American Journal of Epidemiology

Vol. 163, No. 7
DOI: 10.1093/aje/kwj063
Advance Access publication January 12, 2006

Original Contribution

The Inconsistency of ''Optimal'' Cutpoints Obtained using Two Criteria based on the Receiver Operating Characteristic Curve

Neil J. Perkins[1,2] and Enrique F. Schisterman[1]

[1] Division of Epidemiology, Statistics and Prevention Research, National Institute of Child Health and Human Development, Bethesda, MD.
[2] Department of Mathematics and Statistics, American University, Washington, DC.

Received for publication June 17, 2005; accepted for publication October 13, 2005.

The use of biomarkers is of ever-increasing importance in clinical diagnosis of disease. In practice, a cutpoint is required for dichotomizing naturally continuous biomarker levels to distinguish persons at risk of disease from those who are not. Two methods commonly used for establishing the ''optimal'' cutpoint are the point on the receiver operating characteristic curve closest to (0,1) and the Youden index, *J*. Both have sound intuitive interpretations—the point closest to perfect differentiation and the point farthest from none, respectively—and are generalizable to weighted sensitivity and specificity. Under the same weighting of sensitivity and specificity, these two methods identify the same cutpoint as ''optimal'' in certain situations but different cutpoints in others. In this paper, the authors examine situations in which the two criteria agree or disagree and show that *J* is the only ''optimal'' cutpoint for given weighting with respect to overall misclassification rates. A data-driven example is used to clarify and demonstrate the magnitude of the differences. The authors also demonstrate a slight alteration in the (0,1) criterion that retains its intuitive meaning while resulting in consistent agreement with *J*. In conclusion, the authors urge that great care be taken when establishing a biomarker cutpoint for clinical use.

area under curve; biological markers; cutpoints; data interpretation, statistical; epidemiologic methods; ROC curve; statistics; Youden index

Abbreviations: AUC, area under the curve; ROC, receiver operating characteristic.

The proper diagnosis of disease and treatment administration is a task that requires a variety of tools. Through advancements in biology and laboratory methods, a multitude of biomarkers are available as clinical tools for such diagnosis. These biomarkers are usually measured on a continuous scale with overlapping levels for diseased and nondiseased persons. Cutpoints dichotomize biomarker levels, providing benchmarks that label people as diseased or not diseased on the basis of ''positive'' or ''negative'' test results. Biomarker levels of persons with known disease status are used to evaluate potential cutpoint choices and, hopefully, identify a cutpoint that is ''optimal'' under some criterion.

Such a data set would comprise biomarker levels for persons classified as coming from the diseased (D) or nondiseased ($\bar{D}$) population. These levels could then be classified in terms of positive (+) or negative (−) test results on the basis of whether the biomarker levels were above or below a given cutpoint. In most instances, some persons will be misclassified, truly belonging to a population other than the one indicated by their test results. The sensitivity ($q(c)$) and specificity ($p(c)$) of that biomarker for a given cutpoint, c, are the probabilities of correctly identifying a person's disease status (i.e., identifying true positives and true negatives):

$$q(c) = \text{Prob}(\text{test result} = +|D)$$

$$p(c) = \text{Prob}(\text{test result} = -|\bar{D}),$$

Correspondence to Dr. Enrique F. Schisterman, Division of Epidemiology, Statistics and Prevention Research, National Institute of Child Health and Human Development, 6100 Executive Blvd., Bethesda, MD 20852 (e-mail: schistee@mail.nih.gov).

making 1 minus these values the probability of incorrect classification or of obtaining false negatives ($1 - q(c)$) and false positives ($1 - p(c)$).

A receiver operating characteristic (ROC) curve is a mapping of this sensitivity by 1 minus specificity. The ROC curve has become a useful tool in comparing the effectiveness of different biomarkers (1–3). This comparison takes place through summary measures such as the area under the curve (AUC) and the partial AUC, with higher area values indicating higher levels of diagnostic ability (1, 2, 4). A biomarker with an AUC of 1 differentiates perfectly between diseased persons (sensitivity = 1) and healthy persons (specificity = 1). An AUC of 0.5 means that, overall, there is a 50-50 chance that the biomarker will correctly identify diseased or healthy persons as such.

Though useful for biomarker evaluation, these measures do not inherently lead to benchmark "optimal" cutpoints with which clinicians and other health-care professionals can differentiate between diseased and nondiseased persons. Several methods for identifying "optimal" cutpoints using sensitivity, specificity, and the ROC curve have been proposed and applied (4–8). Confidence intervals and corrections for measurement error are some of the supporting statistical developments accompanying cutpoint estimation (9). Applications of these techniques have been demonstrated in several fields, including nuclear cardiology, epidemiology, and genetics (7, 10, 11).

In the "Criteria" section of this article, we describe two criteria for locating this cutpoint that have similar intuitive justifications. In describing the mathematical mechanisms behind these criteria, we demonstrate that one of the criteria retains the intended meaning, while the other inherently depends on quantities that may differ from an investigator's intentions. In the "Example" section, we use data from a nested case-control study carried out in the Calcium for Pre-Eclampsia Prevention cohort (12) to demonstrate how these two criteria identify different cutpoints for the classification of 120 preeclampsia cases and 120 controls based on levels of placenta growth factor, a biomarker of angiogenesis. Next, we discuss the appropriateness of the term "optimal" as it applies to each criterion. This is handled first with equally weighted sensitivity and specificity. Consideration of differing disease prevalences and costs due to misclassification is also presented as a practical generalization (5, 13). We end with a brief discussion.

CRITERIA

The closest-to-(0,1) criterion

If a biomarker perfectly differentiates persons with disease from those without disease on the basis of a single cutpoint, where $q(c) = 1$ and $p(c) = 1$, the ROC curve is a vertical line from (0,0) to (0,1) joined with a line from (0,1) to (1,1) with an AUC of 1. However, for a less-than-perfect biomarker, where $q(c) < 1$ and/or $p(c) < 1$, the ROC curve does not touch the (0,1) point. Here the choice of an "optimal" cutpoint is less straightforward. A criterion by which the point on the curve closest to (0,1) is identified and the corresponding cutpoint is labeled "optimal" has been suggested and utilized (6, 7). The rationale behind this approach is that the point on the curve closest to perfection (i.e., closest to $q(c) = 1$ and $p(c) = 1$) should correspond to the optimal cutpoint chosen from all of the cutpoints available, thus intuitively minimizing misclassification. Mathematically, the point c^* that satisfies the equation

$$\min\{\sqrt{(1-q(c))^2+(1-p(c))^2}\}$$

or

$$\min\{(1-q(c))^2+(1-p(c))^2\} \quad (1)$$

fulfills this criterion and is thus labeled the cutpoint that best differentiates between diseased and nondiseased persons.

This criterion can be viewed as searching for the shortest radius originating at the (0,1) point and terminating on the ROC curve. Reference arcs can be used to visually compare radial distances, with the arc corresponding to c^* being tangent to the ROC curve and thus the minimum and interior of any of the concentric arcs possible. Figure 1 demonstrates this point at which the dotted arc is completely interior to, and thus closer to (0,1) than, the arc formed by the distance to an alternate point on the curve.

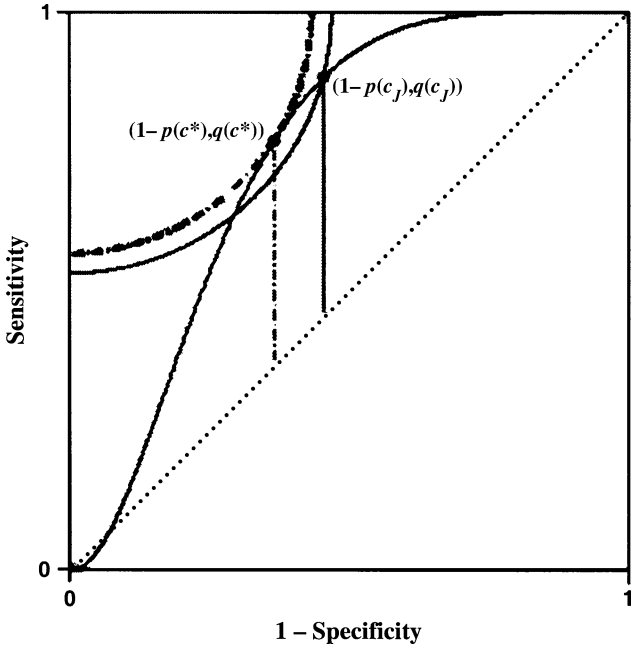


FIGURE 1. Receiver operating characteristic curve based on simulated diseased and nondiseased populations. The vertical lines and reference arcs identify the Youden index, J (solid lines), and the point closest to the (0,1) point (dotted lines) and their corresponding "optimal" cutpoints c_J and c^*, respectively.

The Youden index

Another measure for evaluating biomarker effectiveness is the Youden index (J), first introduced in the medical literature by Youden (14). J is also a function of $q(c)$ and $p(c)$, such that

$$J = \max\{q(c) + p(c) - 1\}$$
$$= \max\{q(c) - (1 - p(c))\} \quad (2)$$

over all cutpoints c, with c_J denoting the cutpoint corresponding to J. On a ROC curve, J is the maximum vertical distance from the curve to the chance line or positive diagonal (figure 1), making c_J the "optimal" cutpoint (5, 15). The intuitive interpretation of the Youden index is that J is the point on the curve farthest from chance. It has also been defined as the accuracy of the test in clinical epidemiology (16).

Agreement/disagreement

The above criteria agree with respect to intuition; they maximize and minimize the rates of people's being classified correctly and incorrectly, respectively. The question "Do they agree on the same 'optimal' cutpoint?" now begs to be answered.

Suppose the biomarker of interest follows continuous distributions for both diseased and nondiseased populations that are known completely, leading to a true ROC curve. Our only distributional restriction is that a ROC curve is generated that is differentiable everywhere. This is intrinsic to the case where diseased and nondiseased persons are assumed to follow any number of common continuous densities (i.e., normal, lognormal, gamma, etc.). Through differentiation, Appendix 1 shows that the two criteria only agree, $c^* = c_J = c$, when $q(c^*) = p(c^*)$ and $q(c_J) = p(c_J)$. When either criterion identifies a point on the curve such that $q(c^*) \neq p(c^*)$ or $q(c_J) \neq p(c_J)$, the criteria disagree on what cutpoint is "optimal," that is, $c^* \neq c_J$.

An investigator with complete knowledge of a biomarker's data distribution could be faced with two different cutpoints labeled "optimal" under two criteria that are intuitively the same. Our motivation here is simply to show that they are different and address the appropriateness of the label "optimal."

EXAMPLE

Preeclampsia affects approximately 5 percent of pregnancies, resulting in substantial maternal and neonatal morbidity and mortality (16). Although the cause remains unclear, the syndrome may be initiated by placental factors that enter the maternal circulation and cause endothelial dysfunction, resulting in hypertension and proteinuria (12). Identifying women suffering from preeclampsia is a very important step in the management of the disease. Placenta growth factor is a promising biomarker for such classification, with an AUC of 0.60 (95 percent confidence interval: 0.53, 0.67); however, at what level would a woman be classified as at risk for the disease? Levine et al. (12) conducted a nested case-control study of 120 women with preeclampsia and 120 normal women randomly chosen from the Calcium for Pre-Eclampsia Prevention cohort study. Placenta growth factor levels were measured from serum specimens obtained before labor. Figure 2 shows the ROC curve generated from the log-transformed placenta growth factor levels. After calculation of the distance to (0,1) and the distance to the diagonal for each point, the cutpoints $c^* = 4.64$ and $c_J = 4.12$, respectively, are identified. Thus, criteria with seemingly identical intuitive intents produce close results but disagree on the "optimal" cutpoint. Again, here it is sufficient to demonstrate that disagreement exists. We will revisit this example after the question of "optimality" has been addressed.

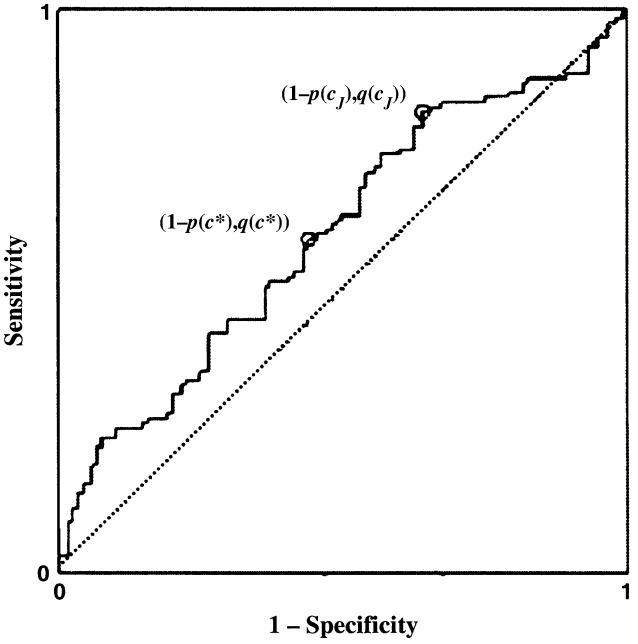


FIGURE 2. Empirical receiver operating characteristic curve obtained using placenta growth factor levels to differentiate between women diagnosed with preeclampsia and those without it. The two points corresponding to cutpoints labeled "optimal" by the closest-to-(0,1) criterion (c^*) and the Youden index (c_J) differ. Data source: Levine et al. (12).

"Optimality"

When attempting to classify people on the basis of biomarker levels, it is always one's intent to do so "optimally." However, the event of interest may intrinsically involve constraints which must, for ethical or fiscal reasons, be considered. These constraints commonly account for the prevalence of the event in both populations and the costs of misclassification, both monetary and physiologic. Thus, mathematical techniques of optimality must now operate within these constraints, but the idea of an "optimal" cutpoint should remain; one still wishes to choose a point that classifies the most people correctly and the fewest incorrectly.

First let us assume the simplest scenario, absent of constraints or weighting. By definition, the c_J found by equation 2 succeeds ideologically by maximizing the overall rate of correct classification, $q(c_J) + p(c_J)$. As a result, the overall rate of misclassification, $(1 - q(c_J)) + (1 - p(c_J))$, is minimized. Thus, we can say that c_J is "optimal" with respect to the total correct and incorrect classification rates and any cutpoint that deviates from it is not.

Under the same scenario, the closest-to-(0,1) criterion in equation 1 minimizes the total squared misclassification rates, quadratic terms for which an ideology does not seem to exist, other than being geometrically intuitive. Equation 1 can be expanded and rewritten as

$$\min\{(1-q(c))+(1-p(c))+(q(c)^2+p(c)^2)/2\} \quad (3)$$

to show that this criterion minimizes the total of the misclassification rates and a third term, the average of squared correct classification rates. Unless a specific justification for this third term exists, its usage results in unwarranted and thus unnecessary misclassification, because it identifies a point $c^* \neq c_J$.

Now, let us consider the circumstance in which cost and prevalence are thought to be factors, as they usually are in practice. Using decision theory, a generalized J can be formed where these factors are represented as a weighting of sensitivity and specificity. The function that minimizes expected loss in classifying a subject can be written as

$$\min\{a\pi(1-q(c))+(1-\pi)(1-p(c))\}, \quad (4)$$

where a denotes the relative loss (cost) of a false-negative classification as compared with a false-positive classification and π is the proportion of diseased persons in the population of interest (prevalence) (17, 18). It is easy to see that minimizing this expected loss over all possible threshold values is the same as

$$J=\max\{q(c)+r\times p(c)-1\}, \quad (5)$$

where $r = (1 - \pi)/a\pi$. For $r = 1$, this is equivalent to J.

Weighting of the (0,1) criterion occurs similarly,

$$\min\{(1-q(c))^2+r\times(1-p(c))^2\}, \quad (6)$$

where r is exactly the same weighting estimate for cost and prevalence. The issue of the quadratic term remains

$$\min\{(1-q(c))+r\times(1-p(c)) + (q(c)^2+r\times p(c)^2)/2\}, \quad (7)$$

only now it is weighted and unnecessary. Comparing this equation to equation 4, it is easy to see that this absolutely does not minimize loss due to misclassification.

Example revisited

To demonstrate this unnecessary misclassification and its possible magnitude, we revisit the example of placenta growth factor levels' being used to differentiate preeclamptic women from those without the disease. Sensitivity and specificity at the cutpoints previously identified are $q(c^*) = 0.592$, $p(c^*) = 0.558$ and $q(c_J) = 0.817$, $p(c_J) = 0.358$, respectively. The overall correct classification rate $(q + p)$ is 1.150 for c^* and 1.175 for c_J out of a possible 2, with a difference of 0.025. Without the justification for the third term in equation 3 and without weighting, this difference can be thought of as one person out of 100 being unnecessarily misclassified. Relative cost and disease prevalence are often difficult to assess, as discussed by Greiner et al. (18) and the references cited therein. Thus, we will not attempt adjustment in this example.

DISCUSSION

In this paper, we demonstrated the intuitive similarity of two criteria used to choose an "optimal" cutpoint. We then showed that the criteria agree in some instances and disagree in others. Placenta growth factor levels used to classify women as preeclamptic or not preeclamptic were used to demonstrate this point and quantify the extent of disagreement.

We addressed both criteria in the context of what an investigator might view as "optimal," with and without attention to misclassification cost and prevalence. Mathematically, J reflects the intention of maximizing overall correct classification rates and thus minimizing misclassification rates, while choosing the point closest to (0,1) involves a quadratic term for which the clinical meaning is unknown. It is for this reason that we advocate for the use of J to find the "optimal" cutpoint.

Since the (0,1) criterion is visually intuitive, we have provided an amended (0,1) criterion in Appendix 2 that is likewise geometrically satisfying while consistently identifying the same "optimal" cutpoint as J. This criterion relies on a ratio of radii originating at (0,1).

Additional motivation for using J is an ever-increasing body of supporting literature (9, 15, 19). Topics such as confidence intervals and correcting the estimate for measurement error have been considered, whereas the (0,1) criterion lacks such support.

Most importantly, cutpoints chosen through less than "optimal" criteria or criteria that are "optimal" in some arbitrary sense can lead to unnecessary misclassifications, resulting in needlessly missed opportunities for disease diagnosis and intervention. We showed above that J is "optimal" when equal weight is given to sensitivity and specificity, $r = 1$, and a generalized J is "optimal" when cost and prevalence lead to weighted sensitivity and specificity, $r \neq 1$. Thus, when the point closest to (0,1) differs from the point resulting in J, using this criterion to establish an "optimal" cutpoint unnecessarily introduces an increased rate of misclassification.

ACKNOWLEDGMENTS

This research was supported by the National Institutes of Health Intramural Research Program, National Institute of Child Health and Human Development.

The authors thank Dr. Richard Levine for allowing them to use the data from the Calcium for Pre-Eclampsia Prevention Study.

Conflict of interest: none declared.

REFERENCES

1. Zhou XH, Obuchowski NA, McClish DK. Statistical methods in diagnostic medicine. New York, NY: John Wiley and Sons, Inc, 2002.

2. Faraggi D. Adjusting ROC curves and related indices for covariates. J R Stat Soc Ser D Statistician 2003;52:179–92.
3. Schisterman EF, Faraggi D, Reiser B. Adjusting the generalized ROC curve for covariates. Stat Med 2004;23:3319–31.
4. Pepe M. The statistical evaluation of medical tests for classification and prediction. New York, NY: Oxford University Press, 2003.
5. Zwieg MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem 1993;39:561–77.
6. Coffin M, Sukhatme S. Receiver operating characteristic studies and measurement errors. Biometrics 1997;53:823–37.
7. Sharir T, Berman DS, Waechter PB, et al. Quantitative analysis of regional motion and thickening by gated myocardial perfusion SPECT: normal heterogeneity and criteria for abnormality. J Nucl Med 2001;42:1630–8.
8. van Belle G. Statistical rules of thumb. New York, NY: John Wiley and Sons, Inc, 2002:98.
9. Perkins NJ, Schisterman EF. The Youden index and the optimal cut-point corrected for measurement error. Biom J 2005;47:428–41.
10. Schisterman EF, Faraggi D, Brown R, et al. TBARS and cardiovascular disease in a population-based sample. J Cardiovasc Risk 2001;8:219–25.
11. Chen R, Rabinovitch PS, Crispin DA, et al. DNA fingerprinting abnormalities can distinguish ulcerative colitis patients with dysplasia and cancer from those who are dysplasia/cancer-free. Am J Pathol 2003;16:665–72.
12. Levine RJ, Maynard SE, Qian C, et al. Circulating angiogenic factors and the risk of preeclampsia. N Engl J Med 2004;350:672–83.
13. Barkan N. Statistical inference on r * specificity + sensitivity. (Doctoral dissertation). Haifa, Israel: University of Haifa, 2001:69–74.
14. Youden WJ. An index for rating diagnostic tests. Cancer 1950;3:32–5.
15. Schisterman EF, Perkins NJ, Aiyi L, et al. Optimal cutpoint and its corresponding Youden index to discriminate individuals using pooled blood samples. Epidemiology 2005;16:73–81.
16. Chmura Kraemer H. Evaluating medical tests: objective and quantitative guidelines. Newbury Park, CA: Sage Publications, 1992.
17. Geisser S. Comparing two tests used for diagnostic or screening processes. Stat Prob Lett 1998;40:113–19.
18. Greiner M, Pfeiffer D, Smith RM. Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. Prev Vet Med 2000;45:23–41.
19. Hilden J, Glasziou P. Regret graphs, diagnostic uncertainty and Youden's index. Stat Med 1996;15:969–86.

APPENDIX 1

For continuous receiver operating characteristic (ROC) curves, we make no distributional assumptions beyond the assumption that the probability density functions f_D and $f_{\bar{D}}$ for biomarker levels of diseased and nondiseased persons, respectively, form a ROC curve that is differentiable everywhere. This is the case when f_D and $f_{\bar{D}}$ are assumed to be any common continuous parametric distributions (i.e., normal, gamma, lognormal).

In order to locate the cutpoints that minimize and maximize equations 1 and 2, respectively, it is first necessary to locate critical values. Thus, differentiating equation 1,

$$\frac{\partial}{\partial c}[(1-p(c))^2+(1-q(c))^2]=2(1-p(c))\left(\frac{\partial(1-p(c))}{\partial c}\right)-2(1-q(c))\left(\frac{\partial q(c)}{\partial c}\right). \tag{A1.1}$$

Then set the derivative equal to zero:

$$\begin{aligned}2(1-p(c^*))\left(\frac{\partial(1-p(c^*))}{\partial c}\right)-2(1-q(c^*))\left(\frac{\partial q(c^*)}{\partial c}\right)&=0\\(1-p(c^*))\left(\frac{\partial(1-p(c^*))}{\partial c}\right)&=(1-q(c^*))\left(\frac{\partial q(c^*)}{\partial c}\right)\\\frac{\partial q(c^*)}{\partial(1-p(c^*))}&=\frac{1-p(c^*)}{1-q(c^*)}.\end{aligned} \tag{A1.2}$$

Now, we differentiate the second criterion,

$$\frac{\partial}{\partial c}[q(c)-(1-p(c))]=\frac{\partial q(c)}{\partial c}-\frac{\partial(1-p(c))}{\partial c}, \tag{A1.3}$$

and then setting the derivative equal to zero,

$$\begin{aligned}\frac{\partial q(c_J)}{\partial c}-\frac{\partial(1-p(c_J))}{\partial c}&=0\\\frac{\partial q(c_J)}{\partial c}&=\frac{\partial(1-p(c_J))}{\partial c}\\\frac{\partial q(c_J)}{\partial(1-p(c_J))}&=1.\end{aligned} \tag{A1.4}$$

The forms of both equation A1.2 and equation A1.4 define the critical points of the criteria in equations 1 and 2, respectively, by the slopes of their corresponding points on the ROC curve. Since these solutions are not necessarily unique, multiple solutions may exist—that is, local maximums or minimums. Therefore, all solutions and endpoints must be evaluated so that c^* and c_J are global solutions.

Equations A1.2 and A1.4 show us that the (0,1) and J methods agree, $c^* = c_J = c$, only when $q(c^*) = p(c^*)$ and thus $(1 - p(c^*))/(1 - q(c^*)) = 1$. When $q(c^*) \neq p(c^*)$, the criteria disagree on what point is optimal ($c^* \neq c_J$).

APPENDIX 2

Equation 1 identifies the point closest to perfection irrespective of the possibilities of imperfection. In other words, this criterion minimizes the distance from (0,1) to the curve but fails to take into account the possible distance to the chance line. To obtain a weighted criterion that accounts for this deficiency, minimize the proportion of the smaller radius (r_2) to the larger radius (r_1), as displayed in appendix figure 1, such that

$$\min\left\{\sqrt{\frac{r_2^{\,2}}{r_1^{\,2}}}\right\} = \min\left\{\sqrt{\frac{(1-p(c))^2+(1-q(c))^2}{\left(\frac{1-p(c)}{1-d}\right)^2+\left(1-\frac{1-p(c)}{1-d}\right)^2}}\right\} = \min\{1-d\}, \tag{A2.1}$$

where $d = q(c) - (1 - p(c))$.

The relation in equation A2.1 can be derived algebraically or by using the proportionality of the triangles in appendix figure 1, such that

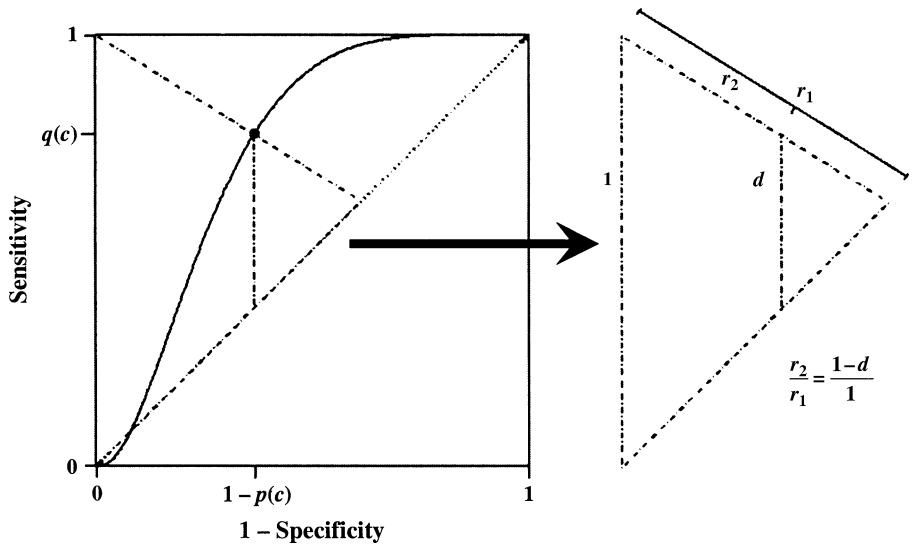
$$\frac{r_2}{r_1} = \frac{r_1-(r_1-r_2)}{r_1} = 1-\frac{(r_1-r_2)}{r_1} = 1-\frac{d}{1}.$$

Appendix figure 1 also provides a visual reference for the proposed weighting, as radii passing through different points on the curve have different distances to the chance line but are treated uniformly in equation 1.

It is now easily seen that the differentiation

$$\frac{\partial}{\partial c}\{1-d\} = -\frac{\partial q(c)}{\partial c}+\frac{\partial(1-p(c))}{\partial c} = 0$$
$$\frac{\partial q(c)}{\partial(1-p(c))} = 1$$

leads to the same critical points on the receiver operating characteristic curve as J and thus to identical cutpoints ($c^* = c_J$).



APPENDIX FIGURE 1. Receiver operating characteristic curve displaying radii extending from the point (0,1) to points on the curve and chance line, denoted by r_2 and r_1, respectively. Through similar triangles, the ratio of radii r_2:r_1 is shown to equal 1 minus the height, d, of the curve from the diagonal or chance line.