THE USE OF DIAGNOSTIC DATA IN CLINICAL DENTAL PRACTICE

Carol Oakley, DDS, MSc, PhD,
and Donald Maxwell Brunette, MSc, PhD

"If it looks like a duck, quacks and waddles like a duck . . . then it probably is a duck!" and "if you hear hoof beats, think of horses not zebras" (unless, of course, you are on the plains of the Serengeti). At first glance, these adages may seem irrelevant to the diagnostic process in clinical dental practice. These adages, however, respectively illustrate the principle of pattern recognition and the effect of prevalence, both of which are important aspects of the diagnostic process.

This article presents the dentist in clinical practice with an evidence-based approach to diagnostic data and tests so that the reader can become a more discriminating user of tests offered by the medical profession and, increasingly, by the pharmaceutic industry for promotional purposes.

This article reviews a few basic principles of biostatistics, discusses test design and test characteristics, and demonstrates how to identify a *good* test and the circumstances in which a test will be useful in the clinical setting. For ease of discussion, this article focuses on dichotomous data that are divided into mutually exclusive categories: positive or negative. Data are presented from the dental literature, and clinical dental examples are used. Texts providing more detailed, comprehensive information regarding biostatistics, clinical epidemiology, and related topics are listed with the references.[4, 22, 29] Much of the following discussion has been summarized from these sources.

From the Department of Oral Biological and Medical Sciences, Faculty of Dentistry, University of British Columbia, Vancouver, British Columbia, Canada

THE DIAGNOSTIC PROCESS

Most dentists have had their height, weight, and blood pressure measured in a physician's office. They may have had blood drawn for complete blood cell count and differential blood series or testing for cholesterol levels, prostate surface antigen, blood glucose, or thyroid hormone levels. They may have undergone tuberculin skin tests, mammography, electrocardiograms, and cardiac stress tests or had suspicious moles removed for histologic examination. They may even have sought the convenience of home pregnancy tests. As dentists, clinicians have probably prescribed dental radiographs and used explorers and periodontal probes to detect caries, defective restorative margins, and periodontal attachment loss. They may have applied electric pulp testers or ice to teeth to determine their vitality. They may have used toluidine blue dye to aid in selecting sites for biopsy of suspicious oral lesions. They may have recorded mandibular excursions, palpated muscles of mastication, and listened for temporomandibular (TM) joint sounds.

As consumers and providers of health care, dentists reasonably expect that the information obtained from diagnostic investigations is reliable and truthful. Moreover, it is generally assumed that the information obtained from these investigations will provide a diagnosis as to the presence or absence of an abnormality or disease and that the diagnosis will direct a subsequent course of management or treatment. The question remains, however: how can patients and clinicians know if the data and subsequent diagnosis are correct?

Beck[2] maintains that dentistry, in contrast to medicine, has deemphasized diagnostic activities and merged them with treatment-planning activities. Nevertheless, the aim of a medical or dental clinician is to arrive at a diagnosis that may direct a subsequent course of management. The diagnostic process is initiated by the patient history and symptoms and is followed by the clinical examination, during which the clinician perceives signs that are manifestations of the disorders. The clinician may also use assays or measurements that are traditionally referred to as diagnostic tests or tools. In reality, symptoms, signs, and assays may all be considered diagnostic tools, because all are sources of information used to generate a diagnosis.[2]

Sacket et al[29] explain that patients, clinicians, and researchers generally agree that the presence of disease indicates a derangement in anatomy, biochemistry, physiology, or psychology. They less often agree, however, on the exact criteria that define the condition that is the target of the diagnostic process.

Wulff[40] distinguished two major principles of disease: (1) the nominalistic or patient-oriented principle, and (2) the essentialistic principle that emphasizes disease as an independent entity. In the nominalistic approach, disease does not exist as an independent entity, and disease classification is really a classification of sick people or patients. Thus, a particular disease is defined by a group of characteristics that occur more often in persons with the disease than in other people. Patients

will have a pattern of similar symptoms and signs, and their prognosis and treatment will have some common features. The nominalistic principle does not require a definition of normality and recognizes that definitions of disease may vary among different societies.[40]

The essentialistic view[40] is closely related to a modern principle of disease termed biochemical fundamentalism.[6] This view is based on the idea that disease can be described in terms of biochemistry and molecular biology. Diseases are assumed to follow regular patterns, and once the underlying biochemical events are understood, the course of the disease can theoretically be predicted. Hence, disease classification becomes a matter of biotechnology, and the need for defining a normal state is avoided by relying upon statistical terms to define the disease state. That is, disease is defined by the distribution of certain features in a particular population and the extent to which that distribution differs from a similar assessment of a group the investigators consider not diseased.[6, 40] This statistical approach forms the basis for using biomarkers as diagnostic or screening tests.

Contemporary clinical medical and dental practice is still an art and a science. Overall, the nominalistic approach may offer a more realistic strategy for coping successfully with the varying manifestations of conditions such as coronary heart disease and temporomandibular disorders (TMD) that can be defined in both essentialistic and nominalistic terms.[24, 29]

DIAGNOSTIC DECISION ANALYSIS

The use of diagnostic data and tests can be considered at three levels: screening, confirmatory, and exclusionary.[13, 29] The objective of screening procedures is the early detection of disease, before symptoms associated with the disease are apparent. Thus, screening tests are conducted on individuals who do not have symptoms associated with the condition for which screening is being conducted. Screening tests classify individuals with respect to their likelihood of having a particular disease, but they do not diagnose disease. Individuals whose screening tests are positive require further evaluation by subsequent tests to rule in or to rule out the presence of the disease.[13, 29]

The use and interpretation of diagnostic data, including signs, symptoms, and diagnostic tests, are based on the four principles of decision analysis[29, 31]:

1. Clinicians should not consider patients as absolutely having a disease but rather as having only the probability of disease. The probability of disease is based on the prevalence of the disease, the patient's history (including risk factors, symptoms, signs, and previous tests), and the clinician's previous experience with similar situations.
2. Clinicians use diagnostic tests to improve their estimates of the

probability of disease, and the estimate of probability following the test may be lower or higher than the estimate of probability before the test. Tests should be selected by their ability or power to revise the initial probability of disease.

3. The probability that disease is actually present, following a positive or negative test result, should be calculated before the test is performed. Application of this principle results in fewer useless tests being performed.
4. A diagnostic test should revise the initial probability of disease. If the revision in the probability of disease does not alter the planned course of management or treatment, then the use of the test should be reconsidered. Unless the test provides information desired for an unrelated problem, tests that will not alter the planned course of management or treatment should not be performed.

Principle 1 states that in the diagnostic context, patients do not have a disease; rather, patients have a probability or likelihood of disease. At the outset, the clinician may assign to the patient a probability of disease that reflects the clinician's level of confidence that the target disease is actually present. This initial probability may be based on the prevalence (see box) of the disease in the population and may be revised, upwards or downwards, based upon the patient's history, symptoms, signs obtained from the clinical examination, previous tests, and the clinician's previous clinical experience with similar situations. If the patient is known to have one or more risk factors for a certain disease, the probability of disease may be increased. Thus, a pretest probability, risk, or likelihood of disease is assigned. Diagnostic tests may then be considered to revise the pretest probability, as per principle 2. That is, by themselves, the measurements, assays or test results do not reflect 100% certainty as to the presence or absence of the disease. Instead, the test results, either positive or negative, are used to revise, upwards or downwards, the initial pretest probability of disease. Moreover, once a test has been carried out, the clinician and patient must accept and deal with the results. That is, the decision that a test provides useful information is independent of the actual result. If the clinician picks and chooses which test results to accept or discard, the clinician opens the door to personal bias and preconceived notions, undermining the principle of objective testing.

On completion of the clinical examination, and before further investigations are considered, the clinician may be confident that a particular disease or condition really is present. In that instance, there is no need for further investigations or tests, and management appropriate for the condition should commence without delay. Likewise, if the clinician is confident that a particular disease is not present, further investigation or treatment of that disease is not warranted. These decisions are based on the threshold approach in decision analysis, shown in Figure 1. For each condition or disease, the clinician sets a threshold for testing known

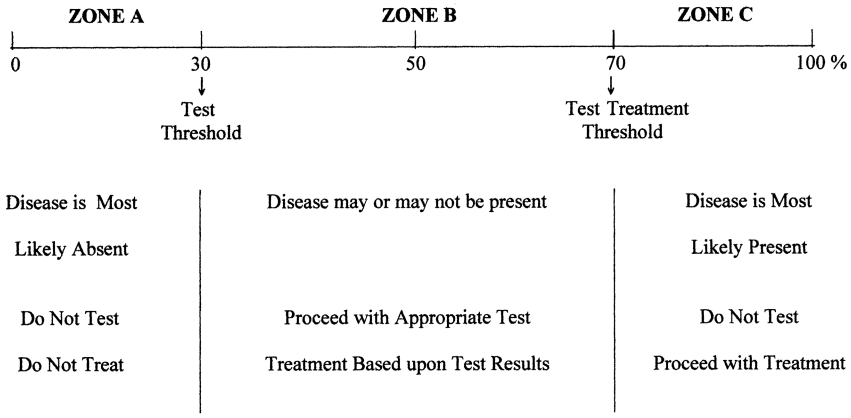

Figure 1. Threshold approach to decision analysis: Examples of threshold approach for disease of pulpal pathology and test of periapical radiograph. *ZONE A*, A patient complains of sensitivity to cold and sweet stimuli. These symptoms are localized to an unrestored tooth with no known history of trauma but with visible cervical abrasion and root recession. Pulpal pathology is most likely absent because root sensitivity caused by exposed dentin is the most probable diagnosis. A radiograph would not be warranted because information obtained from the radiograph would not alter the diagnosis or further management. *ZONE B*, A patient with a poorly maintained dentition describes intermittent and increasing sensitivity to cold and sweet stimuli and occasional spontaneous discomfort lasting over an hour and requiring analgesics for relief. These symptoms are associated with a heavily restored tooth with subgingival restorative margins. Recurrent caries or pulpitis may be present. A radiograph is warranted because it may provide useful information for diagnosis and further management. *ZONE C*, A patient describes severe pain with biting pressure and denies sensation to cold stimuli. These symptoms are localized to a molar with visible gross caries. Radiographs are not required for the clinician to arrive at the diagnoses of caries and a nonvital pulp; however, a periapical radiograph is indicated to guide prognosis and further treatment, such as endodontic therapy or extraction.

as the test threshold and a second threshold for treatment known as the test-treatment threshold.[29] In general, these cutoff threshold probabilities for ruling in or ruling out a disease depend on the particular disease and the subsequent courses of action or follow-up that relate to either ruling in or ruling out the disease. That is, the consequences of false-positive and false-negative results must be weighed in each case. If a test is not powerful enough to alter the pretest probabilities so that a positive or negative test result will not alter the pretest planned course of action, the test should not be performed.[29, 31] The strategies for defining specific test and test-treatment threshold cutoffs are discussed in greater detail by Sacket et al.[29]

Three clinical decisions are depicted in Figure 1. In the first instance, the pretest probability of a disease is below the test threshold (Zone A in Fig. 1). The patient is unlikely to have the disease, and even a positive test result would not alter the posttest probability to a level that would justify treatment. Therefore, neither treatment of the disorder nor further testing for the disorder should proceed. For example, multiple yellowish

spots and plaques are observed bilaterally on the posterior buccal mucosa of an elderly male patient. The spots and plaques cannot be removed with gentle wiping of a gauze across the mucosal surface. The clinician is confident that Fordyce granules are present and that no pathologic condition is present. Therefore, further investigations such as biopsy or further management or treatment are not indicated.

In similar fashion, if the pretest likelihood of disease exceeds the test-treatment threshold (zone C in Figure 1), treatment should proceed without further diagnostic testing. For example, soft white plaques resembling milk curds are observed on the palate and buccal mucosa of an elderly male patient. The plaques may be stripped from the tissue, leaving an intensely erythematous surface with localized bleeding. Oral thrush (candidiasis) is most likely present, and further investigation such as biopsy will not alter the diagnosis or the probable management with antifungal medications.

When the pretest probability falls in between the test and test-treatment thresholds, however (zone B in Figure 1), testing is indicated, and treatment should proceed on the basis of the test results. In general, a diagnostic test is most useful when the pretest probabilities fall between roughly 30% and 70%.[5, 20, 21] For example, an adherent white plaque is observed on the anterior floor of the mouth and ventral left lateral tongue of an elderly adult male. A pathologic condition may or may not be present. Further investigation such as biopsy is indicated to establish a diagnosis and to direct further management.

MEASUREMENT RELIABILITY

Measurement reliability refers to the ability to obtain the same measurement consistently over sequential measures. The reliability of a measurement may be affected by three sources of variability: (1) the system or phenomenon being examined, (2) the examination itself, such as the instruments or equipment used and the examination environment, and (3) the examiners.[4, 29]

Variation in the System or Phenomenon Being Measured

Normal biologic variability may be inherent in the phenomenon being measured. For example, blood pressure and pulse fluctuate throughout the day and under different circumstances such as stress, exercise, and body position; hormonal levels fluctuate with the diurnal and menstrual cycles. Moreover, the very act of measurement may influence or alter the phenomenon being measured so that repeated measurements (test-retest) are not reproducible (not reliable). For example, if persons are asked to bend over and touch their fingers to their toes, they may not be able to do so on the first attempt. After several

attempts, however, the distance between fingers and toes may decrease. In similar fashion, clinical variables for assessment of TMD such as muscle palpation and assessment of joint sounds may not be stable in the short- or long-term, and they may be altered by repeated palpation or repeated mandibular movements.[39] Some phenomena such as blood pressure will demonstrate regression towards the mean by returning to usual levels over time.[4] Therefore, evaluation of some phenomena may require several examinations over time before a diagnosis is finalized.

Variability from Examination Equipment and Environment

In laboratory-based measurements, instruments are typically calibrated against established standards such as those of the American National Bureau of Standards, and the measurements are performed under controlled and specified conditions. The results and variability in these measurements are usually expressed as a standard deviation of the individual values or as confidence intervals around the calculated means.[4, 22]

It is important to distinguish the reliability of a measurement from the precision of the measurement. The precision of a measurement refers to the exactness or degree of refinement with which a measurement is stated. For example, clinicians may measure the anatomic root length on a radiograph to the nearest half-millimeter with a Boley's gauge or measure the depth of a periodontal pocket to the nearest millimeter with a periodontal probe. Alternatively, these measurements could be made electronically using tools with more precision, perhaps facilitating measurements to the nearest hundredth of a millimeter. Such a level of precision, however, may not be clinically relevant and would not necessarily translate to higher reliability scores. That is, just because a measurement is precise does not mean that it is reliable. In fact, the inherent variability of the physical attributes associated with many dental conditions is responsible for the inability to attain higher reliability scores.

Variability may also originate from the incorrect function or use of measuring devices or instruments. For example, reliable periodontal probing requires the use of a calibrated probe, on correct positioning of the probe, and application of appropriate probing pressure.

Variability of Examiners

Examiners may be inexperienced or incompetent. Examiners also differ because of biologic variation in the acuity of their senses (e.g., sight, touch, hearing), which may be further affected by their mood and sleep status. Examiners may also replace evidence by inference, potentially increasing the diagnostic error because a hasty inference may close a clinician's mind to other diagnoses.[29] For example, a middle-aged

female patient describes symptoms of constant aching, throbbing pain that began shortly after a recent lengthy dental appointment. The patient localizes the symptoms to the right submandibular region, right mandibular angle, and the right mandibular molar teeth. The dentist recalls the recent restoration of extensive caries on the mandibular right first molar. No radiographic abnormalities are detected, but irreversible pulpitis is diagnosed, and lengthy endodontic therapy is completed. Unfortunately, the patient returns the following day with increased bilateral pain of the mandibular molar teeth, restricted interincisal opening, and pain radiating from the mandibular molars bilaterally along sides of the face to the preauricular and anterior temporalis regions. Temporomandibular disorders, including referred pain from the masseter muscles to the mandibular molars, are subsequently diagnosed. In this example, the clinician jumped to the conclusion that the initial symptoms were of odontogenic origin and failed to consider the common alternative of referred pain from the masticatory muscles to the teeth.[23]

A clinician's diagnosis may also be affected by the mind set; that is, clinicians tend to diagnose what they expect or hope to find.[29] For example, when pathologists reach a diagnosis, they may be influenced by factors other than the histomorphology of the tissue on the slide. Schwartz et al[33] suggest that the pathologist's knowledge of the patient's clinical presentation may be considered and incorrectly weighted in reaching a diagnosis, so that the clinical data are double counted. If the pathologist knows that a biopsy specimen has been obtained from an area of erythroleukoplakia on the floor of the mouth of a heavy smoker and alcohol drinker, the suspicion of malignancy is raised even before the slide is placed on the microscope stage.[9, 19] In such instances, the dysplasia or carcinoma may be unconsciously graded as more severe than if the clinical information were not available to the pathologist.[33]

Specific biologic assays do not exist for all diseases, and investigators may need to make judgments using criteria that are not very specific or make judgments about subject characteristics that are difficult to evaluate. Because there are no absolute standards, the best that can be done is to determine if the investigators are consistent in their judgments. That is, performance review of the clinician investigators focuses on the likelihood that repeated examinations of the same, unchanged patient by either the same clinician or other clinicians yield identical results.

Comparisons may be made in which the same investigator examines the same subjects two or more times (intraexaminer reliability) or in which different investigators examine the same subjects (inter-examiner reliability). Interobserver variability is minimized when the endpoints are well defined and quantifiable, such as measuring the anatomic root length on a periapical radiograph or measuring overbite or overjet on study models. Interobserver variability is greater when criteria are vague and subjective, as in the clinical diagnosis of TMD[24, 39] or histologic diagnosis of dysplasia.[17, 18, 26]

Reliability measures provide information only about how well the examiners agree, not about whether the conclusions are correct.

Inter- and intraexaminer reliability have been quantitated by such measures as the Pearson correlation coefficient, intraclass correlation coefficient (ICC), or the kappa statistic (κ) (Table 1). For more details, readers are directed to the text by Norman and Streiner.[22]

Correlation Coefficients

The correlation coefficient, more properly called the Pearson product moment correlation coefficient, is used with continuous data. It is based on the extent to which the relationship between two variables can be described by a straight line called the regression line. The Pearson correlation coefficient, *r,* is a measure of the strength of the relationship between two sets of data. The strongest positive correlation has a value of 1.0, no relationship is indicated by 0, and perfect negative correlation has a value of −1.0. Thus, correlation coefficients with values closest to 1.0 demonstrate the greatest relationship between sets of data, but perfect agreement occurs only when the regression line has a slope of 1; that is, the points fall along the line of equality.

In regression analysis, the square of the correlation coefficient, r^2, is known as the coefficient of determination, which is, in effect, the fraction or proportion of the total variance in the dependent variable that can be explained by the relationship between variables; r^2 tends to overestimate the true reliability. In general, *r* values will be higher than, and overestimated in comparison with, the more theoretical sum reliability calculated by the intraclass correlation coefficient (ICC). Bland and Altman[3] have discussed the problems with use of correction coefficients and have developed an alternative method for assessing agreement between two methods of clinical measurements based on graphic techniques.

Intraclass Correlation Coefficient

The ICC is generally derived from analysis of variance calculations. Intraclass correlation coefficient values can range from 0 to 1.0. Unlike *r,* the ICC value indicates what proportion of the total observed variability is caused by variability among the subjects as compared with variability among the examiners. If most of the variability results from discrepancies among examiners, the ICC values are low. Alternatively, if the examiners are reliable (consistent) among themselves, ICC values are high (e.g., between 0.75 and 1.00), and in effect one examiner could be replaced with another.[8] The ICC values may be interpreted in a manner similar to κ scores, which are more commonly used.

Table 1. RELIABILITY OF SOME MEASUREMENTS/TESTS USED IN DENTISTRY

Test	Reference Number	Correlation Coefficient		Intraclass Correlation Coefficient	Kappa	
		Inter-observer	*Intra-observer*		*Inter-observer*	*Intra-observer*
Periodontics						
Probing depth, general	5b, 10a	0.63	0.72		0.26	
Plaque	6b, 11a	0.81	0.32		0.22	
Temporomandibular disorders						
Temporomandibular joint sounds—manual palpation						
Trained examiner	8			0.68	0.62	
Untrained examiner	8			0.35	0.30	
Temporomandibular joint sounds—stethoscope						
Trained examiner	8			0.26	0.61	
Untrained examiner	8			0.32	0.35	
Mandibular kinesiology						
Maximal pain-free vertical opening	8, 10b					
Trained examiner			0.89	0.90		
Untrained examiner	8			0.72		
Dental radiology						
Caries, calibrated examiner	36a				0.73	0.80
Periodontal disease, calibrated examiner	36a				0.80	0.79
Degenerative temporomandibular joint changes on tomography	5a				0.47–0.80	0.58–0.79
Disk displacement on MR imaging	24				0.70	
Oral pathology						
Diagnosis of dysplasia	1	0.30–0.63			0.29–0.48	0.05–0.49
Grading of oral leukoplakia from no dysplasia to carcinoma in situ	16				0.27–0.45	

Kappa Scores

The best approach in evaluating reliability for noninterval data is the κ statistic, which adjusts for the degree of agreement expected by chance. For a perfect association, $\kappa = 1.0$, and for no association $\kappa = 0$. Qualitative interpretation in relation to κ values vary,[16, 29] but Brunette[4] suggests that κ values below 0.4 indicate poor agreement, κ values of 0.4 to 0.75 are fair, and κ values of 0.75 to 1.0 are excellent. A rule of thumb is that clinical studies should not proceed before investigators have been trained and calibrated with demonstrated high κ scores (e.g., $\kappa > 0.6$).

Table 1 lists the reliabilities of some measurements and tests used in dentistry and illustrates the differences between correlation coefficients and κ scores. For example, the interexaminer correlation coefficients for probing depths and plaque assessment are 0.63 and 0.81, respectively; in contrast, the interexaminer κ scores are only 0.26 and 0.22! In similar fashion, Abbey et al[1] calculated correlation coefficients and κ scores for six pathologists whose agreement between their original sign-out diagnoses of dysplasia and subsequent reexaminations of the same slides were compared. Correlation coefficients averaged 0.50; intraexaminer κ scores ranged from 0.05 to 0.49. In the same study, interexaminer κ scores for the presence or absence of dysplasia ranged from 0.29 to 0.48.

MEASUREMENT VALIDITY AND THE REFERENCE STANDARD

Measurement validity refers to the truthfulness of the measurement or technique. In other words, whether the measurement measures what it claims to measure. The determination of measurement validity requires a comparison of the measurement or technique with a reference measure or technique that has been accepted as true and is the acknowledged standard, at the time, for definitive diagnosis of the disease or condition. The principle of measurement validity is crucial to clinical measurements because even if a measurement is highly reliable, the measurement has no diagnostic value if that measure does not accurately reflect the characteristic of interest. For example, a clinician may reliably measure the anatomic root length of an incisor on a periapical radiograph. If however, the bisecting angle technique rather than the parallel technique was used for exposure of the radiograph, the measured root length may not be a true or valid representation of the anatomic root length.

The classification of disease is traditionally based on pathologic anatomy,[40] and therefore the histopathologist's diagnosis is typically regarded as the reference standard. Performing the reference test of autopsy or histopathologic examination is not always feasible, however, because obtaining a specimen is generally an invasive procedure that may also be risky, expensive, and often impossible to perform in a timely

manner. Not all body sites are as readily accessible for biopsy and histologic examination as the oral soft tissues. Therefore, surrogate parameters such as biologic assays or measurements are used as the standard for comparison. For example, in the case of bovine spongiform encephalopathy and its human variant, Creutzfeld-Jacob disease, autopsy is both the reference standard and the only reliable and valid diagnostic tool at this time. If valid and less invasive laboratory techniques were available, earlier diagnosis of the disease would be possible. The assumed benefit of earlier disease detection, such as through screening tests, must be tempered with the possibility that for some diseases earlier detection is unlikely to improve the prognosis. The early detection of disease is assumed to be beneficial, because treatment initiated before the onset of symptoms is assumed to be more effective than later treatment and thereby the development of disease may be reduced or eliminated. For some conditions, such as Creutzfeld-Jacob disease, there is no effective treatment at this time; hence, the earlier diagnosis of some conditions must be weighed against the overall risks and benefits for the individual and society.[13]

Widmer[39] reviewed the measurement validity of TM joint imaging techniques to anatomy. Arthrography demonstrated an 84% true correlation to anatomy,[37] MR imaging had a 73% to 85% true correlation,[7] and tomography had a 63% to 85% true correlation to anatomy.[12] Widmer[39] also reviewed the measurement validity of TM joint sounds by palpation and stethoscope in an arthrographic examination of asymptomatic subjects. Assessment for TM joint sounds by manual palpation revealed that 15% of silent joints had disk displacement.[37] Joint sound assessment by stethoscope revealed 14% of silent joints with disk displacement.[32] These results demonstrate that disk displacements may be present in the absence of joint sounds and that the presence of joint sounds may not offer a valid assessment of disk displacements.

DIAGNOSTIC VALIDITY AND THE REFERENCE STANDARD

Biologic assays do not exist for all disorders, and for some diseases and conditions, a real or practical reference standard does not exist. For example, biologic assays for TMD and fibromyalgia do not exist, and there is no reference standard for the measurement of active periodontal disease. Instead, clinicians use measurements such as probing and attachment levels, which are cumulative indices reflecting the history of disease (in this case, attachment loss) rather than the presence of active disease.[4] In similar fashion, the diagnosis of fibromyalgia relies on the key clinical feature of decreased pain threshold as manifested by tenderness at 18 specified anatomic locations.

Widmer[39] distinguishes measurement validity from diagnostic validity, which is the extent to which diagnostic criteria can be used to classify persons as to the absence or presence of a disorder in regards to the

current reference standard classification system. That is, in the absence of reference standard based on histopathologic or biologic assays, a general nominalistic impression of the diagnostic usefulness of each measure is gained through diagnosis of the presence or absence of the disorder among individuals already known either to have or not to have the disorder of interest. For example, for fibromyalgia, the diagnostic validity of tenderness to muscle palpation is evaluated by the ability of this measurement technique to distinguish between individuals known to have or known not to have fibromylgia. In the future, if laboratory findings are linked to fibromyalgia, this new measurement or test approach must also be assessed for its ability to distinguish between individuals known to have or not to have fibromylgia. Thus, the relative diagnostic abilities of the existing method of muscle palpation and the new laboratory finding can be compared; the more successful method would be regarded as the reference standard until another new test proves superior.

TEST CHARACTERISTICS

Traditionally, a new test, measurement, or technique is evaluated in a sample of patients identified by the existing reference standard either to have or not to have the disease of interest. A general impression of the diagnostic strengths of a measure, test, or technique may then be obtained from characteristics or parameters of the test. Test characteristics are mathematical probabilities that are calculated by direct comparison between a test, measurement, or technique and the reference standard in a 2 × 2 contingency table (Figs. 1–2; Box). Summary statistics such as sensitivity, specificity, and predictive values aid in the comparison and analysis of different tests. Test accuracy is a measure of the agreement between the test and the reference standard, but, as discussed

Gold Standard

New Test	**POSITIVE** Disease is really present	**NEGATIVE** Disease is really absent	
POSITIVE Disease appears present	TRUE POSITIVE **a**	FALSE POSITIVE **b**	**a + b**
NEGATIVE Disease appears absent	**c** FALSE NEGATIVE	**d** TRUE NEGATIVE	**c + d**
	a + c	**b + d**	**a + b + c + d**

Figure 2. Contingency comparison between gold standard and new test. For example, for the disease of caries, the gold standard is histologic examination, and a new test for diagnosis of caries may be direct digital radiography.

Definitions of and Calculations for Test Characteristics

Accuracy is the overall agreement between the test and the reference gold standard. Accuracy may be calculated from a 2 × 2 contingency table as shown in Figure 2 by the formula

$$\frac{a + d}{a + b + c + d}$$

Sensitivity is the proportion of diseased individuals correctly identified by the test. Sensitivity is also known as the true positive rate and may be calculated from a 2 × 2 contingency table as shown in Figure 2 by the formula

$$\frac{a}{a + c}$$

Specificity is the proportion of non-diseased individuals correctly identified by the test and is also known as the true negative rate. Specificity may be calculated from a 2 × 2 contingency table as shown in Figure 2 by the formula

$$\frac{d}{b + d}$$

Prevalence (P) is the overall probability or risk that the disease is present before the test and is also known as the pretest likelihood. Prevalence is the proportion of individuals in a population who have the disease at a specific point in time. Prevalence in a specified population may change over time, and prevalence may change if the definition of the disease changes. Prevalence may be calculated from a 2 × 2 contingency table as shown in Figure 2 by the formula

$$\frac{a + c}{a + b + c + d}$$

Post-test Likelihood of a Positive Test (PTL+) is also known as the positive predictive value. For an individual with a positive test result, PTL(+) is the probability that the disease is actually present. The PTL(+) may be calculated from a 2 × 2 contingency table as shown in Figure 2 by the formula

$$\frac{a}{a + b}$$

When the sensitivity, specificity, and prevalence or pretest likelihood are known, PTL(+) may be calculated by the formula

$$\text{PTL}(+) = \frac{\text{P} \times \text{LR}(+)}{(1.0 - \text{P}) + \text{P} \times \text{LR}(+)}$$

$$\text{where LR}(+) = \frac{\text{true positive}}{\text{false positive}} = \frac{\text{sensitivity}}{1.0 - \text{specificity}}$$

Post-test Likelihood of a Negative Test (PTL[−])

For an individual with a negative test result, PTL(−) is the probability that the disease is actually present. The PTL(−) may be calculated from a 2 × 2 contingency table as shown in Figure 2 by the formula

$$\frac{\text{c}}{\text{c} + \text{d}}$$

When the sensitivity, specificity, and prevalence or pretest likelihood are known, PTL(−) may be calculated by the formula

$$\text{PTL}(-) = \frac{\text{P} \times \text{LR}(-)}{(1.0 - \text{P}) + \text{P} \times \text{LR}(-)}$$

$$\text{where LR}(-) = \frac{\text{false negative}}{\text{true negative}} = \frac{1.0 - \text{sensitivity}}{\text{specificity}}$$

Negative Predictive Value (NPV)

For an individual with a negative test result, the probability that disease is really absent. The NPV may be calculated from a 2 × 2 contingency table as shown in Figure 2 by the formula

$$\frac{\text{d}}{\text{c} + \text{d}}$$

in the section on likelihood ratios, accuracy is not the sole measure or guarantee of a test's clinical usefulness.

Sensitivity is the proportion of individuals who are correctly identified as having the disease. Specificity is the proportion of individuals who are correctly identified as nondiseased. Table 2 illustrates the sensitivities and specificities of some diagnostic tests used in dentistry.

Sensitivity and specificity are typically calculated in defined populations in which the disease status of the individuals is already known and confirmed by the reference standard and in which only extremes of disease (the very sick) and health (the very healthy) are represented. As discussed later, these circumstances do not represent the true clinical situation. If the clinician already knew the disease status of a patient,

Table 2. SENSITIVITIES, SPECIFICITIES, AND LIKELIHOOD RATIOS OF SOME DIAGNOSTIC TESTS USED IN DENTISTRY

Test	Reference Number	Sensitivity	Specificity	LR (+)*	LR (−)†
Caries					
Clinical examination	36b	0.13	0.94	2.2	0.93
Bite-wing radiographs	21a	0.73	0.97	24.3	0.28
Periodontics					
Gingival redness	11a	0.27	0.67	0.82	1.09
Plaque	11a	0.47	0.65	1.3	0.82
Bleeding on probing (2 mm, 5/6 threshold)	18a	0.29	0.88	2.4	0.81
Temporomandibular joint disorders					
Temporomandibular sounds—manual palpation—single click	7a	0.43	0.75	1.7	.76
Disk displacement on MR imaging		0.86	0.63	2.3	.22
Degenerative changes on sagittal tomography	36a	0.47	0.94	7.8	.56

*LR (+) is calculated by $\frac{\text{sesnsitivity}}{1.0 - \text{specificity}}$

†LR (−) is calculated by $\frac{1.0 - \text{sensitivity}}{\text{specificity}}$

there would be no need for further investigation. Instead, the clinician is typically confronted with equivocal cases among a population of healthy and diseased individuals.

The difference between the diagnosis for presence or absence of disease or abnormality depends on the selection of cutoff points. Changes in activity or level of any physiologic, biochemical, or molecular marker are typically reflected by continuous measures. In contrast, the presence or absence of an abnormality or disease is typically a dichotomous diagnosis, such as normal versus abnormal or health versus disease, on occasion gradations of abnormalities are also used, such as mild, moderate, or severe dysplasia, and hypertension, which is classified as stage 1 to stage 4. Continuous measures may be collapsed to dichotomous data by the selection of cutoff points. For example, individuals exhibit a wide range of pain-free unassisted vertical and horizontal mandibular movements, and these mandibular kinesiology measurements are used as diagnostic criteria for TMD.[39] If the cutoff point between non-TMD (health) and TMD (disease) is arbitrarily set at interincisal opening of 40 mm, then theoretically, the patient with a 39-mm opening is eligible for diagnosis of TMD, but another patient with a 41-mm opening is diagnosed as non-TMD. Alternatively, if the cutoff point between non-TMD and TMD is set instead at 35 mm, the same patient with a 39-mm opening would be excluded from diagnosis of TMD. In similar fashion, the number of specified muscle sites that are tender to

palpation and the number and type of TM joint sounds will affect the proportions of individuals diagnosed with TMD.[39]

Ideally, the selection of a cutoff point should be based on what is best for the patients concerned, and the consequences of over- and underdiagnosis must be considered. If the condition is innocuous and neither shame nor anguish is associated with the diagnosis (for example, the diagnoses of linea alba or the common cold), then the cutoff for classification as diseased may be relaxed. Conversely, if there is no advantage in early diagnosis, a positive diagnosis has the potential to produce anxiety in the patient, and there is no effective treatment, the cutoff for disease should be set high ($\geq$ 99%) to exclude the nondiseased.[29]

The selection of the cutoff point will determine the proportion of true-positive, false-positive, true-negative, and false-negative results, which, in turn, will produce different estimates of the sensitivity and specificity of the diagnostic test (see box). A perfect test will yield only true-positive and true-negative results without any overlap or false-positive or false-negative result (Fig. 3).

Criteria for Selection of Test Thresholds

Low Threshold

- selected if it is important that all individuals with the disease or its progression are detected
- provides high sensitivity and high PTL(+)
- results in increased number of false-positive results because of the low specificity
- is useful for screening for serious or life-threatening disease but confirmation testing is required (e.g., dentists perform screening examinations for high blood pressure or for oral cancer in patients who are asymptomatic for these diseases.)

High Threshold

- limits the number of false-positive results
- is required for confirmation testing
- results in high specificity but lower sensitivity. High-specificity values are important for diseases that are not life-threatening such as TMD. High specificity excludes individuals without the disease from pursuing unnecessary, irrelevant, and possibly invasive, irreversible, and expensive treatment.

In general, if a low threshold is selected, the sensitivity is increased, and the specificity is decreased; a high threshold results in high specificity but lower sensitivity. High sensitivity is desirable for screening tests. High specificity is required for exclusionary tests to minimize the number of false-positive results. The highest possible sensitivity and specificity are desirable for confirmatory tests to minimize both false-positive and false-negative results. Unfortunately, high sensitivity and high specificity are rarely found in a single test.

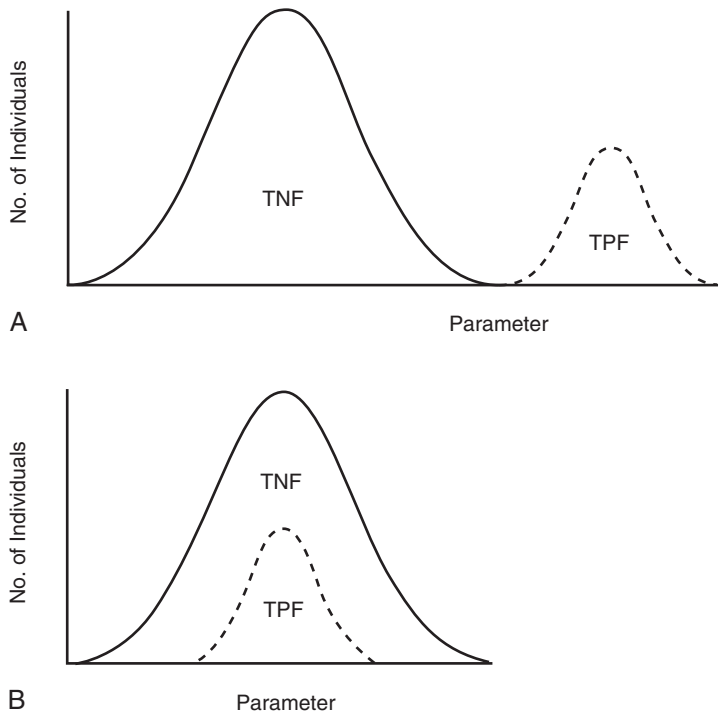

Figure 3. Hypothetical distribution of healthy (true positive fraction [TPF]) and diseased (true negative fraction [TNF]) populations. Test results yield different estimates of sensitivity and specificity *A,* Hypothetical perfect test with 100% sensitivity and 100% specificity. The diseased (TPF, *dashed line*) and healthy (TNF, *solid line*) individuals are identified without false negative (FNF) or false positive (FPF) fractions. *B,* Hypothetical useless test. The diseased and healthy populations are not identified by the test.

Receiver Operating Characteristic Analysis

One of the best methods to evaluate the effect of different cutoff points is the receiver operating characteristic (ROC) analysis (Fig. 4). An ROC analysis plots the true-positive fraction (sensitivity) as a function of the false-positive fraction (1.0 − specificity), and points along the curve can be used to determine the effect of different thresholds for the test. Selection of points towards the left of the curve yields higher specificity, and points to the right yield higher sensitivity. An ROC analysis also permits the comparison of different tests without any selection of upper or lower reference limits or any particular sensitivity or specificity. It is widely agreed that ROC curves are independent of the disease prevalence and therefore reflect the true performance of the diagnostic tests.[11, 29]

In clinical practice, the selection of cutoff points is determined by several factors, including mortality and morbidity of the disease, the

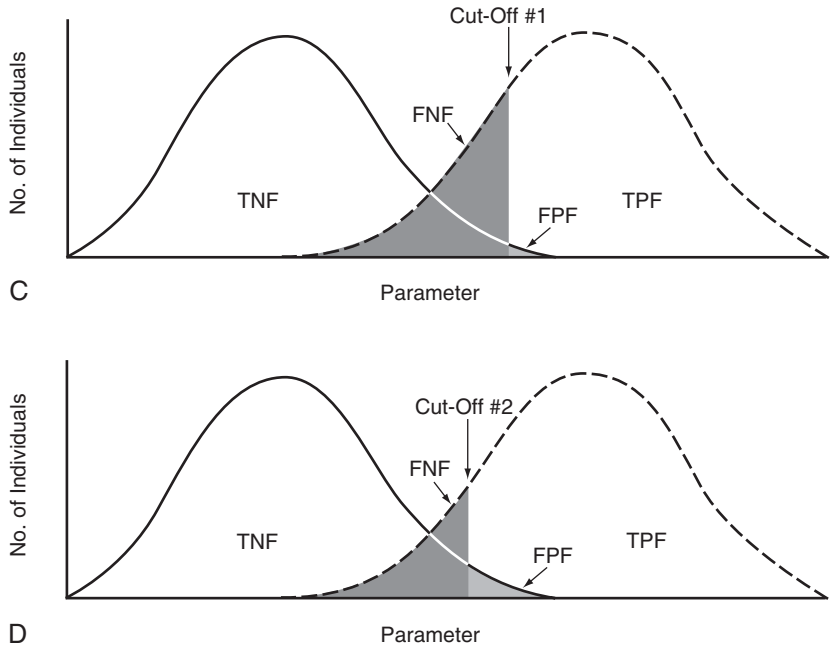

Figure 3 *(Continued). C* and *D,* Hypothetical typical test with overlap of healthy and diseased populations. The selection of the cut-off point to distinguish between healthy and diseased individuals affects the proportion of the FNF and FPF. Sensitivity and specificity are affected by the selection of the cut-off point. In *C,* the cut-off point is located further to the right than the cut-off point in *D.* Therefore, the FPF in *C* is smaller than the FPF in *D.* Conversely, the FNF in *C* is larger than the FNF in *D.*

consequences of over- and undertreatment, and the cost and time required to perform the diagnostic test. Once test thresholds are established, sensitivity and specificity are considered to be stable properties of the test because they are apparently not affected by the prevalence of the target disease. Some evidence, however, indicates that sensitivity and specificity do change from one clinical population to another,[14, 15] especially if the stage of disease varies in different groups of patients.[11, 24]

The Effects of Prevalence

Sensitivity (true-positive rate) and specificity (true-negative rate) are measures of how well the test correctly identifies diseased and healthy individuals, respectively. Sensitivity and specificity do not provide the clinician with any information about whether the test will provide meaningful diagnostic information for individuals whose disease status is not known. Hence, the predictive values (see box on pages 100–101) of a test are required to provide information about how often a test will

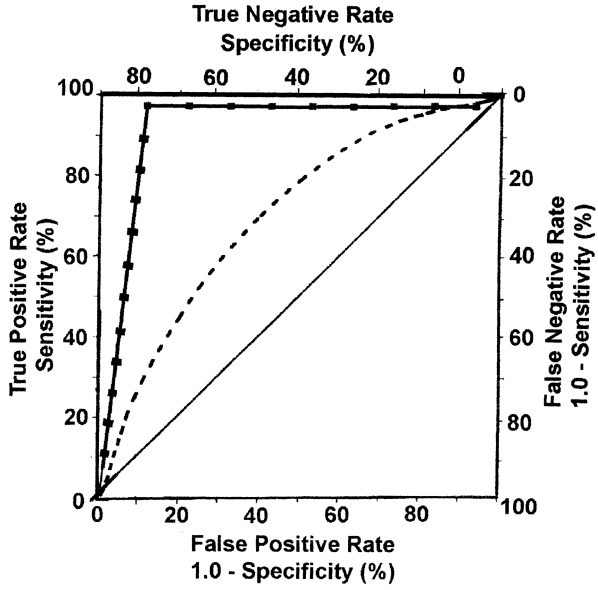

Figure 4. Receiver operating characteristics (ROC) curves plot the TPF (sensitivity) against the FPF (1.0-specificity). ROC curves permit selection of the threshold or cut-off point that provides the best combination of sensitivity and specificity scores. The most discriminating tests cluster in the upper left-hand corner, and the most discriminating test has the greatest area under its ROC curve. ROC curves also permit the comparison of tests without selection of reference limits or sensitivity and specificity. For example, this figure compares the ROC curve for conventional radiographic film evaluation of artificial cortical bone lesions, produced with a size 6 burr in dried mandibles (bulleted line) with the ROC curve for conventional radiographic film evaluation of in vivo periodontal crestal alveolar bone loss (dashed line). In this example, the area under the ROC curve for the detection of in vitro cortical lesions is larger than the area under the ROC curve for the in vivo detection of periodontal crestal bone loss. As expected, conventional radiographic evaluation of in vitro artificial cortical lesions is more discriminating or a more powerful test than conventional radiographic evaluation of in vivo crestal bone loss. Solid line = ROC curve of *noise* or a hypothetical useless test. (*Dashed line, Data from* Nummikoski PV, Steffensen B, Hamilton K, et al: Clinical validation of a new subtraction radiography technique for periodontal bone loss detection. J Periodontol 71:598–605, 2000; *Bulleted line, Data from* Paurazas SB, Geist JR, Pink FE, et al: Comparison of diagnostic accuracy of digital imaging by using CCD and CMOS-APS sensors with E-speed film in the detection of periapical bony lesions. Oral Surg Oral Med Oral Pathol Oral Radiol Endod 89:356–363, 2000.)

provide a correct diagnosis in a mixed population. Three predictive values may be calculated: (1) positive predictive value, (2) negative predictive value and (3) posttest likelihood of a negative test.

The positive predictive value is also known as the post-test likelihood of a positive test (PTL[+]). For a patient who has undergone a diagnostic test and obtained a positive test result, PTL(+) is the probability that disease is actually present. When a negative test result is obtained, the probability that disease is truly absent is known as the

negative predictive value. For a patient with a negative test result, the clinician may need to know the probability that disease is actually present; this probability is known as the post-test likelihood of a negative test (PTL[−]). Although a negative result will reduce the probability of disease being present, typically it will not absolutely eliminate this possibility.

The predictive values of a test vary widely as the prevalence of the disease changes.[11, 29] Prevalence is also known as the pretest likelihood, and it is the overall probability or risk that disease is present before the test is administered.

For example, toluidine blue has been advocated for the detection of oral squamous cell carcinoma (SCC). The sensitivity of toluidine blue ranges from 93.5% to 97.8%, and its specificity ranges from 73.3% to 92.9%.[28] The predictive values of toluidine blue and the conclusions provided by this test will vary, however, depending on the individual patient to whom or the population in which the test is applied. The prevalence of SCC in the general population has been estimated at 3%,[25] and therefore the posttest likelihood of a positive toluidine blue test in the general population is only 6%.[10] In contrast, the prevalence of SCC, either as primary or recurrent disease, is greater in a tertiary care center for oral SCC, where prevalence estimates range from 26%[25, 34] to 33%.[10] Consequently, the posttest likelihood of a positive test in a tertiary care center is also greater (51%).[10] In the high-prevalence setting, the posttest likelihoods of the tests are considerably higher than the pretest probabilities, meaning that there is a considerably increased probability that the disease is actually present. In contrast, the posttest likelihoods of the same test in the general population (low-prevalence setting) are similar to the pretest probabilities, meaning that there is only a slight increase in the probability that the disease is actually present. Nevertheless, the significance of each positive and negative test must be evaluated on an individual basis by the clinician, who must then decide the subsequent course of action.

The example with the toluidine blue test demonstrates that even a test with high sensitivity (93.5%–97.8%) and specificity (73.3%–92.9%)[27] can yield low predictive values when the prevalence (or pretest likelihood) is low. Sacket et al[29] further illustrate this point using a theoretical test with 95% sensitivity and 95% specificity under conditions of variable prevalence. For example, as the prevalence changes from 99% to 1%, the PTL(+) changes from 99.99% to 16%, respectively. Thus, even a test that has excellent specificity and sensitivity will produce a low likelihood of disease being present if it is applied to an individual in a population in which the initial pretest prevalence is low.

The choice of a particular test for a specific disease is determined by the power or ability of the test to revise the pretest probabilities, either upwards to rule in the disease, or downwards to rule out the disease. In general, for a test with a sufficiently high sensitivity, a negative result rules out the disease. In contrast, for a test with a sufficiently high specificity, a positive result rules in the disease.[29] In

other words, the clinician relies on pattern recognition: "if it looks like a duck, quacks like a duck, and waddles like a duck, it probably is a duck."

Different tests for the same disease can be used in combination, either in series, such as screening testing followed by confirmation testing, or in parallel.

In series testing
- tests are used in succession
- if tests A and B are used in series, then either test A or test B can be used first
- a positive result on the first test requires testing with the second test
- is less sensitive in detecting disease than parallel testing, but series testing has greater specificity and is more efficient at confirming the presence of disease
- is used in confirmation testing

In parallel testing
- tests are performed concurrently
- if tests A and B are used in parallel,
  - a positive result requires positive results for either test A or test B
  - a negative result requires that both test A and test B are negative
- is more sensitive than series testing for detecting disease but less efficient at confirming presence of disease

At health fairs, clinician dentists may perform screening tests for oral cancer through a careful visual inspection of the oral soft tissues or a screening test for TMD by evaluating the patient's range of pain-free mandibular movements. With positive results of suspicious oral lesions or a restricted range of jaw movements and associated discomfort, the patients would be referred to their own dentists or to specialists for possible oral biopsy or more detailed TMD evaluation including assessment of joint sounds and TM joint and neck and masticatory muscle tenderness.

When several tests are used in sequence, the posttest likelihood of disease after the first test is used as the pretest likelihood for the subsequent test. A possible problem with this approach is the propagation of errors, because each test can be considered as having some associated error. Therefore, as more tests are performed, the precision of the probability estimate will decline. The posttest probability of disease may also be distorted by the end of the test sequence if the clinician assumes that the tests are independent when the test results are actually dependent. That is, the test result on one test or measure may affect the characteristics of the second test, a phenomenon termed *concordance* or *convergence*.[29] Concordance occurs when patients who are positive on one of the paired tests are likely to be positive on the other one as well, or when patients who are negative on one test are likely to be negative on the other one. For example, the electric pulp-stimulation test is much

more likely to be positive when the thermal (cold) test is positive (i.e., the patient reports sensation upon cold stimulation of the tooth) than when the thermal test is negative (the patient denies sensation to cold stimuli). Conversely, teeth with negative results on one test (either the cold or electric pulp test) are also likely to be negative on the other. Concordance results in an overestimate of disease likelihood. Sacket et al[29] suggest that for short courses of two or three diagnostic tests, convergence is not a serious problem but should be considered. For example, concordance was observed between the use of toluidine blue and visual clinical examination of patients in an oral cancer tertiary care center by a trained and experienced clinician.[10] That is, oral lesions that were classified as suspicious or positive by one these methods were likely to be positive on the other method as well. When the results of both the visual clinical examination and toluidine blue were positive, the pretest likelihood of 33% was raised to a posttest likelihood of 54%, which is greater than the PTL(+) obtained by either toluidine blue application alone (51%) or the visual clinical examination alone (44%).[10] A PTL(+) of 54% calculated with consideration of concordance is a lower but more realistic value than the PTL(+) of 62% that is calculated if the tests are used sequentially and assumed to be independent.

LIKELIHOOD RATIOS AND NOMOGRAMS

Principles three and four of the diagnostic decision analysis require that the interpretation of possible test outcomes precede the ordering of the test and that testing should proceed only if the subsequent management of the patient will be altered as a result of the test result. How can this interpretation be accomplished?

If the sensitivity and specificity of a particular test and the prevalence of the disease of interest are known, the post-test likelihoods of a positive and negative test can be calculated from the formulas for PTL(+) and PTL(−) shown in the box on pages 100–101. These calculations use likelihood ratios that "express the odds that a given level of a diagnostic test result would be expected in a patient with (as opposed to one without) the target disorder."[29] Sensitivity and specificity are probability statements, and they may be converted to odds ratios, which are the ratio of two probabilities. Probabilities and odds contain the same information but convey it differently. Thus, a probability of 50% means even odds of 1:1. Likelihood ratios provide a measure of a test's ability to revise the pretest probabilities, and they are simple to calculate from the sensitivity and specificity of the particular test. Although sensitivity and specificity are used to calculate the likelihood ratios of a test, it is the likelihood ratios, not sensitivity and specificity, that provide information as to the potential power of the test. As a rule of thumb, if the sum of a test's sensitivity and specificity is unity (1.0), the test is useless: the likelihood ratios of the test are also unity (1.0), and therefore the test has no power to revise the pretest probability. In general, power-

ful tests for revising pretest probabilities of disease have positive likelihood ratios with values greater than 10 and negative likelihood ratios less than 0.1.

Likelihood ratios offer diagnostic advantages in that they are less susceptible than sensitivity or specificity to changes in the prevalence or pretest probability of the disease.[29] Likelihood ratios may also be calculated for dichotomous levels of disease and for several levels of the test result. The product of the likelihood ratio for the diagnostic test result and the pretest odds for the target disorder yields the posttest odds for the target disorder.[29] A convenient method for rapidly calculating posttest probability of disease is offered by the use of likelihood ratios for the test and nomograms.

Nomograms (Fig. 5)[30] offer a convenient and fast alternative to the calculation of posttest likelihoods using the formulas shown in the box on pages 100–101. Table 2 illustrates the sensitivities, specificities, and likelihood ratios of some diagnostic tests used in dentistry. Figure 6 demonstrates use of the nomogram in the diagnostic decisions for three examples of potential interproximal caries (the disease) and use of bitewing radiographs (the test). In each case, the clinician detects a small area of discoloration on the distal aspect of the maxillary second bicuspid but is not able to engage the explorer interproximally. For the disease of caries, the clinician has assigned a test threshold of 30% and a test-treatment threshold of 65% (Figs. 1, 6).

Patient A is an adolescent female who aspires to a career in modeling with an unrestored permanent dentition. Patient A practices excellent oral hygiene and is compliant with twice-yearly prophylaxis appointments. Bitewing radiographs taken 2 years ago at the completion of orthodontic treatment do not reveal any abnormalities. The clinician assigns a pretest probability for caries of 1%. The clinician's pretest probability is located well below the test threshold of 30%, and therefore radiographs would not be indicated. In the unlikely event that radiographs (the test) were performed with a positive test result, the probability of caries or PTL(+) can be calculated to be 20%. Despite this positive test result, no further tests or restoration would be indicated, because this probability is still less than the test threshold of 30%. If the test results were negative, PTL(−) can be calculated to be 0.4%, effectively ruling out the presence of caries.

Patient B is a young adult male with a moderately restored posterior dentition. Patient B is a pastry chef apprentice who demonstrates poor oral hygiene and poor compliance with recommended dental recall and prophylaxis appointments. The patient was last seen 3 years ago when bitewing radiographs revealed no sites of interproximal caries in the posterior mandibular dentition. The clinician assigns a pretest probability of 50% to the presence of caries. This pretest probability is located between the test and test-treatment thresholds; therefore, bitewing radiographs are indicated. With a positive test result, treatment is indicated, but a negative test result rules out the disease and treatment.

Patient C is an elderly patient with a heavily restored dentition and

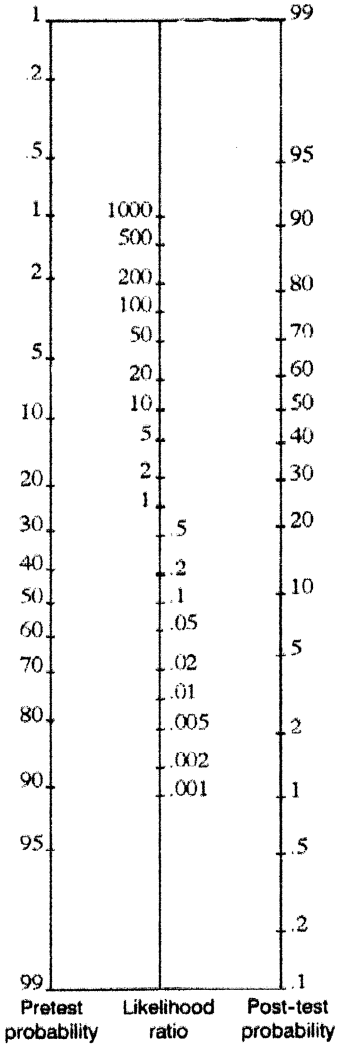

Figure 5. Nomograms have converted pre- and post-test odds to their corresponding probabilities. To use the nomogram, a straightedge is used to align the pretest probability (left column) with the likelihood ratio (center column) of the test being used. The post-test probability is revealed by reading across the straightedge to the right-hand column on the nomogram. (*Data from* Fagan, TJ: Nomogram for Bayes' theorem [letter]. N Engl J Med 293:257, 1975; Sacket DL, Richardson WS, Rosenberg W, et al: Evidence-Based Medicine: How to Practice and Teach EBM. New York, Churchill Livingstone, 1997, p 127.)

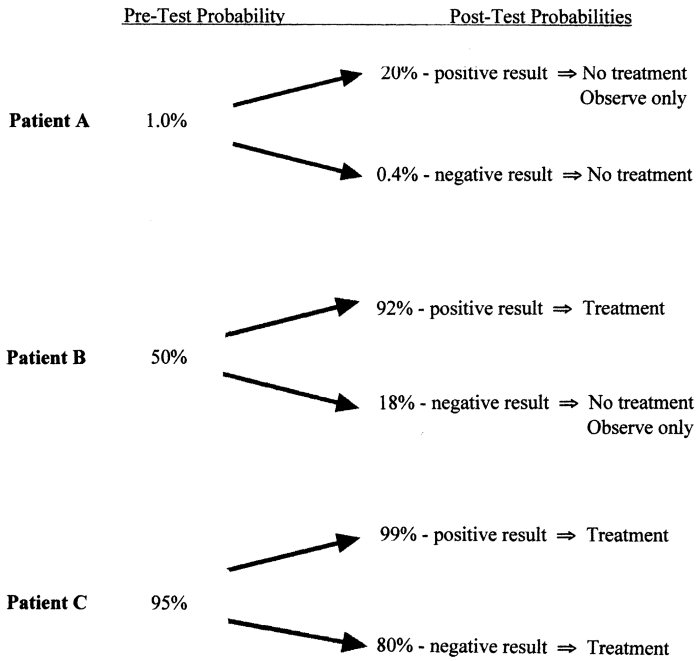

Figure 6. Diagnostic decisions for bitewing radiographs for three patients with possible caries (the disease). *Patient A,* By aligning the straightedge at 1% in the pretest probability column with 24 in the likelihood ratio column, the post-test probability of caries being present is raised to about 20%—a value well below the test-treatment threshold of 65% and below the test threshold of 30%. Despite a positive test result, no further tests or restoration are indicated, and the clinician may feel confident about merely observing the tooth. When the pretest probability of 1% is aligned with the likelihood ratio (LR) of a negative test result (0.28), the post-test probability of disease has been further reduced to about 0.4%, effectively ruling out the presence of caries. *Patient B,* The pretest probability of 50% is located between the test and test-treatment thresholds. Radiographs are indicated. Post-test likelihood of disease (PTL[+]) is raised to 92% and treatment is indicated. PTL(−) is reduced to 18% and treatment is not indicated. *Patient C,* The clinician recognizes that the 95% pretest probability exceeds the established test-treatment threshold; bitewing radiographs are not required for diagnosis and test results would not alter the proposed management (restoration of the tooth). Even a negative test result (no radiographic evidence of caries) would still result in an 80% post-test probability of caries being present. Although 80% is a lesser probability of disease than 95%, it still exceeds the test-treatment threshold and is probably not low enough to change the planned management. LR(+) = 24; LR(−) = 0.28 (see Table 2); test threshold = 30%; test-treatment threshold = 65% (see Figure 1); see Figure 4 for nomogram.

recent past history of recurrent and new caries. Patient C is disabled with rheumatoid arthritis and is xerostomic with poor oral hygiene although she is a compliant patient. The clinician assigns a pretest probability for caries at 95%, and treatment is indicated without further diagnostic testing. That is, radiographs are not required to establish the diagnosis of caries in this case, although radiographs may provide useful information to guide treatment of the caries or the diagnosis or treatment of other pathologic conditions. For patient C, even a negative test result would still result in an 80% posttest probability of caries being present and requiring treatment. This case illustrates that clinicians must be careful not to overestimate the meaning of negative test results when, in fact, the probability of disease is high.

SUMMARY

This article has briefly introduced the dental clinician to the principles and practical application of diagnostic decision analysis. There are trade-offs and uncertainties in the process of arriving at a diagnosis, but they can be understood and controlled. First, the clinician must understand the significance of disease prevalence and assign to the patient an initial probability of disease being present. The clinician must then determine if further diagnostic measurements or tests are warranted. If so, the appropriate test must be selected, based on the ability of the test to revise the initial pretest probability. When a diagnostic test is positive, the clinician must know the probability that disease is actually present. The clinician must also know the probability that disease is actually present if the test result is negative. The astute clinician will calculate the posttest probabilities before proceeding with a test and will base treatment decisions on test results in accordance with predetermined test and test-treatment thresholds.

ACKNOWLEDGEMENTS

The authors are grateful to David Perizzolo for formatting the digital figures, to Lesley Weston for her careful editing, and to Dr. Babak Chehroudi for his critical review.

References

1. Abbey LM, Kaugars GE, Gunsolley JC, et al: Intraexaminer and interexaminer reliability in the diagnosis of oral epithelial dysplasia. Oral Surg Oral Med Oral Pathol Oral Radiol Endod 80:188–191, 1995
2. Beck JD: Issues in assessment of diagnostic tests and risk for periodontal diseases. Periodontology 2000 7:100–198, 1995
3. Bland JM, Altman DG: Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 8:307–310, 1986
4. Brunette DM: Critical Thinking. Understanding and Evaluating Dental Research. Chicago, Quintessence Publishing Co, 1996

5. Choi BCK, Jokovic A: Diagnostic tests. J Can Dent Assoc 62:6–7, 1996
5a. Cholitgul W, Petersson A, Rohlin M, et al: Diagnostic outcome and observer performance in sagittal tomography of the temporomandibular joint. Dentomaxillofacial Radiology 19:1–6, 1990
5b. Clemmer BA, Barbano JP: Reproducibility of periodontal scores in clinical trials. J Periodont Res 9 (suppl 14):118–128, 1974
6. Dabelsteen E, Mackenzie IC: The scientific basis for oral diagnosis. *In* Mackenzie IC, Squier CA, Dabelstein E (eds): Oral Mucosal Diseases: Biology, Etiology and Therapy. Copenhagen, Laegeforeningens Follag, 1987, pp 99–102
7. Drace JE, Young SW, Enzmann DR: TMJ meniscus and bilaminar zone: MR imaging of the substructure–diagnostic landmarks and pitfalls of interpretation. Radiology 177:73–76, 1990
7a. Dworkin SF, LeResche L, DeRouen T, et al: Assessing clinical signs of temporomandibular disorders: Reliability of clinical examiners. J Prosthet Dent 63:574–579, 1990
8. Dworkin SF, LeResche L, DeRouen T: Reliability of clinical measurement in temporomandibular disorders. Clinical J Pain 4:89–99, 1988
9. Ephros H, Samit A: Leukoplakia and malignant transformation. Oral Surg Oral Med Oral Pathol Oral Radiol Endod 83:187, 1997
10. Epstein JB, Oakley C, Millner A, et al: The utility of toluidine blue application as a diagnostic aid in patients previously treated for upper aerodigestive tract cancers. Oral Surg Oral Med Oral Path Oral Radiol Endod 83:537–547, 1997
10a. Fleiss JS, Chilton NW: The measurement of interexaminer agreement in periodontal disease. J Periodont Res 18:601, 1983
10b. Goulet J, Clark GT: Clinical TMJ examination methods. Journal of the California Dental Association 18:25–33, 1990
11. Greenstein G, Lamster I: Understanding diagnostic testing for periodontal diseases. J Periodontol 66:659–666, 1995
11a. Haffajee AD, Socransky SS, Goodson JM: Clinical parameters as predictors of destructive periodontal disease activity. J Clin Periodontol 10:257–265, 1983
12. Hansson LG, Westesson PL, Katzberg RW, et al: MR imaging of the temporomandibular joint: Comparison of joints of autopsy specimens made at 0.3 T and 1.5 T with anatomic cryosections. AJR Am J Roentgenol 152:1241–1244, 1989
13. Hennekens CH, Buring JE: Screening. *In* Mayrent SL (ed): Epidemiology in Medicine. Boston, Little, Brown and Co, 1987, pp 327–347
14. Hlatky MA, Mark DB, Harrell FE, et al: Factors affecting sensitivity and specificity of exercise electrocardiography. Am J Med 77:64–71, 1984
15. Hlatky MA, Mark DB, Harrell FE, et al: Rethinking sensitivity and specificity. Am J Cardiol 59:1195–1198, 1987
16. Karabulut A, Reibel J, Therkildsen MH, et al: Observer variability in the histologic assessment of oral premalignant lesions. J Oral Pathol Med 24:198–200, 1995
17. Kramer IRH: Basic histopathological features of oral premalignant lesions. *In* Mackenzie IC, Dabelstein E, Squier CA (eds): Oral Premalignancy. Iowa City, University of Iowa Press, 1980, pp 23–34
18. Kramer IRH: Prognosis from features observable by conventional histopathological examination. *In* Mackenzie IC, Dabelstein E, Squier CA (eds): Oral Premalignancy. Iowa City, University of Iowa Press, pp 304–311, 1980
18a. Lange JP: Clinical markers of periodontal disease. *In* Johnson NW (ed): Risk Markers for Oral Disease, vol. 3. Periodontal Disease, Markers of Disease Susceptibility and Activity. Cambridge, Cambridge University Press, 1991, pp 179
19. Mashberg A: Clinical features of oral malignancy in relation to prognosis. *In* Mackenzie IC, Dabelstein E, Squier CA (eds): Oral Premalignancy. Iowa City, University of Iowa Press, pp 292–334, 1980
20. Matthews DC, Banting DW: Authors' response. J Can Dent Assoc 62:7, 1996
21. Matthews DC, Banting DW, Bohay RN: The use of diagnostic tests to aid clinical diagnosis. J Can Dent Assoc 61:785–791, 1996
21a. Mileman PA, Vissus T, Pundell-Lewis DJ: The application of decision making analysis to the diagnosis of approximal caries. Community Dental Health 3:65–81, 1985
22. Norman GR, Streiner DL: PDQ Statistics. Toronto, Canada, Decker Inc, 1986

23. Okeson JP: Management of Temporomandibular Disorders and Occlusion. St. Louis, C.V. Mosby Co, 1989, pp 147–300
24. Orsini MG, Kuboki T, Terada S, et al: Clinical predictability of temporomandibular joint disc displacement. J Dent Res 78:650–660, 1999
25. Parker SL, Tong T, Bolden S, et al: Cancer statistics. CA Cancer J Clin 46:5–27, 1996
26. Pindborg JJ, Reibel J, Holmstrup P: Subjectivity in evaluation of oral epithelial dysplasia, carcinoma in situ and initial carcinoma. Journal of Oral Pathology 14:698–708, 1985
27. Rohlin M, Akerman S, Kopp S: Tomography as an aid to detect microscopic changes of the temporomandibular joint. Acta Odontol Scand 44:131–140, 1986
28. Rosenberg D, Cretin S: Use of meta-analysis to evaluate tolonium chloride in oral cancer screening. Journal of Oral Surgery 67:621–627, 1989
29. Sacket DL, Haynes RB, Guyatt, et al: Clinical Epidemiology. A Basic Science for Clinical Medicine, ed 2. Boston, Little, Brown and Co, 1991, pp 3–170
30. Sacket DL, Richardson WS, Rosenberg W, et al: Evidence-Based Medicine: How to Practice and Teach EBM. New York, Churchill Livingstone, 1997, p 127
31. Schechter MT, Sheps SB: Diagnostic testing revisited: Pathways through uncertainty. J Can Med Assoc 132:755–759, 1985
32. Schiffman E, Anderson GC, Fricton J, et al: Diagnostic criteria for intraarticular TM disorders. Community Dent Oral Epidemiol 17:252–257, 1989
33. Schwartz WB, Wolfe HJ, Pauker SG: Pathology and probabilities. A new approach to interpreting and reporting biopsies. N Engl J Med 305:917–913, 1981
34. Silverman SJR: Oral Cancer, ed 3. Atlanta, GA, American Cancer Society, 1990
35. Streiner DL, Norman GR: Health Measurement Scales. Oxford, Oxford University Press, 1989, pp 79–95
36. Tanimoto K, Peterson A, Rohlin M, et al: Comparison of computed with conventional tomography in the evaluation of temporomandibular joint disease: A study of autopsy specimens. Dentomaxillofacial Radiology 19:21–27, 1990
36a. Valachovic RW, Douglass CW, Berkey CS, et al: Examiner reliability in dental radiography. J Dent Res 65:432–436, 1986
36b. Vendonschotsh, Bronkhurst EM, Burgersdijk RCS, et al: Performance of some diagnostic systems in examinations for small occlusal caries. Caries Res 26:59–64, 1992
37. Westesson PL, Bronstein SL, Liedberg J: Temporomandibular joint: Correlation between single-contrast videoarthrography and postmortem morphology. Radiology 160:767–771, 1986
38. Westesson PL, Eriksson L, Kurita K: Reliability of a negative clinical temporomandibular joint examination: Prevalence of disk displacement in asymptomatic temporomandibular joints. Oral Surgery, Oral Medicine and Oral Pathology 68:551–554, 1989
39. Widmer CG: Physical characteristics associated with temporomandibular disorders. *In* Sessle BJ, Bryant PS, Dionne RA (eds): Temporomandibular Disorders and Related Pain Conditions, Progress in Pain Research and Management, vol 4. Seattle, IASP Press, 1995, pp 161–174
40. Wulff HR: Rational Diagnosis and Treatment. Oxford, Blackwell Scientific Publications, 1976

Address reprint requests to

Donald Maxwell Brunette, MSc, PhD
Department of Oral Biological and Medical Sciences
University of British Columbia
2199 Wesbrook Mall
Vancouver, British Columbia
Canada, V6T 1Z3

e-mail: brunette@interchange.ubc.ca