Optimal Thresholds by Maximizing or Minimizing Various Metrics via ROC-Type Analysis

Kelly H. Zou, PhD, PStat(R), Ching-Ray Yu, PhD, Kezhen Liu, MS, Martin O. Carlsson, MS, Javier Cabrera, PhD

Rationale and Objectives: Based on imaging features, the optimal thresholds are typically determined as cutoff points to dichotomize the corresponding measurement scales.

Materials and Methods: Five metrics (ie, the Youden index, Euclidian distance, percent of correct diagnosis, kappa statistic, and mutual information) are individually maximized or minimized to derive the corresponding optimal threshold. These optimal thresholds are estimated under the parametric binormal assumption. Monte Carlo simulation studies are conducted to compare the performances of these different methods. A published radiological example on the choice of treatment outcomes following ureteral stones is used to illustrate and compare the estimated thresholds both empirically and parametrically.

Results: The optimal threshold can be a "moving target" because it would depend on modeling assumptions, metrics, and variability in the data. Even with large samples, disease prevalence has an impact on the robustness of the metrics.

Conclusions: It is recommended that researchers compare different optimal cutoff points using several metrics and select one that is most clinically relevant. The ultimate goal is to maximize diagnostic performances that are clinically meaningful to achieve improved global health.

Key Words: Sensitivity; specificity; receiver operating characteristic analysis; optimal threshold.

The optimal threshold, which is also known as the operating point or cutoff point, is of importance in developing guidelines for clinical decision-making. It can be useful in practice to optimally dichotomize the measurements from an imaging feature or marker (1–3). The reason for maximizing particular metrics is that a practical and useful threshold can be critical in terms of developing medical products and therapeutics (4). An imaging feature can be used as a diagnostic tool for identifying patients with a disease or abnormal condition for determining the stage a disease has reached and for the prediction and monitoring of a clinical response to an intervention. Besides diagnostic imaging, optimal threshold may be useful when analyzing biomarkers (5–7).

In this article dedicated as a special tribute to Professor Charles E. Metz of the University of Chicago, we aim to extend the literature on receiver operating characteristic (ROC) analysis using optimal metrics derived from the sensitivity, specificity, agreement, distance, and information to estimate task-dependent decision criteria (8–13). We will demonstrate and investigate the thresholds that optimize Youden's index (YI) and Euclidian distance (ED) in ROC space as well as percent correct diagnosis (PCdx), kappa (κ), and mutual information (MI).

There have been several methods to determine the optimal thresholds for medical diagnosis. In a seminal article (8), Metz recommended a cost-benefit analysis by considering in terms of the average net benefit of a diagnostic test, defined as the "amount by which using the test can reduce minimum average diagnostic cost." The YI is useful for selecting an optimal threshold value for diagnostic tests (9–12). The cost, benefit, or similar analysis may further be incorporated into our framework using a generalized version by weighting sensitivity and specificity, similar to the concept of the generalized YI (12,13). Distance-based metric such as the Euclidian distance has been proposed (14). When validating automated image analysis by modeling probabilistic segmentations, a spatial agreement metric, Dice similar coefficient, and MI have been maximized to derive optimal thresholds (15). In practice, PCdx is simple to compute (16).

The proposed methods are illustrated using a previously published example to predict treatment options according to an unenhanced helical computed tomography (CT) image

Acad Radiol 2013; 20:807–815

From the Pfizer Inc, 235 East 42nd Street, New York, NY 10017 (K.H.Z., M.O.C., C.-R.Y.) and Rutgers, The State University of New Jersey, Piscataway, NJ 08854 (K.L., J.C.). Received December 17, 2012; accepted February 12, 2013. **Address correspondence to:** K.H.Z. e-mail: Kelly.Zou@pfizer.com

http://dx.doi.org/10.1016/j.acra.2013.02.004

analysis of ureteral stones (17–19). Although this feature analysis is used to exemplify various metrics, the proposed methods may be generalized to other radiological applications, including cancer staging and responder analysis. Therefore, our methods have a wide range of uses in medical imaging and radiological research (20,21).

MATERIALS AND METHODS

Five different metrics, including the YI, ED, PCdx, κ, and MI, are used to derive optimal thresholds. Based on different metrics, we derive solutions and approximations to the optimal thresholds under the binormal parametric modeling assumption and to illustrate these estimation methods empirically and parametrically on an imaging feature example (17–19) as well as via statistical simulations using artificially generated data.

ROC Analysis

Standard notations for ROC analysis may be found in several earlier and recent works (22–25), particularly in a seminal review article that Metz published in 1987 (8). These notations are provided in Appendix 1.

Nonparametrically, the ROC curve based on continuous diagnostic data is constructed using two survival functions, i.e., 1- c.d.f., where c.d.f. represents the cumulative distribution function for each of the two samples. These two samples are drawn according to a binary reference standard (RS) including mutually exclusive healthy (H) population when RS = 0 and the diseased (D) population when RS = 1, respectively. The ROC curve consists of corresponding many pairs of sensitivity versus (1-specificity) values at all possible threshold values. Table 1 depicts a hypothetical two-by-two table of joint probabilities of RS versus diagnosis (dx) at each given threshold. There are m subjects (controls) in the H sample and n patients (cases) in the D sample.

Parametrically, there are two popular ROC methods, binormal and direct parametric fitting. First, the more commonly-adopted maximum likelihood estimation (MLE) and quasi-MLE models developed by Metz and other authors have shown to be quite robust (22–24). Metz and coauthors have developed widely used software programs LABROC4 and LABROC5 (23). According to the Metz et al (23), "After rank-ordering the raw data and then creating categorical data from the resulting truth-state runs, LABROC4 executes a version of our ROCFIT algorithm (26) that has been modified to accommodate the large numbers of categories (that is, truth-state runs) that may be involved."

Dr. John Eng has translated the LABROC4 algorithm into Java calculators JROCFIT and JLABROC4 (27), which are executable directly over the World Wide Web. These software programs are convenient for fitting the binormal model to the two-sample dx data.

Furthermore, "LABROC5 is similar to LABROC4 except that it executes the ROCFIT algorithm only after the number of categories in the rank-ordered data has been reduced by merging truth-state runs in an ad hoc but empirically useful way" (23).

Another way to parametrically model the ROC curve is by directly fitting a suitable parametric form such as normal, Box-Cox–transformed normal with the log transformation as a special case (18,28–31), or beta distributions (15,32). However, the advantages in robustness of Metz and colleagues' LABROC4 and LABROC5 algorithms over the direct fitting method were evident (33).

To evaluate the overall accuracy of an imaging feature, the underlying area under the curve (AUC) is typically of importance. In opposite extremes, the dx is as accurate as the true RS when AUC equals 1, whereas it is as inaccurate as chance (ie, by flip an unbiased coin) when AUC equals 0.5. Generally, the AUC varies between these two extremes, the higher the more accurate the dx becomes (34,35).

Nonparametrically, it can be computed as the Wilcoxon's rank-sum test divided by the product of the two sample sizes, m and n (35). If data have several strata (eg, study sites), stratified nonparametric and parametric inferences may be made by assigning weights across strata (31). Parametrically, it is a simple function of the ROC parameters (22).

In terms of the agreement under the general concept of reliability, an ROC curve completely characterizes agreement in terms of the fundamental measures of diagnostic performance (33).

Optimal Thresholds

We consider five metrics, including YI (9–13) and ED to maximize accuracy (14), PCdx (16), and κ (36) to maximize agreement between RS and dx, and MI (15) to maximize between information shared by RS and dx, as different criteria to derive optimal thresholds. See Appendix 2 for further details on how to obtain these optimal thresholds, respectively. In an earlier application, we have presented our prior research in validating brain tumor resection accuracy using the MI method (15).

It is worth pointing out that the optimal threshold is not invariant to monotone transformations and thus must be mapped back to the original scale. Therefore, it must be cautioned that the optimal threshold on a continuous measurement scale cannot be easily and readily computed if binning is applied using Metz's LABROC4 and LABROC5 algorithms (23), and the binned ordinal scale no longer has a simple linear or nonlinear mapping from the original continuous measurement scale.

An Imaging Example

A total of 100 unenhanced helical CT scans were administered to assess the flank pain in patients with obstructing ureteral stones. A standard protocol was adopted (280 mA;

TABLE 1. A Two-by-Two Table of Joint Probabilities of the Reference Standard Versus Binary Diagnosis at Each Threshold (γ), with Disease Prevalence $\pi = n/N$ Where $N = m + n$

	RS		
Binary Dx Based on Biomarker Measurement	0 (Healthy)	1 (Diseased)	Probability
$\leq\gamma$	$p_{00} = (1-\pi)\mathrm{Sp}(\gamma)$	$p_{01} = \pi[1-\mathrm{Se}(\gamma)]$	$p_{0\bullet} = (1-\pi)\mathrm{Sp}(\gamma) + \pi[1-\mathrm{Se}(\gamma)]$
$>\gamma$	$p_{10} = (1-\pi)\,[1-\mathrm{Sp}(\gamma)]$	$p_{11} = \pi\mathrm{Se}(\gamma)$	$p_{1\bullet} = (1-\pi)\,[1-\mathrm{Sp}(\gamma)] + \pi\mathrm{Se}(\gamma)$
Probability	$p_{\bullet 0} = p_{00} + p_{10} = 1-\pi$	$p_{\bullet 1} = p_{01} + p_{11} = \pi$	$p_{\bullet 0}+p_{\bullet 1} = p_{0\bullet} + p_{1\bullet} = 1$

Dx, diagnosis; RS, reference standard; Sp, sensitivity; Se, specificity.

12 kVp; pitch, 1.0–1.6. The imaging thickness was 5 mm, with images reconstructed at 5-mm increments (17–19).

In-plane stone size, measured in millimeters, is hypothesized to be predictive for intervention. It is thus the main illustrative outcome variable here. A binary treatment option is the RS, including spontaneous passage (m = 71) versus surgical intervention (n = 29), shown in Table 1.

The Web-based JLABROC4 program is used for both nonparametric and parametric ROC analysis of the continuous outcome data (ie, the sizes of ureteral stones), stratified by binary treatment options (27). See the partial output from JLABROC4 in Appendix 3.

Optimal thresholds are derived nonparametrically. Corresponding to each metric, a 95% bootstrap confidence interval is constructed by resampling with replacement from the original ureteral stone size data 5000 times (37). The 2.5th and 97.5th percentile values (ie, the 125th smallest and the 4875th largest) are used as the lower and upper confidence bounds, respectively.

Besides JLABROC4, all remaining statistical programming was conducted using the R software (38).

Monte Carlo Simulations

For simplicity, the H sample is generated by a standard normal distribution using $N(0, 1)$, whereas the D samples are generated using 1 for all of the means under various scenarios and different standard deviations at 1 (equal-variance assumption), 0.5, and 3 (unequal-variance assumptions for the latter two), respectively. Normal distributions $\{N(1,1); N(1, 0.5^2); N(1,3^2)\}$ are specified for the D sample.

In other words, the true binormal model is $(\alpha, \beta) = \{(1,1); (2,2); (1/3, 1/3)\}$, with corresponding AUC = $\{0.76; 0.81; 0.62\}$. These are realistic AUCs commonly encountered in radiological research.

For each combination of the ROC parameters, both equal and unequal samples for the H and D samples are provided in a balanced or unbalanced scenarios in the RS in a large-sample setting, such that $(m,n) = \{(200, 200); (50, 200); (200, 50)\}$.

In this simulation study, because the underline true binormal model is adopted, direct fitting is appropriate using the sample means and standard deviations to estimate the binormal ROC parameters, rather than repeatedly running the JLABROC4 program.

A Monte Carlo procedure using 5000 runs is programmed (39). Within each run, both nonparametric and parametric (via directly fitting using two normal distributions) estimates of the optimal thresholds are computed. The optimization routine in R (38), "optim" or "nlm," is used to compute the optimal thresholds under the direct normal fitting method.

Besides the point estimate based on each optimization criterion, a 95% bootstrap confidence interval, using the 2.5th and 97.5th percentile values (ie, the 125th smallest and the 4875th largest) as the respective lower and upper confidence bounds, is constructed across 5000 Monte Carlo runs.

RESULTS

An Imaging Example

Table 2 reports the descriptive statistics of ureteral stone sizes (in millimeters), stratified by the binary RS for stones, of which m = 71 stones underwent spontaneous passages versus n = 29 stones required interventions, respectively.

The estimated nonparametric and binormal AUCs are $A_{NP} = 0.81$ and $A_{BN} = 0.82$, supporting that the binormal model has yielded an estimated AUC that is greater than its nonparametric counterpart. To generate a binormal ROC curve, the estimated (α, β) binormal ROC parameters for using JLABROC4 are (1.391, 1.145) (see detailed output shown in Appendix 3). The standard errors for these estimates are 0.3200 and 0.2322, respectively, with an estimated correlation of 0.6057.

In Figure 1, both the nonparametric and parametric binormal ROC curves confirm fairly high accuracy by using ureteral stone sizes as a key predictor for interventions.

Figure 2 illustrates different metrics over a range of arbitrary thresholds over the range of all ureteral stone sizes. Coincidentally, as listed in Table 3, all five metrics yield an identical optimal threshold, which is 5 mm, although 95% bootstrap confidence intervals vary slightly. At this particular cutoff value at 5 mm, the corresponding nonparametric specificity is 57/71 = 0.80 and sensitivity is 20/29 = 0.69. Indeed, Fielding and colleagues previously observed that "stones that are larger than 5 mm, located within the proximal two thirds of the ureter, and seen on two or more consecutive CT images are more likely to require endoscopic removal, lithotripsy, or both" (17), which is consistent with the findings using all metrics examined in the present research.

Unfortunately, because the binormal model does not use any direct monotone transformation of the measurement

TABLE 2. Descriptive Statistics of the Ureteral Stone Sizes Measured in Millimeters Stratified by Spontaneous Passage Versus Surgical Intervention

Reference Standard	Sample Size	Mean	SD	Min	Q_1	Median	Q_3	Max
Spontaneous passage	$m = 71$	4.18	2.10	1	3	4	5	10
Surgical intervention	$n = 29$	7.10	2.88	3	5	6	9	16

Max, maximum; min, minimum; Q_1, 25th percentile; Q_3, 75th percentile; SD, standard deviation.

scale, the specificity and sensitivity cannot be directly estimated based on the estimated (α, β) binormal ROC parameters.

Monte Carlo Simulations

Tables 4 to 6 summarize the estimated optimal thresholds using artificially generated data with (α, β) = {(1,1); (2,2); (1/3, 1/3)}, respectively.

Because the underlying binormal model is correct and valid, it is reassuring that both the nonparametric empirical and parametric binomial models tend to agree.

These tables demonstrate that metrics YI and ED appear to be quite robust and stable for each of the underlying set of ROC parameters. However, the results based on PCdx, κ, and MI tend to differ, depending on the prevalence through m and n.

Furthermore, ED produces higher relative precision (eg, reflected by narrower widths of the corresponding 95% confidence intervals).

It is also worth pointing out that for metrics PCdx, κ, and MI, the optimal thresholds varied because of the prevalence. For example, according to Table 4, under equal variance assumption with sample sizes (m, n) from (200, 50), (50, 200), under PCdx, the optimal thresholds are from 0.5, −0.9, to 1.9. Under κ, they are from approximately 0.5, 1, to 0. Under MI, they are from approximately 0.5, 0.7, to 0.3. Obtaining the same threshold for all approaches under the first scenario of equal sample sizes between H and D is not a coincidence, but rather an expected result.

We recommend the use of all of these methods for a robustness check in terms of the optimal thresholds under a priori sample sizes. However, it is not to be expected at all that these different metrics should yield similar results since the original purposes can be different. As stated previously, the objectives for optimization can be accuracy, agreement, and information based.

DISCUSSION

A clearly defined goal for the optimization, according to a priori specified objective or metric, is utterly important.

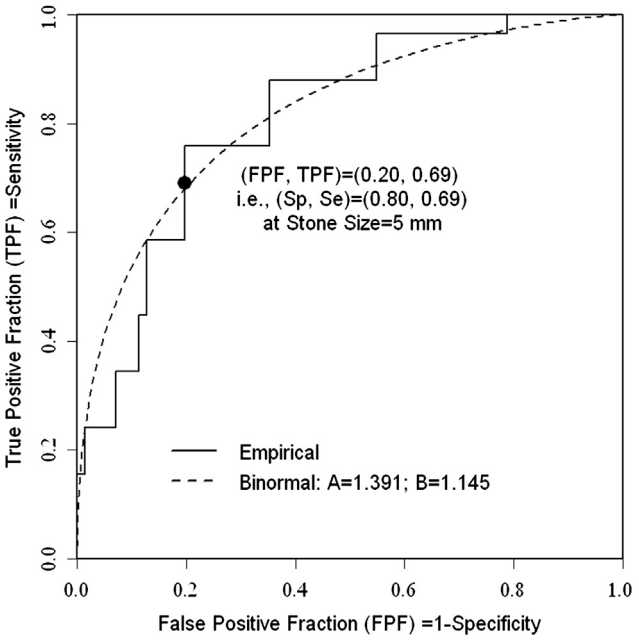

Figure 1. Nonparametric and binormal receiver operating characteristic curves of ureteral stone sizes. On the x axis is the false-positive fraction (FPF) = 1-specificity (Sp) and on the y-axis is the true-positive fraction (TPF) = sensitivity (Se). Nonparametrically, the estimated specificity = 57/71 = 0.80 and sensitivity = 20/29 = 0.69, corresponding to the stone size of 5 mm as the optimal cutoff value.

The goals for deriving an optimal threshold examined in this research included accuracy, agreement, best point, and information. Otherwise, there is no expectation for the thresholds to be the same unless the sizes of H and D samples are equal.

According to our simulation studies, it appears that an optimal threshold can be variable, depending on the distribution of the data and underlying modeling assumptions. This is not at all surprising because such an optimum would depend on various assumptions, metrics, and variability in the data. In particular, thresholds obtained by maximizing indices that depend on relative sizes can be substantially affected by the sample composition and thus the prevalence. Consequently, the resulting thresholds are appropriate only if the sample is representative of the target population, which is often difficult to achieve.

The inconsistency of optimal thresholds would thus suggest their use with great caution, as other authors warned in an earlier publication (11), given the fact that the YI does not prefer sensitivity over specificity, or vice versa, but only maximizes their sum. On the other hand, it is reassuring that all methods yielded identical optimal threshold at 5 mm for the ureteral stones helical CT imaging example.

Our methods may be related and nevertheless have a wide variety of applications such as the validation of image segmentations using the MI-based metric (15). For example, the YI and ED to (0,1) lead to the same result for a symmetric ROC curve (Table 4); the optimal threshold under YI

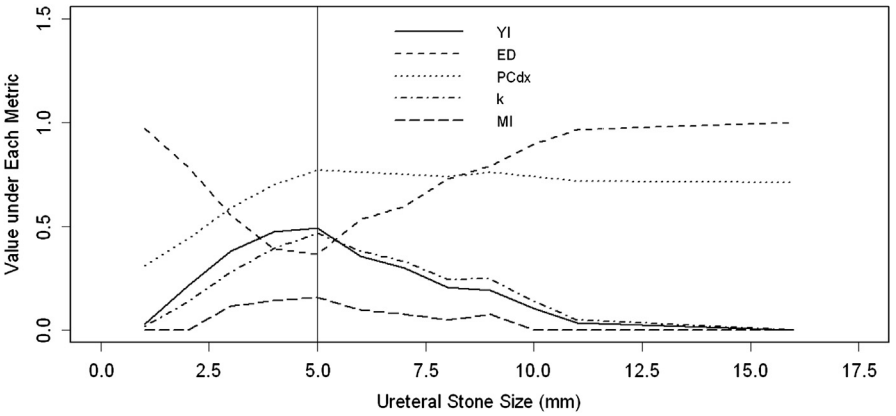

Figure 2. The corresponding values based on five different metrics by varying ureteral stone sizes as thresholds. The optimal threshold is estimated to be 5 mm according to all metrics. ED, Euclidian distance; MI, mutual information; k, kappa; PCdx, percent correct diagnosis; YI, Youden's index.

TABLE 3. Nonparametric Empirical Optimal Thresholds (with 95% Bootstrap Confidence Intervals) for Ureteral Stone Sizes (mm)

Optimization Metric	Optimal Threshold (95% CI), mm
Max (YI)	5 (3–5)
Min (ED)	5 (4–5)
Max (PCdx)	5 (4–6)
Max (κ)	5 (4–6)
Max (MI)	5 (3–6)

ED, Euclidian distance; κ, kappa statistic; MI, mutual information; min, minimum; max, maximum; PCdx, percent of correct diagnosis; YI = Youden index.

All five metrics yielded an identical optimal threshold at 5 mm. Nonparametrically, the estimated sensitivity = 0.80 and specificity = 0.69 corresponds to this cutoff value.

should be smaller for binormal ROC curves with hook in the lower corner (Table 5) and larger when the hook is in the upper corner (Table 6). These proposed metrics may easily be extended to other two-sample problems in the context of ROC analysis (36,40,41).

A major limitation of our simplified approach is that particular weighting factors are not applied across different measures as Metz has recommended (8); thus, generalized versions may further be sought. Not weighing sensitivity and specificity differently may still be a shortcoming. Thus, it must be emphasized that the optimization methods proposed here are purely determined from a statistical perspective. Such considerations must rely on sound clinical relevance. If necessary, these methods may be extended using different weights for measures such as sensitivity and specificity, as in Metz's cost-benefit (8) and the generalized YI approaches (12,13). In reality, we would need to consider either ruling in disease using a high specificity or ruling out disease using a high sensitivity, which are known as the "SPin" or the "SNout" principles (42).

As stated previously, our approaches can be extended to the classical cost-benefit analysis (8), and a generalized version of these methods can be considered based on a weighted sum of sensitivity and specificity, rather than a simple arithmetic sum (12,13). For example, the YI is equivalent to maximizing the cost-benefit tradeoff when the benefits of correct decisions across disease status are equal, incorrect decisions across disease status are equal, and the prevalence is 0.5. This is not the only cost-benefit problem (not the only specification of costs and prevalence) in which the average cost is equivalent to the YI. It is still to be argued whether YI is specifically relevant for whatever clinical task being considered. The metrics can be weighted by various costs associated with false-positive and false-negative fractions. However, such an extension is outside the scope of the current investigation.

Radiologists and clinicians may consider dichotomizing a diagnostic test using the optimal threshold. However, it has been emphasized in the literature that there is uncertainty surrounding a particular point. Therefore, for example, "rather than simply assessing a medical test at a single optimal point, it would be preferable to consider the AUC in a region near the optimal point" (20).

Finally, there are some limitations in automatically running optimization algorithms using continuous two-sample data. For instance, different optimization routines and initial values can yield different results. Non-unique solutions based on the empirical method may occur, making it difficult to compute either the bias or mean squared errors.

In closing, because the ultimate goal of making an accurate dx is to maximize its performance to achieve improved global health, researchers must closely examine the clinical purposes of using an imaging feature or biomarker and guidelines to derive its optimal threshold without living in a vacuum void from clinical considerations. The statistical methods we have presented in this article may nevertheless be used to help researchers achieve better-informed decision-making.

ACKNOWLEDGMENT

This article pays special tribute to Professor Charles E. Metz of the University of Chicago, Professor Harry S. Wieand of

TABLE 4. Optimal Thresholds (with 95% Bootstrap Confidence Intervals) for Simulated Hypothetical Data Using the Equal Variance Assumption with $N(0,1)$ and $N(1,1)$, ie, $(\alpha,\beta) = (1, 1)$ Under Various Sample Sizes

Optimization	$(m, n) = (200, 200)$		$(m, n) = (50, 200)$		$(m, n) = (200, 50)$	
Metric	NP	P	NP	P	NP	P
Max (YI)	0.49 (0.04, 0.95)	0.51 (0.33, 0.67)	0.52 (−0.08, 1.11)	0.51 (0.25, 0.74)	0.47 (−0.14, 1.06)	0.52 (0.20, 0.81)
Min (ED)	0.49 (0.26, 0.74)	0.50 (0.39, 0.62)	0.53 (0.20, 0.87)	0.50 (0.34, 0.67)	0.46 (0.12, 0.77)	0.51 (0.32, 0.71)
Max (PCDx)	0.49 (0.04, 0.95)	0.51 (0.33, 0.67)	−0.86 (−1.94, −0.26)	−0.87 (−1.80, −0.36)	1.82 (1.24, 2.75)	1.89 (1.66, 2.37)
Max (κ)	0.50 (0.04, 0.95)	0.51 (0.33, 0.67)	1.04 (0.46, 1.63)	1.03 (0.71, 1.27)	−0.06 (−0.67, 0.54)	−0.01 (−0.27, 0.27)
Max (MI)	0.50 (−0.30, 1.31)	0.51 (0.16, 0.85)	0.70 (−0.31, 1.76)	0.67 (0.13, 1.19)	0.29 (−0.78, 1.33)	0.36 (−0.18, 0.90)

ED, Euclidian distance; κ, kappa statistic; max, maximum; MI, mutual information; min, minimum; NP, nonparametric; P, parametric; PCdx, percent of correct diagnosis; YI, Youden index.

The underlying area under the curve is 0.76.

TABLE 5. Optimal Thresholds (with 95% Bootstrap Confidence Intervals) for Simulated Hypothetical Data Using the Unequal Variance Assumption with $N(0,1)$ and $N(1,1/2^2)$, ie, $(\alpha,\beta) = (2, 2)$ Under Various Sample Sizes

Optimization	$(m, n) = (200, 200)$		$(m, n) = (50, 200)$		$(m, n) = (200, 50)$	
Metric	NP	P	NP	P	NP	P
Max (YI)	0.39 (0.16, 0.62)	0.38 (0.24, 0.54)	0.37 (0.05, 0.67)	0.39 (0.11, 0.69)	0.41 (0.10, 0.72)	0.39 (0.20, 0.57)
Min (ED)	0.55 (0.40, 0.71)	0.55 (0.43, 0.68)	0.53 (0.29, 0.74)	0.56 (0.33, 0.81)	0.57 (0.37, 0.78)	0.56 (0.40, 0.72)
Max (PCDx)	0.39 (0.16, 0.62)	0.38 (0.24, 0.54)	−0.01 (−0.34, 0.25)	−0.01 (−0.32, 0.29)	1.92 (0.81, 3.48)	1.42 (0.98, 2.87)
Max (κ)	0.39 (0.16, 0.62)	0.38 (0.24, 0.54)	0.11 (−0.19, 0.38)	0.12 (−0.13, 0.39)	0.65 (0.32, 1.01)	0.64 (0.42, 0.85)
Max (MI)	0.17 (−0.11, 0.47)	0.16 (−0.01, 0.34)	0.04 (−0.29, 0.43)	0.04 (−0.26, 0.36)	0.30 (−0.02, 0.70)	0.25 (0.02, 0.48)

ED, Euclidian distance; κ, kappa statistic; max, maximum; MI, mutual information; min, minimum; NP, nonparametric; P, parametric; PCdx, percent of correct diagnosis; YI, Youden index.

The underlying area under the curve is 0.81.

TABLE 6. Optimal Thresholds (with 95% Bootstrap Confidence Intervals) for Simulated Hypothetical Data Using the Unequal Variance Assumption with $N(0,1)$ and $N(1,3^2)$, ie, $(\alpha,\beta) = (1/3,1/3)$, Under Various Sample Sizes

Optimization	$(m, n) = (200, 200)$		$(m, n) = (50, 200)$		$(m, n) = (200, 50)$	
Metric	NP	P	NP	P	NP	P
Max (YI)	1.44 (0.98, 1.96)	1.49 (1.40, 1.58)	1.37 (0.68, 2.09)	1.50 (1.35, 1.64)	1.46 (0.80, 2.20)	1.50 (1.33, 1.63)
Min (ED)	0.78 (0.44, 1.15)	0.79 (0.71, 0.87)	0.73 (0.22, 1.21)	0.79 (0.67, 0.92)	0.81 (0.32, 1.44)	0.79 (0.65, 0.93)
Max (PCDx)	1.44 (0.98, 1.96)	1.49 (1.40, 1.58)	−7.06 (−9.91, −5.27)	−3.50 (−4.33, 0.03)	2.19 (1.61, 2.90)	2.27 (2.17, 2.36)
Max (κ)	1.43 (0.97, 1.96)	1.49 (1.40, 1.58)	1.11 (0.37, 1.84)	1.23 (1.05, 1.39)	1.88 (1.28, 2.53)	1.91 (1.76, 2.04)
Max (MI)	2.01 (1.44, 2.54)	2.12 (2.01, 2.22)	1.64 (0.98, 2.28)	1.99 (1.84, 2.14)	2.12 (1.27, 2.72)	2.32 (2.12, 2.47)

ED, Euclidian distance; κ, kappa statistic; max, maximum; MI, mutual information; min, minimum; NP, nonparametric; P, parametric; PCdx, percent of correct diagnosis; YI, Youden index.

The underlying area under the curve is 0.62.

the University of Pittsburgh, Professor Donald D. Dorfman of the University of Iowa, Dr. Robert F. Wagner of the US Food and Drug Administration, and Professor W. Jackson Hall of the University of Rochester, who made important and extensive contributions in the field of receiver operating characteristic analysis but have unfortunately all passed away. The authors thank the Scientific and Public Affairs Advisory (SPA) Committee of the American Statistical Association. During the 2012 Joint Statistical Meeting held in San Diego, CA, the SPA selected this research to be the first place winner in its annual Statistical Significance ("StatSig") Poster Contest. The views expressed in this presentation are those of the authors and do not necessarily reflect the opinions of the institutions at which they are employed.

REFERENCES

1. Liu A, Schisterman EF, Zhu Y. On linear combinations of biomarkers to improve diagnostic accuracy. Stat Med 2005; 24:37–47.
2. Perkins NJ, Schisterman EF, Vexler A. Receiver operating characteristic curve inference from a sample with a limit of detection. Am J Epidemiol 2007; 165:325–333.
3. Perkins NJ, Schisterman EF, Vexler A. Generalized ROC curve inference for a biomarker subject to a limit of detection and measurement error. Stat Med 2009; 28:1841–1860.

4. Woodcock J. Assessing the clinical utility of diagnostics used in drug therapy. Clin Pharmacol Ther 2010; 88:765–773.
5. Wagner JA, Wright EC, Ennis MM, et al. Utility of adiponectin as a biomarker predictive of glycemic efficacy is demonstrated by collaborative pooling of data from clinical trials conducted by multiple sponsors. Clin Pharmacol Ther 2009; 86:619–625.
6. Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints preferred definitions and conceptual framework. Clin Pharmacol Ther 2001; 69:89–95.
7. Lesko LJ, Atkinson AJ, Jr. Use of biomarkers and surrogate endpoints in drug development and regulatory decision making: criteria, validation, strategies. Annu Rev Pharmacol Toxicol 2001; 41:347–366.
8. Metz CE. Basic principles of ROC analysis. Semin Nucl Med 1978; 8: 283–298.
9. Youden WJ. Index for rating diagnostic tests. Cancer 1950; 3:32–35.
10. Fluss R, Faraggi D, Reiser B. Estimation of the Youden Index and its associated cutoff point. Biomed J 2005; 47:458–472.
11. Perkins NJ, Schisterman EF. The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. Am J Epidemiol 2006; 163:670–675.
12. Schisterman EF, Faraggi D, Reiser B, et al. Youden Index and the optimal threshold for markers with mass at zero. Stat Med 2008; 27:297–315.
13. Nakas CT, Alonzo TA, Yiannoutsos CT. Accuracy and cut-off point selection in three-class classification problems using a generalization of the Youden index. Stat Med 2010; 29:2946–2955.
14. Gönen M. Analyzing receiver operating characteristic curves with SAS®. Cary, NC: SAS Institute Inc, 2007.
15. Zou KH, Wells WM, 3rd, Kikinis R, et al. Three validation metrics for automated probabilistic image segmentation of brain tumours. Stat Med 2004; 23:1259–1282.
16. Davila M, Christenson LJ, Sontheimer RD. Epidemiology and outcomes of dermatology in-patient consultations in a Midwestern U.S. university hospital. Dermatol Online J 2010; 16:12.
17. Fielding JR, Silverman SG, Samuel S, et al. Unenhanced helical CT of ureteral stones: a replacement for excretory urography in planning treatment. AJR Am J Roentgenol 1998; 171:1051–1053.
18. Zou KH, Tempany CM, Fielding JR, et al. Original smooth receiver operating characteristic curve estimation from continuous data: statistical methods for analyzing the predictive value of spiral CT of ureteral stones. Acad Radiol 1998; 5:680–687.
19. O'Malley AJ, Zou KH, Fielding JR, et al. Bayesian regression methodology for estimating a receiver operating characteristic curve with two radiologic applications: prostate biopsy and spiral CT of ureteral stones. Acad Radiol 2001; 8:713–725.
20. McClish DK. Evaluation of the accuracy of medical tests in a region around the optimal point. Acad Radiol 2012; 19:1484–1490.
21. Eng J. Teaching receiver operating characteristic analysis: an interactive laboratory exercise. Acad Radiol 2012; 19:1452–1456.
22. Dorfman DD, Alf E, Jr. Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals - rating-method data. J Math Psychol 1969; 6:487–496.
23. Metz CE, Herman BA, Shen JH. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. Stat Med 1998; 17:1033–1053.
24. Pesce LL, Horsch K, Drukker K, et al. Semiparametric estimation of the relationship between ROC operating points and the test-result scale: application to the proper binormal model. Acad Radiol 2011; 18: 1537–1548.
25. Alemayehu D, Zou KH. Applications of ROC analysis in medical research: recent developments and future directions. Acad Radiol 2012; 19: 1457–1464.
26. Metz CE. Some practical issues of experimental design and data analysis in radiological ROC studies. Invest Radiol 1989; 24:234–245.
27. Eng J. ROC analysis: Web-based calculator for ROC curves. Available at: http://www.jrocfit.org. Accessed December 12, 2012.
28. Box GEP, Cox DR. An analysis of transformations. JRSSB 1964; 26: 211–252.
29. Zou KH, O'Malley AJ. A Bayesian hierarchical non-linear regression model inreceiver operating characteristic analysis of clustered continuous diagnostic data. Biomed J 2005; 47:417–427.
30. O'Malley AJ, Zou KH. Bayesian multivariate hierarchical transformation models for ROC analysis. Stat Med 2006; 25:459–479.
31. Zou KH, Carlsson MO, Yu CR. Comparison of adjustment methods for stratified two-sample tests in the context of ROC analysis. Biomed J 2012; 54:249–263.
32. Zou KH, Warfield SK, Fielding JR, et al. Statistical validation based on parametric receiver operating characteristic analysis of continuous classification data. Acad Radiol 2003; 10:1359–1368.
33. Hanley JA. Receiver operating characteristic (ROC) analysis: the state of the art. Crit Revi Diagnostic Imaging 1989; 29:307–335.
34. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982; 143:29–36.
35. Wieand S, Gail MH, James BR, et al. A family of nonparametric statistics for comparing diagnostic makers with paired or unpaired data. Biometrika 1989; 76:585–592.
36. Zou KH, Liu A, Bandos AI, et al. Statistical evaluation of diagnostic performance: topics in ROC analysis. Boca Raton, FL: Chapman & Hall/ CRC Press, 2011.
37. Efron B, Tibshirani RJ. Introduction to the bootstrap. Boca Raton, FL: Chapman & Hall/CRC Press, 1994.
38. R Development Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Available at: http://www.R-project.org; 2008. Accessed December 12, 2012.
39. Robert CP, Casella G. Monte Carlo statistical methods. New York, NY: Springer Verlag, 2010.
40. Zhou XH, Obuchowski NA, McClish DK. Statistical methods in diagnostic medicine. New York, NY: Wiley & Sons Inc, 2002.
41. Pepe MS. The statistical evaluation of medical tests for classification and prediction. Oxford, UK: Oxford University Press, 2003.
42. Suss R. Sensitivity and specificity: alien edition: a light-hearted look at statistics. Can Fam Physicians 2007; 53:1743–1744.

APPENDIX 1

Notations and Assumptions

For the H sample of size m among subjects whose RS = 0, the i-th measurement from a imaging features is generated by an independent and identical (*i.i.d.*) distribution with the following *c.d.f.*:

$$X_i \sim i.i.d.\ F(\bullet), \forall i = 1, ..., m.$$

Similarly and independently, for the D sample of size n among subjects whose $RS = 1$, the j-th measurement from this imaging features is generated by an *i.i.d.* distribution with another *c.d.f.*:

$$Y_j \sim i.i.d.\ G(\bullet), \forall j = 1, ..., n.$$

The sample size fractions are given by

$$1 - \pi = m/N \text{ for } RS = 0 \text{ and } \pi = n/N \text{ for } RS = 1, \text{with } N = m + n,$$

where π is often called the disease prevalence.

Four underlying mean and variance parameters for the H and D populations according to RS are (μ,σ^2) and (ν,τ^2), which may further be reduced to two ROC parameters to determine the corresponding true ROC curve (22,23):

$$\alpha = (\nu - \mu)/\tau \text{ and } \beta = \sigma/\tau.$$

Nonparametrically, the empirical AUC is proportional to the two-sample Wilcoxon's rank-sum test statistic (W_{NP}) (35).

$$A_{NP} = W_{NP}/(mn).$$

Parametrically, the binormal AUC is a simple function of the two ROC parameters (23),

$$A_{BN} = \boldsymbol{\Phi}\left(\alpha/\left[\left(1+\beta^2\right)^{1/2}\right]\right).$$

APPENDIX 2

Optimal Thresholds Using Different Metrics

Let γ represent an arbitrary threshold. First, the optimal threshold corresponds to the maximum of the YI, the sum of sensitivity and specificity (10). That is,

$$\gamma_{opt,YI} = \text{argmax}_\gamma[YI(\gamma)], = \text{argmax}_\gamma[Se(\gamma) + Sp(\gamma) - 1].$$

A generalized YI is defined as:

$$GYI = Se(\gamma) + RSp(\gamma) - G,$$

where G is a constant with respect to the threshold γ, and $R = [(1- \pi)/\pi]$, a function of the disease prevalence π (12,13,20).

Second, the minimal ED between any ROC point to the most "ideal" upper-left point (1-*Sp*,*Se*) = (0,1). This perfect point means that both sensitivity and specificity are 100%, which are almost unachievable in practice (14).

$$\gamma_{opt,ED} = \text{argmin}_\gamma\left\{[1 - Sp(\gamma) - 0]^2 + [Se(\gamma) - 1]^2\right\}^{1/2}, = \text{argmin}_\gamma\left\{[Sp(\gamma) - 1]^2 + [Se(\gamma) - 1]^2\right\}^{1/2}.$$

Next, we may maximize the agreement between dx and RS using the PCdx across all possible threshold values (16). The PCdx statistic may be computed using the probabilities provided in Table 1, where the subscript $\bullet = 0 + 1$ because $p_{00} = (1-\pi)Sp(\gamma)$ and $p_{11} = \pi Se(\gamma)$. The optimal threshold is given by the following:

$$\gamma_{opt,\kappa} = \text{argmax}_\gamma(p_{00} + p_{11}) = \text{argmax}_\gamma[(1 - \pi)Sp(\gamma) + \pi Se(\gamma)].$$

Similarly, the agreement can also be assessed using the κ statistic across all possible threshold values (36). The κ statistic may also be computed using the probabilities provided in Table 1. For example, $p_{\bullet 0} = p_{00}+ p_{10} = 1-\pi$, $p_{\bullet 1} = p_{01}+ p_{11} = \pi$, and $p_{\bullet 0} + p_{\bullet 1} = 1$. Furthermore, $p_{0\bullet} = p_{00}+ p_{01}$, $p_{1\bullet} = p_{10}+ p_{11}$, and $p_{0\bullet} + p_{1\bullet} = 1$. The optimal threshold is given by the following expression.

$$\gamma_{opt,\kappa} = \text{argmax}_\gamma[2(p_{00}p_{11} - p_{10}p_{01})/(p_{0\bullet}p_{\bullet 1} + p_{\bullet 0}p_{1\bullet})].$$

Finally, the MI may be maximized as follows. The marginal and joint entropies are used to compute MI according to Table 1. Further methodological details have been described earlier (15).

$$\gamma_{opt,MI} = \text{argmax}[p_{0\bullet}\log_2(p_{0\bullet}) + p_{1\bullet}\log_2(p_{1\bullet}) + p_{\bullet 0}\log_2(p_{\bullet 0}) + p_{\bullet 1}\log_2(p_{\bullet 1}) - p_{00}\log_2(p_{00}) - p_{01}\log_2(p_{01}) - p_{10}\log_2(p_{10}) - p_{11}\log_2(p_{11})],$$

where $\log_2$ represents log base 2.

Note that the expressions for κ and MI may indeed be rewritten as a function of the prevalence π and the accuracy measures Se(γ) and Sp(γ). However, it is much better known that κ and MI are relatively simple functions of these joint probabilities (ie, the p's, in a two-by-two table) (Table 1). Because these expressions in terms of Se(γ) and Sp(γ) may not be trivial, only the original functions of the joint probabilities are shown here.

APPENDIX 3

Output Using JLABROC4

JLABROC 4 was used to produce the following output for the ureteral stone example (27).

Maximum likelihood estimation of a binormal ROC curve from continuously distributed test results.

Number of actually negative cases = 71;

Number of actually positive cases = 29.

Operating Points Corresponding to the Input Data

False-negative fractions (FPF): 0.0000, 0.0000, 0.0141, 0.0704, 0.1127;

True-negative fractions (TPF): 0.0000, 0.1034, 0.2069, 0.2759, 0.4138;

FPF: 0.1268, 0.1972, 0.3521, 0.5493, 0.7887;

TPF: 0.4828, 0.6897, 0.8276, 0.9310, 1.0000.

FPF: 1.0000;

TPF: 1.0000.

Final Estimates of Binormal ROC Parameters A and B

A = 1.3910, standard error (A) = 0.3200;

B = 1.1450, standard error (B) = 0.2322;

Correlation (A, B) = 0.6057.

Area Under the ROC Curve

Area under fitted curve (Az) = 0.8199;

Estimated standard error = 0.0444.

Trapezoidal (Wilcoxon) area = 0.8113;

Estimated standard error = 0.0523.