Statistics

Common pitfalls in statistical analysis: Understanding the properties of diagnostic tests – Part 1

Priya Ranganathan, Rakesh Aggarwal[1]

Department of Anaesthesiology, Tata Memorial Centre, Mumbai, Maharashtra, [1]Department of Gastroenterology, Sanjay Gandhi Postgraduate Institute of Medical Sciences, Lucknow, Uttar Pradesh, India

Abstract In this article in our series on common pitfalls in statistical analysis, we look at some of the attributes of diagnostic tests (i.e., tests which are used to determine whether an individual does or does not have disease). The next article in this series will focus on further issues related to diagnostic tests.

Keywords: Biostatistics, predictive values, sensitivity, specificity

Address for correspondence: Dr. Priya Ranganathan, Department of Anaesthesiology, Tata Memorial Centre, Ernest Borges Road, Parel, Mumbai - 400 012, Maharashtra, India.
E-mail: drpriyaranganathan@gmail.com

INTRODUCTION

Diagnostic tests are used to differentiate between individuals with and without a particular disease. However, most diagnostic tests are imperfect, and provide some false-positive (the test is positive though the individual does not have the disease) and false-negative (the test is negative though the individual has the disease) results.

Most diseases have a gold standard diagnostic test, which is used to establish a diagnosis. This concept has some limitations, but let us assume for now that such a "gold standard" does exist for the disease that we are studying. However, such gold standard tests are usually difficult to perform, costly, invasive, time-consuming, or not easily accessible. Hence, we often look to substitute the gold standard with another test, in order to decrease costs, minimize invasiveness or save time, etc. In these cases, we are interested in knowing how the "substitute" test performs in comparison with the gold standard for differentiating between the diseased and the non-diseased individuals. In this article, we explain some of the attributes of diagnostic tests, and some issues related to their clinical interpretation. Another article in the next issue will focus on some additional issues related to diagnostic tests.

Let us look at the example of diagnosis of pulmonary embolism. A perfusion scan is the gold standard for its diagnosis, but is often not available. Also, it is costly and invasive. Hence, we wish to use a blood test (D-dimer level) for the detection of pulmonary embolism. To assess the performance of this test for the diagnosis of pulmonary embolism, one would perform this test in a group of patients suspected to have pulmonary embolism who have also undergone the perfusion scan [Table 1].

SENSITIVITY AND SPECIFICITY

Sensitivity and specificity are the most commonly used measures of the performance of a diagnostic test as compared to an existing gold standard.

Access this article online	
Quick Response Code:	Website: www.picronline.org
	DOI: 10.4103/picr.PICR_170_17

For reprints contact: reprints@medknow.com

How to cite this article: Ranganathan P, Aggarwal R. Common pitfalls in statistical analysis: Understanding the properties of diagnostic tests – Part 1. Perspect Clin Res 2018;9:40-3.

Sensitivity is defined as the proportion of individuals with the disease (as detected by the gold standard test) who have a positive result on the new test. In Table 1, of the ten individuals with pulmonary embolism, seven had a positive result on the D-dimer test; therefore, the sensitivity of D-dimer test for the detection of pulmonary embolism is 7/10 = 70%.

Specificity is defined as the proportion of individuals without the disease (as detected by the gold standard test) who have a negative result on the new test. In Table 1, of the ninety individuals without pulmonary embolism, 77 had a negative result on the D-Dimer test; therefore, the specificity of the D-dimer test for the detection of pulmonary embolism is 77/90 = 85.6%.

If we were to replace the cells in the example above with generic terms, we get a 2 × 2 contingency table [Table 2], which can be used for all diagnostic tests.

A highly sensitive test will be positive in almost everyone with the disease of interest, but may also be positive in some individuals without the disease. However, it would hardly ever be negative in a person with the disease. Thus, if a highly sensitive test is negative, it almost definitely rules out the disease (hence the mnemonic SNOUT: SnNOut = Sensitive…Negative…Rules Out).

A highly specific test will be negative in almost everyone without the disease, but may be negative in some with the disease. However, it would hardly ever be positive in an individual without the disease. If a highly specific test is positive, it almost definitely rules in the disease (hence the mnemonic SPIN: SpPIn = Specific…Positive…Rules In).

Sensitivity and specificity are useful attributes for comparing a new test against the gold standard test. However, these measures have some limitations. First, the sensitivity is calculated based on individuals with the disease and fails to give any information about people without the disease. Similarly, specificity is calculated based on individuals without the disease and does not tell us anything about individuals with the disease. Second, in the clinic, we see a patient with a particular set of symptoms and are unsure whether he has the disease or not. We then do the test and obtain its result. What we need at that point is not what proportion of individuals with the disease have the test positive; instead, we want to predict whether the particular individual has the disease, based on the positive or negative test result. This can be done much better using measures that are referred to as the predictive values of the test result.

Table 1: Number of individuals in whom pulmonary embolism was detected using the perfusion scan (gold standard) versus the results of the blood test for D-dimer

	Pulmonary embolism present	Pulmonary embolism absent	Row total
D-dimer positive	7	13	20
D-dimer negative	3	77	80
Column total	10	90	100

Table 2: 2×2 contingency table for assessing the sensitivity and specificity of a diagnostic test

	Disease present	Disease absent	Row totals
Test positive	a (TP)	b (FP)	a + b
Test negative	c (FN)	d (TN)	c + d
Column totals	a + c	b + d	a + b + c + d

Sensitivity=TP/(TP + FN) = a/(a + c), Specificity=TN/(FP + TN) = d/(b + d), Positive predictive value=TP/(TP + FP) = a/(a + b), Negative predictive value=TN/(TN + FN) = d/(c + d). TP=True positive, FP=False positive, FN=False negative, TN=True negative

POSITIVE AND NEGATIVE PREDICTIVE VALUES

Predictive values refer to the ability of a test result to confirm the presence or absence of a disease, based on whether it is positive or negative, respectively.

Referring to the previous example, of the twenty individuals in whom D-dimer test was positive, only seven actually had pulmonary embolism; therefore, the positive predictive value (PPV; also sometimes more appropriately referred to as the predictive value of a positive test result) of this test is 7/20, or 35%. PPV reflects the probability that an individual with a positive test result truly has the disease.

Similarly, of the eighty individuals with a negative D-dimer test, 77 did not have pulmonary embolism; therefore, the negative predictive value (NPV; or predictive value of a negative test result) is 77/80, or 96%. NPV is the probability that an individual with a negative test result truly does not have the disease.

RELATIONSHIP OF POSITIVE PREDICTIVE VALUE AND NEGATIVE PREDICTIVE VALUE WITH THE PREVALENCE OF DISEASE

It is important to note that sensitivity and specificity are properties of a test and are usually not influenced by the prevalence of disease in the population. By contrast, PPV and NPV are heavily influenced by the prevalence of the disease in the population tested/studied. With all the other factors remaining constant, the PPV increases with increasing prevalence and NPV decreases with increase in prevalence.

To illustrate this, let us look at the use of D-dimer test among three separate groups of individuals : all patients admitted to a hospital (with a hypothetical prevalence of pulmonary embolism of 1%), cancer patients undergoing chemotherapy (10%), and critically ill cancer patients in an intensive care unit (30%). Let us do D-dimer test in 1000 individuals from each of these groups.

Among the 1000 inpatients, the prevalence being 1%, 10 will have pulmonary embolism. By comparison, in the other two groups, 100 and 300 patients, respectively, will have pulmonary embolism. We have already established that the D-dimer test is 70% sensitive and 85.6% specific. Using these numbers, the number of individuals with positive and negative test results in the three groups can be calculated and are shown in Table 3a-c, respectively.

Let us now calculate PPV and NPV in each situation. Thus, when the test is done in all inpatients, its PPV is 7/150 = 4.7% and the NPV is 847/850 = 99.6% [Table 3a]. By comparison, when it is done in all cancer patients, the PPV is 70/200 = 35.0% and the NPV is 770/800 = 96.2% [Table 3b]. Further, when it is done in critically ill cancer patients, the PPV is 210/311 = 67.5% and the NPV is 599/689 = 86.9% [Table 3c]. It is apparent that the values of PPV and NPV are quite different in the three situations.

The above example shows us the importance of likelihood of the disease of interest in the individual in whom the test has been done (also referred to as the pretest probability of disease). Thus, even a test with good sensitivity and specificity has low PPV when used in a population where the likelihood of the disease is low (low pretest probability, as in all inpatients in the example in Table 3a). This is not infrequent. When a test is initially developed, it is costly and is used primarily in those with a high likelihood of disease. However, later, when the test becomes cheaper and more widely available, it is often used more indiscriminately even among those with a low likelihood of disease, resulting in a lower PPV. In view of this phenomenon, it is prudent to apply a diagnostic test only in those with a high pretest probability of the disease (based on symptoms and signs). Similarly, among persons with a strong suspicion of disease (high pretest probability), the NPV of a test may not be high,– i.e., in this situation, even a negative test result may not reliably rule out a disease.

Table 3a: Performance of D-dimer test for pulmonary embolism in 1000 unselected inpatients in a hospital (with hypothetical disease prevalence of 1%)

	Pulmonary embolism present	Pulmonary embolism absent	Total
D-dimer positive	7	143	150
D-dimer negative	3	847	850
Total	10	990	1000

Table 3b: Performance of D-dimer test for pulmonary embolism among 1000 cancer patients in a hospital (with hypothetical disease prevalence of 10%)

	Pulmonary embolism present	Pulmonary embolism absent	Total
D-dimer positive	70	130	200
D-dimer negative	30	770	800
Total	100	900	1000

Table 3c: Performance of D-dimer test for pulmonary embolism among 1000 critically ill cancer patients in an Intensive Care Unit (with hypothetical disease prevalence of 30%)

	Pulmonary embolism present	Pulmonary embolism absent	Total
D-dimer positive	210	101	311
D-dimer negative	90	599	689
Total	300	700	1000

Some tests have a clear dichotomous result – the test is either positive or negative – for example, presence or absence of pus cells in urine or of HBsAg in the blood. For such tests, interobserver variability is negligible. However, when we look at the results of tests such as chest radiographs, interpretation depends on the experience of the assessor, and the sensitivity and specificity of the test can vary, depending on the accuracy of reporting. For tests which report on a continuous scale, for example, random blood sugar for the diagnosis of diabetes, choosing a cutoff point to define disease can change the sensitivity and specificity. We will discuss this in the next article.

EXAMPLES OF DIAGNOSTIC TESTS IN PRACTICE

Enzyme-linked immunoassay (ELISA) tests are generally used as the initial screening tests for HIV infection. This is because they are highly sensitive (and therefore pick up most people with infection). Since the sensitivity is high, a negative ELISA almost certainly rules out infection (recall the SnNOUT mnemonic). However, the problem with highly sensitive tests is that they may also have a number of false-positive results. Therefore, anyone with a positive ELISA should be subjected to another test with a high specificity such as polymerase chain reaction to confirm the presence of HIV infection.

Low-dose computed tomography scan (LDCT) has been recommended as a screening tool for lung cancer. This is a highly sensitive test (sensitivity reported from 80% to 100%) – this means that almost every cancerous lung nodule will be detected on LDCT. The problem here is that the LDCT also picks up benign calcific nodules and therefore has several false positives; the specificity is low (reported around 20%). Since the prevalence of lung cancer in the average population is low, using this test for screening this population will have a high NPV but a low PPV. If we apply clinical criteria and do the test only in individuals at a high probability of lung cancer (e.g., elderly, heavy smokers, and those with hemoptysis), the pretest probability of lung cancer is higher, and the test would have a higher PPV.

SUGGESTED READING

The readers may want to read an article by Kim and colleagues who assessed the use of hip radiographs as an aid to diagnose hip osteoarthritis. This article examines the sensitivity, specificity, PPV, and NPV of this test as a diagnostic tool.[1]

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

REFERENCE

1. Kim C, Nevitt MC, Niu J, Clancy MM, Lane NE, Link TM, *et al*. Association of hip pain with radiographic evidence of hip osteoarthritis: Diagnostic test study. BMJ 2015;351:h5983.