•BIOSTATISTICS IN PSYCHIATRY (45)•

Assessing the Accuracy of Diagnostic Tests

Fangyu LI, Hua HE*

Summary: Gold standard tests are usually used for diagnosis of a disease, but the gold standard tests may not be always available or cannot be administrated due to many reasons such as cost, availability, ethical issues etc. In such cases, some instruments or screening tools can be used to diagnose the disease. However, before the screening tools can be applied, it is crucial to evaluate the accuracy of these screening tools compared to the gold standard tests. In this assay, we will discuss how to assess the accuracy of a diagnostic test through an example using R program.

Key words: AUC; gold standard test; ROC analysis; ROC curve; sensitivity; specificity

[*Shanghai Arch Psychiatry.* 2018; **30**(3): 207-212. doi: http://dx.doi.org/10.11919/j.issn.1002-0829.218052]

1. Introduction

Mental disease is a significant cause of worldwide morbidity, second to cardiovascular disease, based on the estimates from the Global Burden of Disease Study.[1] Among mental diseases, depression is now the leading cause of the global disability burden. In China, depression was one of four leading causes of disability-adjusted life-years (DALYs).[2] The overall prevalence of depression was 37.68% and that of severe depression was 4.08% among all Chinese adults in 2012, and the disease burden is estimated to have increased by 10% in China between 2013 and 2015.[3-4] For diagnosis of depression, the Structured Clinical Interview for DSM-IV (SCID) is widely considered the gold standard in both clinical practice and research, and continues to be commonly used as such.[5-8] However, the use of the SCID is limited due to a number of restraints such as costs and a severe shortage of psychiatrists. Because the SCID cannot be used as an infallible list that automatically provides psychiatric diagnoses, it must be administered by well-trained psychiatrists[9], resulting in high expenses and unaffordable mental health care for some patients. The shortage of psychiatrists in China also limits patients' access to mental health professionals. In addition, going through examinations with some patients such as elderly patients may be difficult and time-consuming. Because of the constraints of SCID, some easily administered screening tools, such as the Hamilton Depression Rating Scale (HAM-D), the Beck Depression Inventory (BDI), and even simpler screening tools such as the Patient Health Questionnaire (PHQ-2, PHQ-9) were developed and admitted to patients for depression diagnosis.[10] For example, HAM-D, the most commonly used instrument, is a multiple item questionnaire used to provide an indication of depression, as well as a guide to evaluate recovery.[11] Similarly, the 21-question multiple-choice self-report inventory BDI is also widely used to measure the severity of depression. However, before the screening tools can be applied to patients, it is crucial to evaluate the accuracy of these screening tools compared to the gold standard, the SCID.[12] If a screening tool can correctly classify diseased subjects as diseased and non-diseased subjects as non-diseased, the screening tool can be advocated for its use in medical practice. Otherwise, its usage should be cautioned. In this essay, we will discuss how to assess the accuracy of a diagnostic test.

2. Accuracy of diagnostic test

Sensitivity and specificity are widely used to assess the accuracy of a diagnostic test when the test result

Department of Epidemiology, Tulane University School of Public Health and Tropic Medicine New Orleans, USA

*correspondence: Hua He. Mailing address: 1440 Canal Street, Suite 2000, New Orleans, LA, USA; Postcode: LA 70122; E-Mail: hhe2@tulane.edu

is binary, such as yes vs no, or positive vs negative. Sensitivity is the probability that the test is positive among diseased subjects, i.e., the probability that the test correctly classifies diseased subjects as diseased, while specificity is the probability that the test is negative among non-diseased subjects, i.e., the likelihood that the test can correctly classify non-diseased subjects as non-diseased.[13] For example, a test with 90% sensitivity correctly classifies 90% of diseased subjects as diseased but misclassifies 10% diseased subjects as the non-diseased. Similarly, a test with 90% specificity can correctly classify 90% of non-diseased subjects as non-diseased but misclassifies 10% non-disease subjects as diseased. Sensitivity and specificity range from 0 to 1, with 1 indicating the test can correctly classify all the diseased subjects as diseased and all the non-diseased subjects as non-diseased. To evaluate the accuracy of the test, the sensitivity and specificity must be considered together. Clearly, both high sensitivity and high specificity are needed for a good diagnostic test. A test with 100% sensitivity and specificity can correctly classify all the diseased and non-diseased subjects and hence is a perfect test.

However, in practice, diagnostic tests with a continuous index are very common. In such cases, Receiver Operator Characteristic (ROC) is used to evaluate the accuracy of the test in discriminating the disease. By dichotomizing the continuous test score, sensitivity and specificity can be calculated at each cut-point. An ROC curve is constructed by connecting all pairs of (1-specificity, sensitivity) at all the possible cut-points of the continuous test. An ROC curve is a way of graphically displaying true positives versus false-positives across a range of cut-offs,[14] and provides a picture of how accurately the test can discriminate the disease. Different cut points result in different sensitivities and specificities. For a given test, there is a tradeoff between the sensitivity and specificity. For example, if a higher test score indicates greater likelihood of being diseased, the lower cutoff will yield higher sensitivity but lower specificity, in which case the test can correctly classify most of the diseased subjects as diseased, but also gives a high chance of misclassifying the non-diseased subjects as diseased, and vice versa for a higher cutoff. An ROC curve provides us with a full picture of how the test discriminates between diseased and non-diseased, with the portion closer to the top left corner being better able to discriminate. The diagonal line shows no ability of discrimination.

The ROC curve is an excellent way to depict the ability of the test in discriminating disease at each cut-point, but in practice, it is also very important to have a single index to summarize the overall performance of the test. The area under the ROC curve (AUC) is an index that evaluates the overall ability. The AUC measures the ability of the test to correctly classify those with and without the disease. The closer a ROC curve is to the upper left corner, the larger the area under the curve is, with a value of 1 indicating perfect discrimination.

Next, we will illustrate how to estimate the accuracy of a test through a real study using R.

3. Illustrating example: screening for depression in aging services clients

3.1 Study sample

377 elderly subjects who spoke English and received an initial home assessment were enrolled in the interview after providing informed consent.

3.2 Measures

SCID: The SCID[15] serves as the gold standard to determine the presence or absence of a current major depressive episode (MDE). In this study, SCID was aministered to all the subjects in the study sample.

PHQ-9/PHQ-2: The Patient Health Questionnaire (PHQ-9) is a nine-item depression scale of the patient health questionnaire given to patients in a primary care setting to screen for the presence and severity of depression. The item scores for each question range from 0 to 3, and the total score of the PHQ-9 ranges from 0 to 27. The PHQ-2 includes only the first two items of the PHQ-9, and thus has a total score ranging from 0 to 6. Even though the PHQ-9 and PHQ-2 total score can be considered continuous, a cut-off at 10 and 3 are usually used for diagnosing depression for PHQ-9 and PHQ-2, respectively. For PHQ-9, subjects whose scores are lower than 10 would be diagnosed negative for depression, others are diagnosed positive. Similarly, for PHQ-2, only those whose scores are lower than 3 would be treated as non-depressed. We will assess how accurate the PHQ-9/PHQ-2 is in discriminating between depression and non-depression in SCID by treating them as both binary and as a continuous test.

3.3 Computation of sensitivity and specificity for a binary test:

We first analyze the accuracy of binary PHQ-9 for which the total score is dichotomized at 10. In this case, subjects with PHQ-9 < 10 have negative test results and those with PHQ-9 ≥ 10 have positive test results. So PHQ-9 positive/negative and SCID depression/non-depression can be summarized in the following 2x2 table with the R codes.

```
temp <- roc
temp$PHQ_9_SCORE <- ifelse(temp$PHQ_9_SCORE>9,1,0)
crosstab<-table(temp$PHQ_9_SCORE, temp$SCID)
```

The data can be summarized in the table 3:

Table 3. Gold standard

		Depression	Non-Depression	Total
Test Result	Positive (PHQ-9>=10)	82	35	117
	Negative (PHQ-9<10)	18	242	260
	Total	100	277	377

Sensitivity and specificity are computed as

Sensitivity =# of positive tests/# of disease subjects

= 82/100

=0.82;

Specificity=# of negative tests/# of non-disease subjects

=242/277

=0.87.

When a cutoff of 10 is used, the test is defined positive for subjects with PHQ-9>=10 and negative for PHQ-9<10, and the sensitivity is estimated as 82%, i.e., 82% of subjects who truly have depression can be successfully classified as depressed, while the specificity is estimated as 87%, that is 87% of non-depressed subjects are correctly classified as non-depressed.

When a cutoff of 3 is used for PHQ-2, i.e., PHQ-2<3 is defined as negative and PHQ-2≥3 is defined as positive, the results are summarized in the table 4:

Table 4. Gold standard

		Depression	Non-Depression	Total
Test Result	Positive (PHQ-2>=3)	80	61	141
	Negative (PHQ-9<3)	20	216	236
	Total	100	277	377

Sensitivity and specificity are estimated as

Sensitivity =# of positive tests/# of disease subjects

= 82/100

=0.82;

Specificity=# of negative tests/# of non-disease subjects

=216/277

=0.78.

When the test positive/negative is defined by the cutoff of 3, 80% of depressed subjects can be correctly classified as depressed and 78% of non-depressed subjects are correctly classified as non-depressed.

With a different cutoff, the sensitivity and specificity will be different. For example, the sensitivity and specificity are 57% and 90% for PHQ-2 if a cutoff of 4 is used. For PHQ-9, the sensitivity and specificity are 0.74 and 0.91 for a cutoff of 11. Table 1 presents the sensitivities and specificities for a range of cutoffs for both PHQ-9 and PHQ-2.

From Table 1, it is obvious that there is a tradeoff between the sensitivity and specificity. A lower cutoff results in higher sensitivity and lower specificity, which means that more depressed subjects can be correctly classified as depressed, but more non-depressed subjects are misclassified as depressed. Because of the tradeoff between sensitivity and specificity, an optimum cutoff is usually used in clinical practice. The optimum cutoff is usually identified as the cutoff which maximizes the summation of sensitivity and specificity. For the PHQ-2, the optimal cutoff is 3.0 as the corresponding sensitivity (80%) and specificity (78%) achieves the largest value. An optimal cutoff for PHQ-9 is 10 as the summation of 82% for sensitivity and 87% for specificity achieves the largest value.

3.4 Construction of receiver operator characteristic curve (ROC)

At each cut point of a continuous diagnostic test, the sensitivity and specificity show how accurately diseased

Table 1. Sensitivities and Specificities at different cut points

Screening Test	Cut Point	Sensitivity	Specificity
PHQ-2	0.0	1.00	0.00
	1.0	0.99	0.26
	2.0	0.95	0.58
	3.0	**0.80**	**0.78**
	4.0	0.57	0.90
	5.0	0.40	0.96
	6.0	0.24	0.99
PHQ-9	7.0	0.93	0.62
	8.0	0.90	0.70
	9.0	0.86	0.82
	10.0	**0.82**	**0.87**
	11.0	0.74	0.91
	12.0	0.66	0.94

subjects and non-diseased subjects are classified. Since different cutoffs yield different sensitivities and specificities, an ROC curve, which is a plot of sensitivity versus (1-specificity) for every possible cut point of the continuous test, is employed to depict the ability of the test in discriminating between diseased and non-diseased. For each cutoff, positive and negative test results can be defined based on whether test scores are greater or less than the threshold, and specificity and sensitivity then can be estimated based on a 2X2 table of the binary positive/negative test results and the true disease status. The ROC curve is constructed by connecting sensitivity (y-axis) versus (1-spicificity) (x-axis) at all the cutoffs. The ROC curves for PHQ_9 and PHQ_2 are presented in Figure 1 and the R codes for constructing ROC curves for PHQ_2 and PHQ_9 are provided below, respectively:

```
roc_PHQ9 <- plot.roc(roc$SCID, roc$PHQ_9_SCORE,
legacy.axes=TRUE, percent=TRUE, auc=TRUE,
col="1",ci=TRUE)

roc2_PHQ2 <- lines.roc(roc$SCID, roc$PHQ_2_
SCORE, legacy.axes=TRUE, percent=TRUE, auc=TRUE,
col="2",ci=TRUE)

testobj<- roc.test(roc_PHQ9,roc_PHQ2)

text(50, 50, labels=paste("p-value =", format.
pval(testobj$p.value)), adj=c(0, .5))

legend("bottomright", legend=c("PHQ-9", "PHQ-2"),
col=c("1", "2"), lwd=2)
```

3.5 The Area under the ROC Curve (AUC) measures and its interpretation

Even though an ROC curve can depict the ability of the test in discriminating disease vs non-disease at each cut-point, it cannot provide an overall index to summarize the overall performance of the test. The area under the ROC Curve (AUC) is such an overall index of the ability of discrimination for continuous test and measures how well people are classified into diseased and non-diseased. AUC ranges from 0.5 to 1. An AUC value of 0.5 corresponds to the diagonal line of the ROC curve and provides no information for classification, while a value of 1 indicates that the test can correctly classify all diseased subject as diseased, and all non-diseased subjects as non-diseased, which is a perfect test. A rough guide for classifying the accuracy of a diagnostic test in is summarized in Table 2. A test with AUC between 0.90 and 1.00 has excellent discrimination ability, AUC from 0.80 to 0.90, 0.70 to 0.80, 0.60 to 0.70 and 0.50 to 0.60 indicates good, fair and poor and fail discrimination ability, respectively. The R codes to obtain AUC for both PHQ-2 and PHQ-9 are provided below. We can also test if two diagnostic tests have the same discrimination ability by testing if there is any significant difference in AUC between the two tests.

```
roc.test(roc1,roc2,paired=TRUE)
```

DeLong's test for two correlated ROC curves

Table 2. A guide for classifying the accuracy of a diagnostic test by AUC

AUC Range	Classification
0.9 -1.0	Excellent
0.8 - 0.9	Good
0.7 - 0.8	Fair
0.6 - 0.7	Poor
0.5 - 0.6	Fail

data: roc1 and roc2

Z = 2.6064, p-value = 0.00915

alternative hypothesis: true difference in AUC is not equal to 0

sample estimates:

AUC of roc1 AUC of roc2

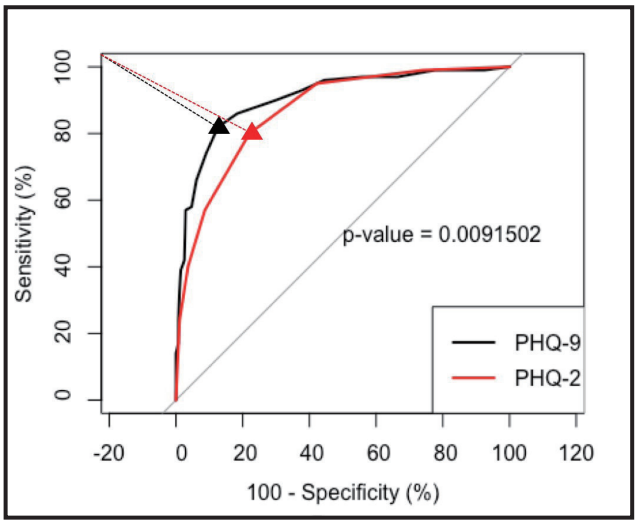
0.9062635 0.8690794

In this case, the AUC for the PHQ-9 is 0.9063 and 0.8691 for PHQ-2. The PHQ-9 achieves excellent accuracy in classifying subjects as depressed and non-depressed, while the accuracy for PHQ-2 is relatively lower, but still pretty good. The p value for testing the difference in AUC between the PHQ-9 and PHQ-2 is 0.00915, which indicates that PHQ-9 and PHQ-2 have different ability for discriminating depression vs non-depression subjects, and the PHQ-9 is more accurate in diagnosing depressed and non-depressed subjects.

4. Discussion

The Structured Clinical Interview for DSM-IV (SCID) has long been accepted as the gold standard for

Figure 1. ROC Curves for PHQ-2 and PHQ-9



diagnosing depression in clinical populations. However, the administration of SCID is not applicable for many reasons and alternative diagnostic tests/screening tools are needed. Before diagnostic tests/screening tools can be applied to the target population, it is critical to assess the accuracy of the diagnostic tests/screening tools.

In practice, in additional to sensitivity and specificity, positive predictive values (PPV) and negative predictive values (NPV) are also widely used. The PPV is the likelihood that subjects with positive tests are also diseased, and NPV is the probability that subjects with negative tests are also non-diseased. Given the prevalence of the disease, PPV and NPV can be determined by the sensitivity and specificity and vice versa.

In clinical practice, where gold standard tests can be invasive, expensive, and carry a higher risk (e.g. angiography, biopsy, and surgery), patients and physicians may be reluctant to undergo such gold standard testing. If the gold standard test is not administered for everyone, the estimates of sensitivity and specificity may be biased since only the subjects with the gold standard testing are used for estimating sensitivities and specificities. This biased is called verification bias. Some methods were developed to correct the verification bias.[19,20]

Funding statement

No funding provided for this study.

Conflicts of interest statement

The authors declare there are no conflicts of interests in this study.

诊断性测试准确性的评估

Li F, He H

概述：金标准测试通常用于诊断一种疾病，但金标准测试可能由于成本、可行性、伦理问题等多个原因而无法实施。在这种情况下，一些仪器或筛选工具可用于诊断疾病。然而，在采用筛选工具之前，与金标准测试对比并评估这些筛选工具的准确性是至关重要的。在本文中，我们将通过 R 程序运行的案例来讨论如何评估诊断性测试的准确性。

关键词：AUC；金标准测试；ROC 分析；ROC 曲线；敏感性；特异性

References

1. Murray C, Lopez A. *The global burden of disease. A comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020*. Cambridge, MA: Harvard School of Public Health; 1996
2. Hoevenaar-Blom MP, Spijkerman AM, Kromhout D, Verschuren WM. Sufficient sleep duration contributes to lower cardiovascular disease risk in addition to four traditional lifestyle factors: the MORGEN study. *Eur J Prev Cardiol*. 2014; **21**(11): 1367-1375. doi: http://dx.doi.org/10.1177/2047487313493057
3. Qin X, Wang S, Hsieh CR. The prevalence of depression and depressive symptoms among adults in China: Estimation based on a National Household Survey. *China Economic Review*. 2016
4. Charlson FJ, Baxter AJ, Cheng HG, Shidhaye R, Whiteford HA. The burden of mental, neurological, and substance use disorders in China and India: a systematic analysis of community representative epidemiological studies. Lancet. 2016; **388**(10042): 376-389. doi: http://dx.doi.org/10.1016/S0140-6736(16)30590-6
5. Lowe B, Kroenke K, Herzog W, Grafe K. Measuring depression outcome with a brief self-report instrument: sensitivity to change of the Patient Health Questionnaire (PHQ-9). *J Affect Disord*. 2004; **81**(1): 61–66. doi: http://dx.doi.org/10.1016/S0165-0327(03)00198-8
6. Zubaran C, Foresti K, Schumacher MV, Amoretti AL, Mullet LC, Thorell MR, et al. Validation of a screening instrument for postpartum depression in Southern Brazil. *J Psychosom Obstet Gynaecol*. 2009; **30**(4): 244–254. doi: http://dx.doi.org/10.3109/01674820903254724
7. Lee DT, Yip AS, Chiu HF, Leung TY, Chung TK. Screening for postnatal depression: are specific instruments mandatory? *J Affect Disord*. 2001; **63**(1–3): 233–238. doi: https://doi.org/10.1016/S0165-0327(00)00193-2
8. Bernstein IH, Wendt B, Nasar SJ, Rush AJ. Screening for major depression in private practice. *J Psychiatr Pract*. 2009; **15**(2): 87–94. doi: http://dx.doi.org/10.1097/01.pra.0000348361.03925.b3
9. Gomes de Matos E, de Matos G, Mello T, de Matos G, Mello G. Importance and constraints of the DSM-IV use in the clinical practice. *Revista de Psiquiatria do Rio Grande do Sul*. 2005; **27**(3): 312-318. Portuguese
10. http://www.thomsoncdc.com/our-services/psychology/importance-early-detection-mental-health-issues/
11. Hedlund JL, Vieweg BW. The Hamilton rating scale for depression: a comprehensive review. *J Operational Psychiatry*. 1979; **10**(2): 149-165

12. Zhu W, Zeng N, Wang N. *Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS implementations*. Maryland, Baltimore: NESUG proceedings: health care and life sciences; 2010. p: 19, 67
13. Šimundić AM. Measures of diagnostic accuracy: basic definitions. *EJIFCC*. 2009; **19**(4): 203
14. Lalkhen AG, McCluskey A. Clinical tests: sensitivity and specificity. *Anaesth Crit Care Pain*. 2008; **8**(6): 221-223
15. First MB, Spitzer RL, Gibbon M. *Structured Clinical Interview for DSM-IV-TR Axis I Disorders - Non-patient Edition*. New York State Psychiatric Institute, NY: Biometrics Research Department; 2001
16. Metz CE. Basic principles of ROC analysis. *Semin Nucl Med*. 1978; **8**: 283–298
17. Fletcher RW, Fletcher SW. *Clinical epidemiology : the essentials (4th ed.)*. Baltimore, MD: Lippincott Williams & Wilkins; 2005. p: 45
18. Liu C, Liu A, Halabi S. A min–max combination of biomarkers to improve diagnostic accuracy. *Stat Med*. 2011; **30**(16): 2005-2014. doi: http://dx.doi.org/10.1002/sim.4238
19. He H, Lyness JM, McDermott MP. Direct estimation of the area under the receiver operating characteristic curve in the presence of verification bias. *Stat Med.* 2009; **28**(3): 361-376. doi: http://dx.doi.org/10.1002/sim.3388
20. He H, McDermott MP. A robust method using propensity score stratification for correcting verification bias for binary tests. *Biostatistics*. 2011; **13**(1): 32-47. doi: http://dx.doi.org/10.1093/biostatistics/kxr020

Fangyu Li obtained a bachelor's degree from the school of pharmacy of Hubei University of Chinese Medicine in 2016, and a Master's degree from the school of Public Health and Tropical Medicine of Tulane University in 2018. She is working as a Ph.D. candidate in the department of epidemiology in the Health Science Center at the University of Texas. Her research interests include nutritional epidemiology, cardiovascular epidemiology, genetic epidemiology, and cancers.

Notice: Shanghai Archives of Psychiatry soon to be renamed General Psychiatry

Shanghai Mental Health Center will publish its last issue of the journal *Shanghai Archives of Psychiatry*, 2018 volume 30 issue 3, on 30th June 2018. The postal code is 4-798. The journal will be renamed *General Psychiatry* and presented to the readers as issue 4 on 30th August 2018.

Changing the name of the journal is a magnificent make-over. We aim to publish a high quality and international journal by cooperating with the BMJ publishing group. The mission of our journal will not change, however the new content and research will be much more innovative and comprehensive. This journal will continue to spotlight important academic exchanges between China and the rest of the world that promote the international development of mental health research.

General Psychiatry Editorial Department