How to evaluate a diagnostic test

Steven E. Eckert, DDS, MS,[a] Gary R. Goldstein, DDS,[b] and Sreenivas Koka, DDS, MS, PhD[c]

Tests are used in dentistry to establish, confirm, or reject the clinical impression of a diagnosis. Not all tests are equal in their ability to establish a diagnosis, with some tests demonstrating positive results when no disease is present (false positive) or negative results when disease is present (false negative). Using simple mathematical computations, it is possible to determine the extent to which a test can reliably establish the presence or absence of disease. This article describes the concepts of sensitivity, specificity, positive predictive value, negative predictive value, and likelihood ratios through the evaluation of a clinically relevant paper on vital staining for oral carcinoma. This article also describes methods for literature evaluation to determine whether a test conveys meaningful diagnostic information.

Medical and dental therapy is provided in an effort to address a specific disease entity. The more specific the diagnosis of that disease, the more predictable the interventive therapy. Thus, precision and accuracy in diagnosis are of paramount importance to the patient and the clinician.

How is a diagnosis established? A patient may have frank dental caries, but it must be remembered that this is already a diagnosis rather than a simple presentation. The diagnosis is based on evidence that is accumulated and analyzed during the examination of the patient. Evidence may take the form of changes in tooth color, softening of tooth surface, or actual cavitation of the tooth. Clinical experience may make the process appear to be simple, almost intuitive, but it is recognized that many factors are considered before the diagnosis is reached. Although this example requires few sophisticated tools or tests, other diagnoses may not be as obvious.

Diagnostic tests become more valuable as the number of potential diseases in the list of differential diagnoses increases. Consider a clinical observation of a white lesion on the buccal mucosa. This description may be associated with a spectrum of diagnoses, ranging from a variation of normal anatomy to the presence of a malignant neoplasm. The clinician clearly needs more information to establish a diagnosis. The history and physical examination provide much data, but definitive tests may be required to reach specific conclusions. Radiographs, cytology, biopsy, needle aspiration, and the patient's body temperature may all be used to assist in the establishment of a diagnosis.

Choosing the appropriate diagnostic tests may provide a more accurate diagnosis earlier in the patient's episode of care and at a lower cost than if inappropriate tests are performed. Methods used to assist the clinician in the identification of the correct test will be of great benefit to all involved. Given the current state of Internet access, literature searches may be performed to determine sources of information on different tests that may yield conclusive answers.

CLINICAL SCENARIO

Imagine that 2 patients present to a practice with a white lesion on the buccal mucosa. The first is a 33-year-old woman. This patient is in good health, exercises daily, and avoids the use of alcohol and tobacco products. The second patient is a 75-year-old retired man with a history of type 2 diabetes. This gentleman successfully underwent a smoking cessation program approximately 10 years ago. Clinically, their lesions appear similar. However, the clinician would prefer not to biopsy either lesion and wonders if there are any noninvasive methods of determining the presence or absence of malignancy. To find an answer, the clinician elects to perform a literature search to determine whether a screening evaluation that may eliminate the need for incisional biopsy exists.

LITERATURE SEARCH

The choice of key words in the literature search limits the number of "hits" encountered in the search. For example, a search of "oral cancer" provides 1919 potential references, whereas the addition of the 1 word "detection" narrows the focus of the search to 713 references. Because the clinician wishes to consider alternatives to biopsy, the search may be further limited by adding the terms "vital staining" to this search. Through a series of searches, each more limited than the previous, the clinician is able to filter out references that are not germane.

Using this search method, a list of 14 articles is established. A review of the titles of these articles allows the clinician to further limit the search until only 1 or 2 articles are considered. In this instance, an article entitled "Toludine blue test for detection of carcinoma

[a] Assistant Professor, Mayo Graduate School of Medicine, Mayo Clinic, Rochester, Minn.

[b] Professor and Director Advanced Education Program in Prosthodontics, College of Dentistry, New York University, New York, N.Y.

[c] Associate Professor, College of Dentistry, University of Nebraska Medical Center, Lincoln, Neb.

of the oral cavity: An evaluation," seems to provide valuable material. However, this article was published in 1972, and the abstract is unavailable.[1] Retrieval of the article through the medical library allows the clinician to determine its true applicability to these specific patients.

The article states that toludine blue, an acidophilic, metachromatic, nuclear stain, has a preferential affinity for carcinoma cells in contrast to normal mucosal cells. Application of toludine blue to suspect tissue results in stain uptake when carcinoma exists. The article evaluates 1190 patients with suspicious lesions within the oral cavity to determine positive and negative results. Biopsy was performed to confirm the clinical impression after the vital tissue stains.

PERTINENT QUESTIONS

Once a pertinent article is reviewed, the clinician is challenged whether or not to "believe" the results of the study. Jaeschke et al[2] suggested that the clinician attempt to answer questions that will help determine how "powerful" the data from the study are, and in turn, how useful they might be in providing care to their patient. The questions relate to validity, results, and meaningfulness of the study.

Are the results of the study valid?

When a diagnostic test is suggested, the results of the test must be true, pertinent to the group of patients in which the test will be used, understood well enough to be performed appropriately, and meaningful to the specific patient for whom the test is used. Before applying a new test, the clinician should determine the value and relevance of that test for the clinician's own practice.

Was there an independent, blind comparison with a reference standard?

For a diagnostic test to be valid, it must be compared with an accepted reference standard. If the clinician cannot accept the standard, then the usefulness of the test is minimized. For the toluidine blue study, the reference standard is a biopsy that is considered the "gold standard" in medicine. Another question the clinician must ask is, if the "gold standard" exists, why replace it? Again, the clinical scenario offers a valid answer. If the new test is less expensive and less invasive, it offers major advantages to the patient. However, the reference standard is often controversial or, in evolving technologies, absent. The reader must understand that lack of a "gold standard" does not obviate the potential of the test. It only means that there is a heavier burden of proof placed on the investigator and that clinicians may need to wait for more confirming studies before adopting the test into their practice.

In the toluidine blue (TB) study,[1] the reader is not informed whether the examiners were "blinded," namely, were the individuals making the histologic diagnosis aware of the results of the staining procedure? Another important concern relates to calibration of examiners. The reader is informed that a number of new clinicians were given an initial training on the design of the study and that they were supervised initially by a more senior colleague. However, the degree to which any one examiner's interpretation of stain results would agree with another examiner's (interexaminer reliability) or whether the same examiner would offer the same assessment on the same stain at 2 times (intraexaminer reliability) is unknown. This information is important in that not knowing to what degree the results of the study are reproducible renders the reader less confident of the results.

To determine whether one can believe the data, the research and data analysis methods need to be assessed. Was the data procured and then analyzed with the intent of minimizing potential bias on the part of the researchers? Two primary guides by Jaeschke et al[2] permit a clinician to evaluate the validity of study results. These guidelines discuss the subject population and the description of methods used in performance and evaluation of the test.

How does the subject population of the study compare with the patients that make up my practice?

If patient populations are similar to the population of the study and the inclusion and exclusion criteria are acceptable, then the reader can be confident that the results of the test will be similar in the study population and in the reader's patient population. If the clinician's patients are not similar, then caution and judgement are mandated.

Were the methods for performing the test described in sufficient detail to permit replication?

A diagnostic test is useless if one does not understand how to use it. It is the author's obligation to provide a detailed description of how the test was performed, how the patient was managed, and how any necessary "hardware" was used. This becomes especially important because subsequent treatment decisions may be made based on the findings.

What are the results of the study?

Study results define the usefulness of the diagnostic test. Clinicians are concerned with the ability of the test to aid in correctly establishing the presence or absence of a malady in a patient. No test is absolutely accurate. Fortunately, there are a number of mathematical methods that define the "diagnostic quality" of a test. These qualitative methods and definitions must be understood by the clinician because they may assist in the

Table I. Guides to calculating sensitivity, specificity, PPV, NPV, and accuracy

Test results	Disease present	Disease absent
Disease present	True-positive (a)	False-positive (b)
Disease absent	False-negative (c)	True-negative (d)

Sensitivit = a/(a + c).
Specificity = d/(b + d).
Positive predictive value (PPV) = a/(a + b).
Negative predictive value (NPV) = d/(c + d).
Accuracy = (a + d)/(a + b + c + d).

Table II. Calculation of sensitivity, specificity, PPV, and NPV for toludine blue staining

	Carcinoma	Other than carcinoma	Total
Positive stain	415 (a)	131 (b)	546 (a + b)
Negative stain	66 (c)	418 (d)	484 (c + d)
Total	481 (a + c)	549 (b + d)	1030

Sensitivity = a/(a + c) = 415/481 = 0.86 (86%).
Specificity = d/(b + d) = 418/549 = 0.79 (76%).
PPV = a/(a + b) = 415/546 = 0.76.
NPV = d/(c + d) = 418/484 = 0.86.

determination of the value of inclusion of the test in the diagnostic workup. The primary methods for determination of the diagnostic quality of a test are sensitivity, specificity, positive value, and likelihood ratios.

SENSITIVITY AND SPECIFICITY

Using Tables I and II, toludine blue had a sensitivity (true positive / [true positive + false negative]) of 86% and a specificity (true negative / [false positive + true negative]) of 76% in determining the presence of squamous cell carcinoma in the oral cavity of patients enrolled in this study. Sensitivity is the proportion of patients with disease that were identified by the test. When sensitivity is high, there is a high true positive rate and a low rate of false negative. Hence, a negative result to a highly sensitive test is a relatively reliable indicator that disease is absent. Specificity is the proportion of patients without disease who were identified by the test. A high specificity means that there is a high number of true negative results and few false positive results. When a test with a high specificity results in a positive response, there is a good chance that disease is present because false positives are rare.[3]

Although sensitivity and specificity values offer a simplified method of assessing the usefulness of a diagnostic test, they are not foolproof. Table I reveals that sensitivity (a / [a+c]) and specificity (b / [b+d]) are calculated using the vertical nature of the table. Sensitivity can be interpreted to mean when the disease is present, how often is a positive test result obtained; specificity can be interpreted to mean when a disease is not present, how often is a negative test result obtained. These measures are useful when the prevalence of a disease, which can be defined as "the proportion of individuals in a population having a disease," is relatively high. When disease prevalence is relatively low, sensitivity and specificity are less useful. However, 2 alternative measures, "positive predictive value (PPV)" and "negative predictive value (NPV)" may be more useful when disease prevalence is low. PPV (true positive / [true positive + false positive]) and NPV (true negative / [false negative + true negative]) are calculated using the horizontal rows of Table I (PPV = a / [a+b] and NPV = d / [c+d]). PPV can be interpreted to mean when a test result is positive, how often is the disease present; NPV can be interpreted to mean when a test is negative, how often is the disease absent. In the toludine study (Tables I and II), the PPV is 0.76 and the NPV is 0.86. To their advantage, positive and negative predictive values will change as disease prevalence varies, and therefore they are potentially more meaningful to the discrimination capacity of a test when a disease has low prevalence in the general population, than when a disease has a high prevalence in the same population.[4] However, the environment in which test results are obtained may diminish the usefulness of PPVs. For example, it appears that PPVs calculated from diagnostic tests administered in the private practice setting are less reliable than when calculated from tests administered in university hospitals.

The toludine blue study demonstrated a number of patients in whom the test was inconclusive. These patients were described as exhibiting "doubtful" results. When sensitivity and specificity are evaluated a 2-by-2 table must be created. However, this article showed data that does not fit discretely into such a table because the inconclusive (doubtful) test result creates another layer of diagnostic classification, namely, there are 3 test results: positive, negative, and doubtful. Another example of such a situation is periodontitis. In dentistry, we might prefer to have periodontal disease classified as mild, moderate, or advanced, again because the therapy for each scenario is different.

LIKELIHOOD RATIOS

The measures of sensitivity, specificity, PPV and NPV require data to be organized into the 2-by-2 table format described above and presented in Tables I and II. Data that do not fit into this 2-by-2 format necessitate a different form of analysis. Jaeschke et al[2] propose that calculation of likelihood ratios (LRs) offers a solution to this problem, in that data of any format, 2-by-2 or 2-by-3 or 2-by-4, can be used. In fact, the authors suggest that the use of LRs offers the clinician a better method of determining whether a positive or negative finding from diagnostic test is meaningful. In essence,

Table III. Calculation of likelihood ratios for toludine blue staining

	Carcinoma	Other than carcinoma	Total
Positive stain	415 (a)	131 (b)	546 (a + b)
Negative stain	66 (c)	418 (d)	484 (c + d)
Doubtful stain	54 (e)	106 (f)	160 (e + f)
Total	535 (a + c + e)	655 (b + d + f)	1190

LR for a positive stain = (a/[a + c + e])/(b/[b + d + f]) = (415/535)/(131/655) = 3.88.

LR for a negative stain = (c/[a + c + e])/(d/[b + d + f]) = (66/535)/(418/655) = 0.19.

LR for a doubful stain = (e/[a + c + e])/(f/[b + d + f]) = (54/535)/(106/655) = 0.62.

Table IV. Guide to interpreting likelihood ratios

Likelihood ratios	Response
Greater than 10 or less than 0.1	Permit a conclusive shift in pretest probability to posttest probability.
Between 5 and 10 and between 0.1 and 0.2	Permit a moderate change in the pretest probability and the posttest probability.
Between 2 and 5 and between 0.5 and 0.2	Lend to only a small shift in probability.
Between 2 and 0.5	Alter the pretest probability to minor (and probably unimportant) degree.

an LR is a measure of how likely any one particular diagnostic finding is to occur in the presence or absence of a disease. Conceptually, the LR represents the ratio between how likely a given test result will occur when the target disorder is present (true-positive) to how likely the same test result will occur in cases in which the target disorder is absent (false-positive).

In regard to the TB study,[1] Table III presents the complete set of data (positive, negative, and doubtful) presented by the authors. The LR for a positive test result is calculated by answering 2 questions. First, how likely is a positive toludine blue stain among lesions that are cancerous? Table III shows that of 535 cancerous lesions, 415 tested positive (415 / 535 = 0.775). Next, how likely is a positive toludine blue stain in lesions that are not malignant? Again, from Table III, 655 lesions were nonmalignant, yet 131 tested positive (131 / 655 = 0.200). The ratio of these 2 likelihoods is the LR for a positive stain and is equal to 0.775 divided by 0.200 or 3.88.

One can calculate LRs for any level of a diagnostic test regardless of the number of levels. In this case, there are 2 other levels of stain, namely, "negative" and "doubtful." Each calculation involves determining a ratio of the likelihood of achieving the particular diagnostic test classification in lesions that are cancerous and the likelihood of achieving the same test diagnostic classification in lesions that are not cancerous (Table III). For a negative stain, the likelihood of achieving a negative test in patients who have cancer is 66 / 535 (0.123) and the likelihood of achieving a negative test in patients who do not have cancer is 418 / 655 (0.638), with an overall LR for achieving a negative stain of 0.19. For a doubtful stain, the likelihood of achieving a doubtful test result in patients who have cancer is 54 / 535 (0.101) and the likelihood of achieving a doubtful test result in subjects who do not have cancer is 106 / 655 (0.162) and the LR for a doubtful stain is 0.62. These ratios actually reveal the possibility that a particular test result comes from someone who actually has the disease for which the test was ordered. LRs allow us to see the relative diagnostic value of the different classifications of a test.

Having calculated LRs, how are the values used to assess the usefulness of any one of the diagnostic findings (positive, negative, or doubtful) as it relates to the presence or absence of a cancerous lesion? One way to interpret an LR is to think of an LR of 1 as meaning that the probability that a patient has a disease for which he or she is being tested, is the same before the test (pretest probability) as after the test (posttest probability). The patient is no better off after the test than before the test, and the clinician is no closer to a diagnosis than before the test was administered. One could certainly question the usefulness of this test for diagnostic purposes. Diagnostic tests that have LRs greater than 1, increase the probability that the target disorder is present and tests with LRs less than 1 decrease the probability that the target disorder is present. For the negative and doubtful findings in the TB study, both LRs are less than 1. In interpreting the negative and doubtful LRs less than 1, the probability that the patient had cancer before the test (pretest probability) is reduced after the TB test results (posttest probability).

When considering the LRs calculated for the TB test, the LR for a positive finding is 3.88. Another way of interpreting this value is to infer that a positive stain is approximately 3.9 times more likely to occur in a cancerous lesion than it is likely to occur in a lesion that is noncancerous. The LR for a negative stain is 0.19 indicating that a noncancerous lesion is approximately 5 times more likely to occur in a noncancerous lesion than it is likely to occur in a cancerous lesion. A negative stain is a better indicator of a noncancerous lesion than a positive stain is an indicator of a cancerous lesion. The LR for a doubtful stain is 0.62, which shows that a doubtful test result does not permit the clinician to alter the pretest probability of the presence of a cancerous lesion (Table IV).

To fully appreciate the use of diagnostic tests and LRs, the clinician must understand the concept of pretest and posttest probabilities. The concept is known to dentists, but the terminology may not be readily recognized. The knowledge that a patient may have a certain malady is derived from making clinical assessments of signs and symptoms. The magnitude of the signs and symptoms, and the clinician's own experience level, formulate a "probability" that the malady exists. In many situations, the greater the probability of the malady, the more likely the clinician will order a diagnostic test to confirm, or exclude the malady. The pretest probability of various diseases is specific for each disease, specific for the signs and symptoms of that disease, and specific to the setting in which it is found.[10]

For instance, a patient who presents with tooth pain in the maxillary right quadrant is clinically assessed to isolate the right maxillary teeth with large restorations, gross caries, and sensitivity to percussion and temperature. As multiple clinical findings become positive for a particular tooth, the probability becomes greater that that particular tooth is the culprit. This high pretest probability is what warrants the subsequent diagnostic test of a periapical radiograph. The clinician may choose to make radiographs of the entire maxillary quadrant, but the pretest probability is highest for a particular tooth. If the radiograph of the symptomatic tooth reveals a periapical radiolucency, the posttest probability is even higher than the pretest probability that this tooth is causing the patient's pain.

When considering the use of TB as a diagnostic tool for oral cancer diagnosis, the pretest probability of the nonsmoking, nonalcohol-consuming young woman is significantly less than the pretest probability of the older, alcohol-consuming, male smoker. Our clinical experience and previous dental literature offers the clinical acumen to estimate pretest probabilities in the low-to-high range, and one may even be able to estimate a numeric probability. Medicine, with more diagnostic tests at its disposal and more demographic data on its diseases, has calculated pretest probabilities for numerous diseases and these values have been published in journals and on the Internet. Having calculated or estimated pretest probability values are necessary to implement LRs in diagnoses, and Jaeschke et al[2] recommend a guide to interpreting LRs as they relate to modifying pretest probabilities to achieve a posttest probability of the target disorder (Table IV). This guide assists the clinician in determining how the magnitude of various LRs should or should not influence clinical decision making.

It is unlikely that any test or combination of tests will unequivocally determine a diagnosis with 100% certainty. The clinician invariably is challenged to draw upon diagnostic information from a variety of sources to decide, in the end, what the probability is of any given diagnosis. For the patient with oral cancer, many other sources are available, for example, the patient's medical and dental histories, family history, smoking status and history, age, occupational circumstances, impressions drawn from the clinician's intraoral examination, radiographic examination, biases from the clinician's training, and the clinician's prior patient experiences. By drawing upon information from all sources, the clinician generates differential diagnoses that represent those diseases / disorders with the highest probability of being present.

It is not the responsibility of the reader to perform mathematical calculations. This task falls to the author(s) of the article. Authors should perform the appropriate computations to allow the clinician to make sound judgements regarding test applicability.

Will the results help me in caring for my patients? Are the results applicable to my patient?

A diagnostic test will only be useful to the degree that it differentiates between different stages of disease / disorder in a population that is relevant to the reader's clinical practice. If the sample population is poorly representative of the panorama of patients seen in the reader's practice, then the results of the study become significantly less meaningful. In the TB study,[1] a relatively large number of patients (1190) made up the sample population. However, additional pertinent information regarding the demographics of the patients is not presented. How many subjects are male and how many are female? What was the mean age of the subjects? Questions such as these and any others that might impact on meaningfulness of results should be considered. To use an extreme example, if 90% of the study population was male subjects over the age of 70, how meaningful are the study results for a young female patient who has a suspected lesion?

Will my patient be better off as a result of the test?

The answer to this question depends on the test. If it is a quick, inexpensive, noninvasive discovery test with a high sensitivity, the answer is yes if an accepted exclusion or confirmation test is available. If a subsequent test is not available, or not possible, then the answer is not as clear. For example, if one uses a staining test for caries, one is not able to do histopathologic confirmation of the finding. Confidence in the test is mandatory, because a false-positive would cause the removal of sound dentin and a false-negative would allow residual caries to remain. In the proposed clinical scenario, an accepted "gold standard" exists. Unfortunately, incisional biopsy is a more invasive, painful, and costly procedure. Also, because the possibility exists that the area biopsied may contain nonmalignant tissue and may miss malignant cells that are present in another area of the lesion, some surgeons prefer to strip the entire lesion in high pretest probability patients.[4] The ultimate value of a test is whether it adds information not readily available and whether that information alters the patient management in a beneficial manner.[2]

SUMMARY

Establishment of an accurate diagnosis allows the clinician to develop treatment plans that, based on the best available evidence, most appropriately meet the patient's needs and desires. Often a diagnosis is established using clinical findings only, but in other situations tests must be performed to aid in determination of a diagnosis. When diagnostic tests are indicated, it is imperative that the clinician understands the value of test results. False results, either positive or negative, make it impossible to rely solely on test results. The clinician must understand the likelihood that specific results truly establish or confirm a diagnosis and must be able to make sound decisions based on risk factors once the results are obtained.

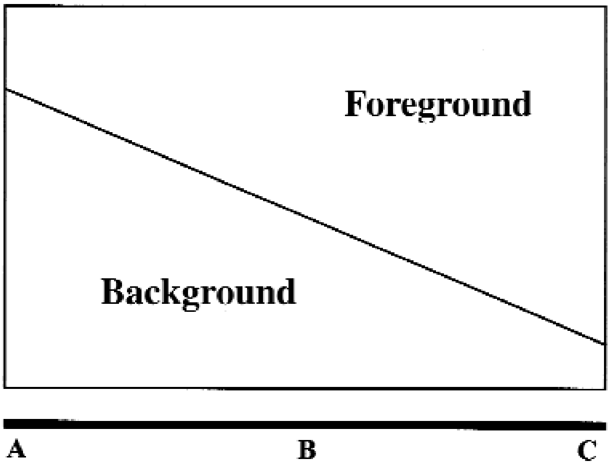
REFERENCES

1. Vahidy NA, Zaidi SH, Jafarey NA. Toludine blue test for detection of carcinoma of the oral cavity: an evaluation. J Surg Oncol 1972;4:434-8.
2. Jaeschke R, Guyatt GH, Sackett DL. Users' guides 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;271:703-7.
3. Brunette DM. Critical thinking: understanding and evaluating dental research. Chicago: Quintessence Books; 1996. p. 99-112.
4. Sackett DL, Richardson WS, Rosenberg W, Haynes RB. Evidence-based medicine: how to teach and practice EBM. New York: Churchill Livingstone; 1997. p. 58, 118.

doi:10.1067/mpr.2000.106734

Correction

In the January issue of the *Journal*, the article by Alan B. Carr and Glen P. McGivney, entitled "Users' Guide To the Dental Literature: How To Get Started," (volume 83, pages 15-20), Figure 1 should be as shown here. The rectangle represents the universe of potentially relevant clinical knowledge, and the diagonal denotes the rough division between background and foreground knowledge. On the scale beneath are 3 levels of experience. "A" is a learner with little clinical knowledge or experience, whose needs are largely on the background type portrayed by the vertical dimension of the rectangle. "B" has increased knowledge and experience, and the needs are more evenly divided. "C" has extensive knowledge and experience; the majority of knowledge needs would be foreground. In this diagram, please notice the diagonal is placed to show that clinicians are never too inexperienced to ask foreground questions (we cannot let out learners off so easily!), or too experienced to ask background questions. It is the condition of the patient that determines the knowledge needs. Clinicians may be at "C" for frequently encountered problems, at "B" for occasional problems, and at "A" for new disorders or those outside their special area of interest.


Foreground
Background
A
B
C