ROC Anatomy—Getting the Most Out of Your Diagnostic Test

Amiran Baduashvili, MD[1], *Gordon Guyatt, MD, MSc*[2], *and Arthur T. Evans, MD, MPH*[1]

[1]Section of Hospital Medicine, Division of General Internal Medicine, Weill Cornell Medical College, New York, NY, USA; [2]Department of Clinical Epidemiology and Biostatistics, Faculty of Health Sciences, McMaster University, Hamilton, Ontario, Canada.

Clinical decision-making relies heavily on making a correct diagnosis. Clinicians have a responsibility to understand the full spectrum of the diagnostic information conveyed by a physical exam finding, laboratory test result, or imaging. Many laboratory tests, such as troponin and B-type natriuretic peptide (BNP), are continuous tests with many possible results. Yet, there is a tendency to dichotomize tests into positive and negative, and use sensitivity and specificity to describe the test characteristics. This approach can lead to waste of important diagnostic information and substandard clinical decision-making. The aim of this paper is to demonstrate the role of ROC curves in developing a more comprehensive understanding of diagnostic information portrayed by continuous tests to augment clinical decision-making.

KEY WORDS: ROC curve; likelihood ratio; diagnosis; Bayesian theorem; troponin; b-type natriuretic peptide.

J Gen Intern Med 34(9):1892–8
DOI: 10.1007/s11606-019-05125-0

CLINICAL SCENARIO

A 65-year-old man presents to the emergency department with chest pain. Based on the history, exam, and electrocardiogram, the clinician estimates the probability of acute myocardial infarction (MI) at 30%. A highly sensitive troponin test (hsT) sent on presentation has a value of 5 pg/mL. In a study evaluating test characteristics of hsT in diagnosing acute MI, the authors reported the test characteristics at two possible cutpoint values—at the 99th percentile (34 pg/mL, sensitivity 82%, specificity 92%) and at the test's limit of detection (LOD, 3 pg/mL, sensitivity 100%, specificity 35%).[1]

The patient's test result (5 pg/mL) lies between these two cutpoints. If the clinician chooses to use the lower cutpoint (3 pg/mL), the associated likelihood ratio (LR) of a positive test (LR+), 1.5, will increase the odds of the target condition by a factor of 1.5, and increase the probability of acute MI from 30 to 39%. If, however, the clinician uses the 99th percentile to define a positive test (34 pg/mL), then the associated likelihood ratio of a negative test (LR−), 1/5, or 0.2, reduces the odds of the target condition by a factor of 5 (to post-test probability of 8%). Although many clinicians would forgo further evaluation for acute MI if, after initial testing, the probability falls below 1%, physicians will act with more urgency—with additional tests, consultations, and presumptive therapies—as the probability of acute MI rises. Given these data, it is not immediately clear which cutpoint—and which post-test probability—is more appropriate to use.

Prior presentations: *6/6/2018 – McMaster Evidence Based Clinical Practice workshop*

Electronic supplementary material *The online version of this article (https://doi.org/10.1007/s11606-019-05125-0) contains supplementary material, which is available to authorized users.*

Received July 26, 2018
Revised February 5, 2019
Accepted May 21, 2019
Published online July 3, 2019

INTRODUCTION

Diagnosis is a central clinical task. Although most clinicians have a gestalt regarding how to adjust the likelihood of competing diagnoses based on test results, many are not skilled in using quantitative information to move from pre-test to post-test probabilities. This may be particularly the case when a test is continuous (such as troponin, procalcitonin, d-dimer), rather than dichotomous (a test with two outcomes—positive and negative, such as a pregnancy test).

Clinicians have an intuitive sense that the magnitude of an abnormal test result matters. For instance, as we will illustrate in the next section, a troponin value 10 times the upper limit of normal increases the probability of cardiac ischemia far more than a result less than 2 times the upper limit of normal. Yet, with few exceptions,[2–4] article authors present quantitative test results with a single threshold—a waste of information and sometimes diagnostically dangerous.

Clinicians seeking to take full advantage of diagnostic test results may consult articles that provide a formal assessment of the test's diagnostic properties, and these articles will frequently include receiver operating characteristic (ROC) curves. In this article, we will describe how to use ROC curves to make valid diagnostic judgments.

USING PRECISE NUMBERS IN DIAGNOSIS

To appreciate the wealth of information embedded in ROC curves, one must understand some of the arithmetic of diagnosis, including the role of LRs.[5, 6] Clinicians are

generally more comfortable thinking in terms of probability—often used interchangeably with "risk"—rather than odds. In the context of diagnosis, probability refers to the proportion or percentage of patients with a target condition (e.g., the pre-test probability is 0.6 or 60%), while odds refers to the relative frequency of a target condition being present versus absent, or probability/1–probability (e.g., 60%/40%, or 1.5). Online supplement Box 1 and online supplement Fig. 1 illustrate risks and odds.

LRs represent the ratio of two probabilities: the probability that patients with the target condition have a particular test result divided by the probability that patients without the target condition have the same test result (online supplement Box 1). Consider, for instance, the diagnosis of chronic HIV, using ELISA for HIV antibodies. The test will be positive in nearly everyone with chronic HIV (100% sensitivity), and negative in approximately 99% of those without HIV (99% specificity). Suppose we test a patient with a pre-test probability of HIV of 1% (probability = 1/100; odds of HIV = 1/99). Since LRs indicate how the odds of HIV changes with a test result, a positive test—with LR+ = 1/(1–0.99) = 100—means that the odds of chronic HIV increases 100-fold, from pre-test odds of 1/99 to post-test odds of 100/99, which is equivalent to a post-test probability of about 50% (100/199). With this test result, the diagnosis of HIV is still uncertain; the patient and clinician will need an additional test, such as the Western Blot, to resolve this uncertainty. Clinicians can get help using LRs to move from pre- to post-test probability using a LR nomogram[7] (online supplement Box 2) or an online calculator (http://getthediagnosis.org/calculator.htm). A previously published primer on precision and accuracy in clinical examination provides an excellent, more detailed overview of the use of LRs.[8]

Now consider two patients presenting with chest pain in the emergency department with the same pre-test probability of acute MI, and both with abnormally elevated values of troponin, a test that can have a wide range of values (continuous rather than dichotomous). The laboratory reports a troponin level of 0.1 ng/mL in one patient, and 10.0 ng/mL in the other (normal is less than 0.05 ng/mL). Most physicians would intuitively, and correctly, estimate the likelihood of acute MI as much higher in the patient with the markedly higher troponin value, even though both test results are abnormal, or "positive."

Instead of assigning the same LR+ to every elevated troponin level, a more appropriate strategy is to assign one LR to those with mildly positive results (LR mildly +) and another to those with markedly abnormal results (LR markedly +) (online supplement table). We can find these different LRs by inspecting the ROC curve (reports describing the diagnostic value of a continuous test routinely display ROC curves).

USING ROC CURVES TO FACILITATE ACCURATE DIAGNOSIS—AN EXAMPLE

We illustrate the diagnostic information embedded in an ROC curve using the example of highly sensitive troponin as a continuous test for diagnosing acute MI.[1] As described in the case scenario, the patient's hsT result is between two suggested cutpoints. Depending on which cutpoint the clinician chooses, the estimated probability of acute MI, and the subsequent management decisions, may differ significantly. Intuitively, this makes no sense.

A more sensible approach to this diagnostic dilemma is to ask: "What is the LR for a hsT level of 5 pg/mL?" And, using that LR: "what is the post-test probability of acute MI?" Clinicians can answer this question by inspecting the ROC curve reported by Keller and colleagues (Fig. 1). Extracting the LR from the ROC curve requires a better understanding of ROC curve anatomy.

ROC CURVE ANATOMY

Two coordinates describe each point on an ROC curve: sensitivity, along the vertical axis, and 1-specificity, along the horizontal axis. Each point represents a potential cutpoint, defining a "positive" test as the value at that point *plus* all values that are more abnormal. And, likewise, a "negative" test would be all less-abnormal values. Thus, ROC curves display the sensitivity and specificity for all possible choices of dichotomizing a continuous test.

Sensitivity describes the test performance in patients with the target conditions, while specificity describes the test per-

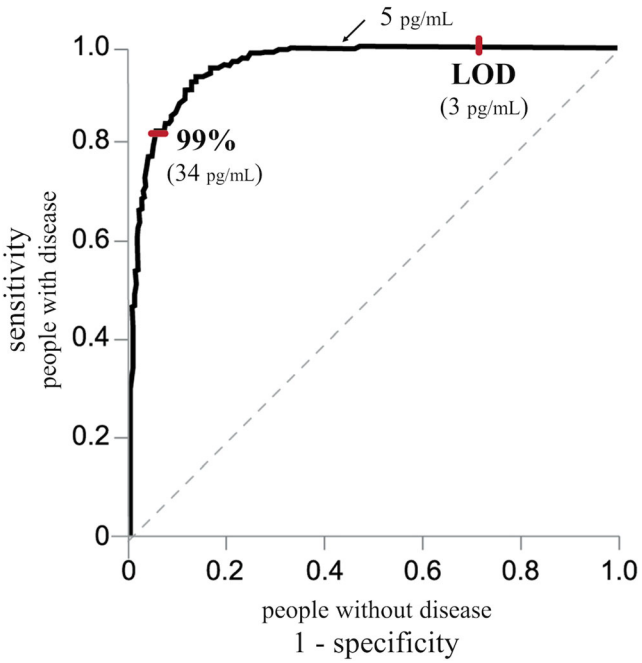

Fig. 1 ROC curve for high sensitivity troponin. LOD, level of detection. 99%: 99th percentile of high sensitivity troponin results in healthy population

formance in patients without the target condition. Thus, the vertical axis on an ROC curve describes the patients with the target condition, and the horizontal axis describes the patients without the target condition.

Consider the ROC curve for a hypothetical test that has three possible outcomes: positive, negative, and indeterminate. Figure 2 depicts the ROC curve of this hypothetical test.

In this figure, the segment of the ROC curve associated with a positive test result is highlighted in red. This test result is found in 60% of all patients with the target condition, and in no patients without the target condition. The LR for any segment of the ROC curve is equivalent to the slope of that segment (that is, the change on the vertical axis divided by the change on the horizontal axis), thus the LR associated with a positive test result is 0.6 divided by 0, which is infinity. Therefore, a positive test result rules in the target condition, increasing the probability of the target condition to 100%.

To further illustrate the logic underlying the calculation we have just made, consider that each line segment has two components—the rise (i.e., the proportion of patients with the target condition who have that test result, or change in sensitivity), and the run (i.e., the proportion of patients without the target condition who have that same test result, or change in 1-specificity). The ratio of those two proportions provides the LR for that segment of the curve. That is, the LR of any line segment on ROC curve is equivalent to the change in sensitivity divided by the change in specificity (same as change in 1 - specificity) for that particular segment. This concept matches the definition of LR, the probability of finding a particular test result in patients with the target condition divided by the probability of finding the same test results in patients without the target condition.

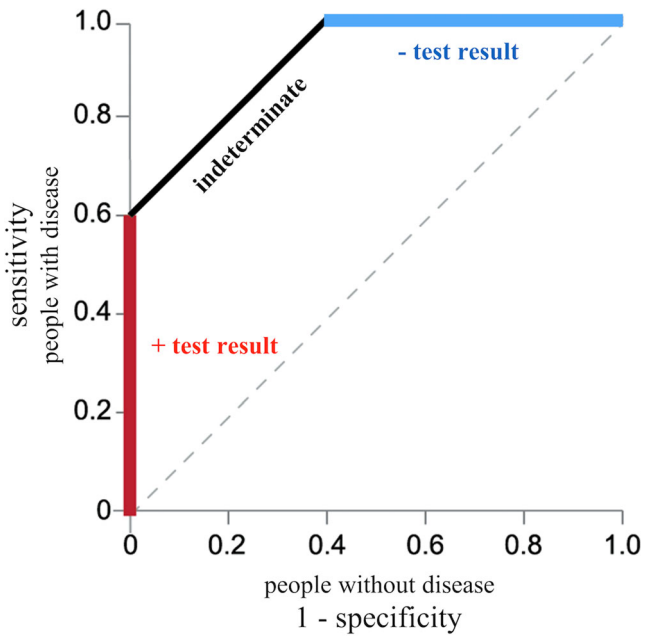

Fig. 2 Hypothetical ROC curve. The highlighted red portion represents the patients with a positive test result; black portion represents the patients with indeterminate test result. And highlighted blue portion represents the patients with negative test result

Consider now the black segment of the ROC curve in Fig. 2 that describes an indeterminate test result. The slope of the segment is 1 (0.4/0.4), for a LR indeterminate of 1. Thus, indeterminate test results will neither increase nor decrease the odds (or probability) of the target condition. This makes sense—an indeterminate test result is just as likely to be found in patients with the target condition (40% of such patients) as it is to be found in patients without the target condition (40% of these patients). Thus, an indeterminate test result does not modify disease odds or probability.

The blue highlighted segment of the ROC curve represents a negative test result. This test result is found in no patients with the target condition, and in 60% of patients without the target condition. The slope of this segment is 0 (0/0.6), which means that the LR− is 0. Thus, a negative test result would rule out the target condition because it decreases the post-test odds (or probability) to zero. This makes intuitive sense, as this test result is not found in any patients with the target condition.

Now imagine that the same ROC curve describes a test that is truly continuous, with one thousand possible values, from 0 to 999. The blue portion depicts results from 0 to 99, the black portion depicts results from 100 to 499, and the red portion depicts results from 500 to 999. The three LRs delineated from this ROC curve are $LR_{0–99} = 0$, $LR_{100–499} = 1$, $LR_{500–999} =$ infinity. Note that there is no LR+ and no LR−, because "positive" and "negative" test results do not exist for this continuous test.

Many investigators feel compelled to dichotomize test results if the test values are continuous—hoping to simplify the diagnostic process by collapsing the continuous test values into "positive" and "negative" results. To illustrate the danger of this approach, let us use the example of the ROC curve just described and try to ascertain the best possible cutpoint.

Assume that for a particular target condition, the harms associated with a false-positive diagnosis are roughly equivalent to the harms associated with a false-negative diagnosis. Under this scenario, many investigators would suggest an optimal cutpoint that was closest to the upper left corner of the ROC curve.[2] In Fig. 3, that point happens to be the test value of 300. With that cutpoint, all "positive" test results (values 300 to 999) would have an associated LR+ equal to the slope of the line segment extending from the origin to that point on the curve (0.8/0.2 = 4). All "negative" test results (0 to 299), would have a LR− equal to the slope of the line segment from that point on the ROC curve to the upper right corner (0.2/0.8 = 0.25).

These two LRs (LR+ of 4 and LR− of 0.25) appear to suggest that the test has only modest ability to increase or decrease the odds (and probability) of the target condition. Using the cutpoint of 300—despite it being the closest point to the upper left corner—has substantially diminished the test's value, a disservice to all patients tested. Patients with test results 0 to 99, as we saw earlier, could

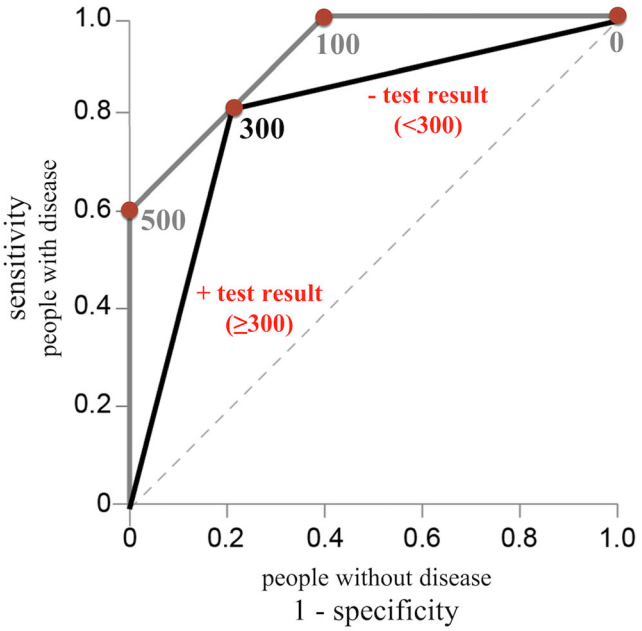

Fig. 3 Hypothetical ROC curve for a test with continuous variable (gray) superimposed with new curve (black) signifying slopes for positive and negative tests when a single cutpoint of 300 is used

have had the target condition definitively ruled out. When using a single cutpoint of 300, however, these patients would have been assigned a LR− of 0.25, which would decrease the probability of the target condition only to a limited extent, potentially exposing them to further unnecessary investigation and treatment. Patients with test results of 500 or higher should have the target condition established as the diagnosis. In contrast, the use of the single cutpoint of 300 would erroneously suggest a limited increase in probability of the target condition (LR+ of 4), potentially delaying treatment and exposing patients to further unnecessary diagnostic testing. Patients with test results between 100 and 499 will have probabilities of the target condition either inappropriately increased or inappropriately decreased, depending on which side of 300 mg/L their test results happen to fall.

To summarize, by dichotomizing this test, we have wasted information and incorrectly estimated post-test probability of the target condition. We incorrectly modified disease probability for many patients (those with values 100–499 mg/L) because we grouped them with patients who should have had their disease either ruled in or ruled out. The fundamental mistake was grouping test results that have very different slopes on the ROC curve, and thus very different LRs. We should only group test results with similar slopes: if the slopes are similar, then the LRs are similar.

Now, let us return to the patient scenario we introduced at the beginning of the article. Using the principles of ROC curve interpretation discussed above, examine the ROC curve produced by Keller (Fig. 1) describing the operating characteristics of the highly sensitive assay for troponin in the diagnosis of acute MI for patients presenting to the emergency department with a complaint of chest pain.[1]

If we dichotomize the hsT using the 99th percentile (34 pg/mL)—perhaps because we seek to minimize false-positive test results—we would modify the ROC curve as depicted in Figure 4. The slope associated with a negative test result (hsT < 34 pg/mL) is approximately 0.2 (0.18/0.91), which would be the LR− we would use for all patients with results in this range. For our patient, whose test result was less than 34pg/mL, the odds of acute MI would decrease to 2/10ths of its pre-test value, or from a probability of 30% to a post-test probability of acute MI of 8% (LR nomogram[7] and/or online calculator can be used to simplify the calculation), a value still too high to reassure the patient and his clinician that he does not have an acute MI.

However, inspection of the ROC curve tells us that dichotomizing values at the 99th percentile (34 mg/dL) wastes information. A portion of the ROC curve has a slope of near zero—at least 60% of the horizontal axis—meaning that the LR for those test results should be close to zero, which would essentially rule out acute MI.

Figure 5 shows that dichotomizing at the 99th percentile (point A) produces a "negative" test result that inappropriately combines 3 disparate groups: (1) the group of patients with values between A and B, which should modestly increase their probability of acute MI; (2) the group with values between B and C, which should modestly decrease the probability of acute MI; and (3) the group with values below C, which could

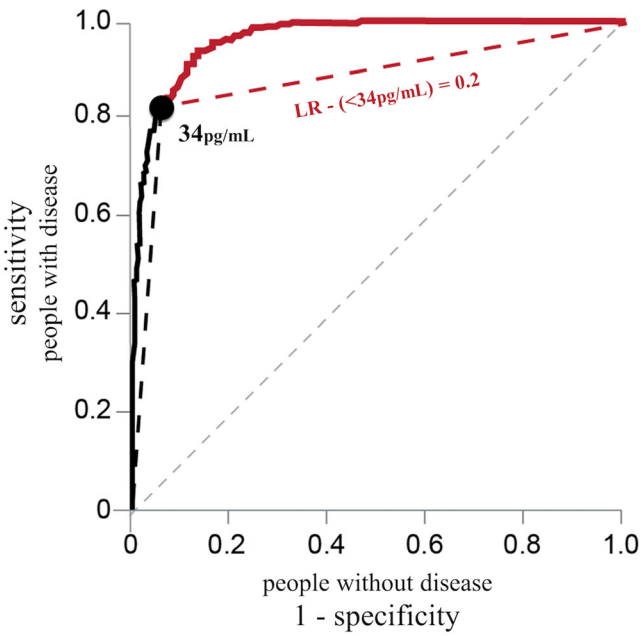

Fig. 4 ROC curve for hsT. The red solid line highlights the "negative" test results if cutpoint at 99th percentile is used. The red dashed line represents the single slope that corresponds to the LR−, while black dashed line represents the single slope that corresponds to the LR+

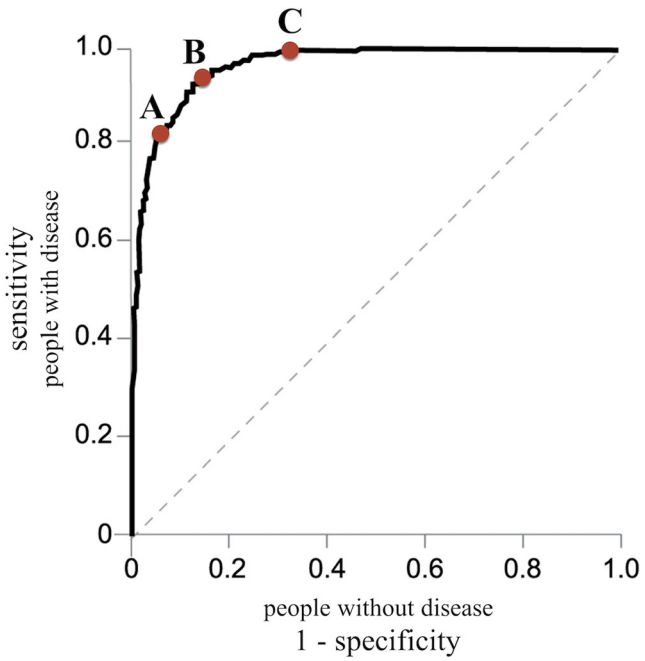

Fig. 5 ROC curve for high sensitivity troponin with cut off points (A, B, and C) placed at inflection points

essentially rule out the probability of acute MI in most patients.

Thus, the ROC curve for hsT has 3 inflection points classifying test results into 4 ranges that share homogeneous LRs. The investigators kindly provided the hsT values corresponding to these inflection points, which allows us to summarize the full diagnostic utility of test in Table 1. There are 4 test ranges defined by the 3 cutpoints, and LRs for the test ranges are equivalent to the slopes of the respective line segments. A table such as this is more consistent with intuition and more helpful to clinicians.

Recall that our patient had a hsT result of 5 pg/mL, which corresponds to a LR of 1/68. Using this LR, the pre-test probability of 30% would be reduced to a 0.6% post-test probability. This is substantially different from the post-test probability we estimated (8%) when using the "negative" test results when dichotomizing at the 99th percentile. For most patients and clinicians, 8% probability of acute MI would call for additional testing with repeat troponin and electrocardiogram. However, as probability of MI drops (in this case to 0.6%), management changes—clinicians would focus more attention on alternative diagnoses. Understanding of ROC curve anatomy can help clinicians make, with minimal effort, more accurate diagnostic judgements in a variety of circumstances, as demonstrated in the next case.

Table 1 Multilevel likelihood ratios based on hsT test characteristics

hsT value (pg/mL)	LR	Clinical implication
≥ 34	12	Large increase in probability of acute MI, likely mandates treatment
15–33	7/5	Very small increase in probability, almost always requiring more testing
7–14	1/4	Modest reduction in probability; most patients likely need further investigation
< 7	1/68	Acute MI ruled out in low- to intermediate-risk patients

hsT highly sensitive troponin, LR Likelihood Ratio, MI Myocardial Infarction

CASE 2: B-TYPE NATRIURETIC PEPTIDE IN SUSPECTED HEART FAILURE

A 60-year-old man with a history of coronary artery disease presents with shortness of breath, orthopnea, and bilateral lower extremity edema. His exam is notable for bibasilar crackles and elevated jugular venous pressure. Chest radiograph reveals cardiomegaly. Based on these clinical findings, you estimate the probability that this patient has heart failure (HF) at 90%, and plan to treat with diuresis. However, the patient's B-type natriuretic peptide (BNP) test result is 90 pg/mL. Your lab describes the upper limit of normal for BNP as 100 pg/mL and references a study by Maisel and colleagues published in 2002.[9] In this article, the test characteristics of BNP for a cutpoint of 100 pg/mL are described as sensitivity 90%, specificity 76%, positive predictive value (PPV) (the probability of heart failure after a positive test result) at 79%, and negative predictive value (NPV) (the probability of no HF after a negative tests) at 89% (corresponding to a post-test probability of HF of 11%). Therefore, with a negative test (LR− = 0.13), the probability of HF decreases to 54% (for calculation, refer to online supplement Box 2). Note that some learners will incorrectly use NPV of 89% to infer that the post-test probability for HF is 11% (1-

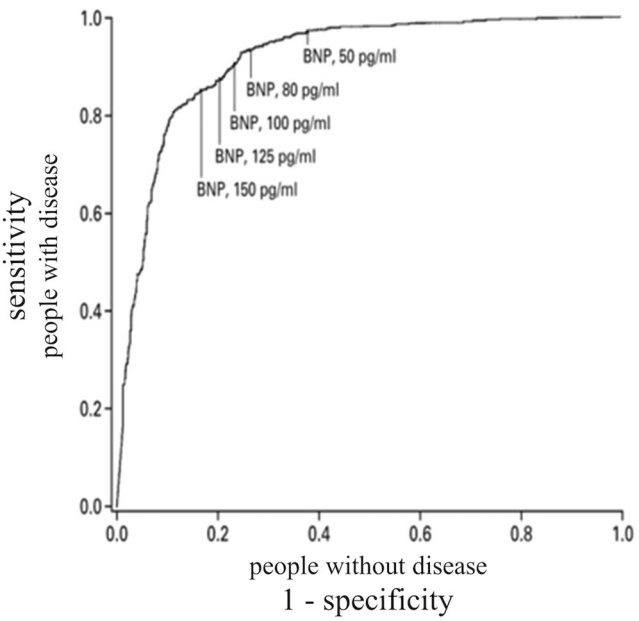

Fig. 6 ROC curve for B-natriuretic peptide for diagnosis of heart failure

NPV, or 1.0–0.89). The predictive values provided by studies reflect the post-test probability of target condition in an individual patient only when the patient's pre-test probability of having the target condition is equal to the disease prevalence in that study. In all other cases, predictive values are going to be misleading.

Given the uncertain diagnosis (HF probability of 54%), one may consider further diagnostic studies, including computerized tomography with angiography (CTA) of the chest to evaluate for pulmonary embolism or an echocardiogram to evaluate for pericardial tamponade.

The publication by Maisel depicts the performance of the BNP test using an ROC curve (Figure 6) with several possible cutpoints labeled. The authors report that a cutpoint of 80 pg/mL has a sensitivity of 93% and a specificity of 74%. Using this cutpoint, your patient's value of 90 pg/mLwould be categorized as a "positive" test, with an associated LR+ 3.6, producing a post-test probability of HF of 97%. This result is very different from the one using a cutpoint of 100 pg/mL. Choosing different cutpoints produces contradictory results. This is another example of how cutpoints that dichotomize tests with continuous results can be misleading.

Inspection of the ROC curve resolves this dilemma. We see that patients with BNP values near 100 pg/mL have LRs that are similar to those with values of 80 pg/mL and 150 pg/mL, because they all share a similar slope. The slope of the line segment between the values of 80 pg/mL and 150 pg/mL is close to 1, which means that the LR for all the values represented by that line segment is roughly 1. Using LR = 1 for a BNP of 90 (through inspection of the ROC curve), our patient's post-test probability remains unchanged at 90%. The implication of this diagnostic reasoning is that administration of furosemide and monitoring response, rather than further testing to look for alternative explanations for the patient's dyspnea, represents the right course of action.

The powerpoint slides on online appendix present multiple cases and a more detailed approach to interpretation and teaching of ROC curves.

ALTERNATIVE APPROACH

There is no exact guidance on how similar the slopes should be for them to be combined. In some cases, because inflection points are not clear, grouping test results based on ROC curve slopes may be challenging. The slope of a tangent line at any point on an ROC curve represents the LR for the test result described by that point. A graphic display of the derivatives (slopes) of all possible tangent lines on the ROC curve will show instantaneous LRs for the full spectrum of test results. An example has been previously been published by this journal (online supplement Fig. 2).[3] This information allows for derivation of a LR for any given test result without the need to combine the results into several groups or levels. When enough data is available, this approach yields the most accurate LRs.

CONCLUSION

Dichotomizing continuous test results (or categorical test results that have more than 2 levels) wastes information and can mislead clinicians. LRs solve the problem, but (sadly) authors of diagnostic test studies often do not provide the relevant LRs, choosing to present sensitivity and specificity at one or more cutpoints. Authors usually present ROC curves, and those graphics can frequently provide a solution to the problem. Often, however, inflection points on ROC curves are not adequately labeled, limiting proper use of diagnostic tests. The correct interpretation of ROC curves requires that test results are collapsed into groups only if they share a similar slope on the curve. Almost invariably, there are at least three regions of the ROC curves (3 ranges of test results) that correspond to 3 or more clinically distinct LRs.

We recommend that clinical investigators who study the diagnostic performance of clinical tests depict test performance with an ROC curve that labels key inflection points, and produce a table describing the multilevel LRs for the different ranges of test results. Alternatively, slopes of all possible tangent lines can be graphed to provide LRs for the full spectrum of test results. These improvements will allow clinicians to adjust the probability of target conditions more accurately and, as a consequence, will lead to more appropriate management decisions. Furthermore, certain laboratory results in electronic medical records can be displayed with multilevel LRs or ROC curves, to improve clinicians' access to more comprehensive information on lab interpretation. Alternatively, we advocate for creation of an online tool that will be easily accessible and searchable to clinicians seeking to find the best LR that fits the test result. A greater appreciation of the diagnostic information inherent in every ROC curve has the potential to improve clinical decision-making.[10]

Corresponding Author: *Amiran Baduashwili, MD; Section of Hospital Medicine, Division of General Internal Medicine Weill Cornell Medical College, 525 E 68th Street, Box 331, New York, NY 10065, USA (e-mail: amb9063@med.cornell.edu).*

Compliance with Ethical Standards:

Conflict of Interest: *The authors declare that they do not have a conflict of interest.*

REFERENCES

1. **Keller T**, **Zeller T**, **Ojeda F**, et al. Serial changes in highly sensitive troponin I assay and early diagnosis of myocardial infarction. JAMA. 2011;306(24) 2684–2693
2. **Guyatt GH**, **Patterson C**, **Ali M**, et al. Diagnosis of iron-deficiency anemia in the elderly. Am J Med 1990;88:205–209
3. **Guyatt GH**, **Oxman AD**, **Ali M**, **Willan A**, **McIlroy W**, **Patterson C**. Laboratory diagnosis of iron-deficiency anemia: an overview. J Gen Intern Med 1992;7:145–153
4. **Raschke RA**, **Curry SC**, **Warkentin TE**, **Gerkin RD**. Improving clinical interpretation of the anti-platelet factor 4/heparin enzyme-linked immunosorbent assay for the diagnosis of heparin-induced thrombocytopenia

through the use of receiver operating characteristic analysis, stratum-specific likelihood ratios, and Bayes theorem. Chest 2013;144(4):1269–1275

5. **Guyatt GH**, **Rennie D**, **Meade MO**, **Cook DJ**. Users' guides to the medical literature: a manual for evidence-based clinical practice. 3rd ed. New York, NY: McGraw-Hill; 2015: Page 345
6. **McGee SR**. Evidence-based physical diagnosis. Philadelphia, PA: Elsevier; 2017: Page 8
7. **Fagan TJ**. Letter: Nomogram for Bayes theorem. N Engl J Med. 1975;293(5):257
8. **Simel DL**, **Rennie D.** The rational clinical examination: evidence based clinical diagnosis. New York, NY : McGraw-Hill;2009. Chapter 1
9. **Maisel SA**, **Krishnaswamy P**, **Nowak RM** et al. Rapid Measurement of B-type natriuretic peptide in the emergency diagnosis of heart failure. N Eng J Med. 2002;347:161–167
10. **Zweig MH**, **Campbell M.** Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clinical Chemistry. 1993;39(4):561–577