Journal of Athletic Training 2012:47(2):233–236

www.nata.org/jat

Application of Statistics in Establishing Diagnostic Certainty

Craig R. Denegar, PhD, ATC, PT, FNATA*; Mitchell L. Cordova, PhD, ATC, FNATA, FACSM†

*Doctor of Physical Therapy Program, Department of Kinesiology, University of Connecticut, Storrs; †College of Health Professions, Florida Gulf Coast University, Fort Myers

The examination and assessment of injured and ill patients leads to the establishment of a diagnosis. However, the tests and procedures used in health care, including procedures performed by certified athletic trainers, are individually and collectively imperfect in confirming or ruling out a condition of concern. Thus, research into the utility of diagnostic tests is needed to identify the procedures that are most helpful and to indicate the confidence one should place in the results of the test. The purpose of this report is to provide an overview of selected statistical procedures and the interpretation of data appropriate for assessing the utility of diagnostic tests with dichotomous (positive or negative) outcomes, with particular attention to the interpretation of sensitivity and specificity estimates and the reporting of confidence intervals around likelihood ratio estimates.

Key Words: sensitivity, specificity, likelihood ratios, confidence intervals

Clinical research strives to evaluate and guide health care delivery across the spectrum of patient care, including prevention, diagnosis, and treatment. The focus of this article is on interpreting the results of diagnostic tests with dichotomous outcomes. We offer a short review of sensitivity, specificity, and likelihood ratios with attention to interpreting estimates of sensitivity and specificity; in addition, we discuss reporting confidence intervals around estimates of likelihood ratios.

Clinical Context

The examination of injured and ill patients and the evaluation of the data obtained from the history and physical examination are central to making decisions about referral, diagnosis, treatment, and return to activity. The extent of the examination and the clarity of the end result vary from patient to patient. For example, an obvious deformity at the ankle joint complex requires only observation for the clinician to conclude that a fracture-dislocation has been sustained and emergency care is warranted. More commonly, however, at the end of the evaluation, uncertainty remains. For instance, consider an athlete with ankle pain that is associated with a reported inversion mechanism of injury. The medical history is unremarkable except for a previous ankle injury 5 months ago. The ankle is swollen and point tender upon palpation. The list of diagnostic possibilities includes sprain and fracture. Radiographs can effectively demonstrate a fracture, but are they warranted? The answer to this question depends largely on obtaining and interpreting additional information, which itself necessitates an understanding of the basic principles of diagnostic statistics. Using the scenario presented, we will introduce statistical concepts and then apply information about selected diagnostic tests to the management of this patient.

Sensitivity and Specificity

The terms *sensitivity* and *specificity* are familiar to many readers and represent the foundation for calculating likelihood ratios and interpreting research into the value of diagnostic procedures in informing clinical decisions and forming diagnoses. These values are easily understood using a 2 × 2 contingency table (Table 1).

Sensitivity and specificity are calculated by comparing a diagnostic test with a reference standard test that is normally referred to as a gold standard. In the case of a traumatic fracture of the ankle, radiography provides a highly accurate reference standard. The results of the reference standard test are then compared with those of a clinical examination procedure. This comparison yields 4 possibilities. In cells A and D, agreement occurs between the results of the reference test and the examination procedure, whereas in cells B and C, the results disagree and diagnostic test error exists. Sensitivity is calculated from the left side of the table (sensitivity = A/[A + C]); specificity is calculated from the right side (specificity = D/[D + B]). A diagnostic test has high sensitivity when it agrees with the reference standard on a high proportion of positive diagnoses. A test has high specificity when it agrees with the reference standard on a high proportion of negative diagnoses. These calculations yield values ranging from 1.0 to 0.0, where 1 represents perfect agreement between the examination procedure under investigation and 0 represents complete disagreement.

Although high sensitivity is associated with positive reference standard findings, examination procedures with sensitivity values approaching 1.0 are very useful in ruling out a condition, whereas those with specificity approaching 1.0 are useful in making a diagnosis. The abbreviations SnNOUT ("a negative examination using a sensitive test rules out") and SpPIN ("a positive examination using a test with high specificity rules

Table 1. Sensitivity, Specificity, and Prediction Values in Diagnostic Testing

Clinical Examination Procedure Result	Reference Standard Result: Condition Present	Reference Standard Result: Condition Absent	Prediction Value
Positive	Cell A True positive	Cell B False positive	Positive prediction value
Negative	Cell C False negative	Cell D True negative	Negative prediction value
	Sensitivity	Specificity	

in") may be helpful in recalling these interpretations when reviewing clinical research.[1] Yet at first, SnNOUT and SpPIN may appear counterintuitive. Consider sensitivity, which is really the proportion of patients who present with a condition that yields a positive test during their clinical examination. If sensitivity equals 0.90, then 90% of patients with the condition of interest will have a positive result on the examination procedure being investigated. However, sensitivity of 0.90 with a positive finding on the examination procedure does not reflect a 90% certainty that the patient has the condition. Sensitivity estimates do not take the number of false-positive findings into consideration. A closer look at Table 1 confirms that a sensitive test is associated with few false-negative results; therefore, if an examination procedure with high sensitivity yields a negative test result, the finding is very likely to be truly negative. In other words, a negative examination finding on a test with high sensitivity is good at ruling out a diagnosis of the disease or condition of concern. The same argument applies to examination procedures with high specificity. Because false-positive findings are rare in tests with high specificity, a positive examination result using such a test may confirm a diagnosis.

The Ottawa Ankle Rules (OAR)[2] have been extensively studied and found to have nearly perfect sensitivity. Therefore, applying these rules to the ankle-injury patient described earlier helps in the clinical decision-making process. If our patient has no tenderness at the posterior-distal 6 cm of the medial or lateral malleolus, the base of the fifth metatarsal, or the navicular and can walk 4 steps on presentation and after injury, the probability that a fracture has occurred is extremely low. Radiographs are not warranted unless some feature in the patient's presentation raises a concern for bony injury. Thus, the list of diagnostic possibilities has been shortened without additional cost or exposure to radiation. Knowledge of sensitivity and specificity, as well as the concepts of SpPIN and SnNOUT, serves athletic trainers well in making sound clinical decisions.

An alternative to the calculation of sensitivity and specificity is the calculation of positive and negative prediction values (PPV and NPV, respectively). The PPV is calculated by dividing the number of true-positive findings by the sum of true-positive + false-positive findings (A/[A + B]); NPV is calculated by dividing the number of true-negative findings by the sum of true-negative + false-negative findings (D/[D + C]). Therefore, the PPV appears to solve the problem of knowing the probability that a positive finding truly indicates that the patient has the condition of concern. Yet the PPV and NPV estimates are affected by the prevalence of a condition. If a condition is very common, the number of true-positive findings will be large, and the number of true-negative findings will be small. In this case, adding a false-positive finding will have a small effect on the PPV estimate because the numerator (true positives) is large, but a false-negative finding will have a greater effect on the estimated NPV. The bottom line is that highly sensitive tests are good at ruling out and highly specific tests are good at ruling in the presence of injury or illness. However, intuitively attractive PPV and NPV values must be interpreted cautiously unless the prevalence of the condition is known.

Likelihood Ratios

Conducting diagnostic procedures with sensitivity or specificity approaching 1.0 clearly guides practice decisions made by clinicians, although few physical examination procedures fall into this category. Therefore, reporting of sensitivity and specificity values alone often leaves the clinician without a clear understanding as to what a positive or negative examination finding really means. A test with a sensitivity of 0.80 is pretty good at ruling out a condition of interest, but "pretty good" is not conclusive. The reality is that the results of many examination procedures, when interpreted in isolation, shift the probability that a condition is present rather than convince the clinician of its presence or absence. However, the question remains: How large a shift in probability is generated by a positive or negative examination finding? The answer does not rest simply with sensitivity or specificity but rather with the values derived from these estimates.

Positive and negative likelihood ratios (LRs) are calculated from sensitivity and specificity values as follows:

$$+\text{LR} = \text{Sensitivity}/(1 - \text{Specificity})$$
$$-\text{LR} = (1 - \text{Sensitivity})/\text{Specificity}$$

A +LR indicates the effect a positive examination finding has on the probability that the condition in question truly exists. A –LR addresses the effect a negative examination has on the probability that the condition in question truly exists. Jaeschke et al[3] summarized LRs (both positive and negative) into broader categories of clinical value (Table 2). From this table, it is apparent that a large +LR and small –LR generate large shifts in the probability that a condition exists, which is helpful for decision making. Diagnostic procedures with +LR and –LR values approaching 1.0 are of little or no value in making sound clinical decisions. For further discussions of the relationships between likelihood ratios and the probability that a patient has or does not have a condition, see Fritz and Wainner[4] and Denegar and Fraser.[5]

Let us return to the patient with the injured ankle and assume that we find exquisite tenderness on palpation of the distal posterior fibula. Based on the application of the OAR, radiography is indicated. Although the sensitivity of the OAR is nearly 1.0, the specificity is much lower.[6] Gravel et al[7] reported OAR specificity at 0.27 in a pediatric population. Thus, the tenderness identified on clinical examination does not necessarily mean

Table 2. Clinical Values of Likelihood Ratios from Jaeschke et al[3]

Positive Likelihood Ratio	Negative Likelihood Ratio	Shift in Probability That Condition Is Present
>10	<0.1	Large, often conclusive
5–10	0.1–0.2	Moderate but usually important
2–5	0.2–0.5	Small, sometimes important
1–2	0.5–1	Very small, usually unimportant

Reprinted with permission from Denegar CR, Fraser M. How useful are physical examination procedures? Understanding and applying likelihood ratios. *J Athl Train*. 2006;41(2):201–206.

that a fracture exists. Radiographs will probably provide the answer, but are there clinical physical examination procedures that provide useful information under these circumstances? The use of a tuning fork and stethoscope in identifying fractures[8] and improving the sensitivity of the OAR[9] has been investigated. Moore[8] reported on a technique with a sensitivity of 0.83 and specificity of 0.92. A specificity value of 0.92 is high, and so we ask whether a positive test is conclusive based on this measure alone. The +LR calculated from these estimates is 10.4, and the –LR is 0.18. Thus, a substantial dampening of the vibration generated by a tuning fork as detected with a stethoscope (positive test) generates a large and perhaps conclusive shift in the probability of a fracture, whereas a normal sound detected through the stethoscope suggests the bone is intact. With a positive finding, radiography is warranted to determine the extent of the injury and to guide treatment. However, the clinician can proceed with a high degree of confidence that a fracture has occurred. Conversely, a negative finding is probably not sufficient to recommend against radiography, but there is a bit more to the story.

Confidence Intervals

In the preceding section, we addressed point estimates of sensitivity, specificity, and likelihood ratios. These data were derived through the study of a sample drawn from a larger population and provide point estimates of the values that would be obtained if the entire population could be studied. Although it is not possible to know the true population values, it is possible to know with a specified degree of confidence (eg, 95% or 99%) the range in which the population values lie. Confidence intervals (CIs) provide this information and assist us in appraising the clinical meaningfulness of the observed treatment effects or test results.[10] The methods used to calculate the CI around an LR and sensitivity and specificity estimates have been detailed by Simel et al.[11] Confidence intervals are not reported in all publications, but it is easy to perform these calculations with public domain software. We recommend the CI calculator available on the PEDro (Physiotherapy Evidence Database) Web site.[12] In the example used earlier, participants were drawn from the population of injured patients requiring evaluation for a fracture. The +LR calculated from the estimate provided in the previous section is 10.4 and the 95% CI is 2.7 to 40.3. The –LR is 0.18 (95% CI = 0.05, 0.65). Thus, in the report by Moore,[8] there is 95% confidence that the true +LR associated with the tuning fork assessment lies between 2.7 and 40.3, whereas the true –LR lies between 0.05 and 0.65. Using the criteria previously referenced, a +LR of 10.4 suggests that a positive test generates a large and often conclusive shift in the probability that a fracture has occurred. This finding might lead a clinician to adopt the tuning fork test in his or her practice. But what if the +LR is only 2.7? A positive test would be of little importance in deciding how to best manage the patient. Similarly, the 95% CI for the –LR ranges from very low (0.05), which is associated with a large shift in the probability of a fracture with a negative examination, to 0.65, which suggests that a negative examination is of little value in ruling out a fracture.

In interpreting these results, the obvious question is, "Why are the confidence intervals so large?" The point estimates (ie, +LR = 10.4, –LR = 0.18) and the accompanying 95% CIs were derived from a sample of 37 patients. Larger samples are likely to better represent the population and offer greater precision for point estimates. The clinician must await studies with larger samples and multiple data sources through a meta-analysis to gain a clearer understanding of the diagnostic utility of a tuning fork in detecting fractures.

When we read research related to diagnostic accuracy, it is important to note that the point estimate does not lie in the middle of the CI for either the +LR or –LR, as one would find when examining the CI around mean values or mean differences. The formula used in the aforementioned calculations[11] generates a CI around a log likelihood ratio. A more detailed explanation of the need for this transformation is provided by Simel et al.[11] It is also necessary to recall that +LR and –LR estimates are positive, whereas a value of 1 reflects a test that has no ability to inform decisions about the patient's condition. Thus, CIs are not anchored to a value of zero or no difference.

Decision Points

The purpose of this short report is to provide the reader with the knowledge to interpret the results of studies of diagnostic tests with dichotomous outcomes. As we have stated, many tests used to evaluate the injured athlete do not resolve the uncertainties of making a diagnosis. With the exception of a few procedures that are very sensitive or very specific, it is apparent that the clinician cannot rely on a single test result in the decision-making process. The reality is that the examination procedure results shift the probability of the presence or absence of a condition. The clinician must then obtain more information to decide how best to proceed. The history and observation components of the examination provide a great deal of information as to the source of a patient's complaint. Fritz and Wainner[4] offered an excellent discussion on the development and application of pretest probability, or the clinician's estimate of the probability that a condition exists before performing specialized testing. The results of one or more diagnostic tests shift this probability toward or away from a specific diagnosis. Ultimately, the clinician reaches a decision point, with some risk of being incorrect in the assessment.

Even "conclusive" shifts in probability must be considered in context. A more serious condition or more costly and complex course of care may warrant greater certainty before treatment begins. The lack of certainty should not discourage

critical appraisal of the clinical research. It is through these efforts that we can eliminate diagnostic procedures that are of little value from practice, so that the results of the procedures conducted can be interpreted appropriately. Over time, more information will become available and allow stronger recommendations to be made, as in the use of the tuning fork test as a standard of practice in diagnosing fractures and improving the sensitivity of the OAR.

Summary

Clinical research can guide the use and interpretation of evaluation procedures and diagnostic tests. The statistics used to report the results of research into diagnostic procedures differ from those used in studies reporting prognosis, prevention, or treatment outcomes. We have provided an overview of sensitivity, specificity, and likelihood ratios and applied these concepts in context to illustrate their use in clinical decision making. Understanding how to use these statistics allows the clinician to apply research findings across a large number of evaluative and diagnostic procedures in clinical practice with full appreciation of the strength and limitations of the diagnostic data.

REFERENCES

1. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical Epidemiology*. 2nd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 1991:69–152.
2. Stiell IG, Greenberg GH, McKnight RD, Nair RC, McDowell I, Worthington JR. A study to develop clinical decision rules for the use of radiography in acute ankle injuries. *Ann Emerg Med*. 1992;21(4):384–390.
3. Jaeschke R, Guyatt JH, Sackett DL. User's guide to the medical literature, III: how to use an article about a diagnostic test. B: What are the results and will they help me in caring for my patients? The Evidence-Based Medicine Working Group. *JAMA*. 1994;271(9):703–707.
4. Fritz JM, Wainner RS. Examining diagnostic tests: an evidence-based perspective. *Phys Ther*. 2001;81(9):1546–1564.
5. Denegar CR, Fraser M. How useful are physical examination procedures? Understanding and applying likelihood ratios. *J Athl Train*. 2006;41(2):201–206.
6. Bachman LM, Kolb E, Koller MT, Steuer J, ter Riet G. Accuracy of Ottawa ankle rules to exclude fractures of the ankle and mid-foot: systematic review. *BMJ*. 2003;326(7386):417.
7. Gravel J, Hedrei P, Grimard G, Gouin S. Prospective validation and head-to-head comparison of 3 ankle rules in a pediatric population. *Ann Emerg Med*. 2009;54(4):534–540.
8. Moore MB. The use of a tuning fork and stethoscope to identify fractures. *J Athl Train*. 2009;44(3):272–274.
9. Dissmann PD, Han KH. The tuning fork test: a useful tool for improving specificity in "Ottawa positive" patients after ankle inversion injury. *Emerg Med J*. 2006;23(10):788–790.
10. Cordova ML. Giving clinicians more to work with: let's incorporate confidence intervals into our data. *J Athl Train*. 2007;42(4):445.
11. Simel DL, Samsa GP, Matchar DB. Likelihood ratios with confidence: sample size estimation for diagnostic test studies. *J Clin Epidemiol*. 1991;44(8):763–770.
12. Physiotherapy Evidence Database. Confidence interval calculator. http://www.pedro.org.au/english/downloads. Accessed February 22, 2011.

Address correspondence to Craig R. Denegar, PhD, ATC, PT, FNATA, Doctor of Physical Therapy Program, Department of Kinesiology, University of Connecticut, Storrs, CT 06269. Address e-mail to craig.denegar@uconn.edu.