OPEN
CME

Diagnostic Testing and Decision-Making: Beauty Is Not Just in the Eye of the Beholder

Thomas R. Vetter, MD, MPH,* Patrick Schober, MD, PhD, MMedStat,† and Edward J. Mascha, PhD‡§

To use a diagnostic test effectively and consistently in their practice, clinicians need to know how well the test distinguishes between those patients who have the suspected acute or chronic disease and those patients who do not. Clinicians are equally interested and usually more concerned whether, based on the results of a screening test, a given patient actually: (1) does or does not have the suspected disease; or (2) will or will not subsequently experience the adverse event or outcome. Medical tests that are performed to screen for a risk factor, diagnose a disease, or to estimate a patient's prognosis are frequently a key component of a clinical research study. Like therapeutic interventions, medical tests require proper analysis and demonstrated efficacy before being incorporated into routine clinical practice. This basic statistical tutorial, thus, discusses the fundamental concepts and techniques related to diagnostic testing and medical decision-making, including sensitivity and specificity, positive predictive value and negative predictive value, positive and negative likelihood ratio, receiver operating characteristic curve, diagnostic accuracy, choosing a best cut-point for a continuous variable biomarker, comparing methods on diagnostic accuracy, and design of a diagnostic accuracy study. (Anesth Analg 2018;127:1085–91)

> *I shall try not to use statistics as a drunken man uses lamp-posts, for support rather than for illumination.*
>
> —Andrew Lang (1844–1912), Scottish poet, novelist, and literary critic

To use a diagnostic test effectively and consistently in their practice, clinicians need to know how well the test distinguishes between those patients who have the suspected acute or chronic disease and those patients who do not.[1] Clinicians are equally interested and usually more concerned whether based on the results of a screening test, a given patient actually (1) does or does not have the suspected disease; or (2) will or will not subsequently experience the adverse event or outcome.[2,3]

Medical tests performed to screen for a risk factor, diagnose a disease, or estimate a patient's prognosis are frequently a key component of a clinical research study—including in anesthesiology, perioperative medicine, critical care, and pain medicine.[4,5] Like therapeutic interventions, medical tests require proper analysis and demonstrated efficacy before being incorporated into routine clinical practice.[6]

However, studies of diagnostic tests are frequently methodologically flawed, and their results are often not well understood or applied in clinical practice.[5] For example, if investigators select clinically inappropriate populations for their study of a diagnostic test, they introduce so-called "spectrum bias," and their study results can be invalid and misinform practicing clinicians.[1,3,7,8] Therefore, rigor must be applied in studying whether and in whom a particular medical test should be performed.[4]

As part of the ongoing series in *Anesthesia & Analgesia*, this basic statistical tutorial, thus, discusses the fundamental concepts and techniques related to diagnostic testing and medical decision-making. This tutorial includes the following concepts and techniques:

- Sensitivity and specificity;
- Positive predictive value and negative predictive value;
- Likelihood ratio;
- Receiver operating characteristic (ROC) curve;
- Diagnostic accuracy;
- Choosing and reporting the cut-point for a continuous variable biomarker;
- Comparing methods on diagnostic accuracy; and
- Design of a diagnostic accuracy study.

SENSITIVITY AND SPECIFICITY

The simplest screening or diagnostic test is one where the results of a clinical investigation (eg, electrocardiogram or cardiac stress test) are used to classify patients into 2 dichotomous groups, according to the presence or absence of a sign or symptom.[9] When the results of such a dichotomous (positive or negative) test are compared with a dichotomous "gold standard" test (eg, cardiac catheterization) that is often costlier and/or more invasive, the results can be summarized in a simple 2 × 2 table (Figure 1).[4]

The validity of such a screening or diagnostic test is its ability to distinguish between patients who have and those who do not have a disease.[2] This validity of a medical test has 2 primary components: sensitivity and specificity. The

From the *Department of Surgery and Perioperative Care, Dell Medical School at the University of Texas at Austin, Austin, Texas; †Department of Anesthesiology, VU University Medical Center, Amsterdam, the Netherlands; and Departments of ‡Quantitative Health Sciences and §Outcomes Research, Cleveland Clinic, Cleveland, Ohio.

Accepted for publication July 3, 2018.

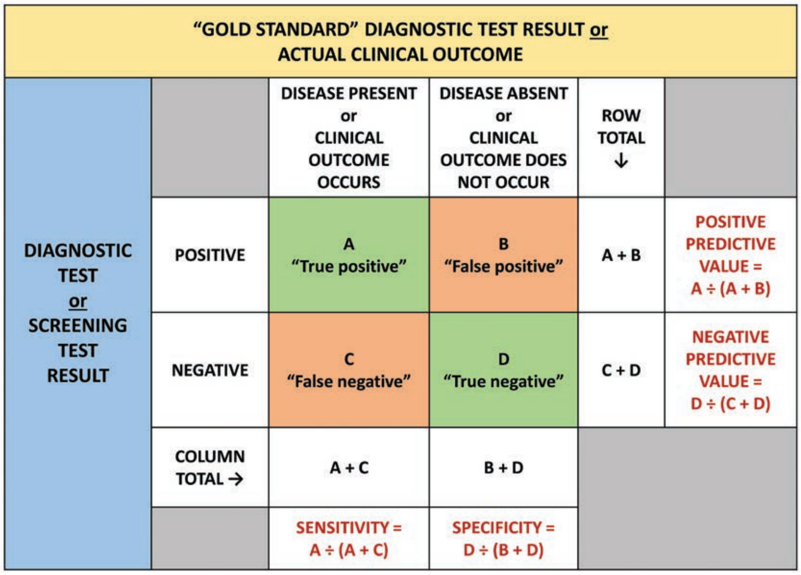
Funding: None.

The authors declare no conflicts of interest.

Reprints will not be available from the authors.

Address correspondence to Thomas R. Vetter, MD, MPH, Department of Surgery and Perioperative Care, Dell Medical School at the University of Texas at Austin, Health Discovery Bldg, Room 6.812, 1701 Trinity St, Austin, TX 78712. Address e-mail to thomas.vetter@austin.utexas.edu.

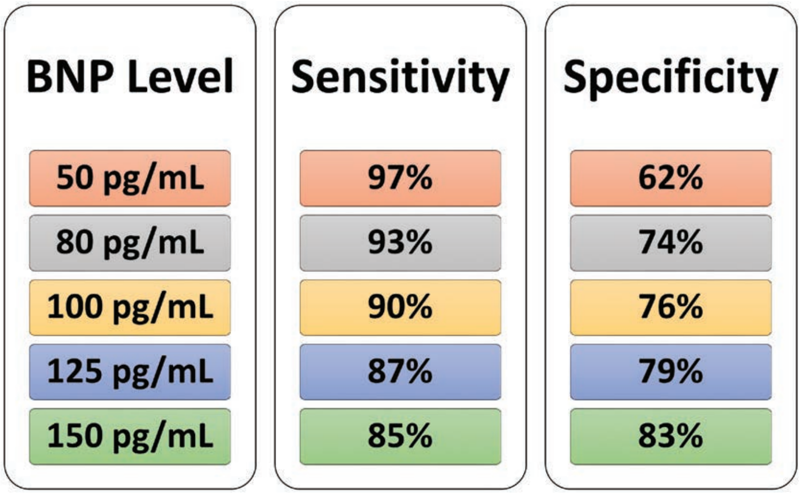
DOI: 10.1213/ANE.0000000000003698



"GOLD STANDARD" DIAGNOSTIC TEST RESULT or ACTUAL CLINICAL OUTCOME

		DISEASE PRESENT or CLINICAL OUTCOME OCCURS	DISEASE ABSENT or CLINICAL OUTCOME DOES NOT OCCUR	ROW TOTAL ↓	
DIAGNOSTIC TEST or SCREENING TEST RESULT	POSITIVE	A "True positive"	B "False positive"	A + B	POSITIVE PREDICTIVE VALUE = A ÷ (A + B)
	NEGATIVE	C "False negative"	D "True negative"	C + D	NEGATIVE PREDICTIVE VALUE = D ÷ (C + D)
	COLUMN TOTAL →	A + C	B + D		
		SENSITIVITY = A ÷ (A + C)	SPECIFICITY = D ÷ (B + D)		

Figure 1. A 2 × 2 table presenting the results (namely, the sensitivity, specificity, positive predictive value, and negative predictive value) from a study comparing a dichotomous diagnostic or screening test with a gold standard test or clinical outcome.[2–4,12]



BNP Level	Sensitivity	Specificity
50 pg/mL	97%	62%
80 pg/mL	93%	74%
100 pg/mL	90%	76%
125 pg/mL	87%	79%
150 pg/mL	85%	83%

Figure 2. Relationship among the chosen cutoff point (cut-point), sensitivity, and specificity, in this example, using BNP for diagnosing congestive heart failure in patients presenting with acute dyspnea.[3,11] BNP indicates B-type natriuretic peptide.

sensitivity of the test is its ability to identify correctly those patients who have the disease, whereas the specificity of the test is its ability to identify correctly those patients who do not have the disease.[2]

Sensitivity is thus defined as the proportion of truly diseased patients who have a positive result on the screening or diagnostic test (Figure 1).[3,10]

Specificity is thus defined as the proportion of truly nondiseased patients who have a negative result on the screening or diagnostic test (Figure 1).[3,10]

Diagnostic Cut-off Point

Ideally, a diagnostic test would display both high sensitivity and high specificity. However, this is often not the case, either for binary or continuous biomarkers.[3] When the clinical data generated by a medical test are not binary (eg, positive or negative) but instead have a range of values (eg, fasting serum glucose for diabetes or B-type natriuretic peptide [BNP] for congestive heart failure), a so-called cutoff point (or cut-point) value is often sought as a means to separate normal (or nondiseased) from abnormal (or diseased) patients.

As discussed in more detail below, the choice of this cut-point value is innately a balance between the sensitivity and the specificity for a diagnostic test.[2,3]

There is typically an inverse relationship between sensitivity and specificity. This is exemplified in a classic study of BNP for diagnosing congestive heart failure in patients presenting with acute dyspnea (Figure 2).[3,11] These authors concluded an acceptable compromise to be a BNP cut-point value (plasma level) of 100 pg/mL, with a corresponding sensitivity of 90% and a specificity of 76%.[11]

POSITIVE PREDICTIVE VALUE AND NEGATIVE PREDICTIVE VALUE

Sensitivity and specificity are the most accepted ways to quantify the diagnostic accuracy and validity of a medical test. However, in clinical practice, even if the sensitivity and specificity of a test are known, all that is reported, and thus known for a particular patient, is the test result. Yet, as noted above, the clinician really wants to know how good the test is at predicting abnormality (ie, what proportion of patients with an abnormal test result is truly abnormal).[9]

In other words, if the test result is positive, what is the probability that this given patient has the disease? Likewise, if the test result is negative, what is the probability that this given patient does not have the disease?[2] Alternatively, what is the probability that a patient with an abnormal test result will subsequently experience the adverse event or outcome of concern (eg, postoperative myocardial infraction)? Or, vice versa, what is the probability that a patient with a normal test result will not subsequently experience the adverse event or outcome of concern?

The positive predictive value is the proportion of patients with a positive test result who truly have the disease, or the probability that a positive test accurately predicts presence of disease or the occurrence of the adverse outcome (Figure 1).[2–4,12]

The negative predictive value is the proportion of patients with negative test results who truly do not have the disease, or the probability that a negative test accurately predicts absence of disease or nonoccurrence of the adverse outcome (Figure 1).[2–4,12]

Effect of Disease Prevalence on Predictive Values

The underlying prevalence of the disease being screened for or diagnosed does not affect the sensitivity or specificity of a medical test, which is why sensitivity or specificity is usually referred to as measures of the intrinsic accuracy of a test. The performance characteristics of the test itself in identifying patients with and without the disease remain the same despite changes in disease prevalence.[1]

However, as the underlying prevalence of the disease of interest increases, the positive predictive value of the test increases and the negative predictive value decreases. The more common the disease in the target population, the stronger the positive predictive value of the test. Similarly, as the underlying prevalence of the disease of interest decreases, the positive predictive value of the test decreases and the negative predictive value increases. The less common the disease in the target population, the stronger the negative predictive value of the test.[1–3]

Due to this relationship between prevalence and predictive values, it is very important to understand that predictive values reported in a study cannot simply be generalized to other settings with different disease prevalence. Particularly, in studies in which the prevalence does not reflect the natural population prevalence, but in which the observed prevalence is determined by the study design (such as in a 1:1 case-control study which artificially sets the prevalence at 50%), any reported diagnostic predictive values are of minimal practical use and must be interpreted carefully.

LIKELIHOOD RATIO

The positive likelihood ratio (LR+) compares the probability of a positive test result in patients with the disease or condition of interest (sensitivity) with the probability of a positive test result in patients without the disease (1 – specificity).[10,13] The LR+ describes how many times more likely a positive test result is to be a "true positive" result compared to "false positive."

Hence, an LR+ >1 indicates that the presence of the disease is more likely than the absence of the disease when the test result is positive; the greater the LR+, the stronger the ability of a positive test result to predict the presence of disease. Positive diagnostic test results with a high LR+ (>10) are considered to provide strong evidence to rule in the diagnosis.[14]

Similarly, the negative likelihood ratio (LR−) compares the probability of a negative test result in patients with the disease or condition of interest (1 – sensitivity) with the probability of a negative test result in patients without the disease (specificity).[10,15] An LR− close to 0 (<0.1) indicates that a negative test result is most likely a true negative result, providing strong evidence to rule out the disease or condition.[14]

Of note, using a nomogram or a simple formula, likelihood ratios can also be used to estimate the probability (or odds) that a positive or negative test result reflects presence or absence of the disease, respectively, for a given pretest probability (or odds) of disease presence.[16] This pretest probability is the assumed probability that a given tested individual actually has the condition, based on the information available before the test is performed.

Unless a particular individual is known or suspected to have a higher or lower risk of having the condition than other patients in the same population undergoing the diagnostic test, the pretest probability can be assumed to be the prevalence of the condition in this population.[13] The posttest probability of presence or absence of the disease can then be estimated by the positive predictive value and negative predictive value of the test, respectively.[13] The positive and negative predictive values, which once again depend on the disease prevalence, can readily be calculated for any given prevalence using the likelihood ratios.

Kruisselbrink et al[17] assessed the diagnostic accuracy of point-of-care gastric ultrasound to detect a "full stomach" (defined as either solid particulate content or >1.5 mL/kg of fluid) in 40 healthy volunteers. The authors reported an LR+ of 40.0 (95% confidence interval [CI], 10.3–∞) and an LR− of 0 (95% CI, 0–0.07), indicating that gastric ultrasound is highly accurate to rule in and to rule out a full stomach. Assuming a pretest probability of 50% for having a full stomach (the prevalence in their study sample), the authors used a nomogram to show that a positive test result increases the probability of having a full stomach to 97%, whereas a negative test result decreases the probability to <.1%.[17]

ROC CURVE

The previous paragraphs focused on diagnostic tests with a dichotomous outcome, in which the test result is either positive or negative. In situations in which a test result is reported on a continuous or ordinal scale, the sensitivity, specificity, and predictive values vary depending on the cut-point value or threshold that is used to classify the test result as positive or negative. Before defining an optimal threshold (as described in a subsequent section), it is useful to first assess the global diagnostic accuracy of the test across various cut-point values.

A ROC curve is a very common way to display the relationship between the sensitivity and specificity of a continuous-scaled or ordinal-scaled diagnostic test across the range of observed test values.[3,18] A ROC curve plots the true-positive rate (sensitivity) on the y-axis against the false-positive rate (1 – specificity) on the x-axis for a range of different cutoff values (Figure 3).[18,19] The ROC curve essentially demonstrates the tradeoff between sensitivity and specificity.

Simple visual inspection of the ROC curve provides useful information on the global diagnostic accuracy. A curve close to the left upper corner of the graph suggests good ability to discriminate patients with and without the condition, whereas a curve close to the diagonal from the bottom left to the upper right corner suggests that the test is only approximately as good as a random guess (Figure 3).[15]

More formally, the area under the curve (AUC) for a ROC curve can be calculated. The closer this AUC is to 1, the stronger the discriminative ability of the test. An AUC of 0.5 suggests that the test is unable to discriminate healthy from nonhealthy subjects, while an AUC <0.5 (not commonly observed in practice) suggests that a positive test result is somewhat predictive of absence of the disease.

Estimates of the AUC should be accompanied by a CI to provide an estimate of plausible values of the AUC in the population of interest.[20] As described below, statistics are available to test the null hypothesis that the AUC is equal to 0.5, and to compare AUC values of different diagnostic tests.

Gastaminza et al[21] studied whether tryptase levels during the reaction (TDR), as well as the ratio of TDR to baseline tryptase levels (TDR/BT), would be useful in discriminating immunoglobulin E (IgE)-dependent from IgE-independent hypersensitivity reactions. ROC analysis was performed to assess the overall diagnostic ability across different cut-points and to compare the AUC of the 2 approaches. The authors observed that the TDR/BT ratio had an overall better diagnostic ability to discriminate IgE-dependent from IgE-independent hypersensitivity reactions than TDR, with the TDR/BT AUC of 0.79 (95% CI, 0.70–0.88), and the TDR AUC of 0.66 (95% CI, 0.56–0.76), respectively, with the difference in AUC of 0.13 (95% CI, 0.05–0.20).[21]

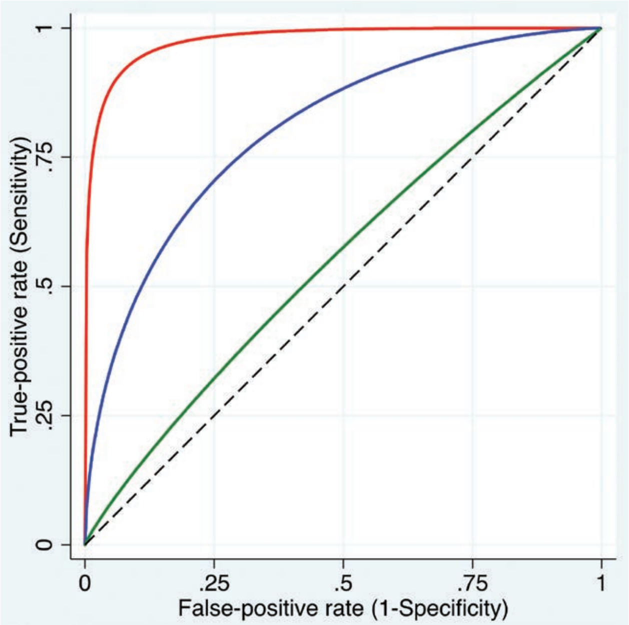


Figure 3. ROC curves are plots of the true-positive rate (sensitivity) against the false-positive rate (1 – specificity) for a range of different cutoff values. Shown are 3 smoothed curves, visually representing high (red curve close to the left top corner), intermediate (blue curve), and low (green curve close to the dashed diagonal line) discriminative ability to distinguish patients with a condition from patients without the condition. More formally, the AUC can be calculated, where an AUC close to 1 indicates high discriminative ability and an AUC close to 0.5 (representing the area under the diagonal line) indicates that the test is no better in predicting the condition than tossing a coin. AUC indicates area under the curve; ROC, receiver operating characteristic.

DIAGNOSTIC ACCURACY

Diagnostic accuracy refers to the discriminative ability of a medical test to distinguish healthy from nonhealthy subjects. The metrics of sensitivity, specificity, and predictive values are often considered measures of accuracy because they provide information on how well a dichotomous test—or a test with a continuous value that is dichotomized at a given cut-point threshold—can distinguish between diseased and nondiseased patients. However, as noted above, some metrics depend on the disease prevalence and are thus actually not measures of the intrinsic accuracy of the test itself. In contrast, sensitivity and specificity do not depend on prevalence and are thus considered measures of intrinsic diagnostic accuracy.[22]

The proportion of correctly classified, true-positive and true-negative patients, sometimes termed "overall diagnostic accuracy" is often reported as a global marker of accuracy.[15] However, overall diagnostic accuracy depends on the prevalence of the condition.[23] Therefore, the overall accuracy obtained from a study sample is not a measure of intrinsic accuracy of the test, and it usually cannot be generalized.

In contrast, the likelihood ratio is particularly useful as a global marker of accuracy because it: (1) combines information from sensitivity and specificity; (2) does not depend on disease prevalence; and (3) allows estimation of the posttest probability of having a particular disease for any assumed pretest probability.[14]

The AUC of the ROC curve is also independent of the prevalence. It is also not influenced by arbitrarily chosen cut-point thresholds. The AUC of the ROC curve is thus often considered the most useful global marker of the diagnostic accuracy of a medical test with continuous values. However, as a summary across all cut-point thresholds (including those that are clinically nonsensical), the AUC of the ROC curve provides very limited information on how well the test performs at a specific threshold as commonly used in clinical practice.[19]

CHOOSING AND REPORTING THE CUT-POINT FOR A CONTINUOUS VARIABLE BIOMARKER

In addition to assessing the overall discriminative ability of a biomarker, it is often of interest to identify the best cut-point for that continuous or ordinal biomarker to be used in practice to classify individual patients as either having or not having the disease or outcome of interest.

First, it is not always prudent or feasible to make a recommendation of a cut-point because there might not exists a cut-point that gives adequate sensitivity and specificity. In their study design phase, researchers should specify the minimal criteria for reporting a cut-point at all—for example, requiring a certain minimum value for each of sensitivity and specificity, such as 70% or 75%, or an AUC of ≥0.75. It is not helpful or prudent to introduce a new cut-point into practice if it does not have sufficient accuracy.

A common method for estimating an optimal cut-point is to choose a threshold that maximizes both sensitivity and specificity (not their sum). One simply calculates sensitivity and specificity for each observed value of the biomarker and identifies the cut-point (or cut-points) that generate the best combination of sensitivity and specificity. This is appropriate when sensitivity and specificity are thought to be equally important for the study at hand, implying that a false-positive or false-negative mistake would be equally costly.

Alternatively, researchers might require a minimum specificity (or sensitivity), which would influence the choice of optimal cut-point. For example, it may be that the cut-point that maximizes sensitivity and specificity yields a sensitivity of 80% and a specificity of 78%. But if ≥90% specificity were required, the optimal cut-point for this study might have a sensitivity of only 60%. The desired balance between sensitivity and specificity should be determined and justified a priori.

Finally, the cut-point that maximizes the sum of sensitivity and specificity could be chosen, as with Youden index.[24] However, this method has the notable disadvantage of not monitoring whether sensitivity and specificity are very different from each other, and it can often identify a cut-point at which they differ markedly. This tends to occur most often when the AUC of the ROC curve is low. When the AUC is very high, Youden index tends to identify cut-points closer to those achieved when maximizing both sensitivity and specificity—the first method described above.[25]

Gomez Builes et al[26] sought to find a cut-point for values of maximum lysis in trauma patients, below which a patient would be less likely to survive 48 hours. Applying Youden index, their chosen cut-point had sensitivity of 42% (95% CI, 27–57) and specificity of 76% (95% CI, 51–88). They note that the discrepancy between specificity and sensitivity was acceptable in their clinical setting because it was important to reduce false positives. However, since sensitivity could just as well have been much higher than specificity, researchers in similar situations might choose the cut-point with the highest sensitivity at a predetermined specificity, or simply maximize both.[25]

Because a chosen cut-point is an estimate, it should be accompanied by a CI. CIs for a cut-point can be estimated using bootstrap resampling.[27] The CI for the estimated cut-point can be interpreted as the estimated range of plausible values of the true optimal cut-point.

The underlying variability in determining a best cut-point to distinguish truly diseased from the truly nondiseased patients can be seen from a different angle using the so-called “grey zone” approach.[28] In this approach, instead of a single cut-point to attempt to discriminate diseased from nondiseased into 2 regions, 2 cut-points are identified to form 3 regions: patients who are believed to be diseased, nondiseased, and indeterminate (ie, not sure, the gray zone). The gray zone is estimated using specified values of LR+ and LR− that indicate the allowable false-positive and false-negative errors. While the details are beyond the scope of this article, in practice, the estimated gray zone often corresponds closely to the region specified by the confidence limits for a best cut-point when maximizing sensitivity and specificity.

COMPARING METHODS ON DIAGNOSTIC ACCURACY

Frequently, researchers undertake a study to assess whether 1 biomarker or laboratory value has better diagnostic accuracy than another. Such situations require formally comparing the biomarkers on AUC, or if there is a specified cut-point, on sensitivity and specificity.

Choice of the appropriate test statistic depends on whether the diagnostic accuracy results for the biomarkers being compared are correlated or not. Results would be correlated if comparing 2 biomarkers measured on all included subjects. They would be independent if different patient groups were being compared (eg, when assessing diagnostic accuracy between males and females).

Comparing independent AUCs is typically done using the method of Hanley and McNeil[29] for independent data. The method of Delong et al[30] or the paired (same case) method of Hanley and McNeil[31] can be used to compare dependent AUCs.

When comparing biomarkers on sensitivity or specificity, the denominator is all the diseased patients or nondiseased patients, respectively. When 2 independent groups such as males versus females are being compared on sensitivity or specificity, a simple Pearson χ^2 test can be used to compare the proportion who tested positive (for sensitivity) or negative (for specificity). For dependent comparisons, the McNemar test for correlated proportions is appropriate.[32] Analogous tests could be conducted for overall accuracy. Mainstream statistical packages include options for most if not all of these methods.[33]

DESIGN OF A DIAGNOSTIC ACCURACY STUDY

Rigorous study design is essential for a diagnostic accuracy study. First, the study objective must be clearly stated. Because the chosen population greatly influences the diagnostic accuracy results, as well as their meaning and applicability, researchers must describe the exact patient population about which they want to make inference.

Questions to consider include the following: Which patients are targeted? Those who already have had a positive result on a certain prescreening test? Those with a certain background predisposing them to have or not have the disease or outcome of interest? Is the goal one of estimation of diagnostic accuracy, comparison between biomarkers, or comparison between populations? Is the biomarker or modality of interest new or well-established?[34]

Choice of the gold standard method used to define diseased versus nondiseased patients should be carefully considered. Many times there is no perfect gold standard—a clear study limitation. An imperfect gold standard raises important questions of how the data will be analyzed and how the results can be interpreted. Nevertheless, statistical methods can attempt to account for an imperfect gold standard.[35] Reliability of the biomarker or medical test being evaluated should also be assessed and reported.

Calculation of the appropriate sample size depends on the goal of the study being either estimation of diagnostic accuracy or comparing biomarkers or groups on diagnostic accuracy. When estimating diagnostic accuracy, the goal is typically to estimate the parameter of interest with a desired precision, measured by the expected width of the CI.[36,37] When comparing biomarkers or groups on diagnostic accuracy, the difference to detect between the biomarkers or populations needs to be specified, and the sample size determined accordingly.

Of note, if the prevalence of the disease is expected to be low in the study sample (<50%), then estimation of or detecting differences in sensitivity would drive the sample size because the truly diseased would have a smaller overall sample size compared to the nondiseased, and sufficient power or precision for the smaller sample (the diseased) would guarantee it for the larger sample (the nondiseased). Likewise, specificity would drive the calculations if prevalence was expected to be >50%.

CONCLUSIONS

Knowing the diagnostic accuracy of a medical test used in clinical decision-making is of paramount importance for clinicians, given that false-positive and false-negative results, and subsequent therapeutic decisions, can have major consequences for patient physical and emotional well-being. The sensitivity, specificity, positive predictive value, negative predictive value, and likelihood ratios of a screening or diagnostic test each have unique merits yet limitations.

For continuous or ordinal-scaled tests, the AUC of the ROC curve provides insight into overall diagnostic accuracy. Choice of cut-point value or threshold should be informed by the relative importance of sensitivity versus specificity of the particular diagnostic test.

Last, a diagnostic accuracy study needs to be carefully designed to obtain valid and useful estimates of diagnostic accuracy. When interpreting results of a diagnostic accuracy study, clinicians should understand that metrics that depend on the disease prevalence—namely, predictive values and the so-called overall diagnostic accuracy—cannot be readily generalized beyond the study population in which they were estimated.

DISCLOSURES

Name: Thomas R. Vetter, MD, MPH.
Contribution: This author helped write and revise the manuscript.
Name: Patrick Schober, MD, PhD, MMedStat.
Contribution: This author helped write and revise the manuscript.
Name: Edward J. Mascha, PhD.
Contribution: This author helped write and revise the manuscript.
This manuscript was handled by: Jean-Francois Pittet, MD.

REFERENCES

1. Montori VM, Wyer P, Newman TB, Keitz S, Guyatt G; Evidence-Based Medicine Teaching Tips Working Group. Tips for learners of evidence-based medicine: 5. The effect of spectrum of disease on the performance of diagnostic tests. *CMAJ*. 2005;173:385–390.
2. Gordis L. Assessing the validity and relaibaility of diagnostic and screening tests. *Epidemiology*. 5th ed. Philadelphia, PA: Elsevier Saunders, 2014:88–115.
3. Fletcher RH, Fletcher SW, Fletcher GS. Diagnosis. *Clinical Epidemiology: The Essentials*. 5th ed. Philadelphia, PA: Wolters Kluwer/Lippincott Williams & Wilkins, 2014:108–131.
4. Newman TB, Browner WS, Cummings SR, Hulley Stephen B. Diagnostic studies of medical tests. In: Hulley Stephen B, Cummings SR, Browner WS, Grady DG, Newman TB, eds. *Designing Clinical Research*. 4th ed. Philadelphia, PA: Wolters Kluwer Health/Lippincott Williams & Wilkins, 2013:171–191.
5. Scott IA, Greenberg PB, Poole PJ. Cautionary tales in the clinical interpretation of studies of diagnostic tests. *Intern Med J*. 2008;38:120–129.
6. Daya S. Study design for the evaluation of diagnostic tests. *Semin Reprod Endocrinol*. 1996;14:101–109.
7. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med*. 1978;299:926–930.
8. Goehring C, Perrier A, Morabia A. Spectrum bias: a quantitative and graphical analysis of the variability of medical diagnostic test performance. *Stat Med*. 2004;23:125–135.
9. Altman DG, Bland JM. Diagnostic tests. 1: sensitivity and specificity. *BMJ*. 1994;308:1552.
10. Straus SE, Glasziou P, Richardson WS, Haynes RB. Diagnosis and screening. *Evidence-Based Medicine: How to Practice and Teach It*. 4th ed. Edinburgh, United Kingdom: Elsevier Churchill Livingstone, 2015:137–167.
11. Maisel AS, Krishnaswamy P, Nowak RM, et al; Breathing Not Properly Multinational Study Investigators. Rapid measurement of B-type natriuretic peptide in the emergency diagnosis of heart failure. *N Engl J Med*. 2002;347:161–167.
12. Altman DG, Bland JM. Diagnostic tests 2: predictive values. *BMJ*. 1994;309:102.
13. Linnet K, Bossuyt PM, Moons KG, Reitsma JB. Quantifying the accuracy of a diagnostic test or marker. *Clin Chem*. 2012;58:1292–1301.
14. Deeks JJ, Altman DG. Diagnostic tests 4: likelihood ratios. *BMJ*. 2004;329:168–169.
15. Eusebi P. Diagnostic accuracy measures. *Cerebrovasc Dis*. 2013;36:267–272.
16. Fagan TJ. Letter: nomogram for Bayes theorem. *N Engl J Med*. 1975;293:257.
17. Kruisselbrink R, Gharapetian A, Chaparro LE, et al. Diagnostic accuracy of point-of-care gastric ultrasound. *Anesth Analg*. 2018 [Epub ahead of print].
18. Altman DG, Bland JM. Diagnostic tests 3: receiver operating characteristic plots. *BMJ*. 1994;309:188.
19. Mallett S, Halligan S, Thompson M, Collins GS, Altman DG. Interpreting diagnostic accuracy studies for patient care. *BMJ*. 2012;345:e3999.
20. Schober P, Bossers SM, Schwarte LA. Statistical significance versus clinical importance of observed effect sizes: what do p values and confidence intervals really represent? *Anesth Analg*. 2018;126:1068–1072.
21. Gastaminza G, Lafuente A, Goikoetxea MJ, et al. Improvement of the elevated tryptase criterion to discriminate IgE- from non-IgE-mediated allergic reactions. *Anesth Analg*. 2018;127: 414–419.
22. Šimundić AM. Measures of diagnostic accuracy: basic definitions. *EJIFCC*. 2009;19:203–211.
23. Alberg AJ, Park JW, Hager BW, Brock MV, Diener-West M. The use of "overall accuracy" to evaluate the validity of screening or diagnostic tests. *J Gen Intern Med*. 2004;19:460–465.
24. Youden WJ. Index for rating diagnostic tests. *Cancer*. 1950;3:32–35.
25. Mascha EJ. Identifying the best cut-point for a biomarker, or not. *Anesth Analg*. 2018;127:820–822.
26. Gomez-Builes JC, Acuna SA, Nasimento B, Madotta F, Rizoli SB. Harmful or physiologic: diagnosing fibrinolysis shutdown in a trauma cohort with rotational thromboelastography. *Anesth Analg*. 2018;127:840-849.
27. Efron B, Tibshirani R. *An Introduction to the Bootstrap*. New York, NY: Chapman & Hall; 1993.
28. Coste J, Pouchot J. A grey zone for quantitative diagnostic and screening tests. *Int J Epidemiol*. 2003;32:304–313.
29. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143:29–36.

30. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44:837–845.
31. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*. 1983;148:839–843.
32. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*. 1947;12:153–157.
33. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12:77.
34. Zhou X, Obuchowski NA, McClish DK. Design of diagnostic accuracy studies. *Statistical Methods in Diagnostic Medicine*. 2nd ed. Hoboken, NJ: Wiley and Sons, 2011:57–102.
35. Zhou X, Obuchowski NA, McClish DK. Methods for correcting imperfect gold standard bias. *Statistical Methods in Diagnostic Medicine*. 2nd ed. Hoboken, NJ: Wiley and Sons, 2011:389–434.
36. Obuchowski NA. Computing sample size for receiver operating characteristic studies. *Invest Radiol*. 1994;29:238–243.
37. Flahault A, Cadilhac M, Thomas G. Sample size calculation should be performed for design accuracy in diagnostic test studies. *J Clin Epidemiol*. 2005;58:859–862.