Published in final edited form as:

Am J Ophthalmol. 2010 June ; 149(6): 878–881.e2. doi:10.1016/j.ajo.2010.01.001.

Diagnostic Tests: Understanding Results, Assessing Utility, and Predicting Performance

ALAIN B. LABRIQUE and **WILLIAM K.-Y. PAN**

Global Disease Epidemiology and Control, Department of International Health and Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland (A.B.L., W.K.-Y.P.).

Clinical practice often is guided by the clever use of diagnostic tests to screen patients for the presence or absence of suspected disease or infection. Test results are woven together with specific patient information to decide on the most appropriate care for each individual. However, the selection of diagnostic procedures are a constant point of contention, as illustrated by the ongoing debate on the appropriate screening strategy for diabetic retinopathy.^{1–4} This type of debate inevitably will continue, given the rapid pace of development of diagnostic tests and procedures, particularly in emerging areas such as genomic and molecular medicine, where tests are being developed to screen not only for preclinical illness, but also for mere disease susceptibility. The selection of diagnostic tests in a clinical context often is not driven as much by the cost of each test, but rather by test performance, patient health, and even the risk of litigation in light of a false-positive or false-negative result. In contrast, widespread population-based screening weighs more heavily the cost effectiveness of a test as well as the public health significance of missed or misdiagnosed cases.⁵ From these points of view, one might assume that the underlying disease prevalence in a population is relevant only for population-based screening and not for clinical use; however, correct interpretation of test results in both contexts benefit from a priori knowledge of disease prevalence.

The commercial development of diagnostic tests pass through several generations to produce progressively better-performing assays over time.⁶ The tests are established using sample populations in which the disease state (ill or healthy) of each member is known through an accepted gold standard test. (A gold standard refers to a test that has 100% sensitivity and 100% specificity. However, most areas of medical science do not have true gold standard tests, and thus many new diagnostic tests undergo debate regarding their usefulness in detecting disease. Recent statistical advances have helped to improve the evaluation and interpretation of diagnostic tests in the absence of a true gold standard.^{7–9}) A test is evaluated based on 2 fundamental descriptive qualities: (1) the ability to classify patients correctly as sick or healthy, and (2) consistency of results across populations. These 2 qualities define a test's clinical utility (the likelihood a test will improve a patient's management or outcome) and clinical validity (consistency and accuracy of a test to predict

Inquiries to Alain B. Labrique, Global Disease Epidemiology and Control, Department of International Health and Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, E5508, 615 North Wolfe Street, Baltimore, MD 21205; alabriqu@jhsph.edu.

a patient's status). Clinical utility and validity are derived from statistical measures of sensitivity, specificity, positive predictive values (PPVs) and negative predictive values (NPVs), false- versus true-positive results, and true-versus false-negative results. These measures are described below.

SENSITIVITY VERSUS SPECIFICITY

SENSITIVITY AND SPECIFICITY REFER TO THE INTRINSIC ability of a test, usually independent of population context, to identify correctly those with the disease as abnormal (sensitivity) and those without a disease as normal (specificity) when compared with the performance of a gold standard.¹⁰ Often, a dichotomous interpretation of a diagnostic test's results of classifying a patient into 1 of 2 distinct groups is the most practical. Figure 1 illustrates a typical evaluation of a new dichotomous diagnostic test being evaluated against a gold standard. The true-positive rate, or sensitivity, is calculated as $[a/(a + c)]$; the true-negative rate, or specificity, is $[d/(b + d)]$. These characteristics also can be explained as conditional probabilities.¹⁰ The test in our example (Figure 1) has an 82.2% sensitivity, which is the percent of truly diseased individuals identified, and a 66.4% specificity, which is the percent of healthy individuals correctly identified. Standard statistical techniques can be used to calculate confidence intervals for sensitivity and specificity.¹¹

SELECTING CUTOFFS: WHEN CONTEXT MATTERS

ALTHOUGH DICHOTOMOUS RESULTS ARE COMMON IN THE laboratory context, many diagnostic tests are continuous, on a scale, accompanied by an indication of whether the result exceeds the bounds of what is considered normal for that biomarker. In both screening and clinical treatment, it is necessary to establish cutoff values to categorize individuals as diseased or nondiseased. However, this is sometimes problematic, because real population variations can result in different normal ranges of some biologic parameters. The degree of acceptable false positives and false negatives will depend on the consequences of a incorrect diagnosis—a process that includes considerations of whether better (often more expensive) assays are available as a second tier of confirmatory testing, the costs (financial and psychological) of such testing relative to immediate treatment, and the risk to the patient if the condition is not identified early. As illustrated in Figure 2, depending on what cutoff value (α) is chosen, a subject may be identified correctly as sick (true positive) or healthy (true negative), and a certain rate of ascertainment error can be expected (false positive, false negative). In our example, a cutoff has been chosen to minimize the rate of false negatives, or individuals incorrectly labeled as healthy, at the expense of a larger number of false positives. This would be appropriate for conditions for which misdiagnosing and treating someone as sick is less egregious than missing truly sick individuals.

Another important consideration in selecting an appropriate cutoff for an assay is to consider the population context for which or in which the assay was developed along with the site of its intended use. Enzyme-linked immunosorbent assays often use colorimetric or fluorescent signals that are correlated with the concentration of the target antibody. In contexts of high endemicity of infection, where antigenic overlap may exist with other cocirculating pathogens, or even epitopic cross-reactivity resulting from particular population genetics, the

cutoff value established in a nonendemic, largely seronaïve population may prove to be inappropriate. In such cases, one may need to redefine appropriate cutoff values to improve test characteristics (sensitivity, specificity) and performance (PPVs and NPVs). Such a process is elegantly described by Laeyendecker and associates for a commercial herpes simplex type 2 enzyme-linked immunosorbent assay developed in a Western setting, but intended for use in an epidemiologic study in Uganda.¹²

RECEIVER OPERATING CHARACTERISTIC CURVES

RECEIVER OPERATING CHARACTERISTIC CURVES WERE adapted in the early 1990s by biostatisticians to help guide the selection of cutoffs for diagnostic tests that maximize sensitivity and specificity in a given population.^{13,14} Receiver operating characteristic plots are developed by calculating the sensitivity and specificity of a test for every data point collected in a validation study, and then by plotting sensitivity (true positive rate, y-axis) against 1-specificity (false positive rate, x-axis). Perfect discrimination of sick and healthy individuals would cluster at the top left corner of the plot, whereas a useless test (equivalent to a coin toss), would generate a perfect diagonal line running from the origin (0,0) to the top right corner of the graph.^{10,14} The ideal cutoff point for a diagnostic test therefore is a point along the receiver operating characteristic curve for a particular diagnostic test that is nearest to the top left corner ($x = 0$, $y = 1$), representing maximum sensitivity and specificity.

PREDICTIVE VALUES: REAL-WORLD PERFORMANCE

MORE IMPORTANT TO MOST CLINICAL PRACTITIONERS than understanding these intrinsic properties of an assay is the diagnostic capability of the test in the real world, when the true disease status of a patient is not known. Although they are the most commonly referred to characteristics of a diagnostic test, a test's sensitivity and specificity do not provide the clinician with the probability that a diagnostic test will provide a correct diagnosis. The PPV and NPV describe the proportion of patients with positive test results or negative test results, respectively, who are identified correctly. Using the example in Figure 1 to illustrate this concept, only 198 of the 236 individuals labeled as abnormal ($a + b$) by the new diagnostic were correct, yielding a PPV of 83.9% ($a/(a + b)$). Among the 118 classified as healthy ($c + d$), only 75 were correct, yielding an NPV of 63.6% ($c/(c + d)$). It is critical to note that PPV and NPV are performance characteristics, and not intrinsic values of the test—they are influenced strongly by the prevalence of the disease or condition in the population of interest. In our example (Figure 1), the prevalence of the condition being evaluated is 68.1%—almost an order of magnitude greater than most conditions seen in a clinical setting.¹⁵ This phenomenon is illustrated in Figure 3, where we see how, even when the sensitivity and specificity of a test are fairly good (90% for both, top line), under conditions of low population prevalence, the PPV of a test will be low. Altman and Bland published some useful formulae for the calculation of PPV and NPV under any estimated prevalence, illustrating how, under conditions of extremely low prevalence, the likelihood of false positives (Figure 1, (b)) will be high, despite excellent sensitivity and specificity of the test being used.¹⁶

$$PPV = \frac{\text{Sensitivity} \times \text{Prevalence}}{\text{Sensitivity} \times \text{Prevalence} + (1 - \text{Specificity}) \times (1 - \text{Prevalence})}$$

$$NPV = \frac{\text{Specificity} \times (1 - \text{Prevalence})}{(1 - \text{Sensitivity}) \times \text{Prevalence} + \text{Specificity} \times (1 - \text{Prevalence})}$$

The NPV of a test, however, will increase as the prevalence of the disease or condition decreases.¹⁵ Another indicator of test accuracy is the overall percentage of correct test results, which reflects the concordant cells in Figure 1 $[(a + d)/(a + b + c + d)]$. However, prevalence also skews the interpretation of this metric, for example, in low prevalence situations, a test with poor sensitivity would result in high concordance.¹⁴ Others have proposed the inclusion of the financial costs of misclassification as part of an improved method to compare and select the most appropriate test.¹⁷

Sensitivity and specificity, intrinsic properties of a diagnostic test, speak to the technical performance of a particular assay, technique, or method in correctly classifying individuals into healthy or ill groups. These are useful benchmarks to evaluate the capability of a new test in relation to an accepted gold standard. PPVs and NPVs, although more useful to guide the clinical interpretation of results, are highly subject to the prevalence of a condition in a population. The use of available contextual data (prevalence, test cost, misclassification consequences) should be incorporated into the selection of an appropriate test as well as into the evaluation of a diagnostic test's results. Depending on the task at hand (population screening vs patient care), different test qualities may be desirable—high specificity for screening, high sensitivity for patient care. However, clinicians should interpret the results of even highly sensitive tests carefully, with some consideration of the prevalence of the disease in the population in which they are working. Many excellent reviews of diagnostic tests provide more detail on the above topics, including guidance for the design and conduct of studies to assess diagnostic test accuracy.^{10,18–19}

Biography

Alain B. Labrique is an Assistant Professor in Global Disease Epidemiology and Control in the Departments of International Health and Epidemiology (Joint) at the Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland. Trained in infectious disease epidemiology, Dr. Labrique conducts large community-based epidemiologic studies focused on reducing maternal and neonatal mortality and morbidity, particularly in South Asia. His expertise ranges from Hepatitis E and bacterial vaginosis to exploring the role of micronutrient status in moderating immunity.

William Kuang-Yao Pan is an Assistant Professor in Global Disease Epidemiology and Control in the Department of International Health at the Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland. Dr. Pan is a biostatistician with training in remote sensing, spatial analysis, mathematical demography, and certain aspects of ecology related to disease transmission. His research seeks to foster a deeper understanding of demographic processes, human health and environmental change using tools from biostatistics, geography, and economics.

REFERENCES

1. Harding SP, Broadbent DM, Neoh C, White MC, Vora J. Sensitivity and specificity of photography and direct ophthalmoscopy in screening for sight threatening eye disease: the Liverpool Diabetic Eye Study. *BMJ*. 1995; 311:1131–1135. [PubMed: 7580708]
2. James M, Turner DA, Broadbent DM, Vora J, Harding SP. Cost effectiveness analysis of screening for sight threatening diabetic eye disease. *BMJ*. 2000; 320:1627–1631. [PubMed: 10856062]

3. Leese G, Broadbent D, Harding S, Vora J. Screening for diabetic retinopathy. Approaching 90% sensitivity with new techniques. *BMJ*. 1995; 311:1230–1231.
4. Klein BE, Davis MD, Segal P, et al. Diabetic retinopathy. Assessment of severity and progression. *Ophthalmology*. 1984; 91:10–17. [PubMed: 6709313]
5. Wildschut HI, Peters TJ, Weiner CP. Screening in women's health, with emphasis on fetal Down's syndrome, breast cancer and osteoporosis. *Hum Reprod Update*. 2006; 12:499–512. [PubMed: 16807275]
6. Constantine, NT. HIV antibody testing.. In: Cohen, PT, Sande, MA, Volberding, PA, editors *The AIDS Knowledge Base*. Lippincott Williams & Wilkins; Philadelphia: 1999. 105–118.
7. Engel B, Swildens B, Stegeman A, Buist W, de Jong M. Estimation of sensitivity and specificity of three conditionally dependent diagnostic tests in the absence of a gold standard. *Journal of Agricultural, Biological & Environmental Statistics*. 2006; 11:360–380.
8. Garrett ES, Eaton WW, Zeger S. Methods for evaluating the performance of diagnostic tests in the absence of a gold standard: a latent class model approach. *Stat Med*. 2002; 21:1307.
9. Joseph L, Gyorkos TW, Coupal L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *Am J Epidemiol*. 1995; 141:263–272. [PubMed: 7840100]
10. Shapiro DE. The interpretation of diagnostic tests. *Stat Methods Med Res*. 1999; 8:113–134. [PubMed: 10501649]
11. Gardner, MJ, Altman, DG. Calculating confidence intervals for proportions and their differences.. In: Gardner, MJ, Altman, DG, editors *Statistics with Confidence*. BMJ Publishing Group; London: 1989. 28–33.
12. Laeyendecker O, Henson C, Gray RH, et al. Performance of a commercial, type-specific enzyme-linked immunosorbent assay for detection of herpes simplex virus type 2-specific antibodies in Ugandans. *J Clin Microbiol*. 2004; 42:1794–1796. [PubMed: 15071053]
13. Altman DG, Bland JM. Diagnostic tests 3: receiver operating characteristic plots. *BMJ*. 1994; 309:188. [PubMed: 8044101]
14. Bewick V, Cheek L, Ball J. Statistics review 13: receiver operating characteristic curves 1. *Crit Care*. 2004; 8:508–512. [PubMed: 15566624]
15. McCarty CA, Mukesh BN, Giampietro PF, Wilke RA. Healthy People 2010 disease prevalence in the Marshfield Clinic Personalized Medicine Research Project cohort: opportunities for public health genomic research. *Personalized Medicine*. 2007; 4:183–190. [PubMed: 29788632]
16. Altman DG, Bland JM. Diagnostic tests 2: Predictive values. *BMJ*. 1994; 309:102. [PubMed: 8038641]
17. Adams NM, Hand DJ. An improved measure for comparing diagnostic tests. *Comput Biol Med*. 2000; 30:89–96. [PubMed: 10714444]
18. 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 5. *JAMA*. 1994; 271:703–707. [PubMed: 8309035]
19. Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? Evidence-Based Medicine Working Group 6. *JAMA*. 1994; 271:389–391. [PubMed: 8283589]

		Gold Standard		
		Abnormal (+)	Normal (-)	
New Diagnostic Test	Abnormal (+)	True Positives 198 (a)	False Positives 38 (b)	236
	Normal (-)	False Negatives 43 (c)	True Negatives 75 (d)	118
		241	113	

FIGURE 1.

An evaluation of a new diagnostic test, compared with a known gold standard assessment. There are 241 individuals in this population known to be ill ($a + c$), and 113 who are considered normal ($b + d$). Using this 2×2 table approach, we can assess the sensitivity, specificity, and predictive strengths of the new test, as described.

FIGURE 2.

Graph showing the consequences of cutoff selection for a continuous diagnostic test, such as antibody levels or low-density lipoprotein (LDL) cholesterol values. Depending on what cutoff value (α) is chosen, a subject may be identified correctly as sick (“True Positives”) or healthy (“True Negatives”), and a certain rate of ascertainment error can be expected (“False Positives,” “False Negatives”). In this example, a cutoff has been chosen to minimize the rate of false negatives, or individuals incorrectly labeled as healthy, at the expense of a larger number of false positives. This would be appropriate for conditions for which mislabeling and treating someone as sick is less egregious than missing truly sick individuals.

FIGURE 3.

Graph depicting the effect of disease prevalence on positive predictive value at 3 levels of test sensitivity (specificity held constant at 90%).