Journal of Reliability and Statistical Studies (ISSN: 0974-8024)
Vol. 2, Issue 2(2009): 53-58

THE CONCEPT OF SENSITIVITY AND SPECIFICITY IN RELATION TO TWO TYPES OF ERRORS AND ITS APPLICATION IN MEDICAL RESEARCH

Devashish Sharma*, U.B. Yadav, Pulak Sharma****
* Deptt. of Statistics and Demography, Post Partum Programme
E Mail: devanita@ymail.com
** Deptt. of Orthopaedics
M.L.N. Medical College, Allahabad, U.P. (India)

Abstract

Sensitivity and specificity are two terms widely used in Medical research and are the statistical measures of performance of a binary classification test. In clinical research the sensitivity of a medical test is the probability of its giving a ‘positive’ result when the patient is indeed positive and specificity is the probability of getting ‘negative’ result when the patient is indeed negative. Wrongly identify a healthy person as sick and a sick person as healthy is closely related to the concept of type I and type II errors of testing hypothesis. It was observed that the sensitivity of a test is equal to power of test in hypothesis testing.

Key Words: True positive, True negative, Fracture neck of femur, Valgus Oseotomy.

1. Introduction

Sensitivity and Specificity are the two terms widely used in medical and epidemiological research, but most of the statisticians in mathematical fields are unaware of it. Sensitivity and specificity are the statistical measures of performance of a binary classification tests. The sensitivity measures the proportion of actual positive which are classified as such (e.g. the percentage of sick people who are identified as having the condition); and specificity measures the proportion of negatives who are correctly identified (e.g. the percentage of well people who are identified as not having the condition). In short, sensitivity refers the probability of true showing up true and specificity to the probability of false showing up false. Sensitivity and specificity are usually expressed in percentage.

In clinical research the sensitivity of a medical test is the probability of its giving a ‘positive’ result when the patient is indeed positive and specificity is the probability of getting a negative result when the patient is indeed negative. A theoretical optimal prediction result can achieve 100% sensitive (i.e. predicts all people from a sick population as sick) and 100% specificity (i.e. not predicts any from the healthy population).

Imagine a scenario, where people are tested for a disease. The test outcome can be positive (sick) or negative (healthy), while the actual health status of a person may be different. Following four conditions may occur:

- Sick people correctly diagnosed sick termed as “True positive”
- Healthy people wrongly identified as sick – “False positive”
- Healthy people correctly identified as healthy – “True negative”
- Sick person wrongly identified as healthy – “False negative”

From the above conditions it is clear that in two cases an error has occurred, when a healthy person is wrongly identified as sick and the other one where a sick person is wrongly identified as healthy. These two types of errors are closely related to the concept of type I and type II errors in hypothesis testing.

Hypothesis testing is a method of making statistical decisions about the population on the basis of experimental data. It is also known as Statistical Significance Testing. In hypothesis testing there is a "null hypothesis" which corresponds to a presumed default "State of nature" (e.g. that an individual is free from disease). Corresponding to null hypothesis is an "alternative hypothesis" which corresponds to the opposite situation. The goal is to determine accurately if the null hypothesis can be discarded in favour of the alternative. A test of some sort is conducted and the data is obtained. The result of this test may be negative (i.e. it does not indicate disease) or it may be positive (it may indicate disease). If the result of test does not correspond with the actual states of nature, then an error has occurred. There are two kinds of error classified as "Type I and Type II errors" depending upon which hypothesis has incorrectly been identified as the true state of nature.

Type I error is known as error of first kind, or "α" error, or a false positive", the error of rejecting null hypothesis when it is actually true. A false positive normally mean that a test claims something to be positive when that is not the case. For example, a test saying that a woman is pregnant when she is actually not pregnant. Type II error, is also known as "error of second kind" or "β" error or a "false negative", the error of accepting null hypothesis when alternative hypothesis is true. The following table illustrates the condition:

		Actual Condition	
		Present	Absent
Test Result	Positive	Condition Present + Positive Result = True Positive	Condition absent + Positive result = False Positive (Type I error)
	Negative	Condition Present + Negative Result = False negative (Type II error)	Condition absent + Negative result = True negative

The probability that an observed prediction result is a false positive (as contrast with an observed positive result being true positive) may be calculated using Bayes's theorem. Bayes theorem is a result in probability theory, which relates the conditional and marginal probability distribution of random variables. The key concept of Bayes's theorem is that the true rates of false positive and false negative are not a function of the accuracy of the test alone, but also the actual rate or frequency of occurrence within the test population; and often, the more powerful issue is the actual rates of the condition within the sample being tested.

Sensitivity is defined as:

$$\text{Sensitivity} = \frac{\text{Number of true positives}}{\text{Number of true positives} + \text{Number of false negatives}}$$

Sensitivity alone does not tell us how well the test predicts the other class (i.e. about negative cases). In the binary classification this is the corresponding specificity test or equivalently the sensitivity for the other classes.

The calculation if sensitivity does not tale into account the intermediate results, the option are either to exclude intermediate samples from analysis (but the number of exclusions should be stated when quoting sensitivity) or alternatively intermediate samples can be treated as false negative.

Specificity is the number of true negative to the number of true negative plus number of false positive.

$$\text{Sensitivity} = \frac{\text{Number of true negatives}}{\text{Number of true negatives + Number of false positives}}$$

A test with high specificity has a low type I error rate. Specificity is sometimes confused with **precision** or the positive predicted value, both of which refer to the fraction of returned positives that are true positives. A test with high specificity can have a very low precision if there are far more true negatives than true positives and vice versa. The relationship among terms can be illustrated as follows:

		Condition as determined by Gold Standard		
		Positive	Negative	
Test result	Positive	True Positive	False Positive (Type I error)	→ Positive Predictive value
	Negative	False Negative (Type II error)	True Negative	→ Negative Predictive value
		↓ **Sensitivity**	↓ Specificity	

The concept of Sensitivity and specificity is very useful in finding out the utilities of new therapies in medical field as compared to standard therapy and also assessing new scaling of assessing improvement in patients as compared to old, well established and widely used criteria.

2. A practical application of sensitivity and specificity

In this paper an attempt was made to find out the reliability of NATH scoring system (2004) which is a new concept of assessing the functional outcome of patients after operation for neglected fracture neck of femur (fracture around hip joint) as compared to well established Askin Bryan Criteria of 1976 (Gold Standard).

2.1 Material and Methods

Twenty two young adult patients of neglected fracture neck of femur were treated in the Department of Orthopaedics of M.L.N. Medical College, Allahabad, from June 2006 to January 2009. In all the patients an operation known as Valgus Osteotomy was performed followed by internal fixation using a 120^{O} double angled dynamic hip screw barrel plate assembly.

The patients were called for follow-up at every month for four months and thereafter every two months. The average follow-up period was 11 months (ranging from 6 months to 18 months).

At each follow-up the assessment of the functional outcome was done by two methods (1) **Askin Bryan criteria** (Gold standard) and (2) by **Nath's scoring system** – a relatively new approach.

According to **Askin Bryan criteria** the functional outcome of the patients were classified into four groups as given below:

Excellent	Full range of movement and strength, little or no pain and essentially normal appearing radiographs
Good	Some limitation of motion, mild discomfort and mild joint space narrowing
Fair	Some limitation of motion and moderate pain with degenerative changes or aseptic necrosis.
Poor	Severe restriction of function and pain requiring salvage procedure.

The new concept of **Nath scoring system** is a hundred point scoring system which takes into consideration four major and four minor criteria, each criteria being assigned a score based on its severity.

Major Criteria **(Total 65 points)**		**Minor Criteria** **(Total 35 points)**	
1.	Pain – 20 points	1.	Walking ability – 10 points
2.	Avascular necrosis – 15 points	2.	Limp – 10 points
3.	Union – 15 points	3.	Movements - 08 points
4.	Shortening – 15 points	4.	Neck shaft angle- 07 points

Details of the Nath's scoring system (NSS) are:
Pain: No pain 20 points, mild tolerate – 15, moderate limiting daily activity – 5 and severe pain-zero; ***Avascular necrosis***: No avascular necrosis- 15 points, increased density of head- 10, segmental collapse -5, severe with arthritic changes zero; ***Union***: Fracture union – 15 points, upto 2.5 cm – 10, 2.5 cm to 5 cm – 5, more than 5 cm –zero; ***Shortening:*** No shortening – 15 points, upto 2.5 cm -10, with single crutches -4 and no weight bearing – zero; ***Walking ability***: without aid – 10 points, with single stick – 7, with crutches – 4, no weight bearing –zero point; ***Limp***: No limp – 10 points, mild – 8 points, mild to moderate – 6, severe – 2 and inability to walk zero point; ***Movement***: > 130^{O} – 8 points,, 110 – 130^{O} – 6, 90 – 100^{O}- 4 points; ***Neck shaft angle***: > 120^{O} – 7 points,, 110-120^{O} 0 5, 100-110^{O} – 3 and less than 100^{O} – 1 point.

Based on the sum total of the scores of major and minor criteria, the functional outcome is classified into the following groups:

Excellent	Patient scoring 90-100 points on Nath scoring system was
Good	80-89 points
Fair	70-79 points
Poor	Below 70 points.

While comparing the final functional outcome of the patients by the two criteria, excellent and good results were treated as success whereas fair and poor results were taken as failure.

2.2 Observations

The procedure was done in young adults ranging from 18-44 years. Mean age of the patients was 32.67± 6.72 years (Mean ± SD). Out of 22 patients 15 were male and seven were females. The injury and operation interval was between 8 to 20 weeks and the average of 9.37±3.69 weeks.

Excellent results were seen in 3 patients by both methods. Thirteen patients showed good functional outcome by Askin Bryan criteria while by Nath's assessment criteria only 12 patients showed good results. Two patients showed poor functional outcome by Askin Bryan as compared to only one by Nath's criteria.

Thirteen patients showed excellent or good functional outcome (i.e. success) by both methods (True positive), while four patients fail to respond treatment (True Negative). Table 1 illustrate the four conditions.

Nath's scoring System	Askin Brya Criteria (Gold Standard)		Total
	Success	Failure	
Success	13 (True Positive)	2 (False Positive)	15
Failure	3 (False Negative)	4 (True Positive)	7
Total	16	6	22

Sensitivity = [TP/(TP+FN)]x 100 = [13 / 16] x 100 = 81.2%

Specificity = [TN/ (TN+FP)] x100 = [4/6] x100 = 66.67%

Related Calculations:

False Positive Rate (α) = [FP / (FP+TN)] x 100 = [2/6] x100 = 33.33%
= 1 – specificity

False negative rate (β) = [FN/ (FN+TP)] x 100 = [3/16] x100 = 18.8%
= 1 – sensitivity

Power = 1 – β = sensitivity

3. Discussion

It was observed that the new scoring system for assessing functional outcome (Naths scoring system) predicts approximately 82% patients correctly which showed excellent or good functional outcome by Askin Bryan criteria, a well established and widely used criteria. The drawback Nath's scoring system is that it has high type I error (33.33%), i.e. a comparatively low specificity. The new system is unable to identify those patients who showed poor satisfactory functional results. About 33.33% patients were wrongly assessed to have a successful operation (i.e excellent of good functional outcome). The large number of false positive is in itself poor at confirming the success of functional outcome.

The false negative rate is 18.8%, which is within the acceptable limits of any clinical trials. It is a usual convention that the type II error in any clinical trial should not exceed 20%.

The advantage of Nath's scoring system is that it gives grading to different parameters which a doctor can asses and summing up these points will reduce the observational bias, whereas Askin Bryan system the functional outcome was estimated solely on the basis of the assessment of the clinician. This may vary from individual to individual.

It will be more meaningful if the assessment of functional outcome of patients were done by both methods because this may help clinicians to assess the patient's condition well and in case of failure of procedure he can timely initiate the necessary action.

4. Conclusions

The sensitivity and specificity of a test is very helpful in medical science. From statistician point of view they are closely related to type I and Type II errors in testing hypothesis. In statistical hypothesis testing type I error is usually denoted by α, and 1- α is defined as specificity. Increasing the specificity of the test lower the probability of type I error. Thus, specificity is a statistical measure how well a binary classification test correctly identifies the negative cases.

Similarly Type II error is denoted by β. In traditional language of statistical hypothesis testing, the sensitivity of a test is called the statistical power of the test, although the word power in that context has a more general usage that is not applicable the present context. A sensitive test have a fewer Type II error.

Thus, we can conclude that the concept of sensitivity and specificity and other related concepts will be very helpful in other fields of applied statistics.

References

1. Altman, D.G. and Bland, J.M. (1994). Statistics Notes: Diagnostic tests 1: sensitivity and specificity, Br. Med. Jour., 308, p. 1152.
2. Askin, S.R. and Bryan, R.S. (1976). Femoral neck fractures in young adults, Clin. Orthop., 114, p. 259-264.
3. Gaddis, G.M. and Gaddis, M.L. (1990). Introduction to biostatistics: Part 3, sensitivity, specificity, predictive values and hypothesis testing. Ann. Emerg. Med., 19(5), p. 591-597.
4. Loong, T.W. (2003). Understanding sensitivity and specificity with the right side of brain. Br. Med. Jour., 327, p. 716-719.
5. Nath, R., Rastogi, S., Gupta, A.K. and Prasad, N. (2006). Reposition osteotomy for fracture neck of femur- A simplified technique of surgery assessment o results, Indian J. of Ortho., 40(4), p. 1-4.