Statistics
in Medicine

Research Article

Received 26 April 2011, Accepted 7 December 2011 Published online 3 February 2012 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4509

Classification accuracy and cut point selection

Xinhua Liu*[†]

In biomedical research and practice, quantitative tests or biomarkers are often used for diagnostic or screening purposes, with a cut point established on the quantitative measurement to aid binary classification. This paper introduces an alternative to the traditional methods based on the Youden index and the closest-to-(0, 1) criterion for threshold selection. A concordance probability evaluating the classification accuracy of a dichotomized measure is defined as an objective function of the possible cut point. A nonparametric approach is used to search for the optimal cut point maximizing the objective function. The procedure is shown to perform well in a simulation study. Using data from a real-world study of arsenic-induced skin lesions, we apply the method to a measure of blood arsenic levels, selecting a cut point to be used as a warning threshold. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: classification accuracy; concordance probability; cut point; sensitivitiy; specificity

1. Introduction

When a quantitative test or biomarker is used for diagnostic or screening purposes, it is necessary to specify a cut-off value above which (or below which) the result is often considered positive (or negative). The positioning of the cut point necessitates a compromise between sensitivity (probability of correctly classifying those with the disease/condition) and specificity (probability of correctly classifying those without the disease/condition). In some cases, sensitivity is emphasized over specificity, for example when the disease is highly infectious or the condition is serious. On the other hand, specificity may be emphasized in situations where subsequent diagnostic testing is risky or costly. Statistically, there are some criteria taking into account both sensitivity and specificity in cut point selection. The optimal cut point, however, is criterion dependent and it may vary. Thus, one may need to choose among the alternative methods for cut point selection.

For a binary outcome variable used to indicate those with and without a particular disease/condition, after defining the sensitivity and specificity of the test or biomarker for each of the possible cut points, the ROC curve shows how sensitivity changes with changes either in specificity or in the proportion of false positives (1–specificity) for all possible cut points [1–3]. One can estimate the ROC curve with or without modeling the distribution of the quantitative measure for each of the two classes, and choose a cut point optimizing a utility function that takes into account both sensitivity and false positive rate at possible cut point. Because a utility function also requires information about cost, which is not always available, it may be necessary to identify the optimal cut point just using the criteria related to the ROC curve. One of the commonly applied criteria uses the Youden index [4], which is defined at a cut point as the sum of the associated sensitivity and specificity minus one or as the difference between sensitivity and the false positive rate. Maximizing the Youden index, one is able to find the cut point that has the largest value in the sum of sensitivity and specificity or in the difference between sensitivity and the false positive rate [5].

Another criterion uses the distance between a point on the ROC curve and an ideal point (0, 1) representing zero false positives and perfect sensitivity [6]. The distance is also a function of cut point. The closest-to-(0,1) criterion selects the cut point that minimizes this distance. These two methods may

Department of Biostatistics, Columbia University, New York, NY 10032, USA
Correspondence to: X. Liu, Department Biostatistics, Columbia University, New York, NY 10032, USA.
[†] *E-mail: xl26@columbia.edu*

not select the same cut point, although they both use quantities related to the ROC curve. Perkins and Schisterman [6] recommended the use of the Youden index and cautioned against the closest-to-(0, 1) criterion.

Youden index-based cut point selection has been studied in many contexts. Confidence interval for the Youden index and associated optimal cut point have been derived based on parametric models for class-specific distributions of the quantitative classifier [7, 8]. In contrast, nonparametric procedures such as the empirical method and the kernel method estimate the Youden index and the optimal cut point without modeling the conditional distributions [5]. The empirical method uses empirical distribution estimators, and the kernel method smoothes the empirical estimate of the Youden index. Alternatively, Klotsche *et al.* [9] used a nonparametric regression approach to find an optimal cut point associated with the Youden index, without using any smoothing technique. Recently, the Youden index-based method of cut point selection has been extended to accommodate various cases including a pooled sample [10], a biomarker subject to measurement error [11], cases where the observations are affected by a lower limit of detection [12], or where a marker has mass at zero [13].

The area under the ROC curve (AUC) is a summary measure derived from the ROC curve, and is often used to evaluate the classification accuracy of the variable for a diagnostic test or a biomarker. When the variable is continuous, the AUC can be interpreted as the concordance probability that the value from a subject in one group is greater than that for a subject in the comparison group [14]. For a variable with integer values or ordinal categories, the concordance probability is no longer equal to the AUC. Liu and Jin [15] showed that a concordance probability defined for a quantitative variable is a useful measure of classification accuracy.

In this paper, the concordance probability used to evaluate the classification accuracy of a dichotomized measure is defined as a function of sensitivity and specificity at the cut point. The cut point that maximizes this objective function is designated as 'optimal'. Like the Youden index and the closest-to-(0, 1) criterion, the proposed objective function is related to the ROC curve, and here a graphic interpretation is presented for comparison. Assuming Normal distribution and Gamma distribution models for the test measure or biomarker, the optimal cut point selected by each of the three criteria is calculated for numerical comparisons. The conditions under which the three criteria will yield the same optimal cut point are discussed, and the shapes of the three objective functions in the neighborhood of the common optimal cut point are compared for a set of Normal distribution models. Then an empirical method is applied to estimate the optimal cut point nonparametrically and the good finite-sample performance of the procedure is demonstrated in a simulation study. Using data from a real-world study of arsenic-induced skin lesions [16], we apply the method to find the cut point on a blood arsenic measure to be used as a warning threshold.

2. Method

Let X be a quantitative test result or biomarker, either ordinal or continuous. Let D be a class indicator with $D = 1$ for the particular disease/condition and $D = 0$ for the class without the disease/condition. Suppose the larger values of X tend to occur more often among subjects in the group having $D = 1$ than among subjects in the comparison group having $D = 0$. To evaluate the classification accuracy of X, one may use the concordance probability $CA = P(X_i^{(0)} < X_j^{(1)})$ with measurement $X_i^{(0)}$ being from the ith subject in $D = 0$ group and $X_j^{(1)}$ from the jth subject in $D = 1$ group [15]. Taking values in the range [0, 1], CA is invariant to rank-preserving transformations of X. If X is continuous, then CA is equivalent to the AUC [15]. Note that if X is binary, taking only the values of 0 or 1, then

$$CA = P(X^{(0)} = 0)P(X^{(1)} = 1)$$

is a product of the specificity $P(X^{(0)} = 0)$ and sensitivity $P(X^{(1)} = 1)$ of the measure.

2.1. A new criterion for cut point selection

Let the conditional distribution of the quantitative variable X in group D be $F_D(x) = P(X \leqslant x|D)$ and $S_D(x) = 1 - F_D(x)$ for $D = 0, 1$. Suppose $D = 1$ for cases and $D = 0$ for controls. Then, dichotomizing X at cut point x, the group-specific indicator function $Z^{(D)}(x) = I(X^{(D)} > x)$ has $EZ^{(D)} = S_D(x)$ for $D = 0, 1$. In particular, at cut point x, specificity $Sp(x) = F_0(x)$ and sensitivity

$Se(x) = S_1(x)$. The classification accuracy of $Z(x)$ is then a function of the cut point,

$$CZ(x) = P(Z^{(0)}(x) < Z^{(1)}(x)) = F_0(x)S_1(x) = Sp(x)Se(x).$$

Using it as an objective function, we choose the cut point x_C such that $CZ(x_C) = \max_{x \epsilon W} CZ(x)$, where W is the set containing all possible values of X. In practice, W could be restricted according to investigator's concern. For example, if a cut point must have sensitivity above 0.5 to be considered meaningful, then W has to contain the values of X that satisfy $Se(x) > 0.5$.

The alternative form $CZ(x) = F_0(x) - F_0(x)F_1(x)$ indicates that $CZ(x)$ can be viewed as the difference between two cumulative distribution functions, $F_0(x)$ and $F_c(x) = F_0(x)F_1(x)$. This form is similar to that of the Youden index with $J(x) = F_0(x) - F_1(x)$ for the distance between $F_0(x)$ and $F_1(x)$.

2.2. *Comparisons with two existing criteria*

The objective function $CZ(x)$ is in the family of ROC curve related quantities, which depend on the sensitivity and specificity at possible cut point x or the conditional distribution functions. Other members of the family include the objective functions: $J(x)$ for the Youden index, and $ER(x)$ for the closest-to-(0, 1) criterion, where

$$J(x) = Sp(x) + Se(x) - 1 = F_0(x) - F_1(x) = S_1(x) - S_0(x)$$

$$ER(x) = \sqrt{[1 - Sp(x)]^2 + [1 - Se(x)]^2} = \sqrt{S_0(x)^2 + F_1(x)^2}$$

Obviously, the three objective functions take values between zero and one.

2.2.1. Geometric interpretation of the three objective functions. By definition, the objective function $ER(x)$ is the distance between a point on a ROC curve and the point (0, 1), which represents perfect classification. The optimal cut point x_E on X minimizes this distance. The objective function of the Youden index $J(x)$ is the difference between the sensitivity and false positive rate at cut point x, which can be expressed as the distance between a point on the ROC curve and the point on the chance line corresponding to the cut point. The optimal cut point x_J on X maximizes this distance. Figure 1 shows a case with $x_E \neq x_J$.

Another interpretation of the Youden index $J(x)$ has to do with the area of a triangle associated with the ROC curve. Shown in Figure 1, the triangle has two lines connecting the point $(1 - Sp(x), Se(x))$ on the ROC curve to the points (0, 0) and (1, 1) on the chance diagonal line. Given that the area of this triangle is equal to $[Se(x) + Sp(x)]/2$ or $[J(x) + 1]/2$, maximizing $J(x)$ is equivalent to maximizing the area of the triangle associated with cut point x.

In contrast, $CZ(x)$, the concordance probability of dichotomized measure at cut point x, can be expressed as the area of a rectangle associated with the ROC curve. In Figure 1, the rectangle for $CZ(x)$ has two sides crossed at the point $(1 - Sp(x), Se(x))$ on the ROC curve, with its height representing

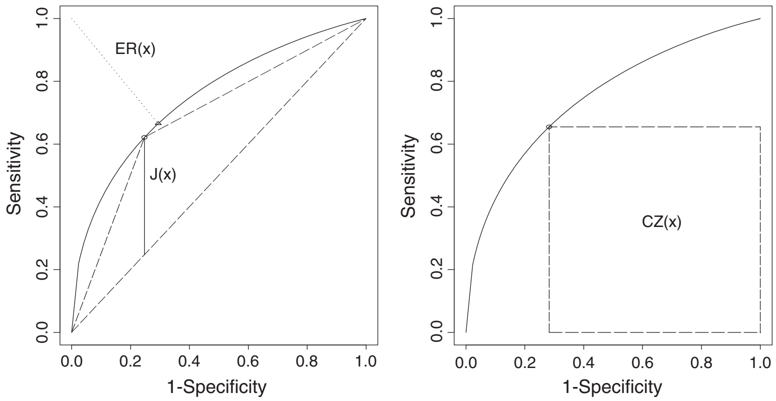

Figure 1. ROC curve related quantities.

the sensitivity $Se(x)$ and its width the specificity $Sp(x)$. The cut point x_C maximizing $CZ(x)$ actually maximizes the area of the rectangle.

2.2.2. Comparisons of optimal cut points selected by the three criteria. Suppose that the conditional distribution functions are differentiable for continuous variable X, and that each objective function has a unique optimum. Then the cut point optimizing objective function $Q(x)$ can be obtained by solving the equation $\partial Q(x)/\partial x = 0$.

Let x_C, x_J, and x_E be the solution to the differential equation with $Q(x) = CZ(x)$, $Q(x) = J(x)$, and $Q(x) = ER(x)$, respectively. Define the likelihood ratio at x as $LR(x) = f_1(x)/f_0(x)$ with $f_D(x) = \partial F_D(x)/\partial x$ for $D = 0, 1$. Then the cut point optimizing each of the three criteria may, respectively, have

$$LR(x_C) = \frac{S_1(x_{\mathrm{C}})}{F_0(x_{\mathrm{C}})}, \quad LR(x_J) = 1, \quad LR(x_E) = \frac{S_0(x_E)}{F_1(x_E)}.$$

Apparently, the optimal cut point is criterion-specific. It is possible, however, to have a value x_m optimizing all of the three objective functions, specifically when

$$1 = \frac{f_1(x_m)}{f_0(x_m)} = \frac{S_1(x_m)}{F_0(x_m)} = \frac{S_0(x_m)}{F_1(x_m)}.$$

In a special case that $X^{(D)}$ follows Normal distribution with mean μ_D and variance σ_D^2 for $D = 0, 1$, when $\sigma_0^2 = \sigma_1^2$, the point $x_m = (\mu_0 + \mu_1)/2$ optimizes all three criteria.

Table I illustrates the differences among the criterion specific optimal cut points based on two families of parametric models. The Normal distribution models assume Normal distribution for each class, with zero mean for $X^{(0)}$, equal and unequal within-class variances, and various positive mean for $X^{(1)}$ set to obtain the specified classification accuracies between 0.60 and 0.90. The Gamma distribution models assume Gamma distribution for each class, with mean of 1.5 for $X^{(0)}$, equal and unequal class variances, and various larger mean for $X^{(1)}$ to yield the specified classification accuracies.

For both sets of the distribution models with unequal within-class variances, the three criterion-specific optimal cut points tend to be somewhat distant from the averages of the two class-means. The cut point x_C, however, takes values between x_J and x_E. With the Normal distribution models, the magnitude of the difference between x_C and x_J decreases as *CA* increases from 0.60 to 0.90, and the difference between x_C and x_E seems small and stable. The Gamma distributions are skewed. There is no common optimal cut point under the settings with Gamma distribution models. In the cases with $\sigma_0{}^2 \geqslant \sigma_1{}^2$, as *CA* increases, the difference between x_C and x_J reduces, and the magnitude of the difference between x_C and x_E seems to increase. In contrast, with $\sigma_0{}^2 < \sigma_1{}^2$, the three cut points tend to be close for *CA* varying between 0.70 and 0.90.

For a cut point x_m satisfying all three criteria, in its neighborhood the shape of the objective functions will not necessarily be the same. To compare the curvatures at the common optimal cut point, we may examine the second derivatives of the convex objective functions of $CZ(x)$, $J(x)$ and $-ER(x)$. Let $f_D'(x) = \partial f_D(x)/\partial x$ be the derivative of density function $f_D(x)$ for $D = 0, 1$. Then for $\mu_1 > \mu_0$, the second derivatives

$$J"(x_m) = \frac{\partial^2 J(x)}{\partial x^2}|_{x=x_m} = f_0'(x_m) - f_1'(x_m) < \quad 0,$$

$$CZ"(x_m) = \frac{\partial^2 CZ(x)}{\partial x^2}|_{x=x_m} = J"(x_m)F_0(x_m) - 2f_0(x_m)^2 < \quad 0,$$

and

$$-ER"(x_m) = \frac{-\partial^2 ER(x)}{\partial x^2}|_{x=x_m} = \frac{J"(x_m)}{\sqrt{2}} - \frac{\sqrt{2}f_0(x_m)^2}{S_0(x_m)} < \quad 0.$$

In the special case of the Normal distribution models with common within-class variance σ^2, the inequalities $-ER"(x_m) < CZ"(x_m) < J"(x_m) < 0$ hold for the optimal cut point $x_m = (\mu_0 + \mu_1)/2$ and weighted group mean difference $\theta = (\mu_1 - \mu_0)/\sigma > 0$. Let $\Phi(x)$ be the standard normal distribution

Table I. The criterion specific optimal cut point based on two parametric models.

CA	μ_1	$\sigma_0{}^2, \sigma_1{}^2$	X_C	X_J	X_E	$X_C - X_J$	$X_C - X_E$
Normal distribution models ($\mu_0 = 0$)							
0.60	0.3583	1, 1	0.1791	0.1791	0.17914	0	0
	0.4388	2, 1	0.0271	−0.4533	0.07946	0.4804	−0.0523
	0.4388	1, 2	0.4117	0.8921	0.359349	−0.4804	0.0523
0.70	0.7416	1, 1	0.3708	0.3708	0.37081	0	0
	0.9083	2, 1	0.3220	0.0741	0.39332	0.2479	−0.0713
	0.9083	1, 2	0.5863	0.8342	0.51497	−0.2479	0.0713
0.80	1.1902	1, 1	0.5951	0.5951	0.59512	0	0
	1.4577	2, 1	0.6654	0.5414	0.74418	0.1240	−0.0788
	1.4577	1, 2	0.7924	0.9163	0.71355	−0.1240	0.0788
0.90	1.8124	1, 1	0.90.63	0.9062	0.90619	0	0
	2.2197	2, 1	1.1384	1.0867	1.21578	0.0515	−0.0775
	2.2197	1, 2	1.0814	1.1330	1.00393	−0.0515	0.0775
Gamma distribution models ($\mu_0 = 1.5$)							
0.60	1.7981	1, 1	1.3882	1.1766	1.4020	0.2115	−0.0138
	1.6710	2, 1	1.0769	0.7374	1.1299	0.3395	−0.0530
	1.9930	1, 2	1.5522	1.9036	1.5310	−0.3514	0.0212
0.70	2.1434	1, 1	1.5355	1.3832	1.5750	0.1523	−0.0395
	2.0989	2, 1	1.3001	1.0788	1.4002	0.2213	−0.1001
	2.3731	1, 2	1.6714	1.7042	1.6646	−0.0284	0.0068
0.80	2.5846	1, 1	1.7536	1.6522	1.8210	0.1014	−0.0674
	2.6871	2, 1	1.6504	1.5139	1.7934	0.1365	−0.1431
	2.8534	1, 2	1.8563	1.8187	1.8765	0.0376	−0.0202
0.90	3.2682	1, 1	2.1350	2.0804	2.2303	0.0546	−0.0953
	3.6705	2, 1	2.3037	2.2366	2.4843	0.0670	−0.1806
	3.5880	1, 2	2.1899	2.1514	2.2446	0.0385	−0.0547

function. The ratio comparing the second derivative of the objective functions of $CZ(x)$ and $J(x)$ can be written as

$$\frac{CZ"(x_m)}{J"(x_m)} = \Phi(\frac{\theta}{2}) + \sqrt{\frac{2}{\pi}} \frac{\exp(-\theta^2/8)}{\theta} \quad ,$$

which is a decreasing function of θ with a lower bound of one that $\lim_{\theta \to \infty} \frac{CZ"(x_m)}{J"(x_m)} = 1$, indicating that in the neighborhood of x_m the shape of $J(x)$ is flatter and becomes closer to that of $CZ(x)$ as the group mean difference gets larger. In contrast, the ratio comparing the second derivative of the objective functions of $-ER(x)$ and $CZ(x)$,

$$\frac{-ER"(x_m)}{CZ"(x_m)} = \frac{\sqrt{2}}{2\Phi(-\theta/2)} \left[\frac{\theta\Phi(-\theta/2) + \sqrt{2/\pi}\exp(-\theta^2/8)}{\theta\Phi(\theta/2) + \sqrt{2/\pi}\exp(-\theta^2/8)} \right] \in (\sqrt{2}, 1.6705).$$

It implies that the shape of $CZ(x)$ is always flatter than that of $-ER(x)$ in the neighborhood of x_m.

2.3. *A nonparametric approach to estimating the optimal cut point*

The objective function of the concordance probability associated with cut point x can be estimated using a parametric method, through modeling the conditional distribution of the quantitative variable X for each class with parametric distribution models. Using independent observations $X_i^{(D)}, i = 1, \ldots, n_D$, from class $D = 0, 1$, the maximum likelihood estimation method can be applied to estimate the model parameters, such that the objective function is a continuous function of x. Consequently, optimization procedures can be applied to search for the optimal cut point, which maximizes the continuous objective function. Thus, the results may depend on how well the parametric conditional distribution models fit the data.

It is more attractive to estimate the objective function without modeling the conditional distribution for $X^{(D)}$. Because the empirical estimation method is easy to implement, it is preferred to be used here. On the basis of the empirical distribution estimator $\hat{F}_D(x)$ for the conditional distribution of $X^{(D)}$ for $D = 0, 1$, a consistent nonparametric estimator of $CZ(x)$ for $x \in \{x_1^{(0)}, \ldots, x_{n_0}^{(0)}, \quad x_1^{(1)}, \ldots, x_{n_1}^{(1)}\}$,

$$\hat{C}Z(x) = \hat{F}_0(x)\hat{S}_1(x) = \frac{1}{n_0 n_1}\sum_{i=1}^{n_0}\sum_{j=1}^{n_1} I(x_i^{(0)} \leqslant x) I(x_j^{(1)} > x),$$

is in the form of a U-statistic [17].

When X has ordinal categories, it is easy to find the optimal cut point maximizing the estimated objective function. For continuous X, on the other hand, the empirically estimated conditional distributions $\hat{F}_D(x)$ are not continuous functions of x; nor is $\hat{C}Z(x)$. To estimate the optimal cut point, we may use a method similar to that of Fluss *et al.* [5]. Suppose that the observations of X in both classes are merged and ordered with n distinct values and the maximum is reached at x_j with $1 < j < n$. Because the value of $CZ(x)$ is constant in the interval $[x_j, x_{j+1}]$, a reasonable estimate of the cut point is $\hat{x}_C = (x_j + x_{j+1})/2$. To evaluate the variation in the cut point estimation, one may use the bootstrapping method [18]. In particular, by applying the procedure of sampling with replacement to the study data set, we can draw a large number of bootstrapping samples and estimate a cut point for each of the samples. On the basis of the empirical distribution of the cut point estimates, we may use the mean estimate as the final estimate of the optimal cut point, and the standard deviation to indicate the variation in the cut point estimates.

3. Simulation study

To examine the finite sample performance of the nonparametric cut point selection procedure, a simulation study was conducted with samples of size 50, 100, 150, and 200 for each of the two groups, that is, $n_0 = n_1$, and with sample size of 100 for one group and 200 for another group. For each sample size, 500 data sets were generated for a continuous variable from specific distribution models with class-specific parameters. The continuous variable $X^{(D)}$ has mean μ_D and variance $\sigma_D{}^2$ for $D = 0, 1$. The mean value μ_0= 0 is set for the Normal distribution models, while $\mu_0 = 1.5$ is for the Gamma distribution models. Values for $(\sigma_0{}^2, \sigma_1{}^2)$ are set to be (1, 1), (2, 1), and (1, 2) to account for the cases with equal and unequal within class variances. For the given values of $(\mu_0, \sigma_0{}^2, \sigma_1^2)$ in the normal distribution models or in the Gamma distribution models, values of μ_1 are chosen to yield classification accuracies of 0.60 and 0.80, respectively, and the values of the optimal cut point x_C are calculated accordingly. With each data set, the empirical method of selecting optimal cut point maximizing $CZ(x)$ is applied to get an estimate of the empirical estimator $\hat{x}_{\mathrm{C}}$. Then from the same data set, sampling with replacement is used to draw 200 bootstrap samples. Each bootstrap sample contributes one cut point estimate so that the mean of the 200 cut point estimates is used as bootstrap estimator $\hat{x}_{\mathrm{C}}^{\mathrm{b}}$ and the standard deviation SD_B of the 200 cut point estimates is used as the bootstrap estimator for the standard deviation of $\hat{x}_{\mathrm{C}}$.

Tables II and III show the bias of the empirical estimator $\hat{x}_{\mathrm{C}}$ and the bootstrap estimator $\hat{x}_{\mathrm{C}}^{\mathrm{b}}$ for the optimal cut point, and the standard deviation of $\hat{x}_{\mathrm{C}}$ and the mean of SD_B, by sample size and values for (μ_1, $\sigma_0{}^2$, $\sigma_1{}^2$). The bias, defined as the difference between mean estimates and true value of x_C, is consistently small with both Normal and Gamma models. On the basis of 200 bootstrap samples, the bootstrap estimator SD_B for standard deviation of $\hat{x}_{\mathrm{C}}$ performs well, because its mean is close to the standard deviation of $\hat{x}_{\mathrm{C}}$estimates, especially when sample size is large, that is, $m \geqslant 100$ for $m = \min(n_0, n_1)$. As expected, both the standard deviation of $\hat{x}_{\mathrm{C}}$ and mean SD_B decrease with increasing sample size. The decrease, however, becomes slower for sample size $m \geqslant 100$. There is evidence supporting the choice of equal group size. Given total sample size $(n = n_0 + n_1)$ of 300, the standard deviation of $\hat{x}_{\mathrm{C}}$ or mean SD_B is larger in the cases with unequal group size than the case with equal group size. The discrepancies, however, are small.

We also note that for a given *CA* and a sample size, the variation in cut point estimates is larger in the cases with $\sigma_0^2 \neq \sigma_1^2$ than in the cases with $\sigma_0^2 = \sigma_1^2$.

When there exists a cut point optimizing all three objective functions, one may estimate the cut point by optimizing anyone of the three empirically estimated objective functions based on the class specific

Table II. Performance of empirical cut point estimator with Normal distribution models ($\mu_0 = 0$).

CA	μ_1	$\sigma_0{}^2, \sigma_1{}^2$	n_0, n_1	Bias($\hat{x}_\mathrm{C}$)	Bias ($\hat{x}_\mathrm{C}^\mathrm{b}$)	$SD(\hat{x}_\mathrm{C})$	Mean (SD_B)
0.60	0.3583	1, 1	50, 50	0.0172	0.0035	0.2741	0.2850
			100, 100	−0.0085	0.0019	0.2118	0.2128
			150, 150	0.0155	0.0064	0.1857	0.1882
			100, 200	−0.0111	−0.0195	0.1991	0.2013
			200, 100	0.0256	0.0301	0.1992	0.1972
			200, 200	0.0057	0.0006	0.1568	0.1683
0.60	0.4388	2, 1	50, 50	−0.0144	−0.0080	0.3111	0.3159
			100, 100	0.0021	−0.0016	0.2550	0.2596
			150, 150	−0.0062	−0.0038	0.2132	0.2131
			100, 200	−0.0121	−0.0231	0.2302	0.2305
			200, 100	0.0080	0.0194	0.2212	0.2180
			200, 200	−0.0122	−0.0092	0.1872	0.1984
0.60	0.4388	1, 2	50, 50	0.0182	0.0121	0.3076	0.3177
			100, 100	−0.0151	−0.0082	0.2593	0.2547
			150, 150	0.0065	0.0108	0.2117	0.2189
			100, 200	−0.0061	−0.0231	0.2254	0.2277
			200, 100	0.0246	0.0286	0.2225	0.2381
			200, 200	0.0167	0.0073	0.1958	0.2030
0.80	1.1023	1, 1	50, 50	0.0084	0.0075	0.2569	0.2532
			100, 100	−0.0076	−0.0044	0.1835	0.1972
			150, 150	0.0054	0.0038	0.1705	0.1747
			100, 200	−0.0080	−0.0169	0.1828	0.1825
			200, 100	0.0132	0.0173	0.1868	0.1805
			200, 200	0.0031	0.0001	0.1580	0.1572
0.80	1.4577	2, 1	50, 50	−0.0169	−0.0033	0.2954	0.2898
			100, 100	−0.0056	−0.0073	0.2347	0.2266
			150, 150	0.0123	0.0060	0.1980	0.2030
			100, 200	−0.0191	−0.0200	0.2124	0.2133
			200, 100	0.0279	0.0299	0.2037	0.2072
			200, 200	−0.0005	0.0003	0.1834	0.1784
0.80	1.4577	1, 2	50, 50	0.0100	0.0034	0.2985	0.2964
			100, 100	−0.0166	−0.0117	0.2294	0.2227
			150, 150	−0.0131	−0.0079	0.1933	0.1983
			100, 200	−0.0267	−0.0303	0.2050	0.2041
			200, 100	0.0160	0.0163	0.2121	0.2087
			200, 200	0.0035	−0.0017	0.1869	0.1821

SD_B is the bootstrap estimate of standard deviation of $\hat{x}_C$.

empirical distributions. A simulation study is conducted to compare the performance of the three empirical estimators of $\hat{x}_\mathrm{C}$, $\hat{x}_\mathrm{J}$, and $\hat{x}_\mathrm{E}$ for the common optimal cut point x_m with Normal distribution models. Assuming that the continuous variable $X^{(D)}$ is from the Normal distribution for $D = 0, 1$, with $\mu_0 = 0$ and $\sigma_0{}^2 = \sigma_1{}^2 = \sigma^2$ taking the value of 1 or 2, then the optimal cut point $x_m = \mu_1/2$ will depend on the value of μ_1, which is set to yield classification accuracy of 0.60 or 0.80. For each sample size described before, 500 data sets are generated. With each dataset, cut point is selected by optimizing each of the three empirically estimated objective functions.

Table IV shows the bias and square root of mean square error for the three empirical estimators of $\hat{x}_\mathrm{C}$, $\hat{x}_\mathrm{J}$ and $\hat{x}_\mathrm{E}$. Not surprisingly, the three empirical estimators all have small biases. Therefore, the mean square error reflects the variance of an estimator. As expected, the square root of the mean square error decreases with increasing sample size and decreasing σ^2.

Because a flatter objective function produces a larger variation in the point estimator, the inequalities $-ER''(x_m) < CZ''(x_m) < J''(x_m)$ at optimal cut point x_m with Normal distribution models implies $\mathrm{Var}(\hat{x}_\mathrm{E}) < \mathrm{Var}(\hat{x}_\mathrm{C}) < \mathrm{Var}(\hat{x}_\mathrm{J})$. This pattern is apparent in the simulation result. For a given sample size and the value for (μ_1, σ^2), the smallest mean square error is of the empirical estimator $\hat{x}_\mathrm{E}$, although not much smaller than that of $\hat{x}_\mathrm{C}$. In contrast, the mean square error of $\hat{x}_\mathrm{J}$ is much larger than the other two estimators, especially when classification accuracy is low.

Table III. Performance of empirical cut point estimator with Gamma distribution models ($\mu_0 = 1.5$).

CA	μ_1	${\sigma_0}^2, {\sigma_1}^2$	n_0, n_1	Bias($\hat{x}_C$)	Bias ($\hat{x}_C^b$)	$SD(\hat{x}_C)$	Mean (SD_B)
0.60	1.7981	1, 1	50, 50	0.0193	0.0355	0.2442	0.2619
			100, 100	0.0065	0.0112	0.1863	0.1939
			150, 150	0.0010	0.0032	0.1655	0.1705
			100, 200	0.0020	−0.0011	0.1767	0.1782
			200, 100	0.0119	0.0264	0.1908	0.1839
			200, 200	0.0074	0.0098	0.1545	0.1541
0.60	1.6710	2, 1	50, 50	0.0046	0.0261	0.2510	0.2571
			100, 100	0.0159	0.0232	0.1930	0.1974
			150, 150	0.0163	0.0153	0.1644	0.1663
			100, 200	−0.0070	−0.0049	0.1721	0.1774
			200, 100	0.0185	0.0314	0.1749	0.1766
			200, 200	0.0209	0.0140	0.1542	0.1558
0.60	1.9930	1, 2	50, 50	0.0388	0.0407	0.3203	0.3152
			100, 100	0.0170	0.0253	0.2510	0.2490
			150, 150	0.0050	0.0130	0.2031	0.2133
			100, 200	−0.0007	−0.0049	0.2204	0.2257
			200, 100	0.0079	0.0320	0.2158	0.2230
			200, 200	0.0090	0.0156	0.1931	0.1949
0.80	2.5846	1, 1	50, 50	0.0165	0.0227	0.2363	0.2306
			100, 100	0.0066	0.0107	0.1755	0.1823
			150, 150	−0.0034	0.0026	0.1514	0.1533
			100, 200	−0.0022	−0.0054	0.1677	0.1681
			200, 100	0.0051	0.0165	0.1561	0.1628
			200, 200	−0.0035	0.0026	0.1395	0.1441
0.80	2.6872	2, 1	50, 50	0.0144	0.0349	0.2427	0.2424
			100, 100	0.0088	0.0125	0.1940	0.1866
			150, 150	0.0049	0.0098	0.1676	0.1715
			100, 200	−0.0131	−0.0113	0.1743	0.1749
			200, 100	0.0205	0.0369	0.1679	0.1742
			200, 200	0.0067	0.0083	0.1454	0.1449
0.80	2.8536	1, 2	50, 50	0.0223	0.0252	0.2823	0.2840
			100, 100	0.0099	0.0139	0.2271	0.2186
			150, 150	−0.0050	0.0096	0.2015	0.1943
			100, 200	−0.0053	−0.0075	0.2029	0.2051
			200, 100	0.0241	0.0314	0.1942	0.1997
			200, 200	0.0046	0.0058	0.1736	0.1764

SD_B is the bootstrap estimate of standard deviation of $\hat{x}_C$.

4. Application

Blood arsenic level is a biomarker of arsenic exposure, which is associated with arsenic-induced skin lesions [16]. To find a threshold in the blood arsenic level that would constitute a warning sign for the development of arsenic-induced skin lesions, we apply the proposed method, Youden index and the closest-to-(0, 1) criterion based approach to data collected from a nested case-control study of Bangladeshi adults with chronic exposure to arsenic [16].

The study includes 274 cases who developed arsenic-induced skin lesions during the two-year follow-up period post-baseline assessment, and 274 controls matched to the cases by gender, age within five years, and drinking well arsenic level within 100 μg/L. Baseline blood arsenic was measured using the blood samples from 261 cases and 196 controls that had sufficient amount of blood for arsenic assessment. Table V shows the sample characteristics. The cases and controls had comparable proportion of males (69.7% vs 67.9%), and the incident cases, with a mean age of 44.7 years, were slightly older than the controls, whose mean age was 43.0 years ($p = 0.0551$). The incident cases had higher drinking well arsenic levels, with a median of 118 μg/L versus 107 μg/L in controls. Although the difference in water arsenic was not statistically significant, the incident cases did have significantly higher levels of blood

Table IV. Comparison of empirical estimators of the common optimal cut point by three criteria with Normal distribution models ($\mu_0 = 0,\ \sigma_0{}^2 = \sigma_1{}^2 = \sigma^2$).

			Sample size	X_C	X_J	X_E
CA	μ_1	σ^2	n_0, n_1	Bias ($\sqrt{}$ MSE)	Bias ($\sqrt{}$ MSE)	Bias ($\sqrt{}$ MSE)
0.60	0.3583	1	50, 50	–0.0022 (0.2719)	–0.0084 (0.5520)	0.0025 (0.2340)
			100, 100	0.0036 (0.2100)	–0.0138 (0.4633)	0.0055 (0.1864)
			150, 150	0.0096 (0.1818)	0.0002 (0.4183)	0.0094 (0.1572)
			100, 200	0.0001 (0.1963)	–0.0081 (0.4590)	–0.0045 (0.1651)
			200, 100	0.0264 (0.2024)	–0.0122 (0.4469)	0.0234 (0.1767)
			200, 200	–0.0001 (0.1725)	0.0064 (0.3948)	–0.0005 (0.1496)
0.60	0.5067	2	50, 50	0.0233 (0.3929)	–0.0098 (0.7549)	0.0108 (0.3420)
			100, 100	0.0153 (0.3068)	0.0329 (0.7304)	0.0135 (0.2728)
			150, 150	0.0186 (0.2602)	–0.0015 (0.6229)	0.0105 (0.2274)
			100, 200	–0.0093 (0.2670)	–0.0127 (0.6210)	–0.0160 (0.2392)
			200, 100	0.0390 (0.2812)	0.0339 (0.6333)	0.0376 (0.2437)
			200, 200	0.0085 (0.2337)	0.0052 (0.5306)	0.0084 (0.1962)
0.80	1.1902	1	50, 50	–0.0029 (0.2607)	–0.0136 (0.3088)	–0.0081 (0.2077)
			100, 100	0.0100 (0.1936)	0.0063 (0.2518)	0.0195 (0.1520)
			150, 150	0.0071 (0.1683)	0.0161 (0.2191)	0.0021 (0.1330)
			100, 200	–0.0144 (0.1719)	–0.0091 (0.2236)	–0.0119 (0.1289)
			200, 100	0.0332 (0.1728)	0.0217 (0.2250)	0.0245 (0.1414)
			200, 200	–0.0124 (0.1566)	–0.0189 (0.1984)	0.0009 (0.1159)
0.80	1.6832	2	50, 50	0.0193 (0.3647)	0.0004 (0.4415)	0.0159 (0.2834)
			100, 100	–0.0015 (0.2821)	–0.0046 (0.3410)	0.0118 (0.2065)
			150, 150	–0.0256 (0.2361)	–0.0424 (0.3094)	–0.0136 (0.1902)
			100, 200	–0.0269 (0.2456)	–0.0441 (0.3265)	–0.0321 (0.1911)
			200, 100	–0.0027 (0.2472)	–0.0204 (0.3198)	–0.0013 (0.1975)
			200, 200	–0.0085 (0.2200)	–0.0140 (0.2944)	–0.0047 (0.1665)

Table V. Sample characteristics of the study participants.

Baseline variable	Skin lesion cases $N = 261$		Controls $N = 196$		p-value
Male	69.7%		67.9 %		0.6682
	Mean ± SD	Median(Range)	Mean ± SD	Median(Range)	
Age	44.7 ± 9.5	45 (21, 69)	43.0 ± 9.1	42 (21, 63)	0.0551
Water arsenic (μg/L)	150.6 ± 141.8	118 (0.1, 790)	130.8 ± 125.7	107 (0.1, 564)	0.3881
Blood arsenic (μg/L)	14.3 ± 10.0	11.8 (1.9, 63.5)	11.0 ± 6.7	9.6 (2.0, 41.4)	0.0003

Chi-square test was used for difference in proportion of male between cases and controls; and t-test was used for difference in age, logarithmic transformed water and blood arsenic measures between cases and controls.

arsenic at baseline ($p = 0.0003$), with mean 14.35 ($SD = 9.98$) μg/L, compared with mean 11.04 ($SD = 6.89$) μg/L among the controls. The AUC with blood arsenic in Figure 2 was 0.5980, indicating an informative classification accuracy of baseline blood arsenic level, which aptly distinguishes controls from the incident cases who developed arsenic-induced skin lesions within 2 years from the baseline.

Table VI presents the selected cut point on blood arsenic by the empirical method with three criteria. The estimated optimal threshold $\hat{x}_{\mathrm{C}} = 10.55$ μg/L is the same as $\hat{x}_{\mathrm{E}}$. Above this cut point, the sensitivity is 0.5747, and at or below the cut point, the specificity is 0.5765. With the 200 bootstrapping samples, the mean of the cut point estimates by maximizing the proposed objective function is $\hat{x}_{\mathrm{C}}^{\mathrm{b}} = 10.79$ ($SD_B = 1.24$) μg/L and by meeting closest-to-(0, 1) criterion is $\hat{x}_{\mathrm{E}}^{\mathrm{b}} = 10.77$ ($SD_B = 1.13$)μg/L, respectively. The standard deviation estimates indicate a moderate variation in the cut point estimates. In contrast, the cut point estimate maximizing Youden index has a larger value than $\hat{x}_{\mathrm{J}} = 12.0\mu g/\mathrm{L}$, with a poor sensitivity of 0.4904 and a better specificity of 0.6735. The bootstrap standard deviation estimate of 2.84 μg/L indicates a large variation in the estimates. When restricting the cut point selection to the set of points with sensitivity above 0.50, the cut point estimate maximizing Youden index is then $\hat{x}_{\mathrm{J}} = 10.55\mu g/\mathrm{L}$, equivalent to $\hat{x}_{\mathrm{C}}$ and $\hat{x}_{\mathrm{E}}$.

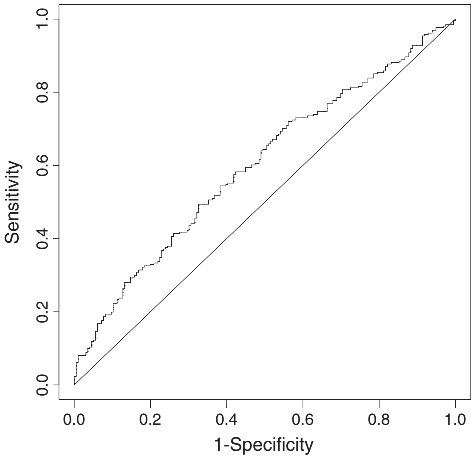

Figure 2. ROC curve with blood arsenic measure.

Table VI. Selected cut point on blood arsenic by empirical method with three criteria.

	New	Youden index	Closest-to-(0,1)
Selected cut point	10.55	12.00	10.55
Sensitivity at the point	0.5747	0.4904	0.5747
Specificity at the point	0.5765	0.6735	0.5765
Bootstrap estimation (200 bootstrap samples)			
Mean (cut point estimates)	10.7891	11.5008	10.7703
SD (cut point estimates)	1.2374	2.8424	1.1307

5. Discussion

To select a threshold for a quantitative diagnostic/screening test or a biomarker, a concordance probability, useful for evaluating the classification accuracy of a dichotomized measure, was used to define an objective function $CZ(x)$. Like the popularly used Youden index $J(x)$ for cut point selection, the function has a simple form with straightforward interpretation, and a relationship with the ROC curve. Maximizing $J(x)$, one is able to find the optimal cut point x_J. Similarly, cut point x_C can be obtained by maximizing $CZ(x)$. When the conditional distributions are parametric, the cut point x_J may have a closed form, which can be explicitly expressed as a function of the model parameters. For example, in the case of Normal distributions, the cut point x_J is a function of four model parameters including two class-specific means and two standard deviations [5, 7]. The cut point x_C optimizing $CZ(x)$, however, does not have a closed form. Although $x_C = x_J$ holds under certain conditions, in general, $x_C \neq x_J$ and $CZ(x_C) \geqslant$ CZ(x_J).

Unlike the concordance probability $CA = P(X_i^{(0)} < X_j^{(1)})$, which is a constant indicating classification accuracy of quantitative measure X, the function $CZ(x) = P(X^{(0)} \leqslant x < X^{(1)})$ evaluates the classification accuracy of the measure dichotomized at cut point x. In the selection of optimal cut point, it seems more meaningful to maximize this objective function than to optimize the other quantities. Therefore, the proposed criterion is recommended for cut point selection.

To estimate the optimal cut point, we propose using an empirical method, a simple nonparametric procedure based on empirically estimated conditional distributions. In a simulation study, the nonparametric estimator of cut point $\hat{x}_C$ showed a very small bias. The bootstrap estimation of standard deviation of empirical cut point estimator based on 200 bootstrap samples seems to work well that the mean bootstrap *SD* estimates is close to the *SD* of the empirical cut point estimator. The simulation study also showed a trend that variation in cut point estimates decreases with increasing sample size. The pattern in the standard deviation of the cut point estimates supports a preference for equal sample sizes for each group. In the case with equal group sizes, the variation reduction is more profound when group size increases from 50 to 100, suggesting use of a group size at least 100. On the basis of the result of simulation study, we also recommend a large sample size for the case when a measure X has different variations in the two classes.

Other nonparametric procedures such as the kernel smoothing method [5, 19] can also be used to estimate the cut point. The kernel method, however, requires not only more computational work but also the choices with regard to kernel function and the bandwidth for smoothing the objective function.

Logistic regression models are popularly used in analyses with multiple predictors for binary classifications. Using the estimated model parameters based on the data from a prospective study and given values for the predictors, one is able to estimate an individual's probability of being in the diseased group. To classify a subject into one of the two groups with and without the disease, it is necessary to establish a cut point on the estimated probability. The proposed method may provide a simple means to select a meaningful threshold for the binary classification.

Acknowledgements

The author gratefully thanks the reviewers for their valuable comments and suggestions, and also thanks Dr. Gamble for providing the data to illustrate the method. Dr. Gamble's study was supported by the National Institute of Environment Health Sciences (grants NOs: R01-ES011601, P42-ES10349 and P30-ES09089).

References

1. Zhou XH, Obuchowski NA, McClish DK. *Statistical Methods in Diagnostic Medicine*. Wiley: New York, 2002.
2. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press: New York, 2003.
3. Krazanowski WJ, Hand DJ. *ROC Curves for Continuous Data*. CRC press: New York, 2009.
4. Youden WJ. Index for rating diagnostic tests. *Cancer* 1950; **3**:32–35.
5. Fluss R, Faraggi D, Reiser B. Estimation of the Youden Index and it's associated cutoff point. *Biometrical Journal* 2005; **47**(4):458–472.
6. Perkins NJ, Schisterman EF. The inconsistency of "optimal" cut-points using two ROC based criteria. *American Journal of Epidemiology* 2006; **163**(7):670–675.
7. Schisterman EF, Perkins NJ. Confidence intervals for the Youden index and corresponding optimal cut-point. *Communications in Statistics - Simulation and Computation* 2007; **36**:549–563.
8. Lai CY, Tian L, Schisterman EF. Exact confidence interval estimation for the Youden index and its corresponding optimal cut-point. *Computational Statistics and Data Analysis* 2010. DOI: 10.1016/jcsda.2010.11.023.
9. Klotsche J, Ferger D, Pieper L, Rehm J, Wittchen H-U. A novel nonparametric approach for estimating cut-offs in continuous risk indicators with application to diabetes. *BMC medical Research Methodology* 2009; **9**(63). DOI: 10.1186/1471-2288-9-63.
10. Schisterman EF, Perkins NJ, Liu A, Bondell H. Optimal cut-point and its corresponding Youden index to discriminate individuals using pooled blood samples. *Epidemiology* 2005; **16**(1):73–81.
11. Perkins NJ, Schisterman EF. The Youden index and optimal cut-point corrected for measurement error. *Biometrical Journal* 2005; **4**:428–441.
12. Ruopp MD, Perkins NJ, Whitcomb BW, Schisterman EF. Youden index and optimal cut-point estimated from observations affected by a lower limit of detection. *Biometrical Journal* 2008; **50**(3):419–430.
13. Schisterman EF, Faraggi D, Reiser B, Hu J. Youden Index and the optimal threshold for markers with mass at zero. *Statistics in Medicine* 2008; **27**:297–315. DOI: 10.1002/sim.2993.
14. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; **143**:29–36.
15. Liu X, Jin Z. Item reduction in a scale for screening. *Statistics in Medicine* 2007; **26**(23):4311–4327.
16. Pilsner JR, Liu X, Ahsan H, Ilievski V, Slavkovich V, Levy D, Factor-Litvak P, Graziano JH, Gamble MV. Folate Deficiency, Hyperomocystinemia and Hypomethylation of Genomic DNA are Risk Factors for Arsenic-Induced Skin Lesions: A Nested Case-Control Study in Bangladesh. *Environmental Health Perspectives* 2009; **117**(2):254–260.
17. Lee AJ. *U-Statistics: Theory and Practice*. Marcel Dekker Inc: New York and Basel, 1990.
18. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Chapman & Hall: New York, 1993.
19. Wand MP, Jones MC. *Kernel Smoothing*. Chapman & Hall: New York, 1995.