Communications for Statistical Applications and Methods
2020, Vol. 27, No. 2, 211–224

https://doi.org/10.29220/CSAM.2020.27.2.211
Print ISSN 2287-7843 / Online ISSN 2383-4757

Positive and negative predictive values by the TOC curve

Chong Sun Hong[1,a], So Yeon Choi[a]

[a]Department of Statistics, Sungkyunkwan University, Korea

Abstract

Sensitivity and specificity are popular measures described by the receiver operating characteristic (ROC) curve. There are also two other measures such as the positive predictive value (PPV) and negative predictive value (NPV); however, the PPV and NPV cannot be represented by the ROC curve. Based on the total operating characteristic (TOC) curve suggested by Pontius and Si (*International Journal of Geographical Information Science*, **97**, 570–583, 2014), explanatory methods are proposed to geometrically describe the PPV and NPV by the TOC curve. It is found that the PPV can be regarded as the slope of the right-angled triangle connecting the origin to a certain point on the TOC curve, while 1 − NPV can be represented as the slope of the right-angled triangle connecting a certain point to the top right corner of the TOC curve. When the neutral zone exists, the PPV and 1 – NPV can be described as the slopes of two other right-angled triangles of the TOC curve. Therefore, both the PPV and NPV can be estimated using the TOC curve, whether or not the neutral zone is present.

Keywords: accuracy, hypotenuse, performance, sensitivity, specificity, threshold

1. Introduction

Assume that the set C that represents patient states consists of two elements, $C = \{0, 1\}$, where elements 0 and 1 represent non-disease / negative and disease / positive states, respectively; and let X be a random variable that measures disease or non-disease. The cumulative distribution function of X, $F(x)$, is assumed to be a linear combination of a distribution function $F_d(x) = P(X \le x|C = 1)$ representing a disease state of sample size p, and a distribution function $F_n(x) = P(X \le x|C = 0)$ representing a non-disease state of the size q, as follows: $F(x) = \pi_1 F_d(x) + (1 - \pi_1)F_n(x)$, where $\pi_1 \in (0, 1)$ is regarded as the total probability of disease, and can be estimated as $p/(p + q)$, the ratio of the sizes of the two groups. $F_d(x)$ and $F_n(x)$ are also considered as probability classifiers with $F_n(x) \ge F_d(x)$ for all x (Metz and Kronman, 1980; Hsieh and Turnbull, 1996; Provost and Fawcett, 2001; Engelmann *et al.*, 2003; Pepe, 2003; Fawcett, 2006).

Classification results can be expressed by a confusion matrix. The true positive (TP) and true negative (TN) in the confusion matrix represent the numbers of correctly classified diseased and non-diseased populations. The true positive rate (TPR, sensitivity) and true negative rate (TNR) have the following relationship: for any threshold (cut-off point) x,

$$\text{TPR} = \frac{\text{TP}}{p} = 1 - \hat{F}_d(x), \qquad \text{TNR} = \frac{\text{TN}}{q} = \hat{F}_n(x).$$

[1] Corresponding author: Department of Statistics, Sungkyunkwan University, 25-2 Sungkyunkwan-ro, Jongno-gu, Seoul 03063, Korea. E-mail: cshong@skku.edu

Published 31 March 2020 / journal homepage: http://csam.or.kr

However, the false positive (FP) and false negative (FN) are the numbers of populations that predicted disease as non-disease, and predicted non-disease as disease, respectively. The false positive rate (FPR, 1 − specificity) and false negative rate (FNR) have the following relationship:

$$\mathrm{FPR} = \frac{\mathrm{FP}}{q} = 1 - \hat{F}_n(x), \qquad \mathrm{FNR} = \frac{\mathrm{FN}}{p} = \hat{F}_d(x).$$

The total number of disease groups, p, is TP + FN in the population, and the number of non-diseased populations, q, is FP + TN. Assume that both p and q are known (Provost and Fawcett, 1997; Fawcett, 2003; Stein, 2005; Hong *et al.*, 2009).

The 'sensitivity' and 'specificity' are the basic accuracy measures of diagnosis. Whereas the 'sensitivity' is the probability that a patient with an actual disease will be positive ($\hat{C} = 1$), 'specificity' is the probability of considering a non-diseased person to be negative ($\hat{C} = 0$). Both the 'sensitivity' and 'specificity' are statistics measured in terms of evaluating the accuracy and performance of the diagnosis that are expressed by the following probability equations for any threshold x on the receiver operating characteristic (ROC) curve (Zhou *et al.*, 2002):

$$\text{sensitivity} = P\left(\hat{C} = 1 | C = 1\right), \qquad \text{specificity} = P\left(\hat{C} = 0 | C = 0\right).$$

When the evaluation result of the disease is positive or negative, the probability of the test performance that an actual health condition is disease or non-disease may be required, rather than how accurate the diagnosis is. There exist two other measures: the positive predictive value (PPV) and negative predictive value (NPV) suggested by Raslich *et al.* (2007) and Zhou *et al.* (2002). The PPV is the probability that the patient has the disease following a positive test result, and the NPV is the probability that the patient does not have the disease following a negative test result. The PPV tells the clinician what percent of those with a positive finding have the disease; however, the NPV reveals what percent of those with a negative result do not have the disease. For any threshold x, the PPV and NPV can be expressed as (Raslich *et al.*, 2007):

$$\mathrm{PPV} = P\left(C = 1 | \hat{C} = 1\right), \qquad \mathrm{NPV} = P\left(C = 0 | \hat{C} = 0\right). \tag{1.1}$$

The neutral zone was proposed to ensure the retesting of diseases with high diagnostic costs. Both the PPV and NPV were also derived when the neutral zone is present (Daniel and Steven, 2016).

Based on statistical decision theory, the ROC curve is a visual tool that can easily identify the classifier's performance in binary classification (Egan, 1975; Zweig and Campbell, 1993; Bradley, 1997; Fawcett and Provost, 1997; Pepe, 2000; Hong and Lee, 2018). Since the ROC curve is implemented with (FPR, TPR) = (1 − specificity, sensitivity) in a unit length square, the 'sensitivity' and 'specificity' can be represented by the ROC curve. In contrast, the PPV and NPV cannot be represented by the ROC curve. Pontius and Si (2014) proposed the total operating characteristic (TOC) curve, which is expressed with the numbers of observations of TP, FN, FP, and TN. The TOC curve is implemented as a curve in a parallelogram with a hypotenuse slope of 45 degrees, belonging to a rectangle, where the lengths of the horizontal and vertical axes are $p + q$ and p, respectively.

In this study, we investigate how the PPV and NPV can be represented by the TOC curve, neither of which can be expressed by the ROC curve. We explain their meanings in terms of geometry, then both the PPV and NPV are examined in relation to the shape of the TOC curve. New explanatory methods are proposed to geometrically describe the PPV and NPV by the TOC curve. In addition, the PPV and NPV are also geometrically described and implemented by the TOC curve, when the neutral

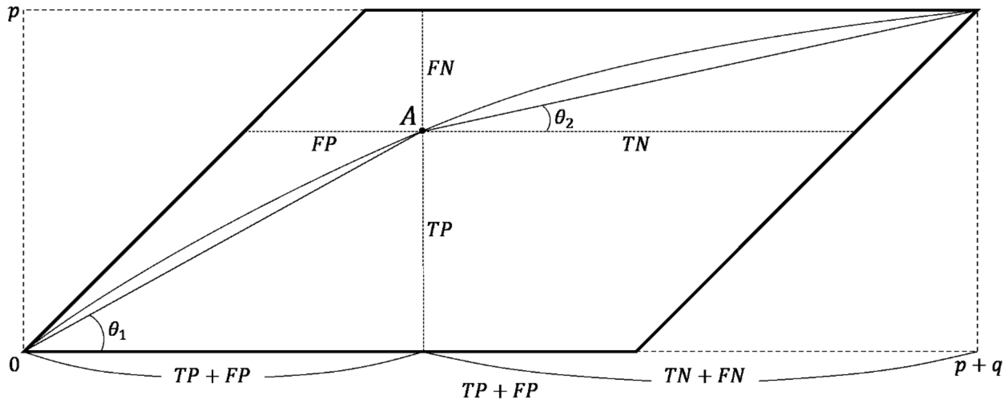


Figure 1: *PPV and NPV by the TOC curve. PPV = positive predictive value; NPV = negative predictive value; TOC = total operating characteristic.*

zone is present. We discuss these relationships between the descriptions of the PPV and NPV by the TOC curve mathematically.

Section 2 of this paper introduces the TOC curve, explains definitions of the PPV and NPV as well as geometrically discusses the representations by the TOC curve of both the PPV and NPV, so that their representations by the TOC curve could be explored. Section 3 discusses and derives the PPV and NPV with probability equations when the neutral zone is present; expresses the PPV and NPV in this case by the TOC curve, and then geometrically compares this description of the TOC curve with those represented in Section 2. Section 4 generates two random samples representing disease and non-disease states with different sample sizes, then obtains and discusses the values of PPV and NPV with interpretations derived in this work. This section also generates two random samples in the presence of the neutral zone with some kinds of two types of errors, and discusses and explores some results based on the neutral zone. Finally, some conclusions are summarized in Section 5.

2. PPV and NPV by the TOC curve

2.1. TOC curve

If two samples are skewed, imbalanced data or very different in size, p and q, the well-known ROC curve has the disadvantage of insufficiently expressing the characteristics of the classification results (Fawcett, 2006; Garcia *et al.*, 2010). Since the ROC curve describes the entire information only in terms of TPR and FPR, it is difficult to explain the frequency of their confusion matrix (Pontius and Si, 2014). To make up for the problems of these ROC curves, Pontius and Si (2014) proposed the TOC curve expressed with cell observations of the confusion matrix.

Figure 1 shows a TOC plot. The unit on both the horizontal and vertical axes of the TOC plot indicates the number of observations. The horizontal axis ranges from 0 to $p + q$, which is the total number of people in the disease and non-disease groups, respectively. The vertical axis ranges from 0 to p, which is the size of the disease group. The TOC curve is implemented as a curve in a parallelogram with a hypotenuse slope of 1 (45 degrees), belonging to a rectangle whose lengths of horizontal and vertical axes are $p + q$ and p, respectively. The coordinate on the horizontal axis indicates the number of diseased in the diagnosis, TP(x) + FP(x), for a given threshold x, while the vertical axis indicates the number of correctly classified diseases TP(x). That is, the coordinate of TOC curve is given by $(\mathrm{TP}(x) + \mathrm{FP}(x), \mathrm{TP}(x))$ for a given threshold x.

From any point A whose coordinate is $(\mathrm{TP}(x) + \mathrm{FP}(x), \mathrm{TP}(x))$ on the TOC curve in Figure 1, the lengths to each side of a parallelogram are represented as TP, FP, TN, and FN of the confusion matrix. That is, the lengths from a point A on the curve to the base and the top sides of the parallelogram are TP and FN, respectively; while the lengths to the left and right sides of the parallelogram are FP and TN, respectively. Therefore, there is the advantage that it is possible to identify and describe information for confusion matrices that cannot be derived from the ROC curve (Pontius and Si, 2014).

2.2. PPV and NPV by the TOC curve

Both conditional probabilities PPV and NPV of the actual disease or non-disease state given a positive or negative diagnosis result in (1.1), can be expressed with frequencies of the confusion matrix as shown in (2.1). For any threshold x,

$$\mathrm{PPV}(x) = \frac{\mathrm{TP}(x)}{\mathrm{TP}(x) + \mathrm{FP}(x)}, \qquad \mathrm{NPV}(x) = \frac{\mathrm{TN}(x)}{\mathrm{TN}(x) + \mathrm{FN}(x)}. \tag{2.1}$$

The denominators of $\mathrm{PPV}(x)$ and $\mathrm{NPV}(x)$ in (2.1) are $\mathrm{TP}(x)+\mathrm{FP}(x)$ and $\mathrm{TN}(x)+\mathrm{FN}(x)$, respectively, so it is difficult to derive and explain their meanings with the ROC curve, where the vertical and horizontal axes are TPR and FPR, respectively. However, since the total length of the horizontal axis in the TOC curve is $p+q$, which equals the sum of the denominators of $\mathrm{PPV}(x)$ and $\mathrm{NPV}(x)$, $\mathrm{PPV}(x)$ and $\mathrm{NPV}(x)$ could be described by the TOC curve. Based on the coordinate $((\mathrm{TP}(x) + \mathrm{FP}(x)), \mathrm{TP}(x))$ of point A on TOC curve in Figure 1, its horizontal and vertical coordinates can be found to be the denominator and numerator of $\mathrm{PPV}(x)$, respectively. Hence, $\mathrm{PPV}(x)$ can be described by the slope from the origin to point A, considering the right-angled triangle connecting the origin and point A. From now on, $1-\mathrm{NPV}(x)$ may be considered, rather than $\mathrm{NPV}(x)$. Then, $1-\mathrm{NPV}(x) = \mathrm{FN}(x)/(\mathrm{TN}(x)+\mathrm{FN}(x))$. Note that the denominator of $1-\mathrm{NPV}(x)$ is $\mathrm{TN}(x)+\mathrm{FN}(x)$, which is the total length of the horizontal axis $p+q$ in the TOC curve minus $\mathrm{TP}(x)+\mathrm{FP}(x)$, and its numerator is $\mathrm{FN}(x)$. Hence, we might regard $1 - \mathrm{NPV}(x)$ as the slope of the right-angled triangle connecting point A and the top right corner of the TOC curve. Therefore, both $\mathrm{PPV}(x)$ and $1 - \mathrm{NPV}(x)$ can also be expressed as the slopes of two right-angled triangles as in Result 1:

Result 1. For any threshold x, both $\mathrm{PPV}(x)$ and $1 - \mathrm{NPV}(x)$ could be described as the angles of the two following right-angled triangles on the TOC curve:

$\mathrm{PPV}(x)$ is the slope of the right-angled triangle connecting the origin to point A,

$1 - \mathrm{NPV}(x)$ is the slope of the right-angled triangle connecting point A to a point $(p + q, p)$,

where the coordinate of A is $((\mathrm{TP}(x) + \mathrm{FP}(x)), \mathrm{TP}(x))$. Therefore,

$$\mathrm{PPV}(x) = \tan\theta_1 \quad \text{and} \quad 1 - \mathrm{NPV}(x) = \tan\theta_2,$$

where θ_1 and θ_2 are the corresponding angles of two right-angled triangles connecting the origin to point A and point A to the top right corner of the TOC curve, respectively, in Figure 1.

Since both the ROC curve and the TOC curve are concave, both $\mathrm{PPV}(x)$ and $1 - \mathrm{NPV}(x)$ are decreasing as the point A moves from the origin to the top right corner of the TOC curve. As the values of X decrease to $-\infty$, point A moves to the top right corner of the TOC curve, so that we may conclude that $\mathrm{PPV}(x)$ is decreasing, and $\mathrm{NPV}(x)$ is increasing.

The slope from the origin to the rightmost coordinate $(p+q, p)$ of the TOC curve is $p/(p+q)$, which turns out to be the weighted average of $\text{PPV}(x)$ and $1 - \text{NPV}(x)$ explained in Result 1. $P(C = 1) = \text{PPV}(x) \times P(\hat{C} = 1) + (1 - \text{NPV}(x)) \times P(\hat{C} = 0)$ with $P(C = 1) = p/(p+q)$; therefore, the following relationship can be obtained:

$$\frac{p}{p+q} = \text{PPV}(x) \times \frac{\text{TP}(x) + \text{FP}(x)}{p+q} + (1 - \text{NPV}(x)) \times \frac{\text{TN}(x) + \text{FN}(x)}{p+q}, \tag{2.2}$$

where $P(\hat{C} = 1) = (\text{TP}(x) + \text{FP}(x))/(p+q)$ and $P(\hat{C} = 0) = (\text{TN}(x) + \text{FN}(x))/(p+q)$. Note that each weight indicates the ratio of the lengths of the horizontal axis divided at point A.

Result 2. For any threshold x, both $\text{PPV}(x)$ and $\text{NPV}(x)$ have the following properties:

$$\text{PPV}(x) \geq \frac{p}{p+q}, \qquad \text{NPV}(x) \geq \frac{q}{p+q}. \tag{2.3}$$

Since $\text{PPV}(x)$ is decreasing and $\text{NPV}(x)$ is increasing as point A moves from the origin to the top right corner of the TOC curve, the minimum value of $\text{PPV}(x)$ can be obtained as $p/(p+q)$, and the minimum value of $\text{NPV}(x)$ is $q/(p+q)$, based on Result 2.

3. PPV and NPV with the neutral zone by the TOC curve

3.1. PPV and NPV with the neutral zone

Sometimes, one of the rates FPR and FNR can be viewed as more important to control. However, it is not possible to choose a point on the ROC curve that achieves the targeted values for both FPR and FNR. In addition, difficult classification problems incur the risk of high FPR and/or high FNR when the group features are not very different between the two groups. In these situations, the most desirable solution is to identify features that have higher discrimination ability. Hence, Daniel and Steven (2016) proposed neutral zone classifiers that add a soft classification outcome, "neutral" in order to handle ambiguous cases and allow the control of both FPR and FNR. Consider the following classifier:

$$\hat{C} = \begin{cases} 1, & \text{disease}, \\ N, & \text{neutral zone}, \\ 0, & \text{non-disease}. \end{cases}$$

Two constants x_0 and x_1 with $x_0 < x_1$ are found by exerting explicit controls of FPR and FNR. Specifically, x_0 and x_1 are the solutions of the following two equations: for two types of errors, α and β,

$$\begin{aligned} \text{FPR}(x) &= 1 - \hat{F}_n(x) \leq P(X \geq x_1 | C = 0) = P\left(\hat{C} = 1 | C = 0\right) = \alpha, \\ \text{FNR}(x) &= \hat{F}_d(x) \leq P(X \leq x_0 | C = 1) = P\left(\hat{C} = 0 | C = 1\right) = \beta. \end{aligned}$$

The solutions for the two constants are $x_0 = \hat{F}_d^{-1}(\beta)$ and $x_1 = \hat{F}_n^{-1}(1-\alpha)$ for fixed values of α and β. A neutral zone exists if and only if $\hat{F}_d^{-1}(\beta) < \hat{F}_n^{-1}(1-\alpha)$ (see Daniel and Steven (2016) for more detail).

Both NZR_0 and NZR_1 are denoted as probabilities for the neutral zone,

$$\begin{aligned}\text{NZR}_0 &= P\left(\hat{C} = N|C = 0\right) = \hat{F}_n(x_1) - \hat{F}_n(x_0) = \text{TNR}(x_1) - \text{TNR}(x_0),\\ \text{NZR}_1 &= P\left(\hat{C} = N|C = 1\right) = \hat{F}_d(x_1) - \hat{F}_d(x_0) = \text{TPR}(x_0) - \text{TPR}(x_1).\end{aligned} \tag{3.1}$$

The NZ_0 and NZ_1 were defined as $\text{NZ}_0 = q \times \text{NZR}_0$ and $\text{NZ}_1 = p \times \text{NZR}_1$, respectively, where areas of NZR_0 and NZR_1 were demonstrated for two probability density functions, and two straight lines were implemented around the ROC curve.

Daniel and Steven (2016, p.2350) defined both PPV $= P(C = 1|\hat{C} = 1)$ and NPV $= P(C = 0|\hat{C} = 0)$ in the presence of the neutral zone. Here, we could derive the following PPV and NPV using (3.1):

$$\begin{aligned}\text{PPV}(x_0, x_1) &= \frac{\text{TP}(x_0) - \text{NZ}_1}{\text{TP}(x_0) - \text{NZ}_1 + \text{FP}(x_1)},\\ \text{NPV}(x_0, x_1) &= \frac{\text{TN}(x_1) - \text{NZ}_0}{\text{TN}(x_1) - \text{NZ}_0 + \text{FN}(x_0)}.\end{aligned} \tag{3.2}$$

Note that both the PPV and NPV in (3.2) depend on x_0, as well as x_1.

3.2. PPV and NPV with the neutral zone by the TOC curve

Suppose that A_0 and A_1 in Figure 2 are the points on the TOC curve corresponding to two arbitrary thresholds x_0 and x_1 ($x_0 < x_1$), respectively. Their coordinates are then assumed to be $A_0 = (\text{TP}(x_0) + \text{FP}(x_0), \text{TP}(x_0))$ and $A_1 = (\text{TP}(x_1) + \text{FP}(x_1), \text{TP}(x_1))$. Considering two coordinates of A_0 and A_1, the horizontal length between A_0 and A_1 can be found to be $\text{NZ}_0 + \text{NZ}_1$. That is, $(\text{TP}(x_0) + \text{FP}(x_0)) - (\text{TP}(x_1) + \text{FP}(x_1)) = \text{NZ}_0 + \text{NZ}_1$.

It is also found that subtracting NZ_1 from the vertical height of A_0 is equal to the vertical height of A_1, so that $\text{TP}(x_0) - \text{NZ}_1 = \text{TP}(x_1)$ can be obtained. The lengths from the two points A_0 and A_1 to the right side of the parallelogram in Figure 2 are also $\text{TN}(x_0)$ and $\text{TN}(x_1)$, respectively. Since the slope of the parallelogram hypotenuse is 1, it can be confirmed that $\text{TN}(x_1) - \text{NZ}_0 = \text{TN}(x_0)$. This relationship can be seen in Figure 1, and the following (3.3) can be derived using (2.1) and (3.1):

$$\text{TP}(x_0) - \text{NZ}_1 = \text{TP}(x_1), \qquad \text{TN}(x_1) - \text{NZ}_0 = \text{TN}(x_0). \tag{3.3}$$

The PPV and NPV in the presence of the neutral zone could be also expressed as in Result 3 by using (3.3).

Result 3. For any two thresholds x_0 and x_1, both PPV and 1 − NPV in the presence of the neutral zone are represented as:

$$\begin{aligned}\text{PPV}(x_1) &= \frac{\text{TP}(x_1)}{\text{TP}(x_1) + \text{FP}(x_1)},\\ 1 - \text{NPV}(x_0) &= \frac{\text{FN}(x_0)}{\text{TN}(x_0) + \text{FN}(x_0)}.\end{aligned}$$

Note that whereas both the PPV and NPV in (3.2) depend on x_0 and x_1 together, the PPV and NPV in Result 3 turn out to be the functions of x_1 and x_0 only, respectively.

We also conclude that both the PPV and NPV in the presence of the neutral zone could be explained as the slopes of two right-angled triangles. Based on the coordinate on the TOC curve,

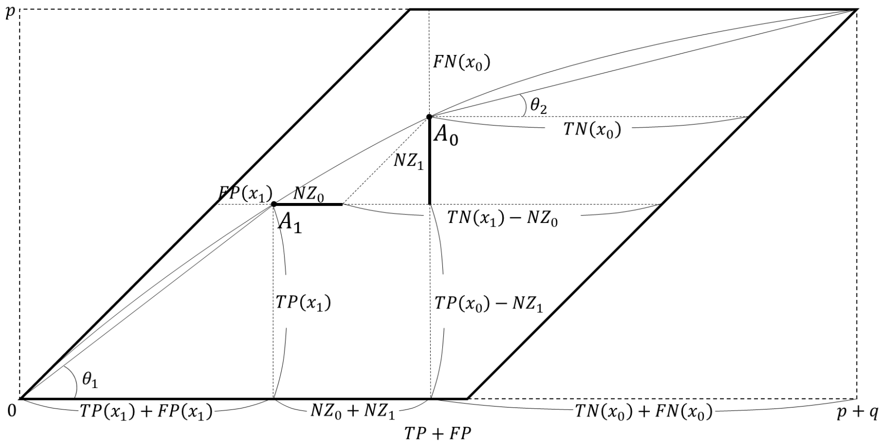


Figure 2: *PPV and NPV with neutral zone by the TOC curve. PPV = positive predictive value; NPV = negative predictive value; TOC = total operating characteristic.*

$\mathrm{TP}(x_1) + \mathrm{FP}(x_1)$ and $\mathrm{TP}(x_1)$ correspond to the denominator and numerator of $\mathrm{PPV}(x_1)$ in Result 3, respectively. Note that $\mathrm{TP}(x_0) - \mathrm{NZ}_1 + \mathrm{FP}(x_1)$ and $\mathrm{TP}(x_0) - \mathrm{NZ}_1$, which are the denominator and numerator of $\mathrm{PPV}(x_0, x_1)$ in (3.2), are identical to $\mathrm{TP}(x_1) + \mathrm{FP}(x_1)$ and $\mathrm{TP}(x_1)$ in $\mathrm{PPV}(x_1)$ in Result 3. Hence, $\mathrm{PPV}(x_1)$ can be explained as the slope of the right-angled triangle connecting the origin and point A_1.

With analogous arguments in Section 2, the denominator of $1 - \mathrm{NPV}(x_0, x_1)$, $\mathrm{TN}(x_1) - \mathrm{NZ}_0 + \mathrm{FN}(x_0)$, is the entire horizontal length, $p + q$, minus the sum of both $\mathrm{TP}(x_0) - \mathrm{NZ}_1 + \mathrm{FP}(x_1)$ and $\mathrm{NZ}_0 + \mathrm{NZ}_1$, where $\mathrm{TP}(x_0) - \mathrm{NZ}_1 + \mathrm{FP}(x_1)$ is the horizontal length from the origin to point A_1, and $\mathrm{NZ}_0 + \mathrm{NZ}_1$ is the horizontal distance of the two points A_0 and A_1. Alternatively, the denominator of $1 - \mathrm{NPV}(x_0)$, $\mathrm{TN}(x_0) + \mathrm{FN}(x_0)$, in Result 3 is also the entire horizontal length minus $\mathrm{TP}(x_1) + \mathrm{FP}(x_1)$. That is, the denominator of $1 - \mathrm{NPV}(x_0)$ means the horizontal distance from point A_0 to the far right of the horizontal axis of the TOC curve. The numerator of $1 - \mathrm{NPV}(x_0)$, $\mathrm{FN}(x_0)$, indicates the vertical distance from point A_0 to the top of the TOC curve. That is $\mathrm{FN}(x_0) = p - \mathrm{TP}(x_0)$. Hence, $1 - \mathrm{NPV}(x_0)$ can be explained by the slope of the right-angled triangle connecting point A_0 to the top right corner of the TOC curve.

Result 4. For any two thresholds x_1 and x_0, $\mathrm{PPV}(x_1)$ and $1 - \mathrm{NPV}(x_0)$ could be described as the angles of the two following right-angled triangles on the TOC curve:

$\mathrm{PPV}(x_1)$ is the slope of the right-angled triangle connecting the origin to point A_1,

$1 - \mathrm{NPV}(x_0)$ is the slope of the right-angled triangle connecting point A_0 to the point $(p + q, p)$,

where $A_0 = (\mathrm{TP}(x_0) + \mathrm{FP}(x_0), \mathrm{TP}(x_0))$ and $A_1 = (\mathrm{TP}(x_1) + \mathrm{FP}(x_1), \mathrm{TP}(x_1))$. Therefore, for any two thresholds x_1 and x_0, both $\mathrm{PPV}(x_1)$ and $1 - \mathrm{NPV}(x_0)$ have the following relationship on the TOC curve.

$$\mathrm{PPV}(x_1) = \tan\theta_1 \quad \text{and} \quad 1 - \mathrm{NPV}(x_0) = \tan\theta_2,$$

where θ_1 and θ_2 are the corresponding angles of the two right-angled triangles connecting the origin to point A_1 and point A_0 to the top right corner of the TOC curve, respectively, in Figure 2.

Table 1: Results of PPV and NPV for Case 1

(a) Results corresponding to A

x	A	PPV(x)	NPV(x)
0.70	(110, 62)	0.56080	0.79871
0.60	(120, 66)	0.54440	0.80815
0.50	(131, 69)	0.52842	0.81759
0.40	(141, 73)	0.51293	0.82698
0.30	(152, 76)	0.49798	0.83627

(b) Results with α and β

α	β	x_1	A_1	PPV(x_1)	x_0	A_0	NPV(x_0)
0.05	0.10	1.6449	(36, 26)	0.7219	−0.2816	(212, 90)	0.8861
0.10	0.05	1.2816	(59, 39)	0.6605	−0.6448	(243, 95)	0.9121

PPV = positive predictive value; NPV = negative predictive value.

Note that the coordinates of A_0 and A_1 are $(\mathrm{TP}(x_0)+\mathrm{FP}(x_0), \mathrm{TP}(x_0))$ and $(\mathrm{TP}(x_1)+\mathrm{FP}(x_1), \mathrm{TP}(x_1))$, which are also dependent on β and α, respectively. Therefore, it can be concluded that whereas PPV(x) and NPV(x) in Section 2 depend on the x value, both PPV(x_1) and $1 - \mathrm{NPV}(x_0)$ in Section 3 depend on x_1 and x_0, which also depend on the fixed values of β and α, respectively.

For $x_0 \leq x \leq x_1$, where $x_1 = \hat{F}_n^{-1}(1-\alpha)$ and $x_0 = \hat{F}_d^{-1}(\beta)$ for fixed values of α and β, we found that $\alpha \geq 1 - F_n(x)$ and $\beta \leq F_d(x)$. Since $\mathrm{PPV}(x_1) \leq \mathrm{PPV}(x)$ and $\mathrm{NPV}(x_0) \geq \mathrm{NPV}(x)$, $p/(p+q) \leq \mathrm{PPV}(x) \leq \mathrm{PPV}(x_1)$ and $q/(p+q) \leq \mathrm{NPV}(x) \leq \mathrm{NPV}(x_0)$ for $x_0 \leq x \leq x_1$ using (2.3). Moreover, Daniel and Steven (2016, p. 2350) derived that $\alpha \leq (1 - F_d(x_1))(1 - F_n(x))/(1 - F_d(x))$ and $\beta \leq F_n(x_0)F_d(x)/F_n(x)$. With these inequalities, as well as $\alpha \geq 1 - F_n(x)$ and $\beta \geq F_d(x)$, we might obtain the following specific inequalities:

$$F_d(x) \leq F_d(x_0) \leq \frac{F_n(x_0)}{F_n(x)} \times F_d(x),$$
$$1 - F_n(x) \leq 1 - F_n(x_1) \leq \frac{1 - F_d(x_1)}{1 - F_d(x)} \times (1 - F_n(x)).$$

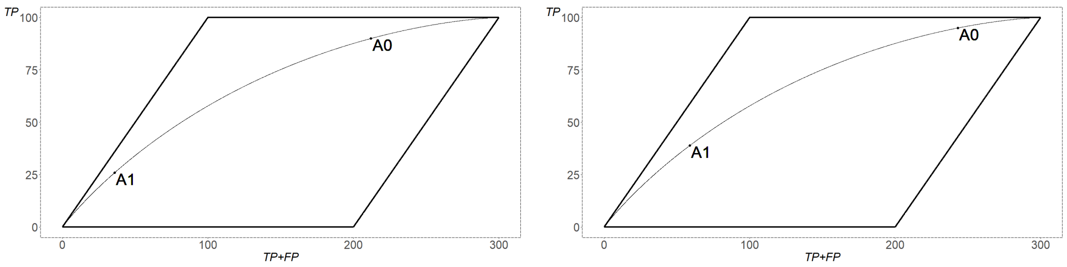
4. Illustrative examples and simulation

Let the distribution function representing a disease state be a normal distribution function with mean μ and unit variance, $X_d \sim N(\mu, 1)$, of sample size p; and the other distribution function representing a non-disease state, be the standard normal distribution function, $X_n \sim N(0, 1)$, of size q. In order to generate two sample data, set $\mu = (1.0, 2.0)$, $p = (100, 200)$, and $q = (200, 100)$. Hence the following four cases are considered:

$$\text{Case 1}: F_d(x) = \Phi(x; 1, 1),\ p = 100,\ F_n(x) = \Phi(x; 0, 1),\ q = 200,$$
$$\text{Case 2}: F_d(x) = \Phi(x; 2, 1),\ p = 100,\ F_n(x) = \Phi(x; 0, 1),\ q = 200,$$
$$\text{Case 3}: F_d(x) = \Phi(x; 1, 1),\ p = 200,\ F_n(x) = \Phi(x; 0, 1),\ q = 100,$$
$$\text{Case 4}: F_d(x) = \Phi(x; 2, 1),\ p = 200,\ F_n(x) = \Phi(x; 0, 1),\ q = 100.$$

For each case, the coordinate of point A in the TOC curve can be obtained corresponding to the random variable X. For values of X near the optimal threshold $\mu/2$, Tables 1 (a)–4 (a) contain the coordinates of A, as well as the values of PPV(x) and NPV(x).

To explore the presence of the neutral zone explained in Section 3, set $\alpha = (0.05, 0.10)$ and $\beta = (0.10, 0.05)$. Then find two thresholds x_1 and x_0. Their corresponding coordinates of points



(a) $\alpha = 0.05$, $\beta = 0.10$

(b) $\alpha = 0.10$, $\beta = 0.05$

Figure 3: *Total operating characteristic curves for Case 1.*

A_1 and A_0 in the TOC curve; in addition, the values of PPV(x_1) and NPV(x_0), are also found and summarized in Tables 1 (b)–4 (b). Figures 3 and 4 show four TOC curves corresponding to the two cases including points A_1 and A_0, respectively. The values of PPV(x), NPV(x), PPV(x_1), and NPV(x_0) for each case are now examined and discussed.

4.1. Case 1 ($\mu = 1.0$, $p = 100$, $q = 200$)

For the first case, area under the curve (AUC) = 0.76025, so that the discriminate power is adequate. Based on Table 1 (a), it can be found that as the value of X decreases, the coordinate of point A moves from the origin to point $(p+q,\ p)$; and as point A moves to $(p+q, p)$, the value of PPV(x) is decreasing, while that of NPV(x) is increasing.

Equation (2.2) is also confirmed for $X = 0.5$, which can be regarded as an optimal threshold (the middle between two means of two distribution functions):

$$\frac{100}{300} = 0.52842 \times \frac{131}{300} + (1 - 0.81759) \times \frac{169}{300}.$$

From Table 1 (b), when α increases from 0.05 to 0.10, the coordinate of A_1 moves to $(p + q, p)$, and the value of PPV(x_1) is decreasing. However, when β decreases from 0.10 to 0.05, the coordinate of A_0 moves to $(p + q, p)$, and the value of NPV(x_0) is increasing. As α increases and β decreases, the distance between points A_1 and A_0 reduces. This means that as the discriminate power is increasing, the two points A_1 and A_0 get closer.

Figures 3 (a) and (b) demonstrate the TOC curve, points A_1 and A_0 for the two kinds α and β. Figures 3(a) and (b) show that as α increases and β decreases, the coordinates of A_1 and A_0 move from the origin to point $(p + q, p)$. Points A_1 and A_0 also become closer to one another.

4.2. Case 2 ($\mu = 2.0$, $p = 100$, $q = 200$)

Case 2 has a different mean value of the distribution function for disease state, compared with Case 1. For Case 2, AUC = 0.9214, which has better discriminate power than Case 1, so that points A_1 and A_0 approach each other closer than for Case 1. Based on Tables 2 (a) and (b), similar results are obtained, except that the values of both PPV(x) and NPV(x) are larger than those in Case 1. It can also be found that equation (2.2) is also satisfied in this case, for example, when $X = 1.0$, which is an optimal threshold, the equality holds:

$$\frac{200}{300} = 0.72614 \times \frac{116}{300} + (1 - 0.91384) \times \frac{184}{300}.$$

Table 2: Results of PPV and NPV for Case 2

(a) Results corresponding to A

x	A	PPV(x)	NPV(x)
1.2	(102, 19)	0.77399	0.89309
1.1	(109, 82)	0.75045	0.90377
1.0	(116, 84)	0.72614	0.91384
0.9	(123, 86)	0.70131	0.92325
0.8	(131, 88)	0.67622	0.93197

(b) Results with α and β

α	β	x_1	A_1	PPV(x_1)	x_0	A_0	NPV(x_0)
0.05	0.10	1.6449	(74, 64)	0.8646	0.7285	(137, 90)	0.9386
0.10	0.05	1.2816	(96, 76)	0.7925	0.3552	(167, 95)	0.9623

PPV = positive predictive value; NPV = negative predictive value.

Table 3: Results of PPV and NPV for Case 3

(a) Results corresponding to A

x	A	PPV(x)	NPV(x)
0.70	(148, 124)	0.83627	0.49798
0.60	(159, 131)	0.82698	0.51293
0.50	(169, 138)	0.81759	0.52842
0.40	(180, 145)	0.80815	0.54440
0.30	(190, 152)	0.79871	0.56080

(b) Results with α and β

α	β	x_1	A_1	PPV(x_1)	x_0	A_0	NPV(x_0)
0.05	0.10	1.6449	(57, 52)	0.9121	−0.2816	(241, 180)	0.6605
0.10	0.05	1.2816	(88, 78)	0.8861	−0.6448	(264, 190)	0.7218

PPV = positive predictive value; NPV = negative predictive value.

4.3. Case 3 ($\mu = 1.0,\ p = 200,\ q = 100$)

Case 3 has the same mean value and different sample sizes of p and q, compared with Case 1. For Case 3, it is obtained that AUC = 0.76025, which has the same discriminate power as Case 1. However, the height of the TOC curve is different. Since the TOC curve in Case 3 is taller than in Case 1, the values of PPV(x) are much larger than those in Case 1, whereas the values of NPV(x) are smaller those in Case 1.

4.4. Case 4 ($\mu = 2.0,\ p = 200,\ q = 100$)

Case 4 has different sample sizes of p and q compared with Case 2, and has a larger mean than Case 3. Based on Tables 4 (a) and (b) for Case 4, AUC = 0.9214, which has the same discriminate power as Case 2, so that both points A_1 and A_0 approach each other closer than for Cases 1 and 3. Also, the values of both PPV(x) and NPV(x) can be found to be larger than those in Case 3 (Figure 4).

5. Real data example

A credit evaluation data was collected by a Korean domestic K bank in June 1918. This sample contains two random variables: one means the risk score (RS) rank which is a risk evaluation variable with 20 grades; the other variable divides each sample into three stages which are the classifications

Table 4: Results of PPV and NPV for Case 4

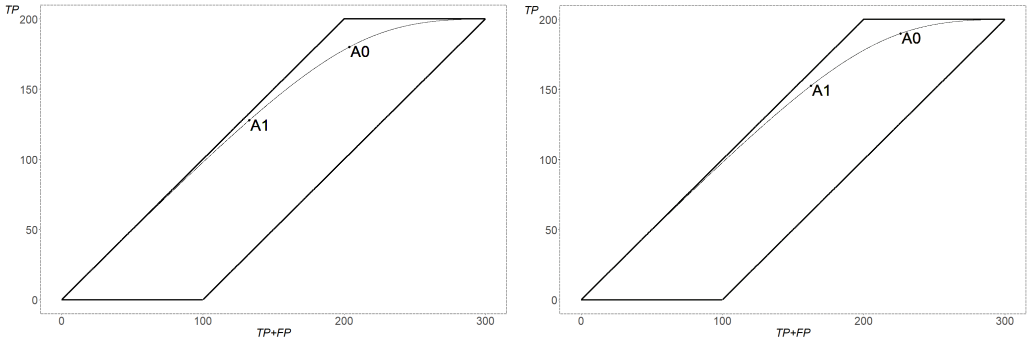
(a) Results corresponding to A

x	A	PPV(x)	NPV(x)
1.2	(169, 158)	0.93197	0.67622
1.1	(177, 163)	0.92325	0.70131
1	(184, 168)	0.91384	0.72614
0.9	(191, 173)	0.90377	0.75045
0.8	(198, 177)	0.89309	0.77399

(b) Results with α and β

α	β	x_1	A_1	PPV(x_1)	x_0	A_0	NPV(x_0)
0.05	0.10	1.6449	(133, 128)	0.9626	0.7185	(204, 180)	0.7925
0.10	0.05	1.2816	(163, 153)	0.9386	0.3552	(226, 190)	0.8646

PPV = positive predictive value; NPV = negative predictive value.



(a) α = 0.05, β = 0.10

(b) α = 0.10, β = 0.05

Figure 4: *Total operating characteristic curves for Case 4.*

of credit risk. The first stage was not used in this work since it is almost a non-default state and has too large sample. Hence, we take the second and third stages of size 6,951 and 1,191, respectively. The second and third stages are regarded as the non-default and default groups, respectively. Note that the total sample size we are concerned with is 8,142. The higher the RS rank, the more likely that bankruptcy will be judged.

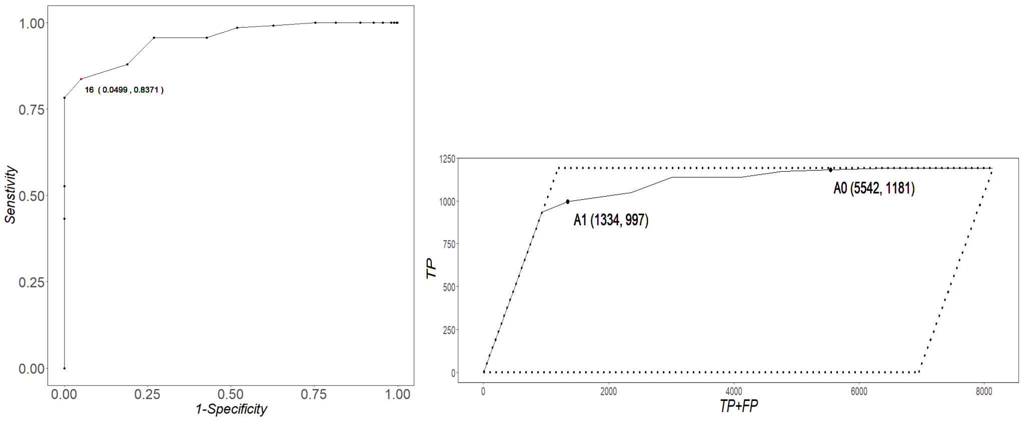
Based on this real data in Table 5, Figure 5 (a) represents the ROC curve. The AUC for this ROC curve is 0.9622. It can be obtained that sensitivity = 0.8371 and specificity = 1 – 0.0499 = 0.9500, corresponding to an optimal threshold at RS rank=16. These values can also be obtained from Table 6 (a): sensitivity = 997/1191 = 0.8371 and specificity = 6604/6951 = 0.9500. All of these values are close to 1; therefore, we might conclude that the values of these statistics that assess the accuracy and performance of the diagnosis are very good.

Figure 5 (b) expresses the TOC curve since the PPV and NPV cannot be explained on the ROC curve. Figure 5 (b) shows that the values of PPV and NPV outside A_0 and A_1 are almost 1, because the PPV is the slope from origin to point A_1 and the NPV is the slope from point A_0 to the top right corner of the TOC curve. We can obtain the value of PPV and NPV from Table 6 (b): PPV = 997/(997+347) = 0.74182 and NPV = 2590/(2590+10) = 0.99615. The points A_1 and A_0 correspond to the RS ranks of 16 and 11, respectively.

The neutral zone can be set from A_1 to A_0 on the TOC curve, and the corresponding RS ranks are x_1 = 16 and x_0 = 11. Two types of errors α and β are also obtained, which are α = 0.0500, and

Table 5: Real data with risk score rank and stages

Risk score rank	Second stage	Third stage	Sum
1	4	0	4
2	1	0	1
3	17	0	17
4	55	0	55
5	56	0	56
6	175	0	175
7	189	0	189
8	279	0	279
9	518	0	518
10	421	0	421
11	875	10	885
12	761	7	768
13	631	34	665
14	1102	0	1102
15	555	92	647
16	965	51	1,016
17	347	64	411
18	0	306	306
19	0	111	111
20	0	516	516
sum	6,951	1,191	8,142



(a) ROC curve (b) TOC curve

Figure 5: *ROC curve and TOC curve for real data. ROC = receiver operating characteristic; TOC = total operating characteristic.*

$\beta = 0.0084$. Therefore, based on the TOC curve in Figure 5, it can be concluded that the value of PPV (the probability that a sample determined as default is actually default) and NPV (the probability that a sample determined as non-default is actually non-default) are 1 almost everywhere with small error probabilities.

6. Conclusion

The 'sensitivity' and 'specificity' are well known measures for evaluating the accuracy and performance of diagnosis. When the evaluation result of the disease is positive or negative, the probability of the test performance that an actual health condition is disease or non-disease may be more impor-

Table 6: Confusion matrix in optimal threshold

(a) At RS rank = 16

	Real non-default	Real default
Non-default	6604	194
Default	347	997

(b) At RS rank = 16 and RS rank = 11

	Real non-default	Real default
Non-default	2590	10
Neutral zone	4014	184
Default	347	997

RS = risk score.

tant than how accurate the diagnosis is. To satisfy this interest, there are two other measures such as the PPV and NPV proposed by Zhou *et al.* (2002) and Raslich *et al.* (2007): the PPV is the probability that the patient has the disease following a positive test result, while the NPV is the probability that the patient does not have the disease following a negative test result.

The 'sensitivity' and 'specificity' can be described on the ROC curve; however, PPV and NPV cannot be represented by the ROC curve. In this paper, some explanatory methods are proposed to describe the PPV and NPV by the TOC curve, which can be geometrically explained by the TOC curve. The coordinate of a certain point corresponding to any threshold x on the TOC curve is $((\text{TP}(x) + \text{FP}(x)), \text{TP}(x))$. If a right-angled triangle is considered connecting the origin to a certain point, $\text{PPV}(x)$ can be regarded as the slope of the right-angled triangle. The denominator of $1-\text{NPV}(x)$ is $\text{TN}(x) + \text{FN}(x)$ and its numerator is $\text{FN}(x)$; therefore, $1 - \text{NPV}(x)$ can be represented as the slope of the right-angled triangle connecting a certain point to the top right corner of the TOC curve.

Therefore, it may be concluded that $\text{PPV}(x) = \tan\theta_1$ and $1 - \text{NPV}(x) = \tan\theta_2$, where θ_1 and θ_2 are the corresponding angles of two right-angled triangles connecting the origin to a certain point, and from the certain point to the top right corner of the TOC curve, respectively.

The neutral zone was suggested to ensure the retesting of diseases with high diagnostic costs. Daniel and Steven (2016) derived definitions of both the PPV and NPV when the neutral zone is present. When the neutral zone is present, the PPV and NPV are also geometrically described and implemented on the TOC curve in this study. The two constants x_0 and x_1 are defined with $x_0 < x_1$ as $x_0 = \hat{F}_d^{-1}(\beta)$ and $x_1 = \hat{F}_n^{-1}(1 - \alpha)$ for fixed values of α and β. Then these $\text{PPV}(x_1)$ and $\text{NPV}(x_0)$, in the presence of the neutral zone, are found to be functions of x_1 and x_0, respectively. It may also be concluded that both $\text{PPV}(x_1)$ and $\text{NPV}(x_0)$, in the presence of the neutral zone, can be explained as the slopes of two right-angled triangles. When the neutral zone exists, $\text{PPV}(x_1)$ can be described as the slope of the right-angled triangle connecting the origin to a certain point that is a function of x_1 on the TOC curve, while $1 - \text{NPV}(x_0)$ can be represented as the slope of the right-angled triangle connecting a certain point that is a function of x_0 to the top right corner of the TOC curve. Therefore, it may be concluded that $\text{PPV}(x_1) = \tan\theta_1$ and $1 - \text{NPV}(x_0) = \tan\theta_2$, where θ_1 and θ_2 are the corresponding angles of two right-angled triangles connecting the origin to a certain point depending on x_1, and from the certain point depending on x_0 to the top right corner of the TOC curve, respectively; therefore, both the PPV and NPV can be estimated by using the TOC curve, whether or not the neutral zone exists.

References

Bradley AP (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms, *Pattern Recognition*, **30**, 1145–1159.

Daniel RJ and Steven S (2016). Maximizing the usefulness of statistical classifiers for two populations with illustrative applications, *Statistical Methods in Medical Research*, **27**, 2344–2358.

Egan JP (1975). *Signal detection theory and ROC analysis*, Academic Press, New York.

Engelmann B, Hayden E, and Tasche D (2003). Testing rating accuracy, *Risk*, **16**, 82–86.

Fawcett T (2003). *ROC graphs: Notes and practical considerations for data mining researchers* (Technical report), Available from: http://www.blogspot.udec.ugto.saedsayad.com/docs/ROC101.pdf

Fawcett T (2006). An introduction to ROC analysis, *Pattern Recognition Letters*, **27**, 861–874.

Fawcett T and Provost F (1997). Adaptive fraud detection, *Data Mining and Knowledge Discovery*, **1**, 291–316.

Garcia V, Mollineda RA, and Sanchez JS (2010). Theoretical analysis of a performance measure for imbalanced data, *20th International Conference on Pattern Recognition*, **2010**, 617–620.

Hong CS, Kim JH, and Choi JS (2009). Adjusted ROC and CAP curves, *Korean Journal of Applied Statistics*, **22**, 29–39.

Hong CS and Lee SJ (2018). TROC curve and accuracy measures, *Journal of the Korean Data & Information Science Society*, **29**, 861–872.

Hsieh F and Turnbull BW (1996). Nonparametric and semiparametric estimation of the receiver operating characteristic curve, *The Annals of Statistics*, **24**, 25–40.

Metz CE and Kronman HB (1980). Statistical significance tests for binormal ROC curves, *Journal of Mathematical Psychology*, **22**, 218–243.

Pepe MS (2000). Receiver operating characteristic methodology, *Journal of the American Statistical Association*, **95**, 308–311.

Pepe MS (2003). *The statistical evaluation of medical tests for classification and prediction*, Oxford University Press, Oxford.

Pontius RG and Si K (2014). The total operating characteristic to measure diagnostic ability for multiple threshold, *International Journal of Geographical Information Science*, **97**, 570–583.

Provost F and Fawcett T (1997). Analysis and visualization of classifier performance comparison under imprecise class and cost distributions, *Knowledge Discovery and Data Mining*, **97**, 43–48.

Provost F and Fawcett T (2001). Robust classification for imprecise environments, *Machine Learning*, **42**, 203–231.

Raslich MA, Markert RJ, and Stutes SA (2007). Selecting and interpreting diagnostic tests, *Biochemia Medica*, **17**, 139–270.

Stein RM (2005). The relationship between default prediction and lending profits: Integrating ROC analysis and loan pricing, *Journal of Banking & Finance*, **29**, 1213–1236.

Zhou XH, Obuchowski NA, and McClish DK (2002). *Statistical Methods in Diagnostic Medicine*, Wiley-InterScience, New York.

Zweig M and Campbell G (1993). Receiver-operating characteristics (ROC) plots: A fundamental evaluation tool in clinical medicine, *Clinical Chemistry*, **39**, 561–577.

Received October 11, 2019; Revised December 10, 2019; Accepted January 18, 2020