Probabilistic Aspects of Medical Testing: Test Results, Test Performance and Medical Decision Making

George R. Bergus, M.D.
Department of Family Practice
University of Iowa
Iowa City, Iowa

Abstract: Clinicians order tests for the diagnostic information contained in test results. Laboratorians focus on the analytical performance characteristics of their tests, but the performance characteristics of concern to clinicians are test sensitivity and specificity. Test results do not directly provide the diagnostic information that clinicians seek, but Bayes' Theorem allows clinicians to use the results to make diagnostic assessments. This probabilistic approach requires appraisal of a patient's pre-test probability of disease and knowledge of a test's likelihood ratio. When the test is interpreted as a dichotomous outcome, the likelihood ratio is calculated from the test's sensitivity and specificity. A further refinement is to use the full information available in a test result by using result-specific likelihood ratios to revise probability assessments.

Estimating test sensitivity and specificity can be biased by methodologic problems which include spectrum bias, test referral bias, reference test bias, and sampling variability. These biases need to be recognized and avoided, although occasionally researchers either ignore or cannot avoid these problems.

The information contained in a test result cannot be appropriately used if clinicians disregard Bayes' Theorem or researchers use biased methodologies to assess a test's performance characteristics. Because of these factors, improved analytic performance in the laboratory might not result in the clinician having greater knowledge about the health state of a patient.

Probability Revision

Because of the inherent error in most clinical tests, using a test result in clinical medicine is a complex procedure. The clinician might attempt to use a test result alone to determine whether the patient is diseased or healthy, but this simplistic approach can lead to incorrect and dangerous conclusions. Instead, the test result should be used to revise the probability of disease that the physician had before testing by the use of Bayes' theorem. The post-test probability of disease is determined by the test result, the probability of disease before testing, and the performance characteristics of the test.[1]

It is possible to have a negative test result despite the presence of disease. Imagine that a clinician estimates a patient has a 90% pretest probability of disease and decides to confirm his/her impression with a test. If the test comes back "negative," the clinician could decide that either there is a laboratory error, or the patient does not have the disease. Instead, the clinician needs to appropriately interpret the test result by asking about the probability of disease given the negative test. This probability is easily calculated once one has an estimate of the test performance characteristics; we will assume that the sensitivity of this test is 90% and the specificity is 80%. Bayes' theorem

indicates that the probability of disease despite the negative test is 53% (appendix 1, calculation 1). Therefore, despite a negative test, the patient has a slightly better than even chance of having the disease. Similarly, a clinician can end up with a non-diseased patient with a positive test. While incongruent results can result with the clinician demanding a "better" test than the one the laboratorian is providing, the clinician should consider a better method of using the test information.

Bayes' theorem permits a new piece of information to be interpreted within the context of prior knowledge. This approach requires that the new piece of information be given an explicit weight known as a likelihood ratio (LR). The LR is the probability of a certain finding in individuals with “disease X” divided by the probability of the same finding in individuals without the disease. Although Bayes' theorem has been available for over 2 centuries, it has not become the standard method by which clinicians interpret a piece of laboratory data. Currently, for clinicians to use Bayes' theorem in their work, they have to take the report from the laboratory, look up the LR for the test result in a textbook or journal article and then calculate the post test probability. This multi-step procedure does not invite probability revision. It is possible that clinicians could be encouraged to use Bayes' theorem if the laboratorian provided on the lab report both the numerical result and its associated LR. Additionally, to ease the computational burden that comes with probability revision, the lab report could incorporate simple Bayesian nomograms.[2,3]

A test result is measured on a continuous scale but frequently used for Bayesian probability revision as a dichotomous (i.e., positive or negative) outcome. The dichotomized results are given one LR if positive (the LR+) and another if negative (the LR-). These LRs can be easily calculated because they are directly derived from the sensitivity and specificity of a test; the LR+ is the sensitivity of the test divided by [1 - specificity of the test] and the LR- is [1 - sensitivity] divided by the specificity. While dichotomizing the test result makes it easier for the clinician to use Bayes' theorem, it also degrades the available information from the test because, regardless of how extreme, there is only one LR for all "positive" results and a single LR for "negative" results.

A simple laboratory test, urine microscopy, can serve as an illustration of how information can be lost.[4] Table 1 is the 2 by 2 table for the urinalysis when 5 or greater WBC per hpf is considered a "positive" result. The LR+ for urine pyuria is 4.0 (appendix 1, calculation 1). Because of the dichotomizing, 5 WBC/hpf, has the same Bayesian weight as 10 WBC/hpf which is both intuitively objectionable and conceptually unsound. A refinement is to increase the number of categories a result can be placed into, so that unique LRs are assigned to narrower ranges of test results. As the number of categories is increased, the data from which the LRs are calculated become increasing sparse. Table 2 contains LRs calculated directly from the data set which has been partitioned into 6 levels. As can be noted, by using these additional levels, 5-9 WBC/hpf now has a different LR than 10-14 WBC/hpf although, because of sampling variability, the LRs do not monotonically increase with increasing number of WBC/hpf. When a stratified analysis is used on this small data set, one could conclude that 5 WBC/hpf is more supportive of urinary tract infection (UTI)

	UTI Present	UTI Absent
WBC ≥ 5	171	32
WBC < 5	67	148
Total People	238	180

Table 1. Microscopic pyuria data from Ferry et al [4] placed into a 2 by 2 table.

Patients with UTI	Leukocyte Count	Patients without UTI	Likelihood Ratio
124	≥ 15 WBC/hpf	14	6.70
14	10-14 WBC/hpf	11	0.96
33	5-9 WBC/hpf	7	3.57
22	3-4 WBC/hpf	22	0.76
21	1-2 WBC/hpf	49	0.32
24	0	77	0.24
238	Total Patients	180	

Table 2. Microscopic pyuria data adapted from Ferry et al. [4] and placed into 6 test-result intervals.

Leukocytes on Micro UA	Calculated LR
15 WBC/hpf	2.23
10 WBC/hpf	1.59
5 WBC/hpf	0.98
3 WBC/hpf	0.62
1 WBC/hpf	0.33

Table 3. Using the microscopic pyuria data found in Table 2, the LRs have been calculated using a MLE algorithm[5] and ROC curve analysis. [6]

than is 10 WBC/hpf!

If the modeling approach is pushed further, an unique LR can be assigned for each and every level of a test result. To deal with the problem of sparse data, the LRs can be determined using statistical techniques and modeling. Table 3 contains result-specific LRs which were calculated from the original data set at 5 different levels of WBC/hpf using a MLE estimator[5] and ROC curve analysis.[6] Alternately, logistic modeling can be used to calculate an LR at any level of WBC/hpf.[7,8] Calculating result-specific LR is beyond the skills of most clinicians but could be provided by a sophisticated laboratory information system and then attached to the lab report sent to the clinician.

Biased assessment of test performance

A second major challenge to the use of Bayes' theorem in clinical medicine is the need for unbiased estimates of a test result's LR. Because post-test probability of disease is directly related to the estimates of test performance, precise and accurate assessment of test performance is essential. Diagnostic test performance is assessed by identifying two groups of diseased and nondiseased patients and then observing how the test classifies these people. Biased estimates of test performance result in biased estimates of the post-test probability. Common biases affecting the assessment of test performance can be divided into two broad categories.[9] The first category pertains to how subjects are selected for assessing the test. The second category of biases are methodologic in origin. Before looking at these biases in greater detail, we first need to focus on two basic definitions: The first is the gold standard test, which defines the truth, as well as we can know it, about a patient's condition. The second is the index test, which we typically use in practice for information about a patient's condition because gold standard test is too expensive, too dangerous or not available.

It has been widely believed that sensitivity and specificity are qualities of a test invariant to the population selected.[10] While this immutability is attractive, it is also a misconception. Severe disease is generally easier to detect than mild disease, and therefore, the sensitivity of a test will, in part, be determined by the severity of disease in the diseased subjects being tested. This bias, known as spectrum bias, is common because many tests are developed in academic medical centers where the spectrum of disease can be very different than in a community hospital.[11] Spectrum bias can also distort the measured sensitivity of a test if researchers attempt to avoid misclassifications by using only patients they are highly certain of having the disease. This approach also gathers very extreme cases of disease.

An example of spectrum bias can be found in research focusing on the sensitivity of the urine dipstick to diagnose UTI. The range reported in the literature is wide, estimated from 66% to 100%,[12] suggesting that some of the variation in the estimates arise from the patients selected to define the sensitivity of the test. In an interesting study, Lachs calculated sensitivity of the dipstick in subgroups of patients stratified by pretest probability of UTI.[13] The dipstick had excellent sensitivity, 92%, in patients with extreme symptoms and a high clinical probability of infection. In contrast, the sensitivity in patients with few symptoms and a low probability of infection was only 56%. Whether the dipstick is a sensitive test for UTI depends on the spectrum of disease

being tested.

Spectrum bias can also impact test specificity because this measure is related to selecting controls. Naturally, if healthy medical students are used as controls, the test usually correctly identifies them as nondiseased and therefore demonstrates a very high specificity. Of greater importance is whether the test correctly identifies nondiseased patients who have signs and symptoms easily confused with the disease. Returning to the example of the urine dipstick, the literature contains a wide range of estimates of specificity for this test, from 60% to 98.4%. The study by Lachs also confirms that the specificity is greatly dependent on the patients in the nondiseased group. In noninfected patients, clinically with a low probability of UTI, the specificity was 78%; but in patients with a high probability of UTI, the specificity was much lower at 42%.

The solution to the problem of spectrum bias is to have the developers of a test clearly define the spectrum of disease in their population and to use reasonable controls for determining the specificity. To help clinicians use the appropriate test characteristics in their clinical populations, test developers could report LRs for well identified subgroups of patients.

A second type of bias, work up/verification bias, arises from researchers using the index test to decide which patients will also undergo the gold standard test. The size of this bias is directly related to how tightly the index test result is used as a selection criterion. If only patients with positive index tests are sent for gold standard tests, the index test will appear to have a sensitivity of 100% because all individuals with a positive gold standard test also have a positive index test. In this situation, the specificity of the index test will appear to be 0% because all individuals with a negative gold standard test also have a positive index test. In reality, work up/verification bias is rarely this extreme or as obvious.

A more subtle case of this bias arises if 100% of persons with a positive index test but only 20% of people with a negative index test were sent for a dangerous biopsy, the gold standard test in this example. Imagine 100 persons with positive index tests, all of whom are sent for biopsy. Eighty of the biopsies return positive. In contrast, another 100 persons have negative index tests, but only 20 are sent for biopsy. These 20 biopsies yield 10 positive results. The sensitivity of the index test appears to be 88.8% because of the 90 patients with positive biopsies 80 had positive index tests. In truth, the sensitivity of the test is much lower because many individuals with negative index tests were not included in the calculation. After realizing that only one fifth of the individuals with negative index tests ended up with biopsy, the clinician should estimate the sensitivity to be 61.5% (appendix 1, calculation 3).

This source of bias in the calculated performance of a test might seem obvious and easy to control. All patients with an index test need to undergo the gold standard test; however, because many gold standard tests are dangerous or very expensive, this control is not always instituted. Alternately, as illustrated in the above calculation, not all index test negative patients need to undergo the gold standard if a random subgroup of these patients need to be referred for this test.

A third common source of bias arises from the gold standard and is known as reference test bias. Although the performance characteristics of an index test

are quoted relative to the true disease state of a patient, they are actually calculated relative to another fallible test, the gold standard. If we assume that the gold standard determines the true health state, then we will ignore any classification errors made by the gold standard. When the index test is discordant with the gold standard, we assume that the index test is imperfect. The misclassification, however, could be on the part of the gold standard.

The relationship between an index test's true performance and its observed performance is predictably related to the prevalence of disease when the index and gold standard tests are conditionally independent (appendix 1, equation 1).[14] The observed sensitivity of the index test approaches its true value when the prevalence of disease approaches 100%; if all subjects in a population are diseased, the gold standard can no longer misclassify nondiseased individuals as diseased. At lower disease prevalence we will observe a lower sensitivity for the index test. The observed specificity of an index test will approach its true value when the prevalence of disease approaches 0% for a similar reason.

In truth, the index and gold standard tests are generally conditionally dependent, causing the relationship between index test characteristics and disease prevalence to be variable (appendix 1, equation 2). The observed sensitivity of an index test can increase, decrease, or remain unchanged with a rise in disease prevalence.[15]

This source of bias is particularly worrisome. While we need to use the best possible gold standard test, it is all too easy to fall into a circular argument about true disease state and gold standard test result. The effect of this bias is predictable and correctable if the index and gold standard tests are conditionally independent.[16] In the face of conditional dependencies, however, the size and direction of the bias cannot be predicted unless the true performance of the index and gold standard tests are known.[9]

The final source of bias to be discussed in this paper arises from sampling variability impacting the sensitivity, specificity and LR reported for a test.[17] These estimates of test performance can be numerically unstable if too few patients have been evaluated. In general, the larger the study the more stable the estimate of performance. For example, a test might have a sensitivity that has been reported to be 80%. When this estimate is based on 50 subjects, the 95% CI of this estimate is quite wide and ranges from 68.9% to 91.1%. When the same estimate of sensitivity is made based on 250 subjects, the 95% CI is narrower, 75.0% to 85.0%. When the estimate is based on 10,000 subjects, the 95% CI is 79.2% to 80.8% and the estimate is quite precise. It is obvious that this source of bias can be controlled by using large sample sizes for quantifying the performance of a test, but this is not always done because of expense, time, or the scarcity of patients with a certain disease. The clinician needs to be aware that although the analytic process behind a test might be very precise, the data on the sensitivity and specificity of the test might be unstable.

Summary

In this paper, we focused on probabilistic aspects of medical testing and presented ways for laboratorian to encourage clinicians to use this approach. The laboratory could attach result-specific LRs to the lab report and integrate computation tools into the lab report in the form of nomograms. We have also detailed four

common sources of bias that impact the LR calculated for a test result and suggested means for limiting their impact.

The clinician uses testing to obtain information about an individual, often in hopes of clarifying a clinical situation. Since tests have inherent error, their results need to be interpreted within the clinical context that the physician is hoping to clarify. It is possible that, by using Bayes' theorem with unbiased result-specific LRs, a clinician could obtain more information from a test result than is currently available.

References

1. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical epidemiology. A basic science for clinical medicine. 2nd ed. Boston/Toronto/London:Little, Brown and Company, 198S:23.

2. Fagan TJ. Nomogram for Bayes's theorem. *N Engl J Med.* 1975;293:2S7.

3. Glasziou PP. Probability Revision. *Pri Care.* 199S;22:235-24S.

4. Ferry S, Andersson S, Burman LG, Westman G. Optimized urinary microscopy for assessment of bacteriuria in primary care. *J fam.. Pact.* 1990;3l:lS3-161.

5. Dorfman DD, Alf E. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals-rating-method data. *J Math Psychol.* 1969;6:487-96.

6. Bergus GR. When is a test positive? The use of decision analysis to optimize test interpretation. *Fam Med.* 1993;S:6S660.

7. Albert A. On the use and computation of likelihood ratios in clinical chemistry. *Clin Chem*. 1982;28:1113-1119.

8. Knottnerus JA. Application of logistic regression to the analysis of diagnostic data: Exact modeling of a probability tree of multiple binary variables. *Med Decision Making.* 1992;12:93-108.

9. Begg CB. Biases in the assessment of diagnostic tests. *Stat in Med.* 1987;6:411-423.

10. Diamond GA. Clinical epistemology of sensitivity and specificity. *J Clin Epidemiol.* 1992;4S:9-13

11. Salive ME. Referral bias in tertiary care: The utility of clinical epidemiology. *Mayo Clin Proc.* 1994;69:808-809.

12. Kellogg JA, Manzella JP, Shaffer SN, Schwartz BB. Clinical relevance of culture versus screens for the detection of microbial pathogens in urine specimens. *Am J Med.* 1987;83:739-4S.

13. Lachs MS, Nachamkin I, Edelstein PH, Goldman J, Feinstein AR, Schwartz JS. Spectrum bias in the evaluation of diagnostic tests: Lessons from the rapid dipstick test for urinary tract infection. *Ann Intern Med.* 1992;117:135-140.

14. Boyko EJ, Alderman BW, Baron AE. Perspectives. Reference test errors bias the evaluation of diagnostic tests for ischemic heart disease. *J Gen Intern Med*. 1988;3:476-481.

15. Bergus GB, Witte DL. Predicting the impact of reference test bias on the observed test characteristics of an index test. *Med Decision Making*. 1994;14:425.

16. Diamond GA, Rozanski A, Forrester JS, et al. A model for assessing the sensitivity and specificity of tests subject to selection bias. Application to exercise radionuclide ventriculography for diagnosis of coronary artery disease. *J Chronic Dis*. 1986;39(5):343-S5.

17. Arkin CF, Wachtel MS. How many patients are necessary to assess test performance? *JAMA*. 1990;263(2):27S-278.

Appendix 1

Calculation 1

$$0.53 = \frac{0.90 * 0.10}{(0.10 * 0.80) + (0.90 * 0.10)}$$

$$\text{post-test probability} = \frac{\text{pre-test probability} * \text{sensitivity}}{(\text{1-pre-test probability}) * (\text{1-specificity}) + (\text{pre-test probability} * \text{sensitivity})}$$

Calculation 2

$$\frac{171 / 238}{32 / 180} = 4.0 = LR+$$

Calculation 3

$$\frac{\text{true positives}}{\text{true positives} + \text{false negatives}} = \text{Sensitivity}$$

$$\frac{80}{80 + (10 * 5)} = 61.5\%$$

Equation 1

$$S_{observed} = \frac{[p\, S_{index\ test} S_{imperfect\ reference} + (1 - p)(1 - SP_{imperfect\ reference})(1 - SP_{index\ test})]}{[p\, S_{imperfect\ reference} + (1 - p)(1 - SP_{imperfect\ reference})}$$

S is the sensitivity of a test, SP is the specificity of a test and p is the prevalence of disease.

Equation 2

$$S_{observed} = \frac{[p\, \beta_1\, S_{imperfect\ reference} + (1 - p)(1 - SP_{imperfect\ reference} - SP_{index\ test} + \beta_2 S_{imperfect\ reference})]}{[p\, S_{imperfect\ reference} + (1 - p)(1 - SP_{imperfect\ reference})]}$$

β_1 is defined as the percentage of actual disease that is positive by the imperfect reference and also positive on the index test, and β_2 is the fraction of people without disease and negative by imperfect reference who are is negative on the index test. These Betas are measures of the dependence between the imperfect reference and the index in diseased and non-diseased populations.