Open Access Publication

How to Evaluate an Article on Diagnosis for Validity

by Michael Turlik, DPM[1]

The Foot and Ankle Online Journal 2 (11): 6

This is the second of four articles discussing the evaluation of diagnostic studies for podiatric physicians. This article deals with the evaluation of internal and external validity of a diagnostic study. Diagnostic articles from the foot/ankle literature will be used to illustrate the concepts of critical analysis.

Key words: Evidence-based medicine.

Accepted: October, 2009 **Published:** November, 2009

The first article of this series[1] discussed two different methods of how clinicians solve diagnostic problems. When clinicians use probabilistic diagnostic reasoning to solve clinical problems they often require diagnostic tests to help refine and revise the diagnostic hypotheses they generate. The use of likelihood ratios (LR) for diagnostic test results was discussed and it was advanced that LRs were based upon the results of published diagnostic studies. The strength of the inference from these studies depended upon the validity of the methods used to obtain the information about the diagnostic test. Different study designs may produce different results for diagnostic accuracy. The purpose of this article is to provide instruction regarding the critical analysis of diagnostic studies. Whether one can believe the results of a study is determined by the methods used to carry it out. This is the second of four articles explaining how to evaluate diagnostic foot/ankle studies.

Internal validity

The ***Consort*** statement is a set of published guidelines for the reporting of randomized controlled trials in the medical literature.[2] The diagnostic counterpart to this document is the **STARD** Statement*****.[3] Most of the major medical journals with high impact factors have adopted these guidelines as requirements for publication. In the first article of this series[1] it was stated that the best study design for a diagnostic test was a cross sectional study of an independent, masked comparison with a reference standard among an appropriate population of consecutive patients. If properly planned and reported this type of study can be considered level I evidence for studies of diagnostic accuracy. The reader is referred to the following reference for a more complete explanation of the various study designs and levels of evidence for diagnostic studies.[4] This paper will confine itself to critically evaluating studies which seek to confirm a diagnosis rather than using the results for a screening test. Articles from the foot/ankle literature will be compared and contrasted to illustrate the principles discussed in this article.

***STA**ndards for the **R**eporting of **D**iagnostic Accuracy studies.

Address correspondence to: Michael Turlik, DPM
Email: mat@evidencebasedpodiatricmedicine.com

[1] Private practice, Macedonia, Ohio.

ISSN 1941-6806 doi: 10.3827/faoj.2009.0211.0006

Was there an appropriate consecutive population of patients enrolled in the study?

It is unacceptable to only enroll patients with severe symptoms and signs of the target disease to be compared to disease free people who are healthy asymptomatic volunteers. Studies[5,6] which use obvious diseased individuals compared with normal individuals do not provide any useful diagnostic information. Large overestimation of diagnostic accuracy has been shown when studies enroll only patients with advanced symptoms compared to normal controls.[7,8] The sensitivity and specificity of a test depends on the characteristics of the population studied. There needs to be a ***spectrum*** of patients similar to the patients we would use the test on in normal clinical practice. The spectrum should contain patients exhibiting mild signs and symptoms who are early in the target disease to late severe patients with the target disease as well as, conditions commonly confused with the target disease. Failure to enroll an appropriate population of patients results in spectrum bias with an overestimation of diagnostic effect. Readers of diagnostic studies should look for a description of the study population to include: definition of the target population, study location, age distribution, sex distribution, summary of presenting clinical symptoms and/or disease stage when evaluating diagnostic studies for spectrum bias.

Ideally patients will be **prospectively** enrolled in the study **consecutively** based upon suspicion of clinically having the target disease. It is important for the reader to review the method by which the patients entered the study. Alternate study designs may produce spectrum bias affecting the results of the study. One study showed[8] that accuracy of the index test was lower in studies that selected patients on the basis of whether they had been referred for the index test rather than on clinical symptoms. Failure to enroll patients **consecutively** (selection bias) and retrospective studies[9] are associated with an overestimation of diagnostic accuracy.[8]

In the first article of this series[1] it was proposed that post-test probabilities and predictive values vary with target disease prevalence. In contrast, sensitivities, specificities and likelihood ratios do not.[10] LRs are affected by disease spectrum. As long as the disease spectrum is the same prevalence will not affect LR, sensitivity or specificity. Studies with a different spectrum of disease will result in different LRs. In general, infected diabetic ulcers are less common and less severe with more subtle changes when seen in a podiatric physician's office setting than in patients with diabetic infected ulcers referred to a tertiary care center. Evaluating the spectrum of patients making up the study is a key and critical decision in judging the validity of diagnostic studies.

The two studies which we will use in this article to illustrate the critical analysis of diagnostic tests both evaluate the usefulness of a probe to bone test for osteomyelitis in diabetic infected ulcerations of the feet.[11,12] Both studies evaluated infected diabetic foot ulcers for osteomyelitis using a metal probe of the wound to detect bone. A positive probe to bone test is thought to aid in the diagnosis of osteomyelitis.

Grayson, et. al.,[11] evaluated prospectively infected diabetic foot ulcers in a tertiary care hospital setting. They describe their patient population as having severe limb-threatening infections. Demographic information was included not in a table format but, in the narrative of the results section. It wasn't clear if the patients were consecutively enrolled. Lavery, et. al.,[12] in a later study prospectively evaluated infected diabetic foot ulcers in an outpatient setting evaluating patients from two large primary care practices. The authors provided a table in the results section clearly describing the demographic features of the patients enrolled in the study. Patients were enrolled in the study based upon clinical findings during an office visit. It wasn't clear if the patients were consecutively enrolled. It should be clear that the spectrum of patients in the two studies is different. Neither study only enrolled patients with severe symptoms and signs of the target disease to be compared to disease free people who are healthy asymptomatic volunteers.

Did the study investigators compare the index test against an appropriate reference standard?

Ideally a consecutive series of patients with clinical findings of the target disease are subject to the index test and then verified by the reference standard. The term reference standard is sometimes referred to as the gold standard. Ideally a reference/gold standard should be able to distinguish between all normal people and people with the target disease. This test would be 100% sensitive and 100% specific with no false-positives or false-negative results. It is unlikely that any test will perform this well. Some common examples of appropriate reference standards are: biopsy, autopsy or long-term follow-up without treatment. While the choice of a reference standard may not be perfect it is essential for the reader of the study to have confidence in the reference standard used to identify the target disorder. It is important that the results of the index test not be part of the decision to perform the reference standard. In addition, there should be a close temporal relationship between the performance of the index test and the reference standard. Failure to perform the tests in an appropriate time sequence may result in an intervention in the care of the patient which would alter results of the reference standard and therefore bias the results.

Grayson, et. al.,[11] chose bone biopsy with histologic determination as the reference standard. Not all patients had a bone biopsy in some cases the reference standard included radiographic changes, clinical identification the bone by a surgeon during surgical debridement and after a short course of antibiotics if the ulceration resolved and did not recur. The authors report that the results of the index test influenced the decision to biopsy the bone. The intervals between the evaluation of the ulcer and biopsy of the bone for evaluation averaged 13.9 days (5-42). Lavery, et al.,[12] also used a bone biopsy as a reference standard however; the authors relied on culture results rather than histologic examination. The authors clearly described that the decision to biopsy the bone was **not** influenced by the results of the index test. It wasn't clearly stated what the time difference was between the performance of the index test and the reference standard. In addition, patients that did not receive a bone biopsy were followed for resolution of the ulcer without complications.

Was the selected reference standard used for all of the patients receiving the index test?

It is essential that all of the patients who received the index test also receive the same reference standard. It has been shown[7,8] that using different reference standards for positive and negative index tests results, produces an overestimation of diagnostic tests effectiveness. The terms verification bias or workup bias are used to describe the process of using different referenced standards.

Grayson, et. al.,[11] enrolled 75 patients with 76 ulcers. Bone was biopsied in 53 patients 46 of which demonstrated histological changes associated with osteomyelitis. Four patients were not biopsied but osteomyelitis was diagnosed clinically by means of radiographs, and/or evaluation the bone clinically by the surgeon during debridement. Nineteen cases were excluded from a diagnosis of osteomyelitis because of clinical follow-up. Lavery, et. al.,[12] biopsied 30 patients of the 247 patients enrolled in the study. It appears the majority of patients were followed forward in time until the wound healed or they required surgical therapy.

Were the investigators evaluating the index test and the reference standard blinded to the results?

Knowledge of the index test can bias the interpretation of the reference standard. This is termed review bias. The index test and reference standard should be **independently** evaluated. Although intuitively this makes sense and one would expect that the diagnostic results would be affected by lack of blinding, this has not been shown to alter the diagnostic results in two large studies.[7,8] The authors of these studies suggest that if the reference standard was subjective this effect would be greater.

Likelihood Ratio	Lavery, et. al.,	Grayson, et. al.,
+ LR	9.4 (6.05-15)	4.29 (1.7-11)
- LR	0.15 (0.06-0.37)	0.40 (0.26-0.61)

Table 1 Likelihood ratios with 95% confidence intervals.

Likelihood ratio values	Effect
> 10	Large and often conclusive
5 to 10	Moderate
2 to 5	Small
1 to 2	Minimal
1	No effect
0.5 to 1.0	Minimal
0.2 to 0.5	Small
0.1 to 0.2	Moderate
< 0.1	Large and often conclusive

Table 2 Interpretation of likelihood ratios.

Grayson, et. al.,[11] clearly stated in their article that the pathologist evaluating the bone biopsy was not aware of the results of the probe to bone test. It is not clear if the physicians interpreting the radiographs, debriding the wounds and following the patients were blinded to results from the probe to bone test. Lavery, et al.,[12] did not state during the methods section if anyone was blinded in the study. The podiatrists who performed the probe to bone test followed the patients who were not biopsied to evaluate their clinical status.

Was there a sample size calculation?

While it is common for randomized controlled trials to report calculations for sample size similar explanations are usually not found when reviewing diagnostic studies.[13]

Failure to calculate sample size estimates prior to beginning the study results in studies which may be too small to provide precise estimates of sensitivity and specificity. A method of how to perform sample size calculations for diagnostic studies has been advanced by Carly, et. al.[14]

Neither probe to bone study[11,12] described a sample size calculation for sensitivity and specificity. Although, neither article reported results in terms of likelihood ratios, using the information in the article point estimates with 95% confidence intervals can be calculated for LRs[15] (Table 1). LRs can be categorized from having a minimal to a large effect on pretest probabilities. [16] (Table 2)

The results for a +LR calculated from Grayson, et. al.,'s study[11] reveals a point estimate for a +LR to be 4.29 indicating a small effect.

However, using the results of the 95% confidence interval the true effect lies between 1.7 and 11 or from a minimal effect to a large conclusive effect. The range of the 95% confidence interval reveals that there is a lack of precision with this measurement likely due to the number of patients enrolled in the study. Similarly in Lavery, et. al.,'s study[12] the point estimate for the -LR is 0.15 indicating a moderate effect. However, the results of the 95% confidence interval indicate the true value could lie anywhere between 0.06 a large conclusive effect to 0.37 consistent with a small effect. Again, this indicates that too few patients were enrolled in the study to allow for accurate estimate of the diagnostic effect.

External validity

Did the authors provide information regarding evaluating and limiting variability of the index test and reference standard?

The authors of the study should clearly describe the index test used in the study to allow for replication by the reader. Failure to do this has been shown to introduce bias into the study.[7] Intraobserver and interobserver variability occurs when interpreting diagnostic tests. Experts may not agree upon the interpretation of radiographs and instruments may need to be calibrated. Differences in test interpretation may bias the results of the study in a systematic manner. The authors should describe the efforts which were made to standardize the evaluation of the index test used in the study. This may include the use of statistical measures to evaluate agreement among experts.

After reviewing the description of index test in the article the reader will need to decide whether or not the test will be reproducible in his or her practice setting. In order to implement the index test it may be necessary for additional costs, training regarding the performance and interpretation of the test.

The authors of the probe to bone studies[11,12] described in sufficient detail the index test used in the study. Neither provided any information regarding statistical measures to evaluate agreement among experts. It does not seem that any special training, cost or education would be necessary to implement the index test in an average podiatric physician's office.

Are the patients in the study similar enough to my patients?

The LR of an index test should vary with the spectrum of the patients enrolled in the study.[17] Podiatric physicians will need to consider if their practice setting is similar to the article and whether their patient under consideration would have been included in the study.

Clearly the spectrum of disease is different in the two studies evaluating the probe to bone test for osteomyelitis.[11,12] Grayson's study[11] was performed in severe limb-threatening infections at a tertiary care hospital. In contrast, Lavery, et. al.,'s study[12] was performed on a less severe population of diabetic patients in an out-patient setting. Since the majority of podiatric physicians will encounter infected diabetic ulcers in an outpatient setting rather than a tertiary care institution the results of Lavery, et. al.,'s study are more relevant for the majority of podiatric physicians.

Will the results of the study change my diagnostic approach?

In order to use likelihood ratios derived from a diagnostic study the podiatric physician should have some idea of the pretest probability of the target disorder in **their** patient population. This can be accomplished by reviewing the published literature on the subject, using the pretest probabilities in the diagnostic study, personal experience and clinical judgment. If the podiatric physician cannot determine pretest probability of the target disease it is unlikely the results of the study will provide any meaningful information.

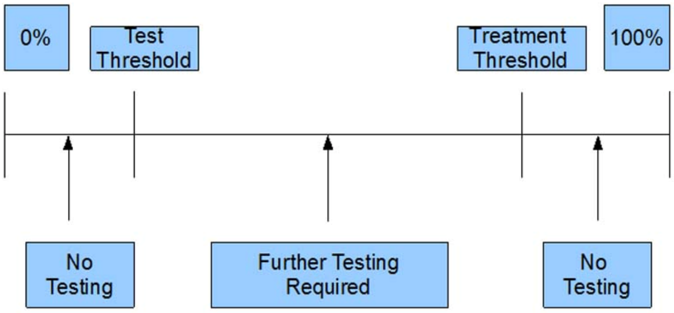

Figure 1 Probability of diagnosis.

Podiatric physicians are likely to have some process by which they arrive at a diagnosis for common conditions. This may include elements of the history and physical examination in conjunction with diagnostic studies. How should the index test be incorporated into the podiatric physician's diagnostic process? Should the index test serve as a replacement for an existing test or an addition to the normal diagnostic process? Will the results of the diagnostic test change the test/treatment threshold? If the answer to this question is no then the podiatric physician needs to recognize that additional testing will be necessary in order to pass the test/treatment threshold. (Fig 1)

Grayson's study[11] demonstrated a prevalence of 66% which was associated with a + LR of 4.4 and a -LR of 0.4. (Table1) The posttest probability with a +LR was 90% and the posttest probability with a -LR was 44%. (Table 3) This should be interpreted that a positive test is above the treatment threshold while a negative result lies within an intermediate range indicating further testing is necessary. (Fig 1) To evaluate the precision of these results we need to consider the 95% confidence intervals about the LR point estimate. The values of the 95% confidence interval for a -LR remained within the intermediate range indicating further diagnostic studies are necessary however, the lower boundary of a +LR is only 77%.
The reader will need to decide if a worst-case scenario for a positive test (77% posttest probability) is above the treatment threshold.

Lavery, et. al.,'s study[12] demonstrated a prevalence of 12% which was associated with a +LR of 9.4 and a -LR of 0.15. (Table 1) The posttest probability with a -LR was 2% while the posttest probability of a +LR was 56%. (Table 3) A positive result falls within in the intermediate range of posttest probabilities indicating that further diagnostic studies are necessary and a negative result falls below the test threshold. (Fig 1) When the values of the 95% confidence intervals are considered about the posttest probabilities it is clear that a positive test still remains within the intermediate range indicating further diagnostic studies are necessary and the 95% confidence intervals for negative test still remain below the test threshold indicating further diagnostic studies are unnecessary. The conclusion that the reader should reach is that a negative probe to bone test in severe diabetic infected ulcers at a tertiary care center and a positive probe to bone test in a less severe diabetic infected ulcer in outpatient primary care setting require additional testing and a negative probe to bone test in an infected diabetic ulcer in a primary care outpatient setting does not require additional testing. A positive probe to bone test in a severely infected diabetic foot ulcer at a tertiary care center may not be definitive for osteomyelitis.

Multiple diagnostic tests may be utilized in two different ways.[18] They may be used in a parallel fashion or a serial fashion. Parallel testing is usually performed in an emergency room or hospital setting when rapid assessment of disease processes are necessary. Typically there will be an increase in sensitivity but a decrease in specificity resulting in higher false positives. This is most useful when the podiatric physician is faced with multiple diagnostic tests of low sensitivity. Serial testing is the process whereby diagnostic tests are ordered one at a time and are dependent upon the results of the previous test. Typically this process is used in outpatient settings where there is not an urgent need to make a diagnosis. This process is useful when the diagnostic studies are expensive or risky.

Study	Posttest probability (+) LR 95% CI	Post test probability (-) LR 95% CI
Grayson, et. al.,	90% (77%-96%)	44% (34%-54%)
Lavery, et. al.,	57% (46%-67%)	2% (1%-5%)

Table 3 Posttest Probabilities with 95% confidence intervals for positive and negative likelihood ratios.[14]

The results are that the overall specificity increases but the overall sensitivity of the tests decrease. This is useful when the podiatric physician is faced with multiple tests with low specificity. With each additional bit of information obtained from the clinical examination or the results of diagnostic studies the probability of the target disease changes. The post test odds of the first test become the pretest odds of the second test. The problem with using sequential tests in the diagnostic workup of the target disorder is that if the diagnostic tests are related the additional information obtained from the second test may not provide any further information about the target disorder. For example, multiple imaging tests are often used in the evaluation of osteomyelitis in an infected diabetic ulcer: radiographs, bone scanning and magnetic resonance imaging (MRI). It is likely that these tests are not independent of each other and the diagnostic information obtained may overlap. Additional information is only gained when the diagnostic tests do not measure the same thing and are independent of each other. For example, probe to bone test and MRI do not appear to be dependent. Multiple regression analysis is used to evaluate different combination of tests to learn about the predictive value of sequential testing. Clinical prediction rules are used to determine the combination of diagnostic studies and their relevance to the target disorder. The most cited clinical prediction rule with regards to the foot and ankle literature is the Ottawa Ankle Rules.[19]

Summary

When evaluating a paper on diagnosis is important for the reader to determine if the authors used methods which would minimize bias and that the results are generalizable to their practice setting.

The spectrum of disease used in Grayson, et. al.,'s study is not as relevant to practicing podiatric physicians as Lavery, et. al.,'s disease spectrum. Although both were prospective studies, neither study clearly described consecutive enrollment of patients. Both studies used a bone biopsy as a reference test however; it was not applied to all patients in either study. Lavery, et. al., clearly indicated that the decision to perform the reference standard was not based upon the results of the index test however; Grayson, et. al., indicated that the results of the index test were used in the determination to perform the reference standard. Grayson provided information with regards to the time between the performance of the index test and reference standard. Lavery did not provide any information regarding the time difference between the index test and reference standard. Grayson, et. al., clearly stated that the pathologist reviewing the bone biopsy was blinded to the clinical information obtained. It was not clear if blinding occurred in any other aspect of either study. Neither study described a sample size calculation.

Both studies provided information describing the index test to allow for replication in practice. The test should be able to be applied in any podiatric physicians practice without any excessive amount of training or cost. Neither study reported statistical analysis of agreement between investigators regarding the test. The results of Grayson, et. al.,'s study are most generalizable to a tertiary care hospital setting. It should be viewed that a negative test in this population requires further diagnostic testing to rule out osteomyelitis. A positive test may confirm osteomyelitis however, the results are not definitive.

Lavery, et. al.,'s study is most generalizable to a podiatric physician's outpatient practice. Lavery, et. al.,'s study demonstrates that a positive test does not confirm osteomyelitis and that additional testing is necessary. The results of a negative probe to bone test in Lavery, et. al.,'s study is definitive for ruling out osteomyelitis in a diabetic ulcer.

References

1. Turlik M: Introduction to diagnostic reasoning. Foot and Ankle Online Journal, 2009.
2 . Consort statement. http://www.consort-statement.org/ Accessed 3/15/09.
3. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Moher D, Rennie D, de Vet HC, Lijmer JG: The STARD Statement for Reporting Studies of Diagnostic Accuracy: Explanation and elaboration. Ann Intern Med. 138 (1): W1 – 12, 2003.
4. Diagnostic levels of evidence http://www.cebm.net/index.aspx?o=1025 Accessed 3/15/09.
5. Gregg J, Silberstein M, Schneider T, Marks P: Sonographic and MRI evaluation of the plantar plate: A prospective study. Eur Radiol. 16 (12): 2661 – 2669, 2006.
6. Sabir N, Demirlenk S, Yagci B, Karabulut N, Cubukcu S: Clinical utility of sonography in diagnosing plantar fasciitis. J Ultrasound Medicine 24 (8): 1041 – 1048, 2005.
7. Lijmer J, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, Bossuyt PM: Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 282 (11): 1061 – 1066, 1999.
8. Rutjes AW, Reitsma J, Di NM, Smidt N, van Rijn JC, Bossuyt PM: Evidence of bias and variation in diagnostic accuracy studies. CMAJ 174: 469 – 476, 2006.
9. Brown R, Rosenberg ZS, Thornhill BA: The C sign: more specific for flatfoot deformity than subtalar coalition. Skeletal radiology 30: 84 – 87, 2001.
10. Deeks J, Altman D: Diagnostic tests 4: likelihood ratios. BMJ 329: 168 – 169, 2004.
11. Grayson M, Gibbons G, Balogh KE, Levin E, Karchmer A: Probing to bone in infected pedal ulcers. A clinical sign of underlying osteomyelitis in diabetic patients JAMA 273: 721 – 723, 1995.
12. Lavery L, Armstrong D, Peters E, Lipsky B: Probe-to-bone test for diagnosing diabetic foot osteomyelitis. Diabetes Care 30: 270 – 274, 2007.
13. Rutten F, Moons K, Hoes A: Commentary: Improving the quality and clinical relevance of diagnostic studies BMJ 332: 1129, 2006.
14. Carley S, Dosman S, Jones S, Harrison M: Simple nomograms to calculate sample size in diagnostic studies. Emerg Med J 22: 180 – 181, 2005.
15. Likelihood Ratio Calculator http://araw.mede.uic.edu/cgi-alansz/testcalc.pl Accessed 3/8/09.
16. Interpretation of likelihood ratios. http://www.poems.msu.edu/InfoMastery/Diagnosis/likelihood_ratios.htm Accessed 3/21/2009.
17. Furukawa T, Strauss S, Bucher HC, Guyatt G: Diagnostic Tests. In: Guyatt G, Drummond R, Meade MO (eds). *Users guides' to the medical literature.*New York: McGraw-Hill, 419 – 438, 2008.
18. Fletcher R, Fletcher S: *Clinical Epidemiology the Essentials*. 4th ed. Philadelphia Pennsylvania: Lippincott Williams and Wilkins, 2005.
19. 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. 4: 384 – 390, 1992.