the OA of test T_1 may not provide the information that it is basically uninformative and in addition it fails to differentiate the agreement of test positives with presence of target/clinical condition and agreement of test negatives with absence of target/clinical condition.

Overall accuracy/agreement as an omnibus agreement measure is sensitive to prevalence, a test with same sensitivity and specificity will give a different OA depending on the prevalence which makes comparison of OA difficult across studies. And additionally, two tests with different sensitivities and specificities may yield the same OA on same study population. Overall accuracy/agreement as an omnibus agreement measure is not acceptable to evaluate a diagnostic test performance/agreement.

2.2 Why not Kappa

Kappa statistics is a chance corrected agreement measure. The Kappa statistics is mathematically defined as below:

$$\kappa = \frac{\Pr(\mathrm{T}=\mathrm{R}) - [\Pr(\mathrm{T}+)\Pr(\mathrm{R}+) + \Pr(\mathrm{T}-)\Pr(\mathrm{R}-)]}{1 - [\Pr(\mathrm{T}+)\Pr(\mathrm{R}+) + \Pr(\mathrm{T}-)\Pr(\mathrm{R}-)]} = \frac{2\mathrm{p}(1-\mathrm{p})[\pi_{\mathrm{se}} + \pi_{\mathrm{sp}} - 1]}{1 - [\mathrm{p}^2\pi_{\mathrm{se}} + (1-\mathrm{p})^2\pi_{\mathrm{sp}} + \mathrm{p}(1-\mathrm{p})(2 - \pi_{\mathrm{se}} - \pi_{\mathrm{sp}})]}$$

Thus, while kappa is a chance corrected agreement measure where $\kappa = 0$ when a test is random (i.e. $\pi_{\mathrm{se}} = 1 - \pi_{\mathrm{sp}}$, the test calls positive with equal probability for subjects with and without the target condition) unlike OA, kappa has same issues of being sensitive to prevalence and that tests with different sensitivities and specificities may produce the same kappa.

κ is sensitive to p, as a test with sensitivity π_{se} and specificity π_{sp} will yield a different κ for different p. As an example we check the effect of prevalence for two tests – one with sensitivity 0.9 and specificity 0.8 and the other with sensitivity 0.8 and specificity 0.9 for prevalence p=0.1-0.9 in increments of 0.1.

Table3: Effect of prevalence on Kappa

P (Prevalence)	Kappa (Test with SE=0.9 SP=0.8)	Kappa (Test with SE=0.8 SP=0.9)
0.1	0.40	0.53
0.2	0.55	0.65
0.3	0.63	0.69
0.4	0.68	0.71
0.5	0.70	0.70
0.6	0.71	0.68
0.7	0.69	0.63
0.8	0.65	0.55
0.9	0.53	0.40

A test with lower sensitivity than specificity (i.e. $\pi_{se} < \pi_{sp}$) will yield a higher Kappa than when the sensitivity and specificity are switched (i.e. $\pi_{se} > \pi_{sp}$) on same study population when prevalence $p < 0.5$. Similarly, a test with higher sensitivity than specificity (i.e. $\pi_{se} > \pi_{sp}$) will yield a higher Kappa on a study population with prevalence $p > 0.5$ than a test where sensitivity and specificity are switched (i.e. $\pi_{se} < \pi_{sp}$).

And two tests with different sensitivities and specificities can yield the same kappa on same sample with prevalence p. Say two tests T1 and T2 have sensitivities 0.7 and 0.99 and specificities 0.82 and 0.68 respectively then $\kappa = 0.45$for both tests.

Thus, Kappa is an omnibus index of agreement is not recommended to evaluate performance or agreement as it is sensitive to prevalence-test with same sensitivity and specificity will give a different kappa depending on the prevalence; and that in addition two tests with different sensitivities and specificities may yield the same kappa on the same study population. Kappa is not independent of the prevalence of the target condition and it is generally not comparable from one study to another as the prevalence may differ (Thompson & Walter, 1988; Feinstein & Cicchetti, 1990). And Kappa may be low although there are high levels of accuracy (sensitivity and specificity).

The statistics above have been constructed for evaluating performance of a new test compared to a clinical reference standard. Often the clinical reference standard is not perfect and is prone to measurement error. Still, the same issues of using the overall agreement or kappa hold.

2.3 Why not McNemar's test

McNemar's test (Fleiss 1981) is well known test for paired binary data and is often used to compare sensitivities and specificities for comparison of two diagnostic tests. Trajman and Luis (2008) recommend comparing sensitivities between two diagnostic tests exclusively among subjects with the target condition and compare specificities of two tests among subjects without the target condition by using McNemar's ch-square test. In comparison studies where the purpose is to assess and compare the agreement between two tests, McNemar's test is not the recommended procedure. It is therefore important to

clarify the purpose of comparison of the tests in order to select an appropriate test related to the study objective.

McNemar's chi-square (continuity corrected) test statistics is given by

$$\text{McNemar Chi square} = (|b-c|-1)^2/(b+c)$$

Where b and c are from the table below:

Table 4: 2x2 table

		Comparator Test	
		R=1	R=0
New Test	T=1	a	b
	T=0	c	d

Thus this statistics is used to compare the marginal probabilities, but not necessarily to check equivalence of two tests. Note that the McNemar's test is only checking for equality and thus the null hypothesis is of equivalence and the alternative hypothesis of difference. However, this is not an appropriate hypothesis, as a failure to find a statistically significant difference is naively interpreted as evidence for equivalence. Alternatively, equivalence in marginal probabilities does not always imply equivalence of two diagnostic tests. Two examples that elicit these issues clearly are when say b and c are almost equal (for e.g. say b=19 c=18 and a=0 and d=0) the p-value is 1 and yet in reality the two tests hardly agree (the overall agreement is 0.0% (0/37)). A second example when say b and c are different but both a and d are very high (for e.g. b=30 and c=5, a=3700 and d=2800), the p-value will indicate a statistically significant difference and yet the tests actually have a very high agreement.

Thus, while comparing two diagnostic tests with binary output, McNemar's chi-square test assumes a null hypothesis that the rates of positive responses by the two tests are equal. The McNemar's Chi-square test could lead to the conclusion that there is not enough evidence to demonstrate that the two medical tests differ, when in truth the two differ. Alternatively, the two medical tests may have very high agreements and yet the McNemar's test rejects that the two are equal.

In summary, overall agreement and kappa statistics do not appropriately measure agreement between two medical tests and are inappropriate as primary measures of agreement and likewise McNemar's chi-square test to evaluate agreement is also not recommended.

3. When the clinical reference standard is imperfect

A common situation arises when one wishes to evaluate a new diagnostic or screening test and there is no perfect clinical reference standard available for comparison (Glasizou et al 2008). Often there is no available clinical reference standard and thus the comparison is against a comparator or an imperfect reference standard. The data is represented as 2 x 2 table as:

Table 5: 2x2 table

Study Population

		Imperfect Reference Standard	
		R=1	R=0
Test	T=1	a	b
	T=0	c	d

The performance of the new test evaluated against an imperfect reference standard is based on the following pair of agreement measures (Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests; issued on: March 13, 2007):
Positive percent agreement (PPA) = P(T=1|R=1)
Negative percent agreement (NPA) = P (T=0|R=0)

And these are estimated as:
$\widehat{PPA}$ = a/(a+b)
$\widehat{NPA}$ = c/(c+d)

Caution should be practiced when interpreting these agreement measures as a high agreement does not necessarily mean that the new test has good diagnostic performance and likewise a poor agreement does not necessarily mean that the new test is worse than the comparator imperfect test (Walter et al 2012).

4. Conclusion

In a 510(k), often a new diagnostic device with dichotomous (or qualitative) output is evaluated for safety and effectiveness by comparing the new test with a comparator device/test. The inappropriate use of common statistics like overall accuracy/agreement and kappa can be misleading, hard to interpret and not comparable across studies. Thus these statistics are not recommended for comparison of two tests. McNemar's chi-square test for paired binary data is again hard to interpret when two tests are compared for agreement and thus is not recommended for evaluating agreement between two tests.

Finally, if a new test is evaluated against an imperfect clinical reference standard, the evaluation is based on two measures- the positive percent agreement (PPA) and the negative percent agreement (NPA). Caution should be practiced while interpreting these agreement measure pairs, as good agreement does not necessarily mean that the new test has good clinical performance and likewise poor agreement does not mean that the new test has poor performance measures. Thus, it is best to have a clinical reference standard, whenever possible, to evaluate the diagnostic performance of a new test. Otherwise evaluation is based on agreement pairs.

References

(1) Bossuyt, P.M., Reitsma, J.B., Bruns, D.E., Gatsonis, C.A., Glasziou, P.P., Irwig, L.M., Lijmer, J.G., Moher, D., Rennie, D., & deVet, H.C.W. (2003). Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD initiative. Clinical Chemistry, 49(1), 1–6.

(2) Bossuyt, P.M., Reitsma, J.B., Bruns, D.E., Gatsonis, C.A., Glasziou, P.P., Irwig, L.M., Moher, D., Rennie, D., deVet, H.C.W., & Lijmer, J.G. (2003). The STARD statement for reporting studies of diagnostic accuracy: Explanation and elaboration. Clinical Chemistry, 49(1), 7–18.
(3) Thompson W.D. and Walter, S.D. (1988). A reappraisal of the kappa coefficient. Journal of Clinical Epidemiology, 41: 949-958.
(4) Feinstein A. R. and Cicchetti D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. J. Clin. Epidemiol,Vol. 43, No. 6, 543-549.
(5) Feinstein A. R. and Cicchetti D. V. (1990). High agreement but low kappa: II. Resolving the paradoxes. J. Clin. Epidemiol , Vol. 43, No. 6, 551-558.
(6) Fleiss, JL, Statistical Methods for Rates and Proportions, John Wiley & Sons, New York (2nd ed., 1981).
(7) Trajman, A and Luiz, R.R. (2008). McNemar's χ^2test revisited: comparing sensitivity and specificity of diagnostic examinations. Scand J Clin Lab Investig, 68(1): 77-80.
(8) Kim, S. and Lee, W. (2014). Does McNemar's test compare the sensitivities and specificities of two diagnostic tests? Statistical methods in medical research, 0(0): 1-13.
(9) Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests; issued on: March 13, 2007 http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071287.pdf
(10) Design Considerations for Pivotal Clinical Investigations for Medical Devices ;issued on November 7, 2013 http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM373766.pdf
(11) Glasziou, P., Irwig, I.M., Deeks, J.J. (2008). When should a new test become the current reference standard? Annals of Internal Medicine, 149: 816-821.
(12) Walter, S.D., Macaskill, P., Lord, S.J., and Irwig, L. (2012). Effect of dependent errors in the assessment of diagnostic or screening test accuracy when the reference standard is imperfect. Stats in Med, 31: 1129-1138.