Published in final edited form as:
Stat Med. 2021 March 30; 40(7): 1767–1789. doi:10.1002/sim.8869.

The Length of the ROC Curve and the Two Cutoff Youden Index within a Robust Framework for Discovery, Evaluation, and Cutoff Estimation in Biomarker Studies Involving Improper ROC Curves

Leonidas E. Bantis[*,1], **John V. Tsimikas**[2], **Gregory Chambers**[3], **Michela Capello**[4], **Samir Hanash**[4], **Ziding Feng**[5]

[1]Dept. of Biostatistics and Data Science, University of Kansas Medical Center, Kansas City, U.S.A.

[2]Dept of Statistics and Actuarial-Financial Mathematics, University of the Aegean, Samos, Greece.

[3]Dept. of Mathematics, Rice University, Houston, U.S.A.

[4]Dept. of Clinical Cancer Prevention, The University of Texas MD Anderson Cancer Center, Houston, U.S.A.

[5]Dept. of Biostatistics, Fred Hutchinson Cancer Research Center, Seattle, U.S.A.

Abstract

During the early stage of biomarker discovery, high throughput technologies allow for simultaneous input of thousands of biomarkers that attempt to discriminate between healthy and diseased subjects. In such cases, proper ranking of biomarkers is highly important. Common measures, such as the area under the ROC curve (AUC), as well as affordable sensitivity and specificity levels, are often taken into consideration. Strictly speaking, such measures are appropriate under a stochastic ordering assumption, which implies, without loss of generality, that higher measurements are more indicative for the disease. Such an assumption is not always plausible and may lead to rejection of extremely useful biomarkers at this early discovery stage. We explore the length of a smooth ROC curve as a measure for biomarker ranking, which is not subject to directionality. We show that the length corresponds to a ϕ divergence, is identical to the corresponding length of the optimal (likelihood ratio) ROC curve, and is an appropriate measure for ranking biomarkers. We explore the relationship between the length measure and the AUC of the optimal ROC curve. We then provide a complete framework for the evaluation of a biomarker

[*]Correspondence to: Leonidas E. Bantis, Dept. 3901 Rainbow Blvd, Dept. of Biostatistics and Data Science, University of Kansas Medical Center, Kansas City, KS, U.S.A. lbantis@kumc.edu.

Supplementary Material
Supplementary material is provided online by the journal's website as a Web Appendix that consists of parts A, B, and C. It includes both our simulation results and the discussed proofs of the theoretical results given in of the main manuscript. MATLAB codes with our approaches are publicly available at the first author's website: www.leobantis.net. The codes include the 'roclength' function that derives the length of an ROC and performs inferences with the proposed methods. Therein we also provide the data that were used for the esophageal squamous cell carcinoma application. Codes for analyzing these data and obtaining all graphs presented in the paper are also available in the same file.

[†]Please ensure that you use the most up to date class file, available from the SIM Home Page at www.interscience.wiley.com/jpages/0277-6715

in terms of sensitivity and specificity through a proposed ROC analogue for use in improper settings. In the absence of any clinical insight regarding the appropriate cutoffs, we estimate the sensitivity and specificity under a two-cutoff extension of the Youden index and we further take into account the implied costs. We apply our approaches on two biomarker studies that relate to pancreatic and esophageal cancer.

Keywords

2 cutoff ROC; ϕ divergence; isoperimetric; kernels; likelihood ratio; sensitivity; specificity; optimal ROC; stochastic ordering; youden

1. Introduction

The most common statistical tool for the evaluation of continuous classifiers that attempt to discriminate between two distinct classes is the Receiver Operating Characteristic (ROC) curve ([1]). It has been a commonplace strategy to evaluate the discriminating ability of various kinds of classifiers such as blood-based tests, urine-based tests, imaging, cognitive tests etc. Typically, in many biostatistical applications biomarkers attempt to discriminate between a non-diseased and a diseased group. An underlying assumption in many ROC studies is that the density ratio is monotone. This implies the traditional setting found in the literature where the higher the marker score, the greater the suspicion for the disease (or the mathematically equivalent claim that the lower the marker score the higher the suspicion for the disease). In such cases the marker scores comply with the stochastic ordering assumption of the form $X < Y$, where X is the random variable that refers to the scores of the non-diseased group and Y is the random variable that refers to the marker scores of the diseased group. Under this scenario we have concavity of the ROC curve and an appropriate and popular summary measure of the discriminatory ability of a continuous marker is the area under the ROC curve (AUC) which is equal to the probability $P(X < Y)$. The popular Mann-Whitney test-statistic ([2]) is simply a two-sided test of the null hypothesis that states that the AUC is equal to 0.5. In cases where the density ratio is not monotone we obtain non-concave (also known as improper) ROC curves. When concavity is severely violated then the use of typical ROC analysis may be very misleading. In some cases, improper ROCs indicate differences in terms of scale apart from differences in terms of location. The vast clinical literature that deals with biomarkers focuses on the traditional settings. However, it may be the case that in many applications there are biomarkers of great promise that exhibit improper ROC curves. They may yield an AUC that is very close to 0.5, potentially misleading a researcher to dismiss such biomarkers as uninformative, in spite of their potentially high discriminatory ability. As we will see in the discussion to follow, it may be the case that a marker is perfect and yet its AUC is equal to 0.5. As a result, when traditional measures such as the AUC are blindly applied in large lists of candidate biomarkers, then potentially excellent biomarkers that are not monotone are doomed to be missed. Modern high-throughput technologies allow us to simultaneously assay hundreds or even thousands of markers, a practice quite common in cancer research. This situation illustrates a need for a robust statistical framework that can reveal all useful markers whether they are monotone or not. Once these are discovered then it is of interest to evaluate them with clinically appealing measures such as the sensitivity and specificity tradeoffs, that need to be redefined for non-monotone markers. Finally, once the useful markers have been identified clinical interest lies in determining decision making cutoffs.

One example of the aforementioned setting refers to assays that involve blood-based biomarkers for the detection of pancreatic cancer. Pancreatic ductal adenocarcinoma (PDAC) is the third leading cause of cancer-related mortality in both men and women in the United States. Overall, PDAC is associated with a dire prognosis and a 5-year survival rate of approximately 8%, which makes PDAC the most deadly of cancers by organ site ([3]). The diagnosis of PDAC at an early stage is uncommon and usually incidental, with the majority of patients (~85%) presenting with locally-advanced or metastatic disease ([4]). Currently, no individual marker has adequate performance characteristics for detecting early stage PDAC in asymptomatic individuals. High throughput technologies like mass-spectrometry and protein microarrays are commonly used for biomarker discovery, often providing a very large set of candidates that initially need to be ranked based on some criterion. Common criteria are the area under the ROC curve or the sensitivity/specificity at a (clinically affordable level. While these criteria enjoy a broad acceptance in both statistical and clinical literature, they fail to capture biomarkers that are of improper behavior. Selection based on the AUC, initially studied by [5], implicitly assumes stochastic ordering of the two distributions (i.e., healthy versus diseased) of biomarker scores. There are cases, however, in which this single-directionality assumption is severely violated. For example, plasma tumor antigen levels may be associated with immune complex formation, leading to depletion of the free autoantibodies against the antigen as well as the antigen itself, which may serve as cancer biomarkers ([6]). Using conventional statistical techniques for discriminatory ability rankings might eliminate further consideration of promising or even excellent biomarkers without allowing researchers to question and explore them further. An additional example is studies associating lung cancer with levels of pro-surfactant protein B (ProSFTPB) ([7]), for which both non-detectable (very low) and higher levels of ProSFTPB are being associated with the disease.

When a stochastic ordering does not hold, we cannot rely on the regular definitions of the sensitivity and specificity for ranking the candidate biomarkers. The notion of sensitivity and specificity needs to be extended to take into account this improper behavior. When stochastic ordering is violated, the AUC can be equal to 0.5, even when the underlying biomarker is perfect. The partial AUC (pAUC) can also be subject to that artifact. In Figure 1 we provide such an example, which also is discussed in [8]. Some authors have considered alternative criteria to capture such biomarkers in similar situations. Parodi (2008) [9] et al. propose the area between the ROC curve and the reference diagonal (named as ABCR). It is defined as $ABCR = \int_0^1 | ROC(t) - t | dt$ and can be easily shown to be equal to $\int_{-\infty}^{\infty} |F_Y(z) - F_X(z)| f_X(z) dz$. Even though such a measure does manage to attribute discriminatory ability to markers that may be rejected by the AUC, it is still a hostage of directionality. For example, consider the two perfect markers shown in Figure 1. The ABCR for the upper panels is equal to 0.5 while for the lower panels it is equal to 0.25. Another interesting measure discussed in the same paper is the TNRC which is a useful measure to detect improper markers. We note that all these criteria refer to the two-sample problem which is also the focus of this paper.

When the stochastic ordering assumption is violated, essential differences between the two groups are frequently implied in terms of dispersion. As a result, goodness-of-fit tests, such as the Kolmogorov Smirnov and the two-sample extension of the Anderson Darling test, might provide useful alternatives ([10], [11]). Another measure proposed in [12] is the projected length of the ROC curve. Even though this measure is immune to directionality and can be linearly transformed to attain an interpretation as a probability, it is geometrically difficult to visualize. In addition, its asymptotic behavior is not yet thoroughly studied.

One measure that was very recently studied under a binormal setting is the length of the ROC curve ([13]). Therein, the utility of the length of the binormal model-based ROC curve as an accuracy index is discussed. In this paper, we study its use as a measure for ranking biomarkers in a general setting that involves discrimination of cases and controls that includes non-monotone biomarkers under general parametric as well as non-parametric frameworks. The length successfully captures useful markers, which would have been rejected using traditional measures. We illustrate that the length of the ROC is directly related to a ϕ divergence and we provide a test statistic for inference. We show that the length of the common ROC curve is identical to the length of the optimal, likelihood ratio-based, ROC curve. In other words, the length does not change when one applies the likelihood ratio transformation to the marker values. As we discuss, this is true for any ϕ divergence. Hence, a ϕ divergence reflects the potential discriminatory ability of a biomarker. Among possible ϕ divergences one might employ for inference, the length has the advantage of being geometrically appealing to practitioners. Furthermore, we apply isoperimetric methods to explore its relationship to the AUC of the optimal ROC curve. In other words, given the length of an ROC curve we obtain lower and upper bounds of the AUC of the optimal (likelihood ratio) ROC. We study the asymptotic behavior of the estimated length and its related test statistic assuming a general parametric family. Furthermore, in a non-parametric framework we estimate the length using kernel-densities and explore its statistical properties via bootstrap and permutation tests. Based on simulations, we observe that when the stochastic ordering assumption is violated, then the test based on the length outperforms traditional tests, such as the Wilcoxon, the t-test, and the Kolmogorov-Smirnov test.

Once the initial ranking is done, interest lies in clinically relevant measures that will provide insights regarding the performance of a selected biomarker. While the traditional ROC curve is the most popular statistical tool for the evaluation of a biomarker, it should be avoided when it is improper. We present an appropriate analogue of the ROC graph that visualizes the sensitivity and specificity in the ROC space under a framework that uses two cutoff points. We define an optimal ROC curve, under the notion of a unimodal likelihood ratio ordering, which defines a dispersion ordering. This optimal ROC curve is identical to the generalized ROC curve (gROC) first introduced in [14]. A parametric estimator has been considered in [15] and its Youden index in [16]. Optimality of the gROC under the normality assumption has been shown in [17]. In this paper, we show that the gROC is the optimal ROC under the assumption of a unimodal density ratio. We further discuss the notions of sensitivity and specificity in the presence of two cutoffs which we estimate based on an extension of the Youden index initially discussed in [18] and provide inferences around them that can also accommodate costs.

This paper is organized as follows: In Section 2 we present the proposed measure for biomarker ranking. We discuss its properties and provide results that relate the length of the ROC curve to the length and area of the optimal (likelihood ratio transformed) ROC curve. In addition, we explore the estimation and statistical inference based on this measure. In Section 3 we explore the appropriate ROC analogous when two cutoffs are considered and discuss the conditions under which the two-cutoff ROC is optimal. We present an example that illustrates the fact that a common ROC curve does not reveal the discriminatory ability of an improper marker and the appropriateness of using the length as a measure of discriminatory ability. In Section 4 we discuss estimation and inference for the sensitivity and specificity of such markers under different parametric and non-parametric settings. In Section 5 we discuss our simulations and in Section 6 we employ our approaches in two data sets related to cancer.

2. Length of the ROC Curve and ϕ Divergence

Let Y denote a continuous biomarker score in a study of n_0 healthy individuals and n_1 patients of the diseased group. Denote with $f_0(\cdot)$ and $f_1(\cdot)$ the densities of the healthy and the diseased group, respectively. Also denote the corresponding cumulative distribution functions as $F_0(\cdot) = 1 - S_0(\cdot)$ and $F_1(\cdot) = 1 - S_1(\cdot)$. We assume that the two distributions have common support. The traditional ROC curve is defined as $ROC(t) = S_1\left(S_0^{-1}(t)\right)$, $t \in (0, 1)$. The length of a differentiable ROC curve, l_{ROC}, is given by the following two equivalent expressions:

$$l_{ROC} = \int_0^1 \sqrt{1 + \left(\frac{dgROC(t)}{dt}\right)^2} dt = \int_{-\infty}^{+\infty} \sqrt{\left(\frac{dFPR(c)}{dc}\right)^2 + \left(\frac{dTPR(c)}{dc}\right)^2} dc \tag{1}$$

where $FPR = S_0(c)$ and $TPR = S_1(c)$ for a given cutoff c. Using the result in [1] (pg. 70), $\frac{dgROC(t)}{dt} = \frac{f_1\left(S_0^{-1}(t)\right)}{f_0\left(S_0^{-1}(t)\right)}$ we obtain

$$l_{ROC} = \int_{-\infty}^{+\infty} \sqrt{1 + \left(\frac{f_1(x)}{f_0(x)}\right)^2} f_0(x) dx = \int_{-\infty}^{+\infty} \sqrt{f_1^2(x) + f_0^2(x)} dx \,. \tag{2}$$

We see that $\sqrt{2} \le l_{ROC} < 2$. The minimum value, $\sqrt{2}$, is attained if and only if the distribution of the diseased is identical to that of the non-diseased, which implies a non-informative marker. The quantity $l_{ROC} - \sqrt{2}$, the difference between the length of the ROC curve and the length of a non-informative marker, is a ϕ divergence and is equivalent to the perimeter divergence measure studied by [19], as it is simply equal to the difference in perimeter of the set underneath the ROC curve between the marker studied and a non-informative marker. As shown in [19], the square root of this divergence is a metric divergence.

The following theorem shows that the length (and the corresponding divergence measure) remains the same when one considers the likelihood ratio ROC curve based on the marker Y, which is the length of the ROC curve associated with the random variable $W = \frac{f_1(Y)}{f_0(Y)}$. It is well known that the likelihood ratio ROC is concave and is the optimal ROC one can construct based on a given marker.

Theorem 1

Denote the length of the ROC curve obtained by using the untransformed Y by l_{ROC}^{Y} *and the length of the likelihood ratio ROC curve by* l_{ROC}^{W} *(obtained by using the transformation* $W = \frac{f_1(Y)}{f_0(Y)}$. *Then* $l_{ROC}^{W} = l_{ROC}^{Y}$.

Proof: The proof follows from the fact that the likelihood ratio of the likelihood ratio equals the likelihood ratio (see [1], pg. 94), that is $\frac{g_1(w)}{g_0(w)} = w$, where g_0 and g_1 denote the densities of W for the healthy and diseased respectively. Let $\phi(u) = \sqrt{1 + u^2}$. We have

$$l_{ROC}^{W} = E_{G_0}\phi\left(\frac{g_1(W)}{g_0(W)}\right) = E_{G_0}\phi(W) = E_{F_0}\phi\left(\frac{f_1(Y)}{f_0(Y)}\right) = l_{ROC}^{Y}\square\,. \tag{3}$$

We note here that it is evident from the proof that the above result is valid for *all* ϕ divergences. Thus, one can argue that ϕ divergences are well suited for making fair comparisons between markers based on a single ROC summary measure, as they are invariant to the optimal (likelihood ratio) transformation of the marker values. This is especially obvious in the case when no directionality of disease is assumed by the researchers. The length has the advantage of being a geometrically appealing measure. Moreover, we can relate the length to the AUC of the optimal (likelihood ratio) ROC. If the likelihood ratio and its distribution were known, then the AUC of the optimal ROC could be computed directly. This is almost never the case in practice. However, one can use the following theorem, which provides bounds on the AUC of the optimal ROC curve, given the length of the original ROC curve. These bounds can be used by practitioners to assess biomarkers.

Theorem 2

Let $l_{ROC}^{Y} = l$ *and denote the AUC of the optimal ROC by* AUC^{W}. *Then 2a) For* $\sqrt{2} \le l \le \frac{\pi}{2}$ *we have*

$$\frac{1}{2} + \frac{\left(l^2 - 2\right)}{4(l - 1)} < AUC^{W} \le \frac{1}{2} + S, \tag{4}$$

where S equals the area of a circular segment with chord length equal to $\sqrt{2}$ *and corresponding arc length equal to l. 2b) For* $\frac{\pi}{2} < l < 2$ *we have*

$$\frac{1}{2} + \frac{\left(l^2 - 2\right)}{4(l - 1)} < AUC^W < 1 - \frac{(4 - 2l)^2}{4(4 - \pi)}. \tag{5}$$

Proof: Note that we are dealing here with isoperimetric problems. Since the likelihood ratio ROC curve is concave, and as we have shown earlier has length l, we consider the mathematically equivalent problems of maximizing and minimizing the area among all convex sets within the right isosceles triangle ABC (see Figure 2) formed by a rope of length l tied to the edges B and C.

The maximization of the area given the rope's length, l, is a variant of the famous Dido problem, and is dealt with in [20] (pg. 4, 5, and 292) using methods of optimal control. Although the maximization problem is typically formulated without the convexity constraint, its solution forms a convex set, and hence is the solution to our problem. It turns out that if the rope length, l, does not exceed $\pi/2$, then the solution is to place the rope so as to form a partially circular region (see Figure 3). The area of the convex set formed can be found numerically (the solution coincides with that of the Dido problem, since, in this case, the circular arc stays within the triangle ABC). When l is larger than $\pi/2$, we lay the rope on the two sides AB and AC, moving along the sides an equal distance until we have enough rope remaining to draw the arc of a quarter circle inside the triangle ABC. It can be shown that the distance we need to move up on each side equals $\frac{(2l - \pi)}{(4 - \pi)}$, and the quarter circle arc corresponds to a circle with radius equal to $\frac{(4 - 2l)}{(4 - \pi)}$. The area of the maximizing convex set is then computed using elementary geometry. The corresponding optimal ROC curve in this case is not attained in our framework, due to our assumption that the densities for diseased and non-diseased have common support. An ROC curve in our setting, however, can be infinitely close to the optimal one.

There are two solutions to the minimization problem, which yield a lower bound for AUC^W. One is the triangle BCD, with the length of CD, let γ, equal to $\frac{\left(l^2 - 2\right)}{2(l - 1)}$. The other is a triangle symmetric to the above around a line passing through point A and perpendicular to BC. These two solutions, along with the corresponding minimum area, are obtained by considering the dual problem of maximizing the perimeter of a given area. We spare the reader of the mathematical details of the minimization problem and provide the related Theorem in Web Appendix A □.

Table 1 gives the lower and upper bounds of the optimal (likelihood ratio) ROC for select values of the length.

Inference on the length of the ROC curve is rather straightforward, if we assume that the distributions of diseased and non-diseased belong to the same parametric family, with density $f(x;\theta)$, where θ is an $M \times 1$ parameter vector. Let $f(x;\theta_0) = f_0(x)$ and $f(x;\theta_1) = f_1(x)$ denote the densities of the non-diseased and diseased populations, respectively.

The quantity $l_{ROC} - \sqrt{2}$ can be written as an (*h, ϕ*) divergence measure (see [21]). Since the two distributions have densities, an (*h, ϕ*) divergence is defined as

$$D_{\phi}^{h}(\theta_0, \theta_1) = \int_{\Lambda} h_{\alpha}\left(\int_{-\infty}^{+\infty} \phi_{\alpha}\left(\frac{f(x;\theta_1)}{f(x;\theta_0)}\right) f(x;\theta_0)dx - \phi_{\alpha}(1)\right) d_{\eta}(\alpha), \tag{6}$$

where $h = (h_{\alpha})_{\alpha \in \Lambda}$ and $\phi = (\phi_{\alpha})_{\alpha \in \Lambda}$ are real valued C^2 functions, with $h_{\alpha}(0) = 0$ and η is a σ-finite measure on the measurable space (Λ, β). Furthermore, for every α, either ϕ_{α} is convex and h_{α} is increasing or ϕ_{α} is concave and h_{α} is decreasing. The divergence measure $l_{ROC} - \sqrt{2}$ is obtained by letting $\Lambda = 1$, η a point mass at 1, $h_1(u) = u$, and $\phi_1(t) = \sqrt{1 + t^2}$.

Let $\hat{\theta}_0$ and $\hat{\theta}_1$ be the maximum likelihood estimators of θ_0 and θ_1, and consider the divergence statistic

$$\hat{l}_{ROC} - \sqrt{2} = \int_{-\infty}^{+\infty} \sqrt{1 + \left(\frac{f\left(x;\hat{\theta}_1\right)}{f\left(x;\hat{\theta}_0\right)}\right)^2} f\left(x;\hat{\theta}_0\right)dx - \sqrt{2} \tag{7}$$

We obtain the following result, presented as a theorem, to test the null hypothesis $H_0 : \theta_1 = \theta_0$.

Theorem 3

Under the null hypothesis $H_0 : \theta_1 = \theta_0$, *and assuming* $\frac{n_0}{n_0 + n_1} \to \lambda \in (0, 1)$

$$\frac{4\sqrt{2}n_0n_1}{n_0 + n_1}\left(\hat{l}_{ROC} - \sqrt{2}\right) \to \chi_M^2 \tag{8}$$

Proof: The proof is an immediate consequence of the Corollary 2.(b) in [21] □.

The theorem above allows us to test, within a parametric framework, whether a marker is informative. It is a divergence-based test of the equality of two distributions within a parametric family and is closely related to the likelihood ratio test. We explore the power and size of this test in our simulation studies where we consider also randomization (permutation based) versions of it under different parametric assumptions (results are presented in Web Appendix C and discussed in our simulation section). If the null hypothesis is rejected, then the following theorem, which is an immediate consequence of Corollary 2.(a) in [21], allows the construction of a confidence interval for the length of the ROC curve.

Theorem 4

Under the alternative hypothesis $H_A: \theta_1 \neq \theta_0$, *assuming* $\frac{n_0}{n_1 + n_0} \to \lambda \in (0, 1)$

$$\sqrt{\frac{n_0 n_1}{n_0 + n_1}}\left(\hat{l}_{ROC} - l_{ROC}\right) \to N\left(0, \lambda \mathbf{t}^T I^{-1}(\theta_1)\mathbf{t} + (1 - \lambda)\mathbf{s}^T I^{-1}(\theta_0)\mathbf{s}\right), \qquad (9)$$

where $I(\theta_1)$ *and* $I(\theta_0)$ *are the Fisher information matrices associated with* θ_1 *and* θ_0, *and the elements of the* $M \times 1$ *vectors* $\mathbf{t}$ *and* $\mathbf{s}$ *are given by*

$$t_i = \int_{-\infty}^{+\infty} \frac{\left(\frac{f(x;\theta_1)}{f(x;\theta_0)}\right)}{\sqrt{1 + \left(\frac{f(x;\theta_1)}{f(x;\theta_0)}\right)^2}} \frac{df(x;\theta_1)}{d\theta_{1i}} dx$$

$$s_i = \int_{-\infty}^{+\infty} \left[\sqrt{1 + \left(\frac{f(x;\theta_1)}{f(x\mid;\theta_0)}\right)^2} \frac{df(x;\theta_0)}{d\theta_{0i}} - \frac{\left(\frac{f(x;\theta_1)}{f(x;\theta_0)}\right)}{\sqrt{1 + \left(\frac{f(x;\theta_1)}{f(x;\theta_0)}\right)^2}} \frac{df(x;\theta_0)}{d\theta_{0i}} \frac{f(x;\theta_1)}{f(x;\theta_0)}\right] dx$$

Theorems 3 and 4 assume a common parametric family for the two distributions and are well suited for modeling data sets with large sample sizes using flexible parametric families, which is typically not the case in biomarker discovery. An alternative is to obtain smooth nonparametric estimates for the two distributions and estimate the length of the ROC curve based on equation (2). Testing can then be performed using a randomization test and bootstrap confidence intervals for the length can be obtained.

It should finally be noted that the length of the empirical ROC curve cannot be used for statistical inference, since it can easily be seen that, in the absence of ties, the length of the empirical ROC equals 2. So, in practice the length is estimated either by plugging in estimates of the parameters involved when assuming a parametric family or by first obtaining smooth nonparametric estimates of the densities involved. Further note, that under the classical binormal setting, in the case of equal variances knowledge of the AUC implies knowledge of the length of the ROC and vice versa.

3. The ROC with two cutoffs

Under the usual setting where it is assumed that the higher the measurement the more likely the existence of disease, the common ROC analysis plots the sensitivity $Se(c) = TPR = 1 - F_1(c) = S_1(c)$ over the false positive rate $1 - Sp(c) = FPR = 1 - F_0(c) = S_0(c)$ for all possible cutoffs c, with $c \in (-\infty, +\infty)$. In the case where we consider very high and very low measurements as indicative of the disease, while measurements of the healthy are considered to lie in a bounded interval, let (c_1, c_2), then a proper analogous to an ROC plot can be considered that may be more appropriate to assess the biomarker's performance. In this case the sensitivity, the specificity, and the false positive rates are defined as follows:

$$Se(c_1, c_2) = F_1(c_1) + S_1(c_2) \;\; Sp(c_1, c_2) = F_0(c_2) - F_0(c_1) \;\; Fp(c_1, c_2) = F_0(c_1) + S_0(c_2),$$

where $c_1 < c_2$. To perform an ROC analysis one can plot two surfaces in the unit cube, one that refers to the sensitivity and one that refers to the specificity for all $c_1 < c_2, c_i \in (-\infty, +\infty)$, $i = 1{,}2$ (see Figures 4 and 5). Proceeding further, one can construct an equivalent definition, analogous to the one in the common settings in which $ROC(t) = S_1\left(S_0^{-1}(t)\right)$, $t \in (0, 1)$. In our case, since we have two cutoffs, there is the implication that for different pairs of c_1 and c_2 we may get the exact same sensitivity and specificity. Observe that by setting the false positive rate equal to an argument $t \in (0, 1)$ we derive:

$$Fp(c_1, c_2) = t \Rightarrow c_2 = S_0^{-1}(t - F_0(c_1)), \quad F_0(c_1) < t < 1 \tag{10}$$

and by substitution to the sensitivity we derive:

$$Se(t; c_1) = F_1(c_1) + ROC(t - F_0(c_1)), \quad F_0(c_1) < t < 1. \tag{11}$$

Obviously, as $c_1 \to -\infty$ we obtain the classical ROC setting. Note that as $t \to 0$ $Se(t, c_1) \to 1$. Formula (11) above implies that for a given cutoff c_1 we can plot a projection of the sensitivity surface on the unit rectangle for all $t' \in (0, 1 - F_0(c_1))$ where $t' = 1 - t$ expresses the specificity. One could derive this plot by scanning c_1 restricted on a straight line parallel to the c_2–axis (i.e. taking an infinitely thin slice) of the sensitivity surface. It can be easily shown (since the derivative of the ROC curve equals the density ratio evaluated at the t-th upper quantile of the healthy distribution) that the derivative of the sensitivity w.r.t. c_1 is

$$\frac{dSe(t; c_1)}{dc_1} = f_0(c_1)\left\{r(c_1) - r\left(S_0^{-1}(t - F_0(c_1))\right)\right\}. \tag{12}$$

If one is interested in forcing desired sensitivity levels to derive the corresponding specificity values, then using the argument t^* to avoid confusion, we get:

$$Se(c_1, c_2) = t^* \Rightarrow c_2 = F_1^{-1}(F_1(c_1) + 1 - t^*) \quad F_1(c_1) < t^* < 1 \tag{13}$$

and by substitution to the specificity we derive:

$$Sp(t^*; c_1) = F_0\left(F_1^{-1}(F_1(c_1) + 1 - t^*)\right) - F_0(c_1), \quad F_1(c_1) < t^* < 1. \tag{14}$$

Similar expressions for given c_2 can be also derived. Given two cutoffs we utilize the dominant generalized ROC curve (gROC):

$$\begin{aligned} gROC(t) &= \sup_{F_0(c_1) + S_0(c_2) = t} \{F_1(c_1) + S_1(c_2)\} \\ &= \sup_{c_1 \le F_0^{-1}(t)} Se(t; c_1) = \sup_{c_1 \le F_0^{-1}(t)} \{F_1(c_1) + ROC(t - F_0(c_1))\} \;\; t \in (0, 1). \end{aligned} \tag{15}$$

The above ROC curve was first introduced in [14]. They refer to it as the generalized ROC (gROC) curve. We follow this terminology even though one could hypothesize that a generalized ROC would, in principle, refer to settings with more than two cutoff points. Whereas considering more than two cutoff points may be useful in general classification problems with multimodal distributions, we find it hard to imagine practitioners entertaining this idea when dealing with biomarkers.

It is easily seen that $gROC(t) \to 0$ as $t \to 0$, $gROC(t) \to 1$ as $t \to 1$ and is non-decreasing. Furthermore, by observing that $Se(t; -\infty) = ROC(t)$ and $Se\left(t; F_0^{-1}(t)\right) = 1 - ROC(1-t)$, we have that $gROC(t) \geq ROC(t)$ and $gROC(t) \geq 1 - ROC(1-t)$. We note here that $1 - ROC(1-t)$ is the appropriate ROC curve in the case where smaller measurements are indicative of the disease.

The two cutoff ROC curve is optimal under the assumption that the density ratio $r(y) = \frac{f_1(y)}{f_0(y)}$ is unimodal in the sense that there exists y^* (possibly $-\infty$ or $+\infty$) such that $r(y)$ is strictly decreasing in $(-\infty, y^*)$ and strictly increasing in $(y^* = +\infty)$. This is easy to see since the ROC curve based on the density (likelihood) ratio transformed measurements results in a two cutoff decision rule. The assumption of a unimodal density ratio is a dispersion ordering of the two distributions involved (the distribution of healthy individuals is less dispersed than the distribution of the diseased) and is closely related to the uniform conditional variability ordering studied in [22]. Note that in the case of a multimodal density ratio one would need to consider more than two cutoffs to define an optimal ROC.

We note that when $y^* = -\infty$ then we have an increasing density ratio and $gROC(t) = ROC(t)$ for $0 < t < 1$, whereas when $y^* = +\infty$ then we have a decreasing density ratio and $gROC(t) = 1 - ROC(1-t)$ for $0 < t < 1$. Also, when y^* is a real number and $r(-\infty) = r(\infty)$ (as is the case with two Normal distributions with the variance of the diseased greater than the variance of the healthy) for each t we obtain two cutoffs that are real numbers. In this case, by setting the expression in equation (16) equal to 0, we obtain the unique solution for the lower cutpoint, $c_1^*(t)$, that satisfies

$$r(c_1) = r\left(S_0^{-1}(t - F_0(c_1)\right). \tag{16}$$

Obviously the second cutoff equals $c_2^*(t) = S_0^{-1}\left(t - F_0(c_1^*)\right)$. Thus $gROC(t) = F_1\left(c_1^*(t)\right) + S_1\left(c_2^*(t)\right)$. A straightforward application of the envelope theorem (see for example [23], pg. 603–609) yields

$$\frac{dgROC(t)}{dt} = r\left(c_1^*(t)\right) = r\left(c_2^*(t)\right) \tag{17}$$

Furthermore, the Youden index of $gROC(t)$, which is commonly used to select cutoff points in an ROC setting, is given by

$$\sup_{0 < t < 1} \left| gROC(t) - t \right| = \sup_{0 < t < 1} (gROC(t) - t), \tag{18}$$

since, in this case, $gROC(t)$ is the optimal ROC curve and thus concave. It is easy to see that maximization of $gROC(t) - t$ leads to selecting the two cutoff points that satisfy

$$\frac{dgROC(t)}{dt} = r\left(c_1^*(t)\right) = r\left(c_2^*(t)\right) = 1 \tag{19}$$

This method of obtaining cutoffs can be extended to the situation where costs are available. For example, as discussed in [1] (pg. 31,32,72), denote by C the cost of performing the test, $C_D^{(1)}$ and $C_D^{(0)}$ the costs of treatment and morbidity for diseased subjects that test positive and negative respectively, $C_{\bar{D}}^{(1)}$, the cost of work-up, stress and unnecessary treatment for the non-diseased that test positive. Then minimization of the overall cost of disease per subject in the population in the presence of testing leads to selecting the two cutoff points that satisfy

$$\frac{dgROC(t)}{dt} = r\left(c_1^*(t)\right) = r\left(c_2^*(t)\right) = \left(\frac{1 - \rho}{\rho}\right) \frac{C_{\bar{D}}^{(1)}}{C_D^{(0)} - C_D^{(1)}}, \tag{20}$$

where ρ denotes disease prevalence. In the following section we discuss, among other things, the estimation of cutoffs in various settings and the associated inference.

Example:

Consider two markers M_1 and M_2 that both follow a standard normal distribution for the non-diseased population. Assume the measurements of the diseased population for the first marker, M_1, follow a $N(1, 1^2)$ distribution, whereas the measurements of the diseased population for the second marker, M_2, follow a $N(0, 4^2)$ distribution (see Figure 6 for the discussion that follows). The ROC curves for these two markers are: $ROC_1(t) = \Phi\left(1 + \Phi^{-1}(t)\right)$ (with AUC_1 = 0.76025 and length $l_{ROC_1} = 1.5465$) and $ROC_2(t) = \Phi\left(\frac{1}{4}\Phi^{-1}(t)\right)$ (with AUC_2 = 0.5000 and length $l_{ROC_2} = 1.6701$). By examining the two ROC curves one will be tempted to conclude that M_1 is superior to M_2. The same conclusion would follow by comparing the two AUCs. However, whereas $ROC_1(t)$ is the optimal curve for M_1, $ROC_2(t)$ is not the optimal curve for M_2. The density ratios for the two markers are $r_1(y) = e^{\frac{1}{2}(2y - 1)}$ and $r_2(y) = \frac{1}{4}e^{\frac{15}{32}y^2}$. Since an ROC curve is invariant to monotone transformations, the optimal curve (likelihood ratio ROC) for M_1 is based on the measurements themselves whereas the optimal curve for M_2 is the curve based on the squared values of the measurements, which here yields $gROC_2(t) = 1 - G\left(\left(G^{-1}(1 - t)\right)/16\right)$, where G is the cdf of a χ_1^2 distribution. We have $AUC_{gROC_2} = 0.8440$, $l_{gROC_2} = 1.6701$. Note that the length of the optimal ROC for the second marker is the same as that of ROC_2. By comparing the two markers based on their optimal ROC curves one would conclude that the second marker is superior, which is the conclusion one would arrive at when using the length as a summary measure.

To obtain cutoffs based on the Youden index criterion we simply set the density ratios to one and thus obtain a single cutoff for M_1 ($c = 0.5$) and two cutoffs for M_2 ($c_1 = -1.71972$ and $c_2 = +1.71972$). For M_1 this results in a false positive rate of 0.308538 and a sensitivity equal to 0.691462. For M_2, using two cutoffs, the false positive rate is 0.0854834 and the sensitivity is equal to 0.667246.

4. Estimation and inference for the sensitivity and specificity of the selected markers

In a given set of the top candidates, both proper and improper ROC curves might appear. For those biomarkers that exhibit proper ROC curves, well defined and studied measures are available in the literature in order to evaluate them. For the evaluation of those biomarkers that exhibit an improper ROC curve, caution is required. The AUC bounds of the optimal ROC curve for a given marker (see Theorem 2) provide a rough indication of the discriminatory potential of the marker. Obviously, the full potential of the marker can be assessed by using the likelihood ratio transformation (see also [24]). However, in practice the likelihood ratio is not known. Martínez-Camblor et al. (2018) ([17]) explore functional transformations to deal with this problem but as they note this practice can produce classification regions with no practical interpretation. An alternative would be to consider nonparametric estimation of the density ratio under the assumption of unimodality, but we leave this as the goal of future research. Furthermore, given a large number of candidate biomarkers to be ranked, exploring a suitable functional transformation for each would be a daunting task. We present below a simplified framework with which we can estimate the sensitivity, specificity, Youden-based optimal cutoffs (possibly modified to account for costs), and the appropriate construction of confidence intervals. Below we explore parametric and non-parametric approaches.

4.1. Normality assumption

4.1.1. Estimation and inference for the sensitivity and specificity when given a pair of cutoffs—For given cutoffs of an improper marker the corresponding sensitivity and specificity under the normality assumption (binormal setting) we have:

$$Se(c_1, c_2) = \Phi\left(\frac{c_1 - \mu_1}{\sigma_1}\right) + \Phi\left(\frac{\mu_1 - c_2}{\sigma_1}\right), \quad Sp(c_1, c_2) = \Phi\left(\frac{c_2 - \mu_0}{\sigma_0}\right) - \Phi\left(\frac{c_1 - \mu_0}{\sigma_0}\right)$$

and the estimated sensitivity and specificity for given c_1, c_2 are given by plugging in maximum likelihood estimates of the corresponding means and variances: $\hat{\mu}_0$, $\hat{\sigma}_0$, $\hat{\mu}_1$, and $\hat{\sigma}_1$. For the construction of the underlying confidence intervals, one can simply apply the delta method: $Var(\widehat{Se}) \approx \left(\frac{\partial Se}{\partial \mu_1}, \frac{\partial Se}{\partial \sigma_1}\right)\Sigma_1\left(\frac{\partial Se}{\partial \mu_1}, \frac{\partial Se}{\partial \sigma_1}\right)'$ where an estimate of the 2 × 2 diagonal covariance matrix Σ_1 can be derived by using $Var(\hat{\mu}_0) = \frac{\hat{\sigma}_1^2}{n_1}$ and $Var(\hat{\sigma}_1) = \frac{\sigma_1^2}{2(n_1 - 1)}$.

The partial derivatives can be derived in closed form. A similar expression is derived for $Var(\hat{S}p)$, for which the involved partial derivatives also can be derived in closed form. In order for the proposed confidence intervals to avoid exceeding the bounds of (0,1), following the ideas presented in [25], we first transform the sensitivity using $\Phi^{-1}(\cdot)$, then derive based on the delta method a 95% confidence interval for the transformed sensitivity, and finally back transform the endpoints of that confidence interval with $\Phi(\cdot)$. Thus, the proposed confidence interval is: $\Phi\left(\Phi^{-1}(\hat{S}e) \pm 1.96\sqrt{Var\left(\Phi^{-1}(\hat{S}e)\right)}\right)$, where $Var\left(\Phi^{-1}(\hat{S}e)\right) \approx \left(\frac{\partial\Phi^{-1}(Se)}{\partial\mu_1}, \frac{\partial\Phi^{-1}(Se)}{\partial\sigma_1}\right)\Sigma_1\left(\frac{\partial\Phi^{-1}(Se)}{\partial\mu_1}, \frac{\partial\Phi^{-1}(Se)}{\partial\sigma_1}\right)'$. The derivation for the corresponding confidence interval of the specificity is similar.

4.1.2. Estimation and inference for the sensitivity and specificity when the pair of cutoffs is estimated—

In the absence of any clinical insight regarding where to locate the two cutoff points, we can estimate them under the optimization of some clinically relevant criterion. Such a criterion is the Youden index initially suggested in [26] for the single cutoff setting and extended to the dual cutoff framework given in [18]:

$$J = max_{c_1, c_2 : c_1 < c_2}\left(Se(c_1, c_2) + Sp(c_1, c_2) - 1\right) \tag{21}$$

Under the normality assumption, the pair of cutoffs can be derived in closed form by:

$$J = max_{c_1}\left(\Phi\left(\frac{c_1 - \mu_1}{\sigma_1}\right) + \Phi\left(\frac{\mu_0 - c_1}{\sigma_0}\right)\right) + max_{c2}\left(\Phi\left(\frac{\mu_1 - c_2}{\sigma_1}\right) + \Phi\left(\frac{c_2 - \mu_0}{\sigma_0}\right) - 1\right) \tag{22}$$

To maximize J, we need to separately maximize both terms of (22), which are both equivalent to the maximization of the regular Youden index in the common single cutoff setting of the binormal case. Hence, under the necessary assumption that $\sigma_0 < \sigma_1$, the pair of cutoffs can be derived in closed form by:

$$c_{1,2}^{*} = \frac{\sigma_0^2\mu_1 - \sigma_1^2\mu_0 \pm \sigma_0\sigma_1\sqrt{(\mu_0 - \mu_1)^2 + \left(\sigma_0^2 - \sigma_1^2\right)\log\left(\frac{\sigma_0^2}{\sigma_1^2}\right)}}{\sigma_0^2 - \sigma_1^2} \tag{23}$$

We note that in the binormal setting the density ratio is unimodal and the two cutoffs are obtained by solving equation (19). The estimated cutoffs can be derived by simply plugging in (23), the maximum likelihood estimates of the means and variances. After the pair of cutoffs is estimated, then the corresponding sensitivity and specificity (at the estimated pair of cutoffs) are correlated, as opposed to the previous section where the cutoffs were given. Following the ideas in [25] we denote $\delta_e = \Phi^{-1}\left(Se\left(c_1^*, c_2^*\right)\right)$ and $\delta_p = \Phi^{-1}\left(Sp\left(c_1^*, c_2^*\right)\right)$. We can derive the estimates $\hat{\delta}_e$ and $\hat{\delta}_p$ by simply plugging in the maximum likelihood estimates of the means and variances i.e., $\hat{\mu}_0$, $\hat{\sigma}_0$, $\hat{\mu}_1$, $\hat{\sigma}_1$). Then, based on the delta method we derive: $Var\left(\hat{\delta}_e\right) \approx \left(\frac{\partial\hat{\delta}_e}{\partial\mu_0}, \frac{\partial\hat{\delta}_e}{\partial\sigma_0}, \frac{\partial\hat{\delta}_e}{\partial\mu_1}, \frac{\partial\hat{\delta}_e}{\partial\sigma_1}\right)\Sigma\left(\frac{\partial\hat{\delta}_e}{\partial\mu_0}, \frac{\partial\hat{\delta}_e}{\partial\sigma_0}, \frac{\partial\hat{\delta}_e}{\partial\mu_1}, \frac{\partial\hat{\delta}_e}{\partial\sigma_1}\right)$, where

Σ is the diagonal matrix $\Sigma = diag\{Var(\hat{\mu}_0), Var(\hat{\sigma}_0), Var(\hat{\mu}_1), Var(\hat{\sigma}_1)\}$ and $Var(\hat{\mu}_0) = \frac{\sigma_0^2}{n_0}$, $Var(\hat{\sigma}_0) = \frac{\hat{\sigma}_0^2}{2(n_0 - 1)}$, $Var(\hat{\mu}_1) = \frac{\sigma_1^2}{n_1}$, and $Var(\hat{\sigma}_1) = \frac{\sigma_1^2}{2(n_1 - 1)}$. The corresponding estimates, and thus $\sum^{\wedge}$, can be derived by simply plugging in the maximum likelihood estimates of the means and variances. The expression is similar for $Var\left(\hat{\delta}_p\right)$. The covariance can be derived by: $Cov\left(\hat{\delta}_e, \hat{\delta}_p\right) \approx \left(\frac{\partial \hat{\delta}_e}{\partial \mu_0}, \frac{\partial \hat{\delta}_e}{\partial \mu_0}, \frac{\partial \hat{\delta}_e}{\partial \mu_1}, \frac{\partial \hat{\delta}_e}{\partial \sigma_1}\right)\Sigma\left(\frac{\partial \hat{\delta}_p}{\partial \mu_0}, \frac{\partial \hat{\delta}_p}{\partial \mu_0}, \frac{\partial \hat{\delta}_p}{\partial \mu_1}, \frac{\partial \hat{\delta}_p}{\partial \sigma_1}\right)'$. All partial derivatives can be found in closed form. An approximate joint confidence region for (δ_e, δ_p) can be obtained by the ellipse defined by $(\mathbf{x} - \mathbf{a})'\hat{\Sigma}^{-1}(\mathbf{x} - \mathbf{a}) = q$, where $\mathbf{a} = \left(\hat{\delta}_e, \hat{\delta}_p\right)$ and q is the 95%th percentile of a χ_2^2. To derive a joint confidence region of the estimated sensitivity and specificity within the ROC space, we back-transform the elliptical confidence region by applying $\Phi(\cdot)$ on its coordinates, and hence derive an egg-shaped confidence region for (*Se*, *Sp*). Marginal Wald type 95% confidence intervals for the sensitivity and specificity can be also derived within the ROC space by $\Phi\left(\hat{\delta}_e \pm 1.96\sqrt{Var\left(\hat{\delta}_e\right)}\right)$ and $\Phi\left(\hat{\delta}_p \pm 1.96\sqrt{Var\left(\hat{\delta}_p\right)}\right)$, respectively. Based on the Bonferroni correction, once the marginal confidence intervals are derived, we also can proceed with rectangular confidence regions that are more easily communicated to clinicians. By constructing the rectangle with sides based on $\Phi\left(\hat{\delta}_e \pm 2.24\sqrt{Var\left(\hat{\delta}_e\right)}\right)$ and $\Phi\left(\hat{\delta}_p \pm 2.24\sqrt{Var\left(\hat{\delta}_p\right)}\right)$ that correspond to adjusted 97.5% marginal confidence intervals, we obtain a 95% confidence rectangular region for the (*Se*, *Sp*). However, such a rectangular region does not accommodate the implied correlation of the estimated sensitivity and specificity at the Youden-based cutoffs, and for that reason it consistently provides a larger area compared to the egg-shaped proposed region previously described. This is observed in our simulation studies (Web Appendix C). In case normality is violated, transformations to normality might be useful alternatives for a selected marker. A discussion about the Box-Cox transformation ([27]) in this setting is given in the Web Appendix B.

4.2. Estimation and inference using kernel density estimates

4.2.1. Estimation and inference for the sensitivity and specificity when given a pair of cutoffs—

When no parametric assumptions can be justified by the data at hand, then we need to proceed non-parametrically. In such a case, we employ the kernel density estimate of the form:

$$\hat{f}_0(y) = \frac{1}{n_0 h_0} \sum_{i=1}^{n_0} K\left(\frac{y - Y_{0i}}{h_0}\right), \tag{24}$$

where h_0 is the bandwidth for which one can use a simple plug-in expression like the one given by Silverman's rule ([28]): $h_0 = 0.9min(sd(Y_0), iqr(Y_0)/1.34)n_0^{-0.2}$, where Y_{0i}, $i = 1, \ldots, n_0$ are the marker scores for the healthy group. Similarly for $\hat{f}_1(y)$. Given the cutoffs we obtain for the sensitivity and specificity: $\hat{S}e(c_1, c_2) = \hat{F}_1(c_1) + 1 - \hat{F}_1(c_2)$, $\hat{S}p(c_1, c_2) = \hat{F}_0(c_2) - \hat{F}_0(c_1)$ where $\hat{F}_0(x) = \int_{-\infty}^{x} \hat{f}_0(z)dz$, and similarly for $\hat{F}_1(y) = \int_{-\infty}^{y} \hat{f}_1(z)dz$.

A common choice that we employ in practice is the normal kernel. This allows us to write $\widehat{F}_0(y) = \frac{1}{n_0}\sum_{i=1}^{n_0}\Phi\left(\frac{y - Y_{0i}}{h_0}\right)$ and $\widehat{F}_1(y) = \frac{1}{n_1}\sum_{j=1}^{n_1}\Phi\left(\frac{y - Y_{1i}}{h_1}\right)$. From standard kernel theory (see [29] and [30]) it can be shown that kernels are asymptotically equivalent to empirical estimates (Theorem 2.1. of [31]), and thus we have:

$$Var\left(\widehat{S}e(c_1, c_2)\right) \sim \frac{1}{n_1}F_1(c_1)(1 - F_1(c_1)) + \frac{1}{n_1}F_1(c_2)(1 - F_1(c_2)) - \frac{F_1(c_1)(1 - F_1(c_2))}{n_1}.$$

Similarly for the specificity. An estimate can be derived by simply plugging in the estimated kernel-based distribution functions. Using the delta method we have $Var\left(\Phi^{-1}\left(\widehat{S}e(c_1, c_2)\right)\right) = \frac{\partial\Phi^{-1}\left(\widehat{S}e(c_1, c_2)\right)}{\partial\widehat{S}e(c_1, c_2)} \times Var\left(\widehat{S}e(c_1, c_2)\right)$, and thus the proposed 95% confidence interval for the sensitivity is given by $\Phi\left(\Phi^{-1}\left(\widehat{S}e(c_1, c_2)\right) \pm 1.96\sqrt{Var\left(\Phi^{-1}\left(\widehat{S}e(c_1, c_2)\right)\right)}\right)$. Similarly for the specificity. For a joint confidence interval, one can readily obtain a rectangular region based on the Bonferroni adjustment, since the estimated sensitivity and specificity are uncorrelated, as the cutoffs are given.

4.2.2. Estimation and inference for the sensitivity and specificity when the pair of cutoffs is estimated—

Since the kernel based cutoffs cannot be derived in closed form, a numerical procedure is needed to obtain the estimates of the optimal pair of cutoffs, $\widehat{c}_1^*$ and $\widehat{c}_2^*$. Once the corresponding $\widehat{S}e\left(\widehat{c}_1^*, \widehat{c}_2^*\right)$ and $\widehat{S}p\left(\widehat{c}_1^*, \widehat{c}_2^*\right)$ are derived, then we consider again: $\widehat{\delta}_e = \Phi^{-1}\left(\widehat{S}e\left(\widehat{c}_1^*, \widehat{c}_2^*\right)\right)$, $\widehat{\delta}_p = \Phi^{-1}\left(\widehat{S}p\left(\widehat{c}_1^*, \widehat{c}_2^*\right)\right)$. To derive estimates of $Var\left(\widehat{\delta}_e\right)$ and $Var\left(\widehat{\delta}_p\right)$ and $Cov\left(\widehat{\delta}_e, \widehat{\delta}_p\right)$, we employ the following bootstrap scheme:

- Step 1: Calculate the bandwidths h_0 and h_1 based on the scores of the healthy and diseased group respectively. Derive the corresponding estimates of the cutoffs using the kernel density estimates $\widehat{f}_0$ and $\widehat{f}_1$. Obtain numerically the Youden-based optimal pair of cutoffs $(\widehat{c}_1^*, \widehat{c}_2^*)$ as well as $\widehat{S}p\left(\widehat{c}_1^*, \widehat{c}_2^*\right)$ and $\widehat{S}e\left(\widehat{c}_1^*, \widehat{c}_2^*\right)$.

- Step 2: Sample with replacement from the scores of the healthy and diseased group separately, let $Y_0^{(b)}$ and $Y_1^{(b)}$.

- Step 3: Set $Y_{0i}^{(s)} = Y_{0i}^{(b)} + e_{0i}$, where $e_{0i} \sim N\left(0, h_0^2\right)$, and $Y_{1i}^{(s)} = Y_{1i}^{(b)} + e_{1i}$, where $e_{1i} \sim N\left(0, h_1^2\right)$. Using $Y_0^{(s)}$ and $Y_1^{(s)}$ obtain numerically the Youden-based optimal pair of cutoffs $(\widehat{c}_1^*, \widehat{c}_2^*)$, as well as $\widehat{S}p\left(\widehat{c}_1^*, \widehat{c}_2^*\right)$ and $\widehat{S}e\left(\widehat{c}_1^*, \widehat{c}_2^*\right)$.

- Step 4: Use the $\Phi^{-1}(\cdot)$ transformation and derive the current (for the current bootstrap sample) $\Phi^{-1}\left(\widehat{S}p\left(\widehat{c}_1^*, \widehat{c}_2^*\right)\right)$ and $\Phi^{-1}\left(\widehat{S}e\left(\widehat{c}_1^*, \widehat{c}_2^*\right)\right)$.

- Step 5: Repeat Steps 2 and 4 B times, and based on these B estimates derive the bootstrap-based estimates of $Var\left(\Phi^{-1}\left(\hat{S}p\left(\hat{c}_1^*, \hat{c}_2^*\right)\right)\right)$, $Var\left(\Phi^{-1}\left(\hat{S}e\left(\hat{c}_1^*, \hat{c}_2^*\right)\right)\right)$ and the underlying $Cov\left(\Phi^{-1}\left(\hat{S}p\left(\hat{c}_1^*, \hat{c}_2^*\right)\right), \Phi^{-1}\left(\hat{S}e\left(\hat{c}_1^*, \hat{c}_2^*\right)\right)\right)$.

- Step 6: Construct the 95% elliptical confidence region for $\left(\Phi^{-1}(Sp(c_1, c_2)), \Phi^{-1}(Se(c_1, c_2))\right)$ and then construct the egg-shaped confidence region in the ROC space by transforming back the coordinates of the elliptical confidence region using $\Phi^{-1}(\cdot)$.

In practice, we use $B = 400$ bootstrap samples. An approximate joint confidence region for (δ_e, δ_p) can be obtained as described in the previous sections, using the construction of an ellipse defined by $(\mathbf{x} - \mathbf{a})'\hat{\Sigma}^{-1}(\mathbf{x} - \mathbf{a}) = q$, where q is the 95%th percentile of a χ_2^2. The resulting joint confidence region of $\mathbf{a} = (Se, Sp)$ in the ROC space is obtained by back-transforming. Note that in Step 3 we add a generated error term to the bootstrapped sample of each group. Alternatively, one could simply proceed with the drawn bootstrap samples instead. However, as we point out later in our simulation study, we find the presented smooth bootstrap scheme to perform better than a regular bootstrap algorithm that would not involve the addition of these error terms. Such a finding is in-line with similar conclusions under a different ROC setting presented in [32].

5. Simulations

To evaluate our approaches, we consider Monte Carlo simulations under various scenarios. Our simulations refer to the evaluation of the following:

- The size and the power of the test statistic related to the proposed length of the ROC curve (thoughout our simulations the significance level considered is $a = 0.05$).

- The coverage of the egg-shaped versus the rectangular confidence region, when the two underlying cutoffs are estimated by the data.

In our simulations we have generated data from normal distributions, gamma distributions, as well as a mixture of normals to explore the robustness of our approaches. In terms of sample size, we considered scenarios with (n_0, n_1) = (50,50),(100,100),(200,200), and (500,500). All tables of our simulation results are given in Web Appendix C.

5.1. Size and power of the proposed test based on the length of the ROC curve

Regarding the size and power of the proposed test, we consider both traditional scenarios in which the stochastic ordering $X < Y$ holds (or traditional setting), as well as scenarios in which this assumption is violated (or non-traditional setting). In both cases we compare the size and power of our statistic to the well-known Wilcoxon (Mann-Whitney), t-test, and the Kolmogorov-Smirnov test (KS). For the traditional setting, we consider 12 different scenarios in terms of distributions for each sample size: 6 related to normal distributions and 6 to gamma distributions. We observe that, in terms of size, our test provides satisfactory performance, just as the competing well-known tests (see Tables 1-2 of the Web Appendix).

In terms of power, for the normal related scenarios, we observe that the t-test and the Wilcoxon test exhibit the best performances, as expected. However, even though our performance is inferior, as anticipated, the differences compared to the traditional tests are not dramatic. For example, after applying a Box-Cox transformation our test exhibits almost equivalent performance with the Kolmogorov Smirnov test, even in the gamma-related scenarios. Our performance is even better if we assume normality, and the generated data are indeed normally distributed. This is not a surprise, as we can accommodate such an assumption. We observe that the Box-Cox based length-test works similarly well in all scenarios, in spite of the estimated extra parameter. Regarding our kernel based approach, we observe nice size performance, as well as satisfactory power performance throughout.

Regarding the non-traditional scenarios, we consider two settings for the power. One that refers to two underlying normal distributions with unequal variances, and one that refers to two gammas (the selected parameters are given in the Web Appendix). Under the non-traditional scenarios, the differences are dramatic in favor of our methods. More specifically, we observe that the $t-test$ and the Wilcoxon test completely collapse, as expected. The KS test is dramatically outperformed by our approaches throughout, even in the cases for which we do not make any distributional assumption (see Tables 1-2 of the Web Appendix). Furthermore, even in cases for which we make the wrong distributional assumptions, since a gamma model does not lie in the Box-Cox family, we still observe better performance in favor of our test (see Table 2 of the Web Appendix).

5.2. Coverage and areas of egg-shaped and rectangular confidence regions of the sensitivity and the specificity

For this setting we consider both cutoffs to be unknown, and we estimate them based on the extension of the Youden index. Thus, we compare the approaches of $egg(normal)$, $egg(BC)$, and $egg(kernels)$, depending on our underlying assumption of the distribution of the two groups. We compare these confidence regions in terms of coverage and area to their rectangular counterparts: $rect(normal)$, $rect(BC)$, and $rect(kernels)$. We observe that the rectangular regions are more conservative throughout, and that they are outperformed by the proposed egg-shaped confidence regions in terms of their area. The reason is that the egg-shaped regions take into account the underlying correlation of the estimated pair of sensitivity and specificity, whereas the rectangular confidence regions do not. All relevant results are presented in Table 3. We observe that a smooth bootstrap ([32], [33]) scheme is more preferable in attaining the desired coverage as opposed to a traditional bootstrap based algorithm.

5.3. A note on bandwidth selection

As pointed out by a referee different kernel based strategies may provide improved results with regards to our non-parametric strategy. This is true as presented in the vast literature regarding bandwidth selection when kernel density estimates are involved. For our methods and simulations, we have used a simple closed form bandwidth [28]. Other more sophisticated bandwidths used in kernel density estimation are available in the literature. A standard textbook that contains such strategies is given by [29] among others. Another kernel technique of interest is presented in [34] and is based on diffusion. We have compared the latter with the standard plug-in bandwidth of Silverman's rule previously discussed in section 4. Our comparisons are focused on the bias, the standard error, and the MSE that relate to the estimation of the length of the ROC curve. We have considered our aforementioned gamma related scenarios as well as the bimodal mixture (for the values of the parameters considered see Web Appendix C) and used 1000 Monte Carlo iterations. Our results are presented in Table 4 of the Web Appendix C where we compare the two kernel strategies. We observe that for the gamma related scenarios the diffusion is the clear winner while results are similar in the bimodal setting. Generally, the optimal bandwidth strongly depends on the underlying setting. For that reason, inferences based on smoothed estimations are generally complex. Further comparisons with cross validation bandwidths are also of interest but such issues lie beyond the scope of this study and are left for future research.

6. Applications

6.1. Pancreatic Cancer

We investigated the contribution of autoantibodies as potential circulating biomarkers for PDAC early detection using a case-control study design. To assess autoantibody levels in early stage PDAC patient samples, we used microarray slides coated with a nitrocellulose-surface and spotted with 121 proteins, previously described as tumor-associated antigens across different solid tumor types from multiple prior studies ([35], [6], [36]). We tested three independent plasma sample sets: set 1 consisted of stages IB to IIB PDAC cases ($n = 10$), healthy controls ($n = 10$), and chronic pancreatitis cases ($n = 10$); set 2 consisted of early-stage (IA to IIA) PDAC cases ($n = 42$), healthy controls ($n = 50$), and chronic pancreatitis cases ($n = 50$); and set 3 consisted of resectable PDAC cases ($n = 21$) and benign pancreatic cyst cases ($n = 14$). Plasma samples were individually hybridized, and immunoglobulin G reactivity against each spotted antigen was quantified using an indirect immunofluorescence protocol. In our analysis, we consider the performance of the autoantibodies for group separation of the healthy versus the PDAC patients, after merging the data of all three sets. This suggests a total sample size for the control group of $n_0 = 60$, and a sample size for the PDAC group of $n_1 = 73$. In Figure 7, we plot the AUC of all 121 autoantibodies versus the kernel-based length, and we observe that the highest length is achieved by the autoantibody HNRNPUL1. The empirical and the kernel-based estimates of the traditional ROC of HNRNPUL1 are shown in Figure 8. The empirical AUC of that autoantibody with the corresponding 95% confidence interval is 0.4468(0.3392,0.5544). For the kernel-based length, we derive an estimate of 1.8255 with a 95% confidence interval of (1.7846–1.8748). We note that, based on Theorem 2, a length of 1.8255 corresponds to an area under the optimal ROC curve between 0.9035 and 0.9645, which implies that the marker has high potential. We provide the estimated kernel-based dominant generalized ROC (gROC) curve in Figure 8 which is obtained by simply using the ratio of the kernel based estimators. Its AUC equals to 0.9154. We have to stress that in theory the gROC is the optimal ROC under the condition of a unimodal likelihood ratio and thus not always optimal. Moreover, the estimation of the gROC is computationally intensive, especially if one wants to obtain the gROC estimates for a large number of biomarker candidates. On the other hand, estimates of the length are computationally trivial to obtain and at the same time it provides estimated bounds of the optimal and unknown optimal AUC. the theoretical gROC is not always the optimal ROC and is computationally intensive. The surfaces of the sensitivity and specificity of HNRNPUL1 are given in Figure 9. This plot helps an investigator visualize on the original scale of the marker the tradeoff of the sensitivity and specificity for various pairs of c_1 and c_2. The top panel of Figure 9 provides the egg-shaped confidence region around the Youden-based optimal pair of sensitivity and specificity. Remember that a perfect marker always yields a length of 2, while a useless one would yield a length of $\sqrt{2}$. The kernel density estimates for the healthy and the PDAC group are presented in Figure 9. We observe a bimodality for the PDAC-group related density that extends to the right and to the left of the estimated density of the controls. Namely, PDAC patients do exhibit either very high or very low measurements of this autoantibody compared to the healthy group. The reason is that the increase of tumor antigen levels in the plasma is associated with immune complex formation, leading to the depletion of free autoantibodies against the antigen ([6]).

We proceed by further evaluating this autoantibody in terms of sensitivity and specificity. As there are no available clinical insights regarding the underlying cutoffs, we estimate them by the data at hand. In this data set, as the diseased group exhibits bimodality, we do not proceed with the parametric-based approaches, but with our kernel-based approaches throughout. The kernel-based estimated cutoffs are derived to be (−2.1336, 1.8944) and are visualized in Figure 9. The underlying Youden-based sensitivity and specificity, along with their marginal 95% confidence intervals, are 0.8429(0.7946, 0.8831) and 0.9502(0.8498, 0.9880). The corresponding egg-shaped and rectangular 95% confidence regions are given in the first panel of Figure 9, in which we observe that the area of the egg-shaped region is 0.0151, and smaller than the rectangular one that yields an area of 0.0164. The surfaces of the sensitivity and specificity within the ROC space, which now is the unit cube, are given in Figure 9 (middle panel). A contour plot of those are given in the bottom two panels of Figure 9. By those figures, it is evident that the specificity achieved based on the Youden-based cutoffs is impressively large for any pair of cutoffs for which $c_1 < -2.1336$, $c_2 > 1.8944$. In the contour plots of Figure 9, we also visualize by a dot the corresponding pair of the Youden-based optimal pair of cutoffs. Overall, our results show a very promising performance of this autoantibody, which would have been rejected had we relied on the traditional empirical-based AUC.

6.2. Esophageal squamous cell carcinoma

Esophageal squamous cell carcinoma (ESCC) is an aggressive form of cancer with poor prognosis [37]. Biomarker discovery for its early detection is important since esophageal cancer is usually detected only after an advanced stage. Su et al. (2011) ([37]) focus on profiling global gene expression and their study revealed 159 genes that showed statistically significant difference between cases and controls based on fold-changes. Out of those 116 are derived to be up-regulated while 43 down-regulated. Their study involves 53 cases and 53 controls, and their data are publicly available from Affymetrix U133A expression arrays. These data are also discussed in [18] where it is also pointed out that non-monotone (or 'non-traditional') markers are also available in this data set.

For probe 209644 the empirically estimated AUC is 0.5421 (0.4175–0.6667). This indicates that probe 207039 would be discarded as uninformative. By employing the proposed approaches, we derive the length of the ROC that corresponds to this marker to be 1.6518 and statistically significant (95% CI is 1.5938–1.7403). The two cutoff Youden yields a pair of Spec and Sens equal to (0.8754, 0.6659). In addition, the area under the gROC for this marker is 0.8238. The sensitivity and specificity surfaces along with the confidence region around their Youden based optimized values are given in Figure 10.

Results are analogous for probe 207039. The empirical AUC is equal to 0.5180(0.3963–0.6397). Namely, this marker too was discarded as uninformative while in fact the length of its ROC is equal to 1.7442 and statistically significant with a 95% CI (1.6807–1.8101). The two cutoff analysis results to a pair of (Spec, Sens) equal to (0.9180, 0.7648). The area under the gROC of this marker is equal to 0.8858 which implies a potentially promising marker that would have been missed with traditional ROC methodology. The sensitivity and specificity surfaces along with the confidence region around their Youden based optimized values are given in Figure 11. The corresponding gROCs for both probes are shown in Figure 12.

7. Discussion

Nowadays, high throughput technologies allow for the simultaneous input of hundreds of markers. Hence, there is a crucial need of proper biomarker ranking. Traditional measures involve the AUC, as well as affordable levels of sensitivity and specificity that may be ill posed under a potential improper behavior of a biomarker, implying a severe violation of the underlying stochastic ordering assumption of the healthy and cancer patients.

In this paper we explore the properties of the length of the ROC curve both under a parametric and a non-parametric setting. We illustrate that the length of the ROC can be expressed as a ϕ–divergence and for parametric models we study its asymptotic distribution. We provide a link of the length of the ROC with the area of the optimal ROC. Namely, given the length, we provide the upper and lower bounds of the AUC of the optimal ROC which reveals the full potential of a marker. We focus on settings where the density ratio is unimodal and thus the gROC is optimal. For cutoff derivation we explore the two cutoff Youden index and illustrate that under this context the sensitivity and the specificity are surfaces that can be visualized in the unit cube to reveal the usual sens/spec tradeoff. We further provide joint confidence regions for the achieved sensitivity and specificity. We provide both parametric and non-parametric frameworks.

Our approaches can have a broad application in the biomarker field as they may be safely applied to any candidate list, ensuring that all four kinds of markers will be captured: 1) markers of the usual monotone ordering (i.e. the higher the marker score the greater the suspicion of the disease), 2) markers of the reverse ordering (the lower the marker score the greater the suspicion of the disease), 3) a non-monotone behavior which is commonplace in many mundane tests and may sometimes be present in cancer biomarkers. That is, very high and very low scores tend to correspond to diseased individuals while the healthy scores tend to lie in a zone in between. Finally, 4) more complex non-monotone behaviors than in case (3) that could imply more than two cutoffs (when the density ratio has two or more modes). We note that under a biomarker setting, it is hard to think of plausible situations where more than two cutoffs are necessary. Regarding case (3) above, we have illustrated three examples of such markers in our application section. The non-monotone behavior of HNRNPUL1 can be explained biologically. For example, we have previously observed that, with tumor development and progression, elevated levels of antigen in the plasma lead to the formation of circulating immune complexes, which deplete plasma of free autoantibodies against the antigen ([6]). This could explain the improper behavior of autoantibodies against the protein HNRNPUL1, implying that very high and very low measurements are indicative of cancer. We have observed that immune complexes for HNRNP family members can be identified in the plasma of cancer patients, and that their levels are, similarly, either elevated or decreased in patients, compared to matched controls. These results suggest that autoantibodies and immune complexes for the same antigen may exhibit an inverse relationship, leading to the improper behavior which is the focus of our paper. Based on previously discussed traditional biomarker ranking approaches, an improper behavior of the autoantibody HNRNPUL1, in terms or recognizing early PDAC stages, would naively lead to its rejection as uninformative. In this paper, we provide a complete framework to discover and then evaluate such promising biomarkers. Our approaches fill in an essential literature gap, which may have caused the rejection of very useful biomarkers of this nature.

There are several points for future research. One practical topic that needs to be further investigated is the efficiency of different versions of bandwidth especially since this setting may involve bimodal densities. Further comparisons based on other measures can also be of interest. One such measure is the overlap coefficient (OVL) discussed in [38], [39], [40], and [41]. Graphically, the OVL is the area between the densities of the non-diseased and the diseased group when those are plotted on the same axes and is related to Kullback-Leibler divergence ([42]). An interesting measure that can be used in conjunction with our approaches is the TNRC (see [9]) to classify which of the detected markers are improper. As discussed above, it is important to note that our approach may reveal markers for which the use of more than two cutoffs is suggested. A smooth graph of the density ratio could be used for revealing such markers. A graph suggesting a monotone density ratio indicates a proper ROC curve whereas a graph suggesting a unimodal density ratio indicates the need of two cutoffs. Other type of configurations may indicate the necessity of more than two cutoffs which, as stated above, is hard to justify when dealing with biomarkers.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The research of Dr. Leonidas E. Bantis is supported in part by the COBRE grant (NIH) P20GM130423 and by the NIH Clinical and Translational Science Award (UL1TR002366) to the University of Kansas. The authors would like to thank two anonymous referees and the Associate Editor for their valuable comments that improved the presentation of this paper.

References

1. Pepe MS (2003). The Statistical Evaluation of Medical Diagnostic Tests for Classification and Prediction. Oxford: Oxford University Press.
2. Mann HB, Whitney DR (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. Annals of Mathematical Statistics. 18(1): 50–60.
3. Siegel RL, Miller KD, Jemal A. (2017) Cancer Statistics. CA: A Cancer Journal for Clinicians 67(1): 7–30. [PubMed: 28055103]
4. Partensky C. (2013). Toward a better understanding of pancreatic ductal adenocarcinoma: glimmers of hope? (2013) Pancreas 42(5): 729–739. [PubMed: 23648843]
5. Bamber D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. Journal of Mathematical Psychology 12(4): 387–415.
6. Ladd JJ, Chao T, Johnson MM, et al. (2013). Autoantibody signatures involving glycolysis and splicesome proteins precede a diagnosis of breast cancer among postmenopausal women. Cancer research 73(5): 1502–1513. [PubMed: 23269276]
7. Sin DD, Tammemagi CM, Lam S, Barnett MJ, Duan X, Tam A, Auman H, Feng Z, Goodman GE, Hanash S, Taguchi A (2013). Pro–Surfactant Protein B As a Biomarker for Lung Cancer Prediction. Journal of Clinical Oncology 31(36):4536–4543. [PubMed: 24248694]
8. Hilden J. (1991). The area under the ROC curve and its competitors. Medical Decision Making 11: 95–101. [PubMed: 1865785]
9. Parodi S, Pistoia V, and Muselli M. (2008). Not proper ROC curves as a new tool for the analysis of differentially expressed genes in microarray experiments. BMC Bioinformatics 9(1): 410. [PubMed: 18834513]
10. Engmann S, Cousineau D. (2011). Comparing Distributions: The two sample Anderson-Darling test as an alternative to the Kolmogorov-Smirnov test. Journal of the Applied Quantitative Methods 6(3).
11. Nakas CT, Yiannoutsos CT, Bosch RJ, Moyssiadis C. (2003). Assessment of diagnostic markers by goodness-of-fit tests. Statistics in Medicine 22(15): 2503–2513. [PubMed: 12872305]
12. Lee WC, Hsiao CK (1996). Alternative summary indices for the receiver operating characteritic curve. Epidemiology : 605–611. [PubMed: 8899386]
13. Franco-Pereira AM, Nakas CT, Pardo CM (2020). Biomarker assessment in ROC curve analysis using the length of the curve as an index of diagnostic accuracy: the binormal model framework Advances in Statistical Analysis 10.1007/s10182-020-00371-8: 1–23.
14. Martínez-Camblor P, Corral N, Rey C, Pascual J, Cernuda-Morollon E (2017). Receiver operating characteristic curve generalization for non-monotone relationships Statistical methods in medical research 26(1): 113–123. [PubMed: 24986857]
15. Martínez-Camblor P, Pardo-Fernandez JC (2019). Parametric estimates for the receiver operating characteristic curve generalization for non-monotone relationships Statistical Methods in Medical Research 28(7): 2032–2048. [PubMed: 29243554]
16. Martínez-Camblor P, Pardo-Fernandez JC (2019). The Youden index in the generalized receiver operating characteristic curve context The international journal of biostatistics 15(1)
17. Martínez-Camblor P, Peres-Fernandez S., Diaz-Coto S. (2018). Improving the biomarker diagnotic capacity via functional transformations. Journal of Applied Statistics 46(9): 1550–1556.
18. Dawson P. (2012). Beyond traditional biomarkers: Methods for identifying and evaluating non-traditional biomarkers. Ph.D. thesis.
19. Österreicher F. (2013). Distances Based on the Perimeter of a Risk Set of a Test Problem Austrian Journal of Statistics 42: Number 1, 3–19.
20. Alekseev VM, (1987). Optimal Control. Springer US.
21. Salicru M, Morales D., Menendez ML, Pardo L. (1994) On the applications of divergence type measures in testing statistical hypotheses. Journal of Multivariate Statistics 51: 372–391.
22. Whitt W. (1985). Uniform Conditional Variability Ordering of Probability Distributions. Journal of Applied Probability 22(3): 619–633.

23. Carter M. (2001). Foundations of Mathematical Economics. Cambridge: MIT Press ISBN 0–262-53192–5.

24. McIntosh MW, Pepe MS. (2002). Combining Several Screening Tests: Optimality of the Risk Score Biometrics 58: 657–664 [PubMed: 12230001]

25. Bantis LE, Nakas CT, Reiser B. (2014). Construction of confidence regions in the ROC space after the estimation of the optimal Youden index-based cut-off point. Biometrics 70: 212–223. [PubMed: 24261514]

26. Youden WJ. (1950). An index for rating diagnostic tests. Cancer 3: 32–35. [PubMed: 15405679]

27. Box GEP, Cox DR (1964). An analysis of transformations. Journal of the Royal Statistical Society, Series B 26: 211–252.

28. Silverman BW (1998). Density Estimation for Statistics and Data Analysis. London: Chapman & Hall/CRC.

29. Wand MP, Jones MC (1995). Kernel Smoothing. Boca Raton.

30. Nadaraya EA (1964). Some New Estimates of Distribution Functions, Teor. Veroyatnost. i Primenen. 9(3) : 550–554; Theory Probab. Appl. 9(3) : 497–500

31. Wang J, Cheng F, Yang L. (2013). Smooth simultaneous confidence bands for cumulative distribution functions. Journal of Nonparametric Statistics 25: 395–407.

32. Bantis LE, Nakas CT, Reiser B, Daniel M, Dalrymple-Alford JC. (2017). Construction of joint confidence regions for the optimal true class fractions of Receiver Operating Characteristic (ROC) surfaces and manifolds. SMMR 26(3): 1429–1442.

33. Yin J and Tian L. (2014). Joint inference about sensitivity and specificity at the optimal cut-off point associated with Youden index. Comput Stat Dat Anal 77: 1–13.

34. Botev ZI, Grotowski JF, Kroese DP (2010). Kernel density estimation via diffusion. Annals of Statistics 38(5).

35. Qiu J, Choi G, Li L, et al. (2008). Occurrence of autoantibodies to annexin I, 14–3-3 theta and LAMR1 in prediagnostic lung cancer sera. Journal of Clinical Oncology: official journal of the American Society of Clinical Oncology 26(31): 5060–5066. [PubMed: 18794547]

36. Katayama H, Boldt C, Ladd JJ, et al. (2015). An Autoimmune Response Signature Associated with the Development of Triple-Negative Breast Cancer Reflects Disease Pathogenesis. Cancer Research 75(16): 3246–3254. [PubMed: 26088128]

37. Su H, Hu N, Yang HH, Wang C, Takikita M, Wang QH, Giffen C, Clifford R, Hewitt SM, Shou JZ. et al. (2011). Global gene expression proffling and validation in esophageal squamous cell carcinoma and its association with clinical phenotypes Clinical Cancer Research 17(9): 2955. [PubMed: 21385931]

38. Silva-Fortes C, Amaral Turkman MA, Sousa L. (2012). Arrow plot: a new graphical tool for selecting up and down regulated genes and genes differentially expressed on sample subgroups BMC Bioinformatics 13(147).

39. Bradley EL. (1985). Overlapping Coefficient. Encyclopedia of Statistical Sciences, New York: Chapman and Hall 6: 546–547.

40. Inman HF, Bradley EL (1989). The overlapping coefficient as a measure of agreement between two probability distributions and point estimation of the overlap of two normal densities Commun Statist Theory Methods 18(10): 3851–3872.

41. Samawi HM, Yin J., Rochani H, Panchal V. (2017). Notes on the overlap measure as an alternative to the Youden index: How are they related? Statistics in Medicine 36(26): 4230–4240. [PubMed: 28809042]

42. Dhaker H, Ngom P, Ibrahimouf B, Mbodj M. (2019). Comparing Distributions: The two sample Anderson-Darling test as an alternative to the Kolmogorov-Smirnov test. Overlap Coefficients Based on Kullback-Leibler of Two Normal Densities: Equal Means Case doi:10.5539/jmr.v11n2p114.

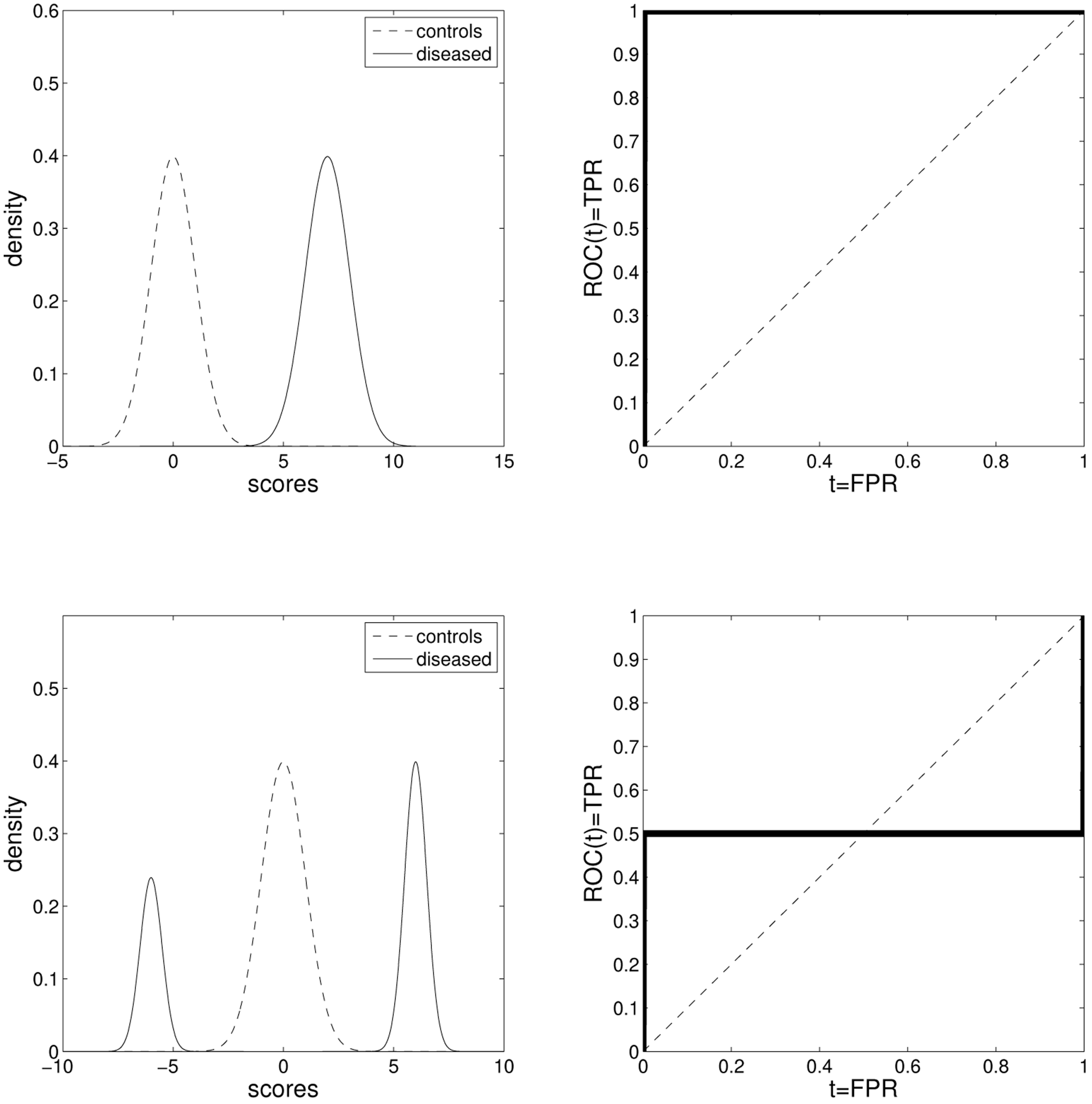


Figure 1.

Upper panels: An example of a perfect biomarker where the densities of the controls and the diseased (left) comply with a traditional framework, under which the stochastic ordering $X < Y$ holds, and the corresponding ROC curve is in the right panel. Corresponding metrics: $AUC = 1$, $pAUC(t_1, t_2; t_2 > t_1) = t_2 - t_1$, area between the ROC and the reference diagonal= 0.5, proposed length= 2. Lower panels: An example of a perfect biomarker in which the stochastic ordering $X < Y$ cannot be assumed. Corresponding metrics: $AUC = 0.5$, $pAUC(t_1, t_2; t_2 > t_1) = 0.5(t_2 - t_1)$, area between the ROC and the reference diagonal= 0.25, proposed length = 2 (the proposed length is immune to directionality and characterizes the underlying biomarker as perfect in both cases, as opposed to all other measures.)

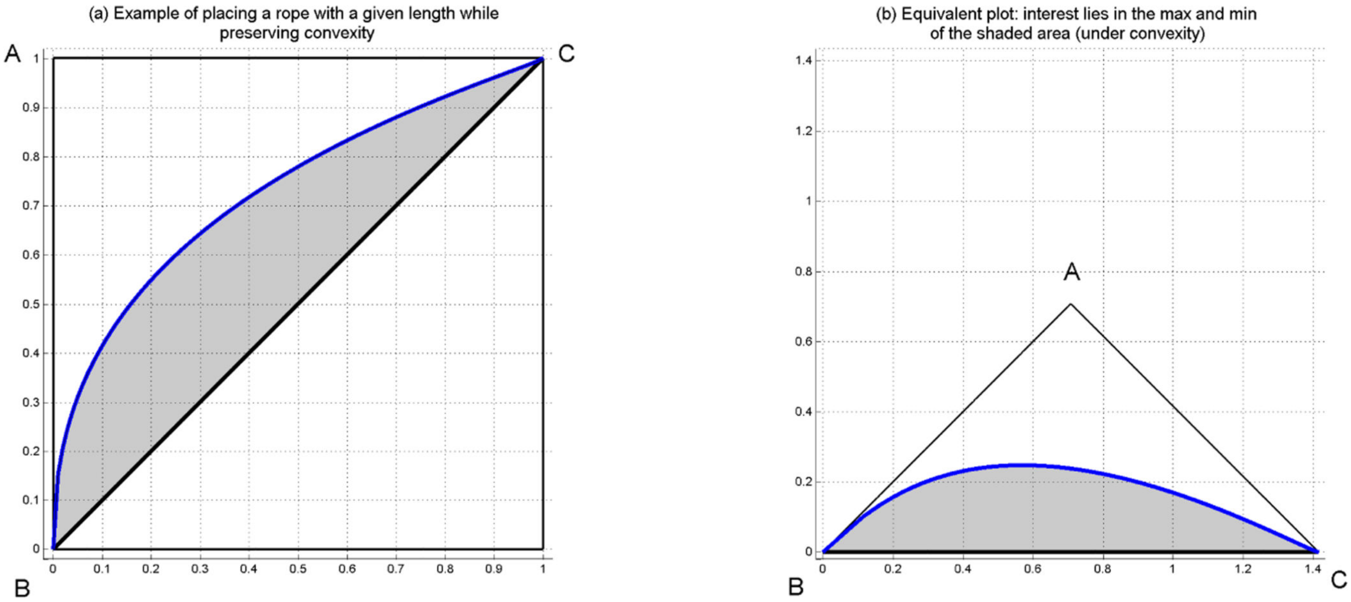


Figure 2.

Geometric representation of the problem of finding the max and min area under the curve, when assuming convexity for a given length of the rope (ROC). Under convexity, it is enough to study the area between the reference diagonal and the rope. Hence, we only can focus on the right panel (b), which is equivalent to the left panel (a).

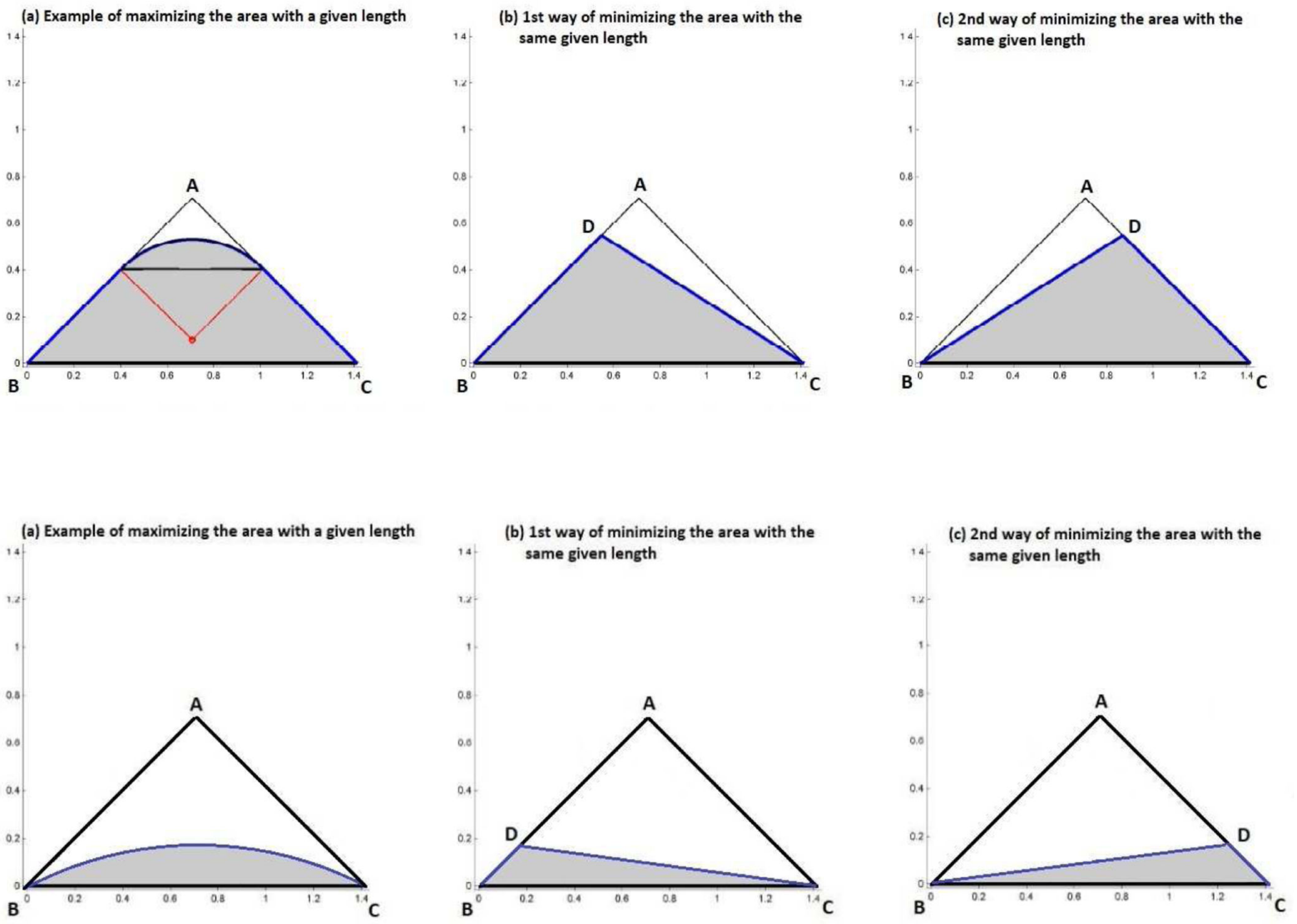

(a) Example of maximizing the area with a given length
(b) 1st way of minimizing the area with the same given length
(c) 2nd way of minimizing the area with the same given length
A
B
C
D
0
0.2
0.4
0.6
0.8
1
1.2
1.4
(a) Example of maximizing the area with a given length
(b) 1st way of minimizing the area with the same given length
(c) 2nd way of minimizing the area with the same given length
A
B
C
D

Figure 3. Upper panels:

Geometric representation of the maximum area for a given length when the length is larger than $\frac{\pi}{2}$ (panel (a)), and the minimum area for the same given length (panels (b) and (c)). Both panels (b) and (c) correspond to the same area. All three panels correspond to the same length. **Lower panels:** Geometric representation of the maximum area for a given length when the length is smaller than $\frac{\pi}{2}$ (panel (a)), and the minimum area for the same given length (panels (b) and (c)). Both panels (b) and (c) correspond to the same area. All three panels correspond to the same length.

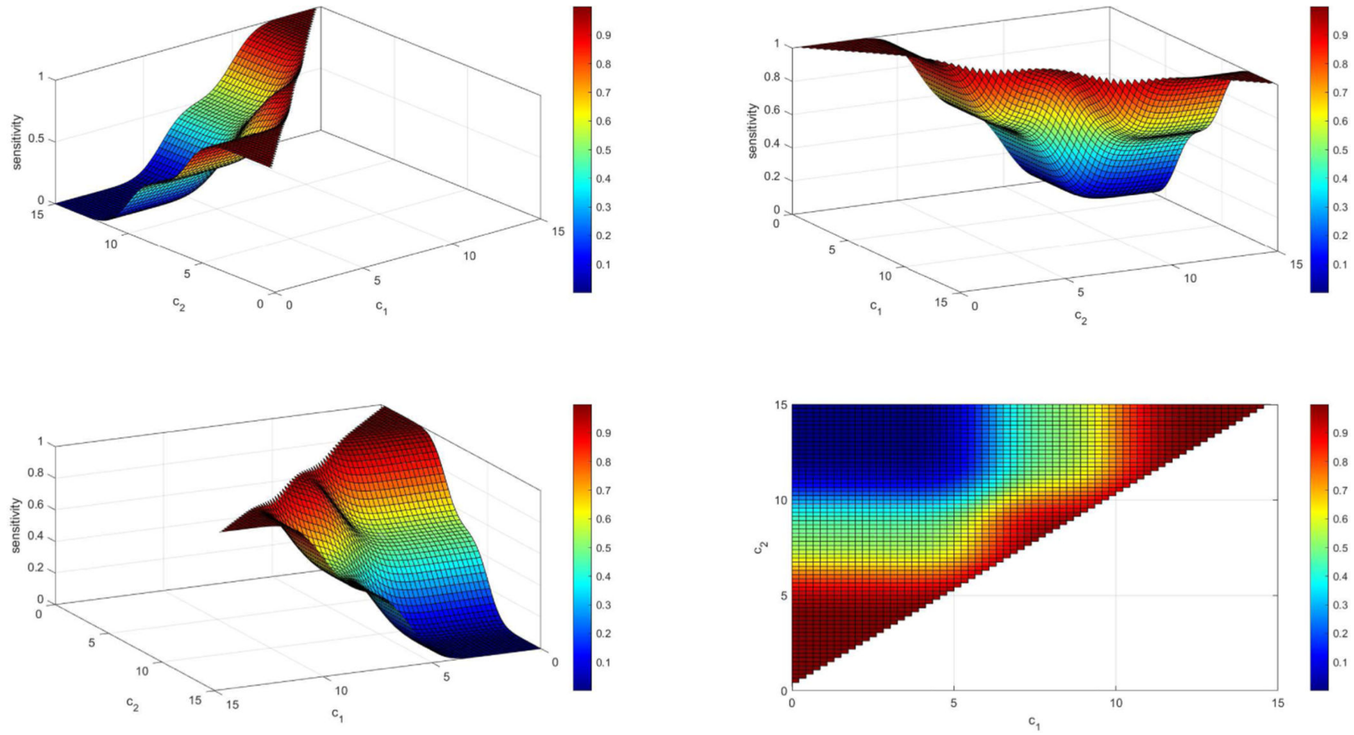


Figure 4.

Example of a sensitivity surface when the density of the controls is $N(8,1)$ and the density of the cases is a two component normal mixture: $0.5N(6, 1) + 0.5N(10, 1)$. The sensitivity plot is given under four different angles (panels above) for better visualization. We observe that when c_1 is very small and c_2 is very large then the sensitivity yields very low values since the mass of both densities is low below c_1 and beyond c_2. In addition, if c_1 is very close to c_2 then we expect that almost all individuals that are diseased will be categorized as such since it is very likely that they will not be between c_1 and c_2 and thus the sensitivity for those cutoffs is very high (red regions).

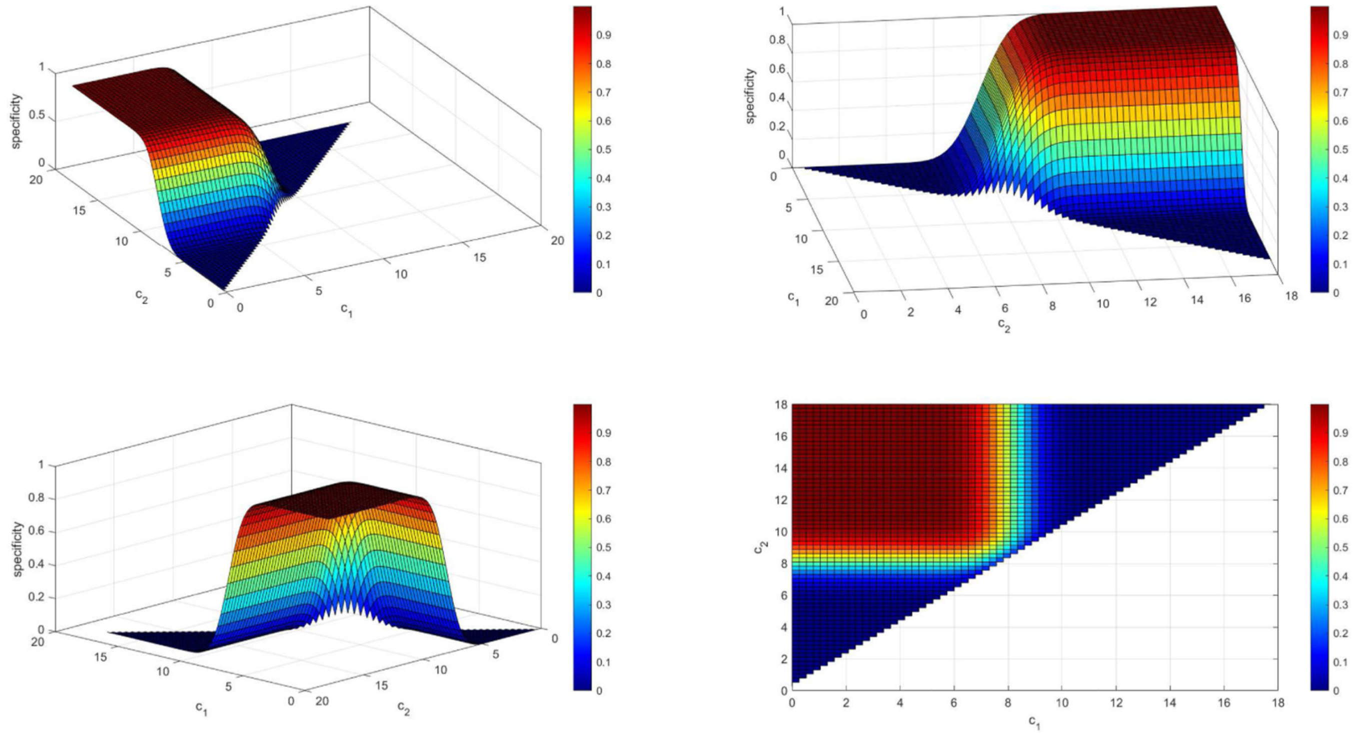


Figure 5.

Example of a specificity surface when the density of the controls is $N(8, 1)$ and the density of the cases is a two component normal mixture: $0.5N(6, 1) + 0.5N(10, 1)$. The specificity plot is given under four different angles (panels above) for better visualization. We observe that when c_1 is very small and c_2 is very large then the specificity yields very high values as expected since most healthy individuals will lie within (c_1, c_2). In addition, if c_1 is very close to c_2 then we expect that almost all individuals that are healthy will be categorized as diseased and since it is very likely that they will not be between c_1 and c_2 and thus the specificity for those cutoffs is very low (blue regions).

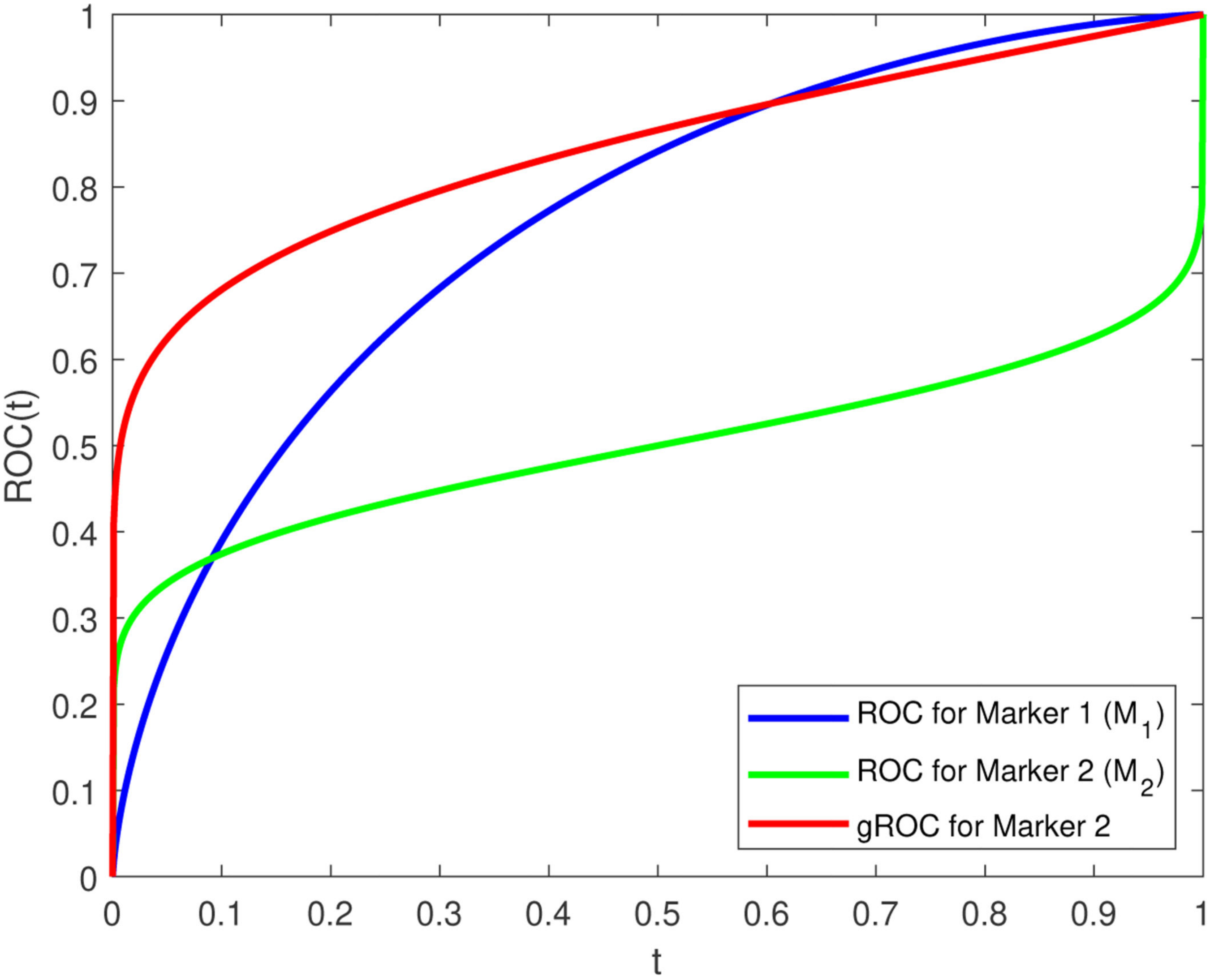


Figure 6.

ROC curves for the hypothetical example with markers M_1 and M_2. Both ROC curves are illustrated and would tempt us to consider M_1 as an overall better biomarker. The gROC of marker 2 is presented as well showing its potential in actually outperforming M_1. The gROC of M_2 and the ROC of M_2 have exactly the same length that is larger than the length of M_1.

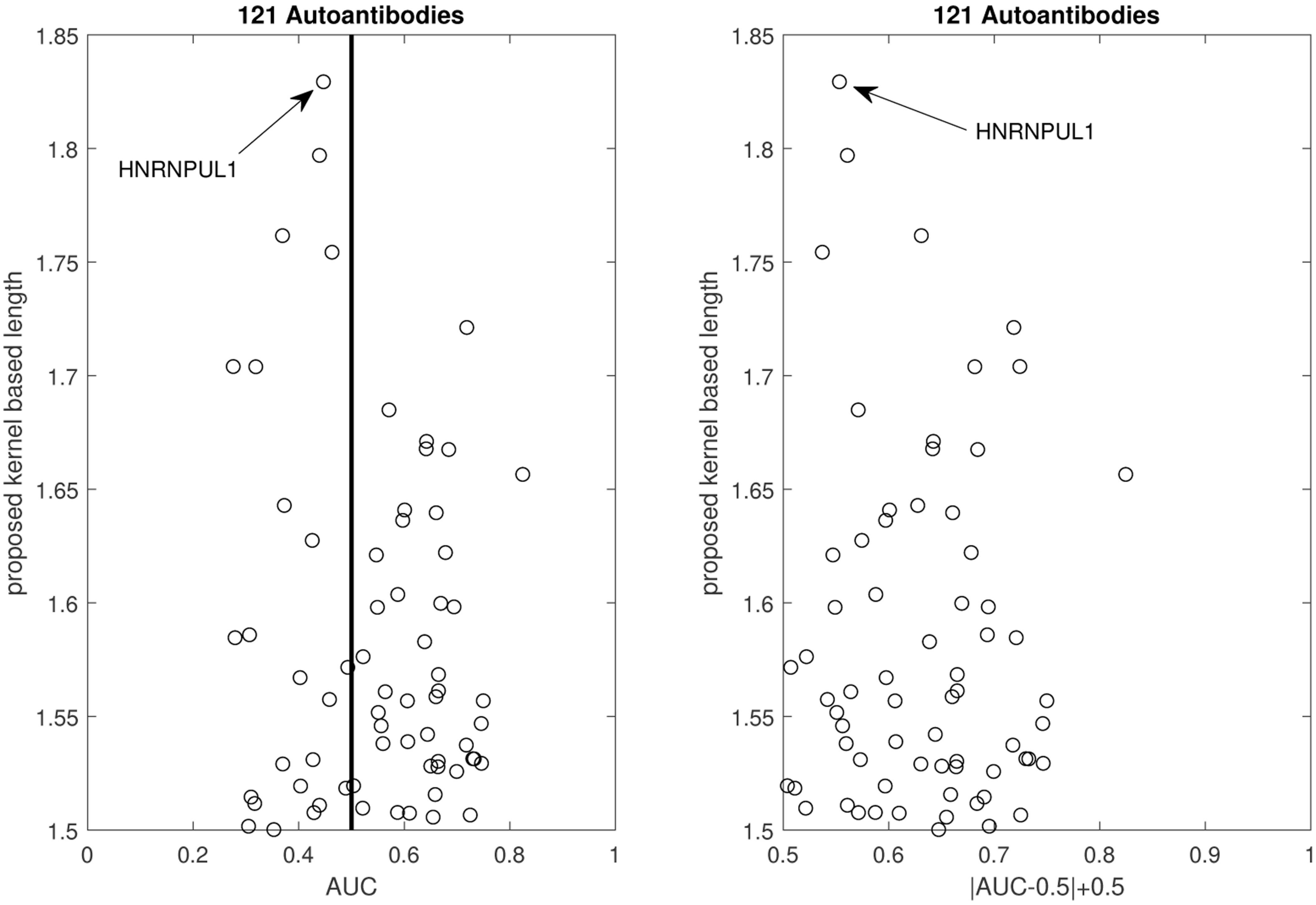


Figure 7.

Scatterplot of the AUCs and proposed lengths of all 121 autoantibodies.

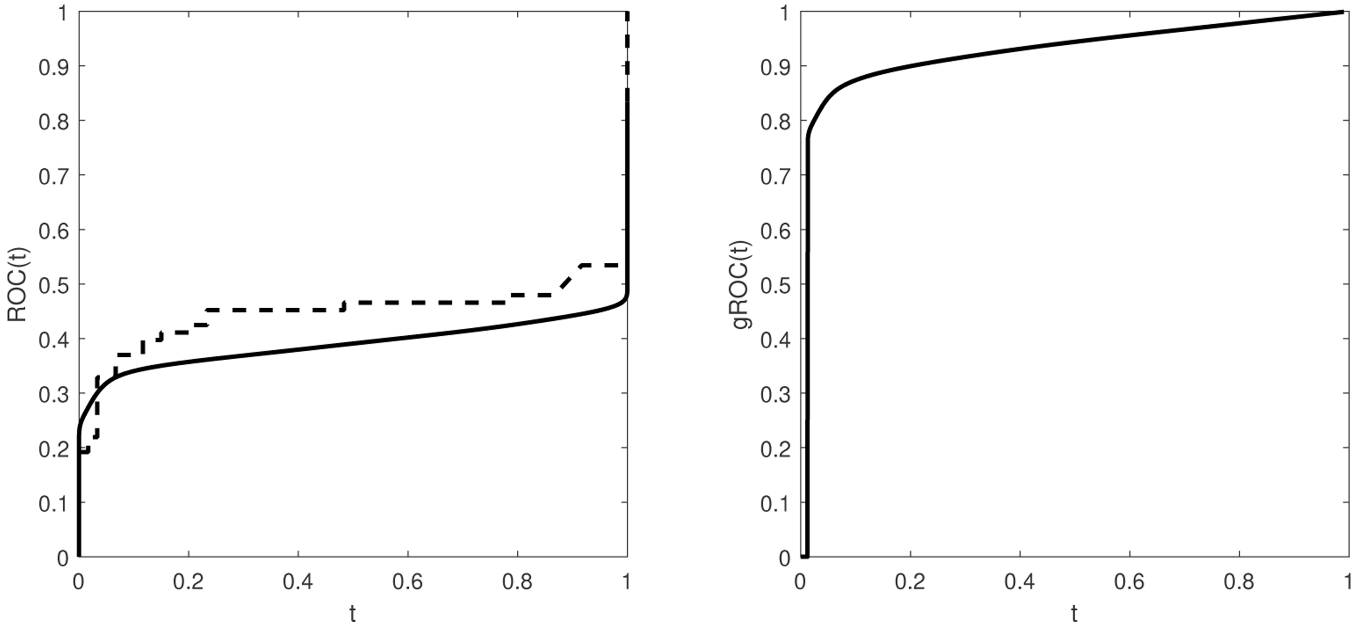


Figure 8.

Left Panel: Traditional ROC curves for the top performing autoantibody: dashed line refers to the empirical, and solid line refers to the kernel-smoothed based one. The corresponding AUCs and 95% CIs are 0.4468(0.3392, 0.5544) and 0.3892(0.3055,0.4730), respectively. The estimated kernel-based length with the corresponding 95% CI is 1.8255 (1.7846–1.8748). Right panel: The corresponding kernel-based estimated gROC. The area under this kernel based estimate of gROC is 0.9154.

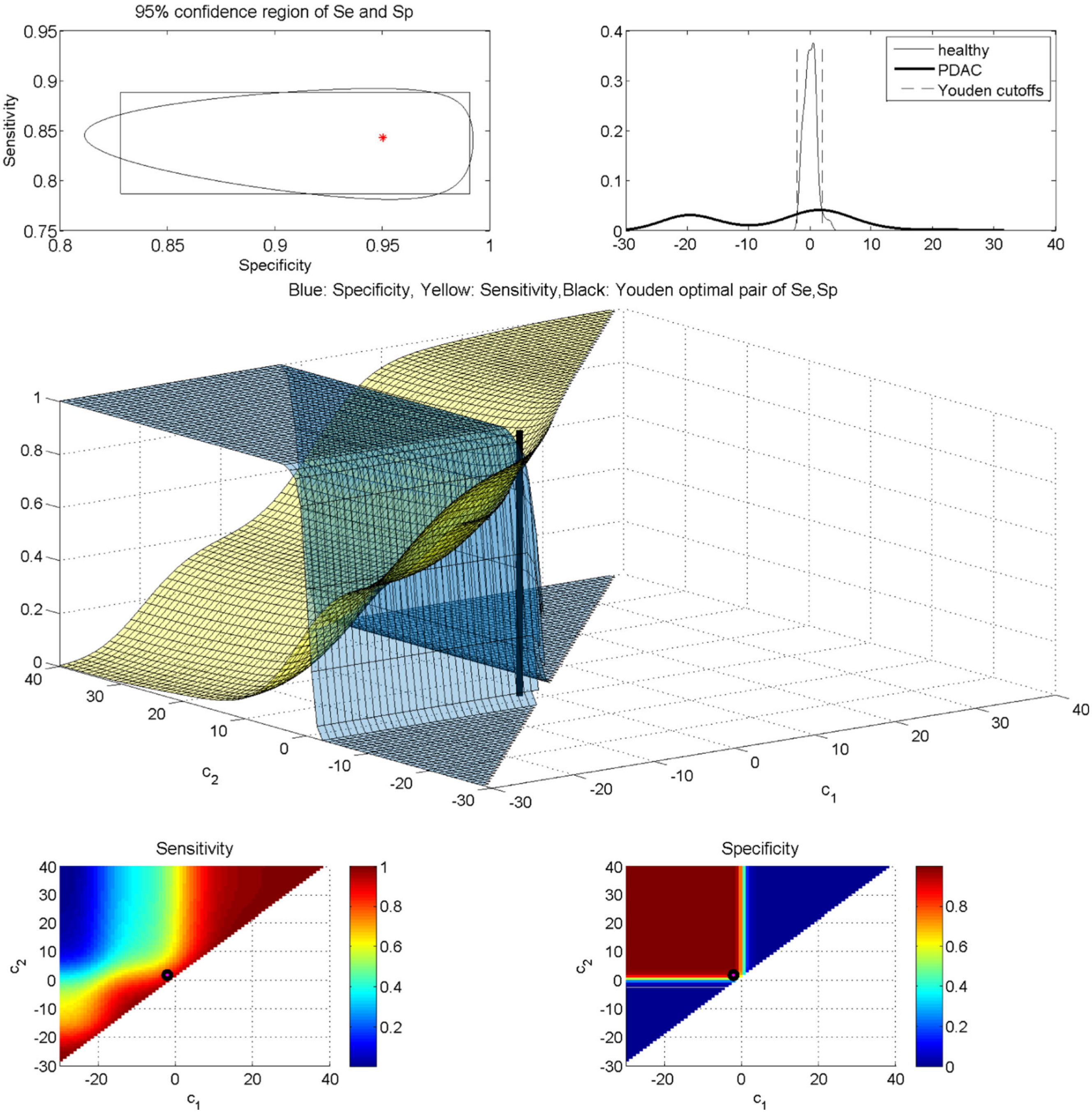


Figure 9.

Top left: Egg shaped 95% confidence region for the sensitivity and specificity at the Youden-based estimated cutoffs. Top right: Underlying kernel densities of the healthy and the diseased. Middle: Surfaces of sensitivity and specificity along with the estimated Youden-based optimal pair of cutoffs. Bottom left: Contour plot of the surface of sensitivity for all possible cutoffs, along with the estimated Youden-based optimal pair of cutoffs (black dot). Bottom right: Contour plot of the surface of specificity for all possible cutoffs, along with the estimated Youden-based optimal pair of cutoffs (black dot).

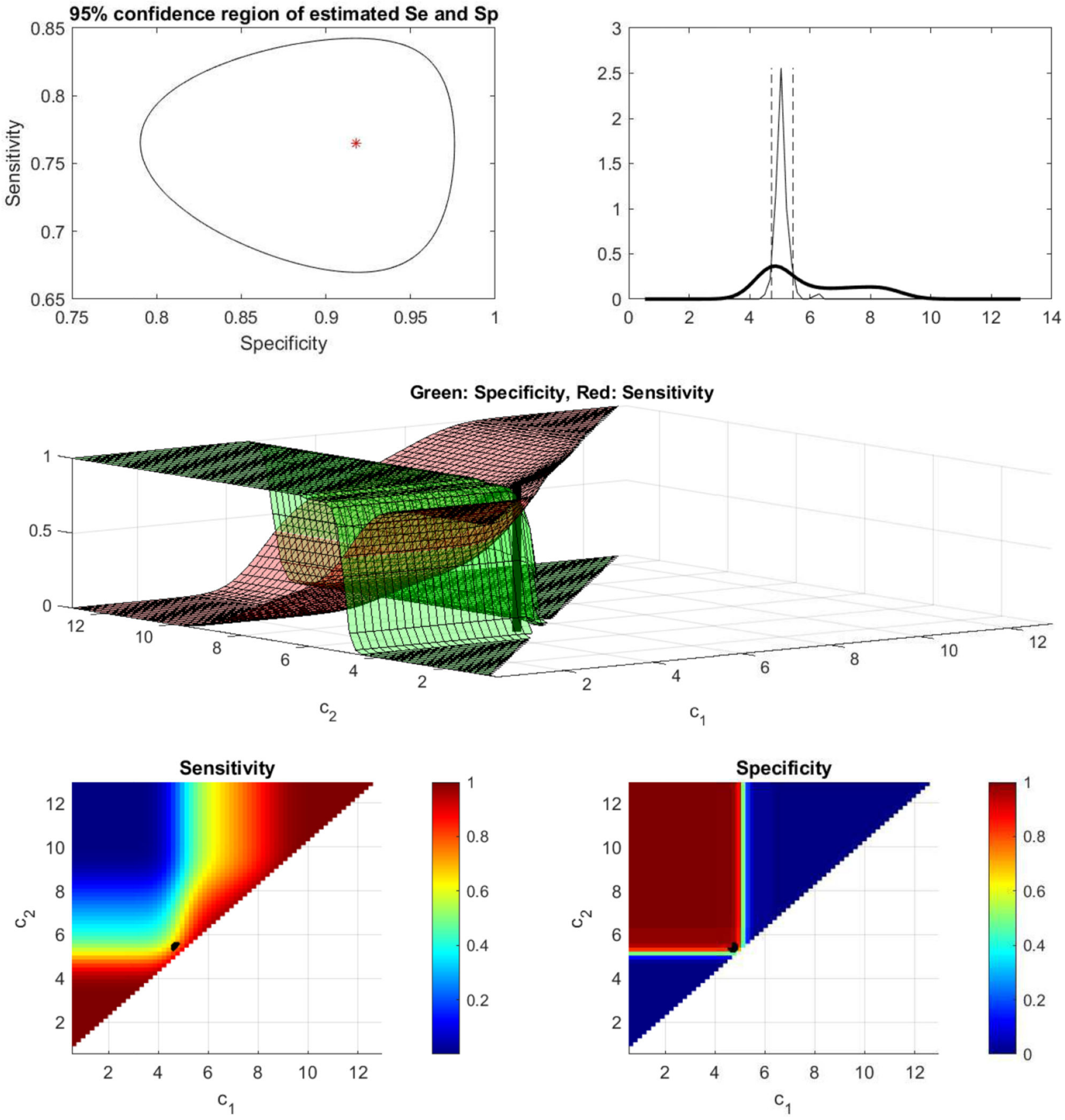


Figure 10.

ESCC data (probe 207039).Top left: Egg shaped 95% confidence region for the sensitivity and specificity at the Youden-based estimated cutoffs. Top right: Underlying kernel densities of the healthy and the diseased. Middle: Surfaces of sensitivity and specificity along with the estimated Youden-based optimal pair of cutoffs. Bottom left: Contour plot of the surface of sensitivity for all possible cutoffs, along with the estimated Youden-based optimal pair of cutoffs (black dot). Bottom right: Contour plot of the surface of specificity for all possible cutoffs, along with the estimated Youden-based optimal pair of cutoffs (black dot).

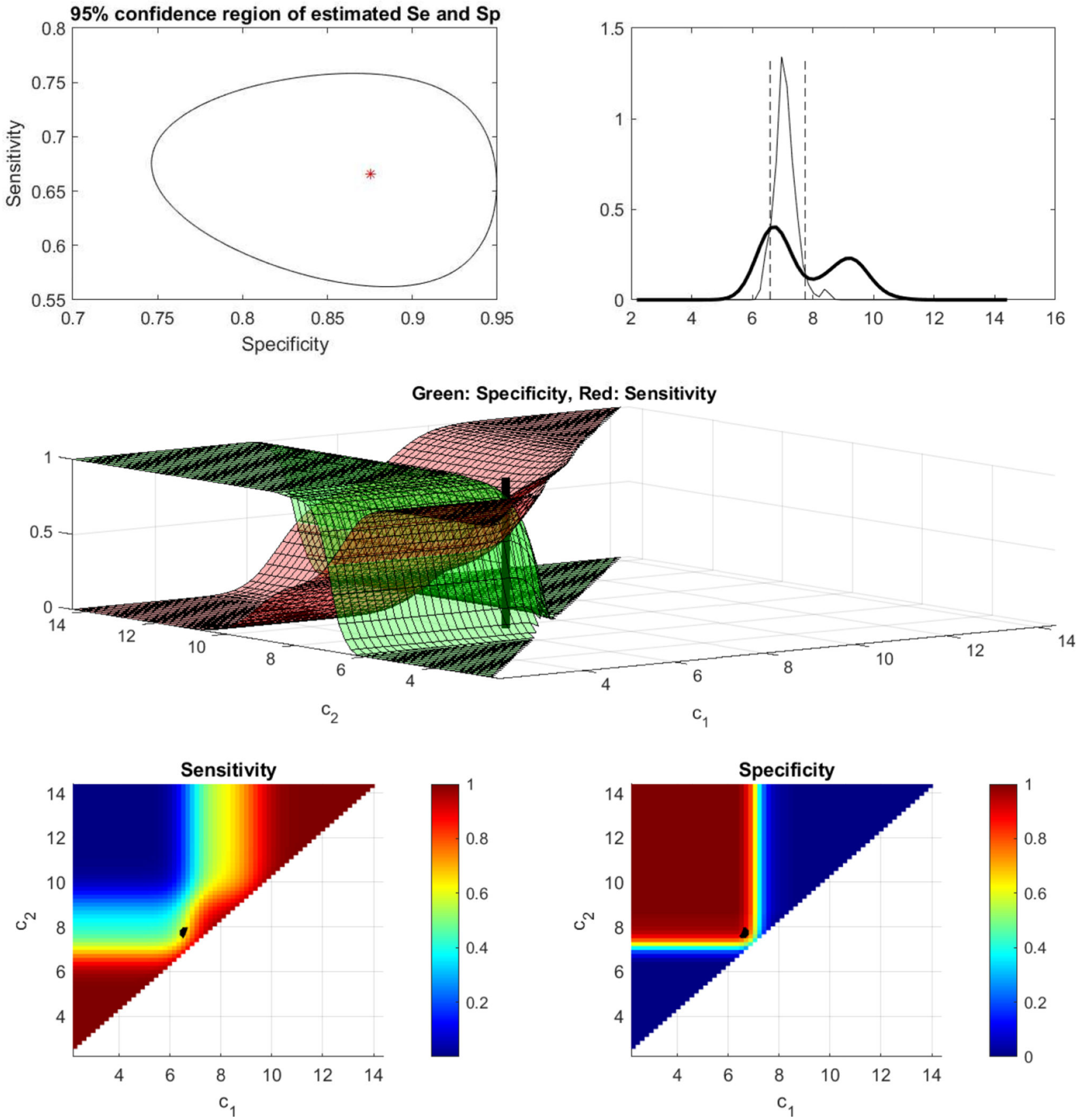
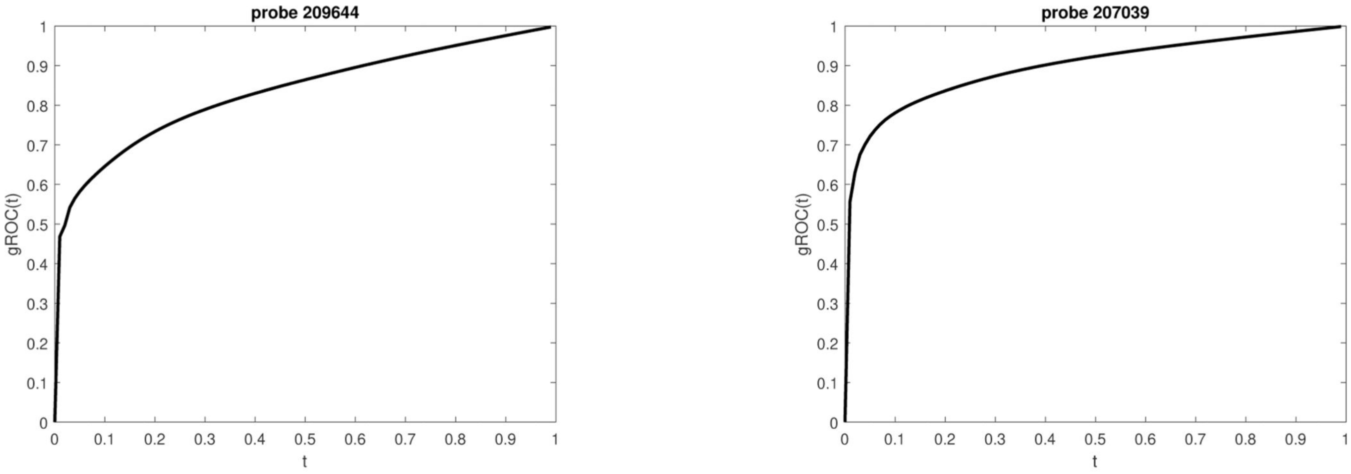


Figure 11.

ESCC data (probe 209644). Top left: Egg shaped 95% confidence region for the sensitivity and specificity at the Youden-based estimated cutoffs. Top right: Underlying kernel densities of the healthy and the diseased. Middle: Surfaces of sensitivity and specificity along with the estimated Youden-based optimal pair of cutoffs. Bottom left: Contour plot of the surface of sensitivity for all possible cutoffs, along with the estimated Youden-based optimal pair of cutoffs (black dot). Bottom right: Contour plot of the surface of specificity for all possible cutoffs, along with the estimated Youden-based optimal pair of cutoffs (black dot).



(a) Probe 209644

(b) Probe 207039

Figure 12.

gROCs for both probes discussed in the application that refers to the esophageal data. **Left panel:** gROC for probe 209644 with an area under it equal to 0.8238. The two cutoff Youden-based sensitivity and specificity are derived to be (0.8754, 0.6659). The area under the usual empirical ROC estimate is 0.5421 (0.4175–0.6667) which implies that this marker would have been discarded as uninformative in spite of its discriminatory ability. **Right panel:** gROC for probe 207039 with an area under it equal to 0.8858. The two cutoff Youden-based sensitivity and specificity are derived to be (0.9180, 0.7648). The area under the usual empirical ROC estimate is 0.5185 (0.3963–0.6397) which implies that this marker would have been discarded as uninformative in spite of its discriminatory ability.

Table 1.

Attained lower and upper bounds of the AUC for some given values of the length.

length	lower bound of AUC	upper bound of the AUC
$\sqrt{2}$	0.50000	0.50000
1.65	0.77788	0.85729
1.75	0.85417	0.92719
1.85	0.91838	0.97379
1.95	0.97434	0.99709
2.00	1.00000	1.00000