Journal of Data Science 0 (0), 1–18
??? 2026
DOI: 10.6339/26-JDS1218
Statistical Data Science

VAROC: Value Added Receiver Operating Characteristics Curve

Danielle Brister[1] and Yunro Chung[2,3,*]

[1]*College of Liberal Arts and Sciences, Arizona State University, U.S.A.*
[2]*College of Health Solutions, Arizona State University, U.S.A.*
[3]*Virginia G. Piper Center for Personalized Diagnostics, Biodesign Institute, Arizona State University, U.S.A.*

Abstract

The receiver operating characteristics (ROC) curve has been widely used to evaluate the discrimination performance of biomarkers, but it has been criticized for overlooking their underlying distributions. In this paper, we propose a continuous version of the ROC curve that can assess not only the discrimination performance of biomarkers but also their continuity performance. Our method summarizes the biomarker values as conditional tail expectations at varying thresholds and compare them with true positive and false positive rates. The proposed method is particularly useful for an early phase of biomarker study that enrolls heterogeneous disease populations. Analysis of data from an ovarian cancer biomarker study illustrates the practical utility of the proposed method over the standard ROC curve analysis. The proposed methods are implemented in the R package *varoc*.

Keywords *area under the ROC curve; biomarker; classification; outlier sum statistics; truncated mean analysis*

1 Introduction

Early detection of disease is essential for alleviating disease burden, increasing treatment success rate and decreasing mortality rate. Over the past decade, many candidate biomarkers have been sought to improve diagnostic accuracy using molecular biology or image analysis techniques. These candidate biomarkers are evaluated by multiple studies, and those that perform consistently well across all studies hold potential for clinical application (Pepe et al., 2001, 2008b). For these clinical applications, a continuous biomarker, denoted as X, is converted to a binary biomarker for a threshold θ, where $X > \theta$ (or $X \leqslant \theta$), called a positive (or negative) biomarker test result, is to rule in (or rule out) a disease of interest. It is thus crucial to choose the optimal threshold to accurately classify individuals with and without the disease. The receiver operating characteristics (ROC) curve analysis has been widely accepted for this problem, where it visualizes X's discrimination performance over all possible threshold values and determines the optimal threshold by considering a trade-off between true positive fraction (TPF or sensitivity) and false positive fraction (FPF or one minus specificity) or maximizing Youden's J statistics (Youden, 1950).

Despite the widespread use of the ROC curve analysis, a number of authors have pointed out its limitations and proposed alternatives to better evaluate X. Parmigiani (2019) proposed

*Corresponding author. Email: yunro.chung@asu.edu.

Received January 6, 2025; Accepted January 15, 2026

fuzzy ROC curve to address situations where an intermediate or gray zone leads to inconclusive biomarker test results. Lorenz curve, which was originally developed to represent inequality in wealth distribution (Lorenz, 1905), was extended for evaluating X (Lee, 1999; Lee and Hsiao, 1996). Decision curve analysis was proposed to consider costs and benefits associated with X (Vickers and Elkin, 2006). Bangdiwala et al. (2008) developed the agreement chart to include not only TPF and FPF but also predictive values. Predictive plot was suggested for evaluating X while taking disease prevalence into account (Huang et al., 2007; Pepe et al., 2008a). These methods can capture another aspect of X that the ROC curve does not and can therefore be useful for biomarker studies. However, all of them, including the ROC curve, focuses on visualizing and quantifying the discrimination performance of X, which results in information loss by transforming the continuous biomarker to a binary one, i.e., X to $X > \theta$. When the value of X itself is important, the standard techniques, such as the histogram or two-sample t-test, can be used to visualize or compare the continuity performance of X between groups with and without disease, but it makes no consideration about θ. As the threshold θ plays a key role in evaluating X for biomarker studies, a special technique that incorporates it becomes imperative.

To address this issue, we suggest the value-added ROC (VAROC) curve as an alternative of the ROC curve. Specifically, we introduce new evaluation measures to summarize X values exceeding a threshold θ, formulated as tail expectations. We then propose new visualization tools by incorporating the proposed measures into the ROC curve. Consequently, the VAROC curve assesses not only the discrimination performance of X but also its continuity performance, whereas the ROC curve only evaluates the former. Importantly, both the ROC and VAROC curve analyses share the same goal of evaluating the performance of X under the common assumption that higher values of X are associated with greater likelihood of diseases. However, they differ in their definitions of "high performance" — high discrimination probabilities in the ROC curve versus high tail expectations in the VAROC curve. Assessing biomarkers from both perspectives is useful particularly for an early stage of biomarker research, such as preclinical or biomarker discovery studies, further discussed in Subsection 3.4.

The truncated or tail mean analysis has been extensively studied in literature, primarily through two distinct approaches. The first aimed to improve normality and get more robust results by excluding outliers (Tukey and McLaughlin, 1963; Bickel, 1965; Huber, 1972; Stigler, 1973; Yuen and Dixon, 1973; Yuen, 1974). The second utilizes outliers for specific applications, such as detecting oncogenes that consistently exhibit high levels of gene expression in a subset of samples (Tibshirani and Hastie, 2007; Wu, 2007; Chen et al., 2010). While our approach is more closely aligned with the latter, it differs in important ways. The previous methods relied on a single, fixed threshold, e.g., the third quartile plus 1.5 times the interquartile range, to identify outlier samples. In contrast, our approach a continuum of thresholds and assesses the continuity performance of X, alongside the trade-offs between the corresponding TPFs and FPFs.

The remainder of the paper is organized as follows. In Section 2, we review the ROC curve analysis and discuss its limitations especially for heterogeneous diseases. In Section 3, we introduce our proposed method and demonstrate its advantage over the ROC curve analysis via simulated data. In Section 4, we conduct simulation studies. In Section 5, we analyze a real ovarian cancer biomarker study. We conclude with a discussion in Section 6.

2 Review: ROC Curve

Let's denote Y as a binary random variable, where $Y = 1$ and $Y = 0$ indicate case (or disease) and control (or non-disease), respectively, and X as a biomarker measured on a continuous scale. By

convention, we assume higher X is associated with a greater probability of $Y = 1$. The observed data consist of n independent and identically distributed replicates of (Y, X), denoted by (Y_i, X_i) for $i = 1, 2, \ldots, n$.

The TPF and FPF at a threshold or cutoff θ are defined as

$$\text{TPF}(\theta) = \Pr(X > \theta \mid Y = 1)$$

and

$$\text{FPF}(\theta) = \Pr(X > \theta \mid Y = 0),$$

respectively. The ROC curve plots $\text{TPF}(\theta)$ versus $\text{FPF}(\theta)$ for $\theta \in (-\infty, \infty)$. There are three popular ROC-based summary measures for evaluating the discrimination performance of X. The area under the ROC curve (AUC) is a global measure of X across all thresholds, which is defined as

$$\text{AUC} = \int_{-\infty}^{\infty} \text{TPF}(\theta) d\theta.$$

When it is of interest to evaluate the discrimination performance of X at a specific threshold θ, $\text{TPF}(\theta)$ or (normalized) partial AUC (PAUC)(θ) can be used, where

$$\text{PAUC}(\theta) = \frac{\int_c^{\infty} \text{TPF}(u) du}{\text{FPF}(\theta)}, \tag{2.1}$$

and $\text{FPF}(\theta)$, which is used in the denominator of the above equation, is the normalization factor. The threshold θ is commonly determined by considering an acceptable $\text{FPF}(\theta) \equiv t$, and θ is set to $\text{FPF}^{-1}(t)$ for a fixed $t \in [0, 1]$ (Pepe, 2003). Let θ denote an arbitrary threshold value, whereas θ_t specifically refers to the threshold corresponding to a FPF of t. The ROC curve is then equivalently defined as $\{(t, \text{TPF}(\theta_t)) | t \in (0, 1)\}$, where $t = \text{FPF}(\theta_t)$. Consequently, $\text{TPF}(\theta_t)$ is interpreted as the TPF corresponding to a FPF of t, and AUC (or $\text{PAUC}(\theta_t)$) is interpreted as an average of TPFs corresponding to a FPF between 0 and 1 (or 0 and t). Since the perfect (or useless) biomarker X is the one connecting points (0,0), (0,1) and (1,1) (or (0,0) and (1,1)) on the ROC curve, its AUC, $\text{TPF}(\theta_t)$, and $\text{PAUC}(\theta_t)$ is 1, 1, and 1 (or 0.5, t, and $t/2$), respectively.

For a given t, the empirical estimators of TPF and FPF are

$$\widehat{\text{TPF}}(\hat{\theta}_t) = \frac{\sum_{i \in \mathbb{Y}_1} I(X_i > \hat{\theta}_t)}{|\mathbb{Y}_1|},$$
$$\widehat{\text{FPF}}(\hat{\theta}_t) = \frac{\sum_{j \in \mathbb{Y}_0} I(X_j > \hat{\theta}_t)}{|\mathbb{Y}_0|},$$

where $\mathbb{Y}_1 = \{i \mid Y_i = 1, i = 1, 2, \ldots, n\}$ and $\mathbb{Y}_0 = \{j \mid Y_j = 0, j = 1, 2, \ldots, n\}$ are disjoint sets of case and control samples, respectively, and $|\mathbb{Y}|$ is a size of set $\mathbb{Y}$, and

$$\hat{\theta}_t = \underset{\theta \in \{-\infty, X_1, X_2, \ldots, X_n, \infty\}}{\arg\min} \{\widehat{\text{FPF}}(\theta) \leqslant t\}. \tag{2.2}$$

The empirical estimator of PAUC is

$$\widehat{\text{PAUC}}(\hat{\theta}_t) = \frac{\sum_{i \in \mathbb{Y}_1} \sum_{j \in \mathbb{Y}_0} \psi(X_i, X_j) I(X_j > \hat{\theta}_t)}{|\mathbb{Y}_1||\mathbb{Y}_0|} / t, \tag{2.3}$$

where $\psi(X_i, X_j) = I(X_i > X_j) + I(X_i = X_j)/2$. When $t = 1$, $\hat{\theta}_1$ is estimated as $-\infty$, and the empirical estimator of AUC is computed as $\widehat{\text{PAUC}}(\hat{\theta}_1)$ because

$$\widehat{\text{PAUC}}(\hat{\theta}_1) = \frac{\sum_{i\in\mathbb{Y}_1}\sum_{j\in\mathbb{Y}_0}\psi(X_i, X_j)}{|\mathbb{Y}_1||\mathbb{Y}_0|} \equiv \widehat{\text{AUC}}.$$

Alternatively, we can compute PAUC using a numerical integration technique. For given t and K, we define a sequence of FPFs as $t = t_0 > t_1 > \cdots > t_{K-1} > t_K = 0$ such that the interval between adjacent FPFs is uniform: $\triangle t_k = t_{k-1} - t_k = t/K$ for $k = 1, 2, \ldots, K$. For example, if $t = 0.3$ and $K = 6$, then $t_0 = 0.3$, $t_1 = 0.25, \ldots$, $t_5 = 0.05$, $t_6 = 0$ with $\triangle t_k = 0.05$ for each $k = 1, 2, \ldots, K$. Threshold values corresponding to these FPFs are estimated to $\hat{\theta}_{t_0} \leqslant \hat{\theta}_{t_1} \leqslant \cdots \leqslant \hat{\theta}_{t_{K-1}} \leqslant \hat{\theta}_{t_K}$ via (2.2). Since the empirical estimator of the ROC curve $\{(\hat{t}, \widehat{\text{TPF}}(\hat{\theta}_{\hat{t}}))|\theta \in \{-\infty, X_1, X_2, \ldots, X_n, \infty\}\}$ is equivalent to $\{(t_k, \widehat{\text{TPF}}(\hat{\theta}_{t_k}))|k = 1, 2, \ldots, K\}$ for large K, where $\hat{t} = \widehat{\text{FPF}}(\theta)$, (2.3) is alternatively computed as

$$\widehat{\text{PAUC}}(\hat{\theta}_t) = \left\{\sum_{k=1}^{K} \triangle t_k \widehat{\text{TPF}}(\hat{\theta}_{t_k})\right\}/t = \left\{\frac{t}{K}\sum_{k=1}^{K}\widehat{\text{TPF}}(\hat{\theta}_{t_k})\right\}/t = \frac{\sum_{k=1}^{K}\widehat{\text{TPF}}(\hat{\theta}_{t_k})}{K}. \tag{2.4}$$

Here, $\widehat{\text{TPF}}(\hat{\theta}_0)$ is excluded from this numerical integration above because the empirical estimator of the ROC curve is a right-continuous step-function.

3 Value-Added ROC Curve

3.1 Summary Measures of Continuity Performance

As continuous versions of TPF and FPF, we introduce true positive mean (TPM) and false positive mean (FPM), where

$$\text{TPM}(\theta) = E(X|X > \theta, Y = 1)$$

and

$$\text{FPM}(\theta) = E(X|X > \theta, Y = 0)$$

are the tail expectations of X that are higher than θ for the case and control groups, respectively. By combining these two, we introduce a composite metric, called the tail mean difference (TMD), where

$$\text{TMD}(\theta) = \text{TPM}(\theta) - \text{FPM}(\theta)$$

is the difference of the two tail expectations conditioning on Y and $X > c$.

Similar to AUC = PAUC($-\infty$) and PAUC(θ) that integrate TPF(θ) over thresholds of $(-\infty, \infty)$ and (c, ∞), respectively, we propose integrated TMD (ITMD) and (normalized) partial ITMD (PITMD), where

$$\text{PITMD}(\theta) = \frac{\int_c^{\infty}\text{TMD}(u)du}{\text{FPF}(\theta)}$$

and

$$\text{ITMD} = \text{PITMD}(-\infty)$$

are integrations of TMDs over thresholds above θ and all thresholds, respectively. Similar to (2.1), FPF(θ), used as the denominator of PITMD(θ), serves as the normalization factor. When

θ is set to $\text{FPF}^{-1}(t)$, $\text{TMD}(\theta_t)$ is interpreted as the TMD corresponding to a FPF of t, and ITMD (or $\text{PITMD}(\theta_t)$) is interpreted as an average of TMDs corresponding to a FPF between 0 and 1 (or 0 and t).

Fog a given t, we estimate

$$\widehat{\text{TMD}}(\hat{\theta}_t) = \widehat{\text{TPM}}(\hat{\theta}_t) - \widehat{\text{FPM}}(\hat{\theta}_t),$$

where

$$\widehat{\text{TPM}}(\hat{\theta}_t) = \frac{\sum_{i \in \mathbb{Y}_1(\hat{\theta}_t)} X_i}{|\mathbb{Y}_1(\hat{\theta}_t)|}, \tag{3.1}$$

$$\widehat{\text{FPM}}(\hat{\theta}_t) = \frac{\sum_{j \in \mathbb{Y}_0(\hat{\theta}_t)} X_j}{|\mathbb{Y}_0(\hat{\theta}_t)|}, \tag{3.2}$$

where $\mathbb{Y}_1(\theta) = \{i : X_i > \theta, i \in \mathbb{Y}_1\}$ and $\mathbb{Y}_0(\theta) = \{j : X_j > \theta, j \in \mathbb{Y}_0\}$ are disjoint sets of the case and control samples whose X's are higher than θ, respectively. Here $\mathbb{Y}_1(\theta)$ and $\mathbb{Y}_0(\theta)$ are disjoint because they are subsets of the disjoint sets $\mathbb{Y}_1$ and $\mathbb{Y}_0$, respectively. If $\hat{\theta}_t$ is set too high, (3.1) (or (3.2)) is not well defined because $\mathbb{Y}_1(\hat{\theta}_t)$ (or $\mathbb{Y}_0(\hat{\theta}_t)$) is an empty set. In this case, we set $\widehat{\text{TPM}}(\hat{\theta}_t) = \widehat{\text{TPM}}(\tilde{c}_t)$ (or $\widehat{\text{FPM}}(\hat{\theta}_t) = \widehat{\text{FPM}}(\bar{c}_t)$), where $\tilde{c}_t$ (or $\bar{c}_t$) is the maximum threshold value among all possible thresholds that satisfies $|\mathbb{Y}_1(\tilde{c}_t)| > 0$ (or $|\mathbb{Y}_0(\bar{c}_t)| > 0$). Similar to (2.4), for large K, we estimate

$$\widehat{\text{PITMD}}(\hat{\theta}_t) = \left\{ \frac{t}{k} \sum_{k=1}^{K} \widehat{\text{TMD}}(\hat{\theta}_{t_k}) \right\} / t = \frac{\sum_{k=1}^{K} \widehat{\text{TMD}}(\hat{\theta}_{t_k})}{K}, \tag{3.3}$$

and $\widehat{\text{ITMD}} = \widehat{\text{ITMD}}(\hat{\theta}_1)$. Thus, (3.3) is interpreted as an average of TMD due to the normalization factor t.

Given the same classification rule (e.g., classifying individuals with X values greater than θ_t as having the disease), the performance of X can be evaluated by either the standard metrics, which include TPF, FPF, AUC, and PAUC, or the proposed metrics, which include TPM, FPM, TMD, ITMD, and PITIMD, respectively. We refer to the former as assessing the discriminatory performance of X, and the latter as assessing its continuity performance.

3.2 Hypothesis Testing

The useful (or useless) biomarker has an $\text{TMD}(\theta_t)$ greater than (or equal to) zero or equivalently has a higher $\text{TPM}(\theta_t)$ than $\text{FPM}(\theta_t)$ (or the same value of $\text{TPM}(\theta_t)$ as $\text{FPM}(\theta_t)$). We thus state the null and alternative hypotheses as

$$\text{H}_0 : \text{TPM}(\theta_t) \leqslant \text{FPM}(\theta_t) \text{ versus } \text{H}_A : \text{TPM}(\theta_t) > \text{FPM}(\theta_t). \tag{3.4}$$

When $t = 1$ or $\theta_t = -\infty$, $\text{TPM}(\theta_t)$ and $\text{FPM}(\theta_t)$ are population means of the case and control groups, and Welch's two-sample t-test is directly applicable to test (3.4). When θ_t is higher, (3.1) and (3.2) are computed using truncated data that is possibly small or skewed. We propose a nonparametric bootstrap technique (Efron, 1979) to test (3.4), in the following steps:

Step 1. Set an acceptable FPF of t.

Step 2. Compute $\widehat{\text{TMD}}(\hat{\theta}_t)$ using the original data, where $\hat{\theta}_t = \widehat{\text{FPF}}^{-1}(t)$.

Step 3. Draw B bootstrap samples with replacement from the original data.

Step 4. For each bth bootstrap sample, $b = 1, 2, \ldots, B$, assign the first $|\mathbb{Y}_1|$ resampled observations the label $Y = 1$ and the remaining observations $Y = 0$ and compute $\widehat{\text{TMD}}_b(\hat{\theta}_{b,t})$, where $\hat{\theta}_{b,t} = \widehat{\text{FPF}}_b^{-1}(t)$.

Step 5. Calculate the bootstrap p-value as $B^{-1}\sum_{b=1}^{B} I\{\widehat{\text{TMD}}_b(\hat{\theta}_{b,t}) \geqslant \widehat{\text{TMD}}(\hat{\theta}_t)\}$.

Similarly, the useful (or useless) biomarker has an PITMD(θ_t) greater than (or equivalent to) zero or equivalently has a higher ITPM(θ_t) than IFPM(θ_t) (or the same value of ITPM(θ_t) as IFPM(θ_t)) because ITMD is expressed as integrated TPM (ITPM) minus integrated FPM (IFPM), where $\text{ITPM}(\theta_t) = \int_{\theta_t}^{\infty} \text{TPM}(u)du$ and $\text{IFPM}(\theta_t) = \int_{\theta_t}^{\infty} \text{FPM}(u)du$. We state the null and alternative hypothesis as

$$\text{H}_0 : \text{ITPM}(\theta_t) \leqslant \text{IFPM}(\theta_t) \text{ versus } \text{H}_A : \text{ITPM}(\theta_t) > \text{IFPM}(\theta_t). \tag{3.5}$$

The nonparametric bootstrap technique as stated above is directly applicable to test (3.5) by replacing TMD(θ_t) with PITMD(θ_t) (or ITMD) in Steps 3–5.

There are two practical considerations when conducting the proposed bootstrap hypothesis test. First, although the total sample size n may be large, the number of observations satisfying $X > \theta_t$ can be relatively small, particularly when the FPF of t is set to a low value. It is therefore important to examine the conditional distributions of X given $X > \theta_t$. If these distributions are skewed, a transformation of X may be considered to improve normality. Second, both the TMD estimate and corresponding p-value can vary depending on whether raw and transformed values of X values are used. This dependence is an undesirable property in comparison to the standard ROC curve, which is invariant to the monotonic transformation of X. Accordingly, we recommend applying a transformation to X only when it is scientifically justified and results in a meaningful improvement in normality, rather than for the purpose of achieving statistical significance.

These considerations are primarily relevant in settings with finite sample sizes. As n approaches infinity, the number of truncated observations is sufficiently large regardless of the FPF of t, and the proposed bootstrap test remains appropriate for non-normal data, provided that TMD(θ_t) (or PITMD(θ_t)) has finite second moments. However, if this condition is violated, for example, when the underlying distribution is a Cauchy distribution with infinite mean and variance, the proposed test becomes invalid.

3.3 Visualizations

In this subsection, we illustrate visualization methods under a finite Gaussian mixture ROC curve model (Gönen, 2013), assuming X given $Y = 1$ and $Y = 0$ are $\pi N(\mu_1, \sigma_1^2) + (1-\pi)N(\mu_0, \sigma_0^2)$ and $N(\mu_0, \sigma_0^2)$, respectively. Here, μ_1 and σ_1^2 (or μ_0 and σ_1^0) are mean and variance of the case (or control) group, respectively, and $N(\mu, \sigma^2)$ is a normal distribution with a mean μ and variance σ^2. The mixture proportion $\pi \in (0, 1)$ represents a level of heterogeneity, indicating that X is effective for only a fraction π of the case population. This modeling framework is particularly relevant for heterogeneous diseases such as cancer, which often comprise multiple molecular subtypes. In such contexts, one can only find a biomarker that can perform only for a specific subtype corresponding to only a subset (i.e., π fraction) of the case population rather than entire population.

We consider the following three scenarios: (i) $\mu_1 = 3$, (ii) $\mu_1 = 6$, and (iii) $\mu_1 = 9$, while fixing the remaining parameters at $\pi = 0.3$, $\mu_0 = 0$, $\sigma_1 = \sigma_0 = 1$ across the all three scenarios. Under each scenario, we generated simulated datasets consisting of 50 case and 50 control samples. We then computed $(\widehat{\text{TPF}}(0.97), \widehat{\text{PAUC}}(0.97), \widehat{\text{AUC}}) = (0.42, 0.36, 0.70)$, $(0.44, 0.38, 0.72)$, $(0.44, 0.38, 0.72)$ and $(\widehat{\text{TMD}}(0.97), \widehat{\text{PITMD}}(0.97), \widehat{\text{ITMD}}) = (1.59, 1.33, 1.34)$, $(3.27, 3.26, 2.48)$,

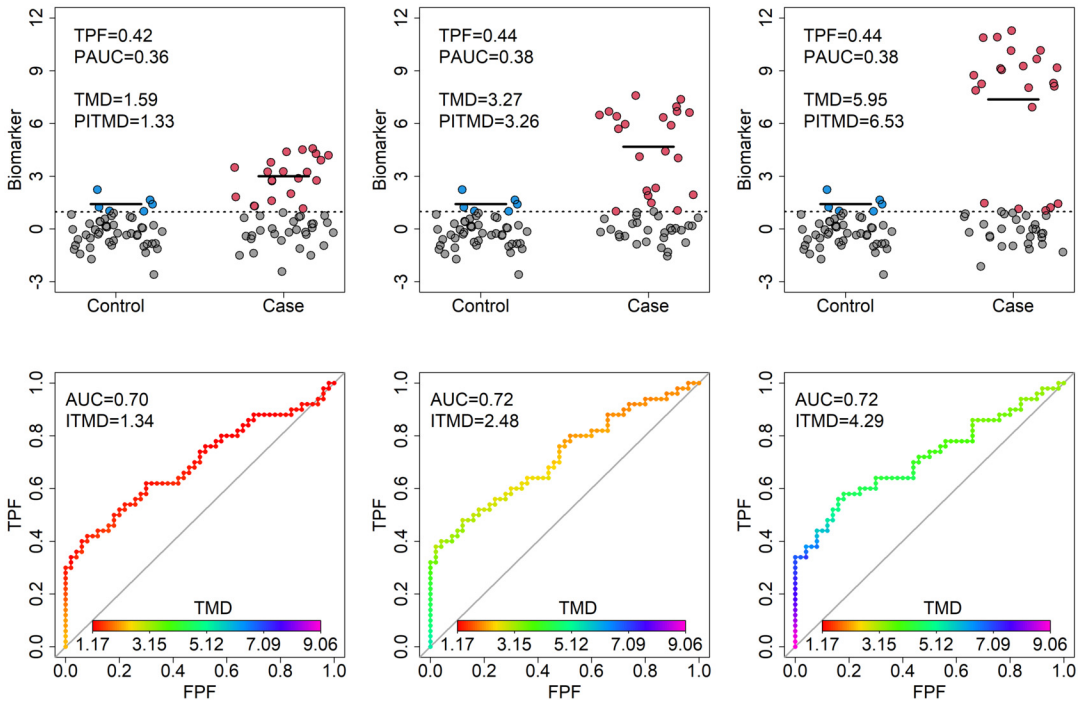

Figure 1: JDPs for the scenarios (i), (ii), and (iii) (upper left, middle, and right), and corresponding VAROC curves for each scenario (bottom left, middle, and right). In the JDPs, the dotted horizontal line represents the threshold $\hat{\theta}_t = 0.97$, which corresponding to the FPF of $t = 0.1$.

(5.95, 6.53, 4.29) for scenarios (i), (ii), (iii), respectively, where $\hat{\theta}_t$ is estimated to 0.97 by fixing FPF of 0.1.

To visualize these, we propose a jittered dot plot (JDP) by plotting jittered X values over case and control groups, as displayed in the first row of Figure 1. Here, the first, second, and third columns correspond to the scenarios (i), (ii), and (iii), respectively. On each JDP, $\hat{\theta}_t = 0.97$ is represented by a horizontal line, and thus, $\widehat{\text{TPF}}(\hat{\theta}_t)$ and $\widehat{\text{FPF}}(\hat{\theta}_t)$ are visualized as the numbers of blue and red dots; and $\widehat{\text{TPM}}(\hat{\theta}_t)$ and $\widehat{\text{FPM}}(\hat{\theta}_t)$ are visualized as their averages. Next, we propose the VAROC curve, which incorporate $\widehat{\text{TMD}}(\hat{\theta}_t)$ to the ROC curve, as illustrated in the second row of Figure 1. All pairs of $\widehat{\text{FPF}}(\theta)$ and $\widehat{\text{TPF}}(\theta)$ are colored according to $\widehat{\text{TMD}}(\theta)$ for $\theta \in (-\infty, \infty)$, thereby enabling simultaneous visualization of both the discrimination and continuity performance of X. Here, the VAROC curves are presented as two-dimensional plots. Additionally, three-dimensional VAROC curve, with x-, y- and, z-axes representing TPF, FPF, and TMD, respectively, are also provided in Appendix A.

In this example, the VAROC curves clearly differentiate among the scenarios (i), (ii), and (iii), whereas the ROC curves do not. These distinctions facilitate a more comprehensive evaluation of X, as further discussed in the following subsection.

3.4 Characteristics of VAROC Curve Analysis

In general, the summary measures of continuity and discrimination performance are positively correlated because X is considered informative when higher values are associated with an in-

creased probability of $Y = 1$. As a result, high discrimination performance typically coincides with a high continuity performance, and vice versa, as illustrated in Appendix A. However, exceptions do occur: X may exhibit strong discrimination performance while showing either high or low continuity performance. In such cases, reporting both continuity and discrimination metrics is particularly useful in the context of biomarker development.

Biomarker development typically proceeds through multiple phases (Pepe et al., 2001), broadly categorized into early (discovery) and late (validation) phases. These phases are commonly separated in study organization and study design (Feng and Pepe, 2020). For example, discovery laboratories employ high-throughput assays to screen a large number of candidate biomarkers using relatively small sample sizes, aiming to identify promising candidates for further evaluation. Pharmaceutical companies may then acquire these biomarkers and conduct large-scale trials to validate their effectiveness. These validation studies often use clinical-grade assays to measure biomarkers with greater precision and consistency. Although the discovery and validation assays are expected to be correlated, they do not always yield concordant results. Even when using similar technologies, variations in biological samples, collection sites, or timing can introduce substantial study-to-study heterogeneity. Therefore, successful validation requires prioritizing biomarkers that not only demonstrate strong discrimination performance but also exhibit robustness and reproducibility across diverse experimental settings.

To illustrate this, we revisit the Gaussian mixture model introduced in the subsection above. Its AUC is formulated as

$$\text{AUC} = \pi\,\Phi\left(\frac{\mu_1}{\sqrt{\sigma_1^2+1}}\right) + (1-\pi)0.5, \tag{3.6}$$

where Φ is the cumulative standard normal density. Due to the technical or study-to-study variations, we consider potential degradation in biomarker performance, assuming (μ_1, σ_1) is changed to $(\mu_1 - \delta_\mu, \sigma_1 + \delta_\sigma)$ for $\delta_\mu, \delta_\sigma > 0$. We consider three levels of variation: weak, moderate, or strong variations if $(\delta_\mu, \delta_\sigma) = (1,1)$, $(2,2)$, or $(3,3)$, respectively. The resulting reductions in AUC are shown in Figure 2, where AUC decreases from 0.64 to 0.59, 0.54, or 0.50 for (i); from 0.65 to 0.65, 0.62, or 0.58 for (ii); and from 0.65 to 0.65, 0.65, or 0.63 for (iii). We thus conclude (iii) exhibits the greatest robustness to technical or study-to-study variation, followed by (ii), with (i) being the most sensitive biomarker. Furthermore, it usually requires complicated techniques

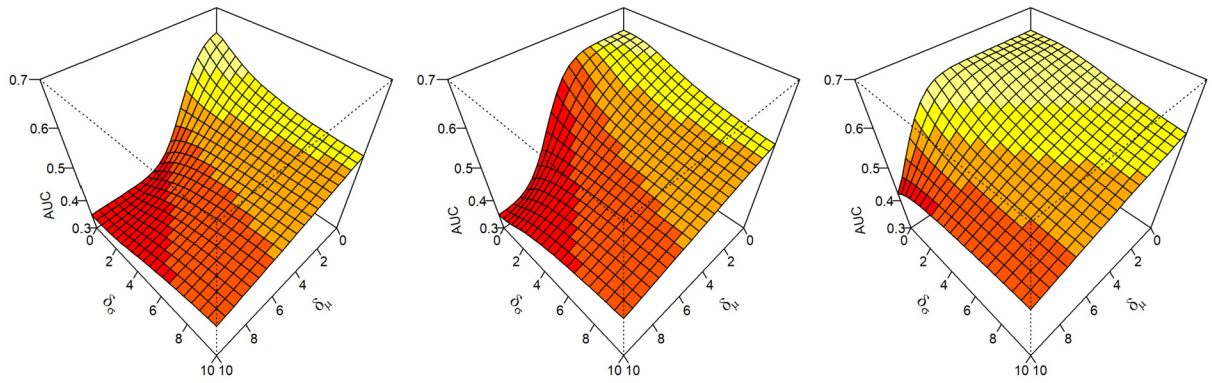

Figure 2: Robustness of AUC values under study-to-study variations over δ_μ and δ_σ for the scenarios (i) (left), (ii) (middle), and (iii) (right), respectively.

to quantify biomarkers from small molecules present in the human body, and (iii) that has the higher μ_1 is generally easier to measure and more reproducible than (i) and (ii). This aspect of reproducibility and robustness is captured by the VAROC curve, not by the conventional ROC curve.

In summary, while the ROC curve captures discrimination ability of X, it is sensitive when study-to-study or technical variation exists. In contrast, the VAROC curve integrate both its discrimination and continuity performance, thereby supporting more informed decisions in the biomarker development process.

4 Simulations

We conducted simulation studies to assess the performance of our proposed method. We considered the two scenarios, as below:

Scenario 1 (Binomial ROC curve): $X|Y = 1 \sim N_1$ and $X|Y = 0 \sim N_0$,

Scenario 2 (Gaussian mixture ROC curve): $X|Y = 1 \sim 0.3N_1 + 0.7N_0$ and $X|Y = 0 \sim N_0$.

Scenario 3 (Gamma mixture ROC curve): $X|Y = 1 \sim 0.3\Gamma_1 + 0.7\Gamma_0$ and $X|Y = 0 \sim \Gamma_0$,

where $N_1 = N(\mu, 1)$, $N_0 = N(0, 1)$, $\Gamma_1 = \Gamma(\mu + 1, 1)$, and $\Gamma_0 = \Gamma(1, 1)$, respectively. Here, $\Gamma(\mu, \beta)$ is a gamma distribution with shape and rate parameters of μ and β, respectively. Thus, μ represents the mean difference of N_1 and N_0 or Γ_1 and Γ_0. Across all scenarios, we set $\mu = (0, 0.5, 1, 1.5, \ldots, 5)$, where $\mu = 0$ implies a useless biomarker with $\text{TMD}(\theta_t) = \text{PITMD}(\theta_t) = 0$; and when μ_1 increased, the biomarker became more useful with increased $\text{TMD}(\theta_t)$ and $\text{PITMD}(\theta_t)$.

For each scenario, we generated $R = 200$ datasets with sample sizes of 50, 100, 200, 500, and 1000, respectively, with equal allocations for the control and case groups. For each rth dataset, $r = 1, 2, \ldots, R$, we computed TMD and PITMD at a FPF of 0.2 and ITMD and performed the bootstrap hypothesis tests with $B = 2000$ replicate samples. We then estimated the power function as an average of the p-values less than $\alpha = 0.05$ over the $R = 200$ replicates.

The upper panels of Figure 4 displayed the simulation results for Scenario 1. The proposed methods performed well across all settings, with statistical power increasing as either μ or n increased. When n was small, type I error rates were slightly inflated for TMD and PITMD due to the small number of samples beyond a FPF of 0.2. However, these error rates approached the nominal level of 0.05 as n increased. This issue did not arise for ITMD which used information from all samples as a global measure. As a result, ITMD consistently demonstrates higher power compared to TMD and PITMD.

Similar patterns were observed in Scenarios 2–3, shown in the middle and bottom panels of Figure 3, respectively. However, statistical power in these scenarios was lower due to the greater skewness in the data-generating distributions. In Scenario 1, the data were normally distributed, and the truncated data (i.e., truncated normal) were skewed but still relatively close to a normal distribution. In contrast, the Gaussian mixture model in Scenario 2 produced more skewed truncated data, and the Gamma mixture data in Scenario 3 were the most skewed. Consequently, Scenario 3 yielded the lowest power, followed by Scenario 2, with Scenario 1 achieving the highest power when comparing across scenarios with the same sample size.

In summary, the proposed method controlled the type I error rate well, regardless of whether the truncated distributions were skewed. However, statistical power decreased as the underlying truncated distributions deviated further from normality. Therefore, when biomarker data are expected to be skewed, a larger sample size should be considered to ensure adequate statistical power.

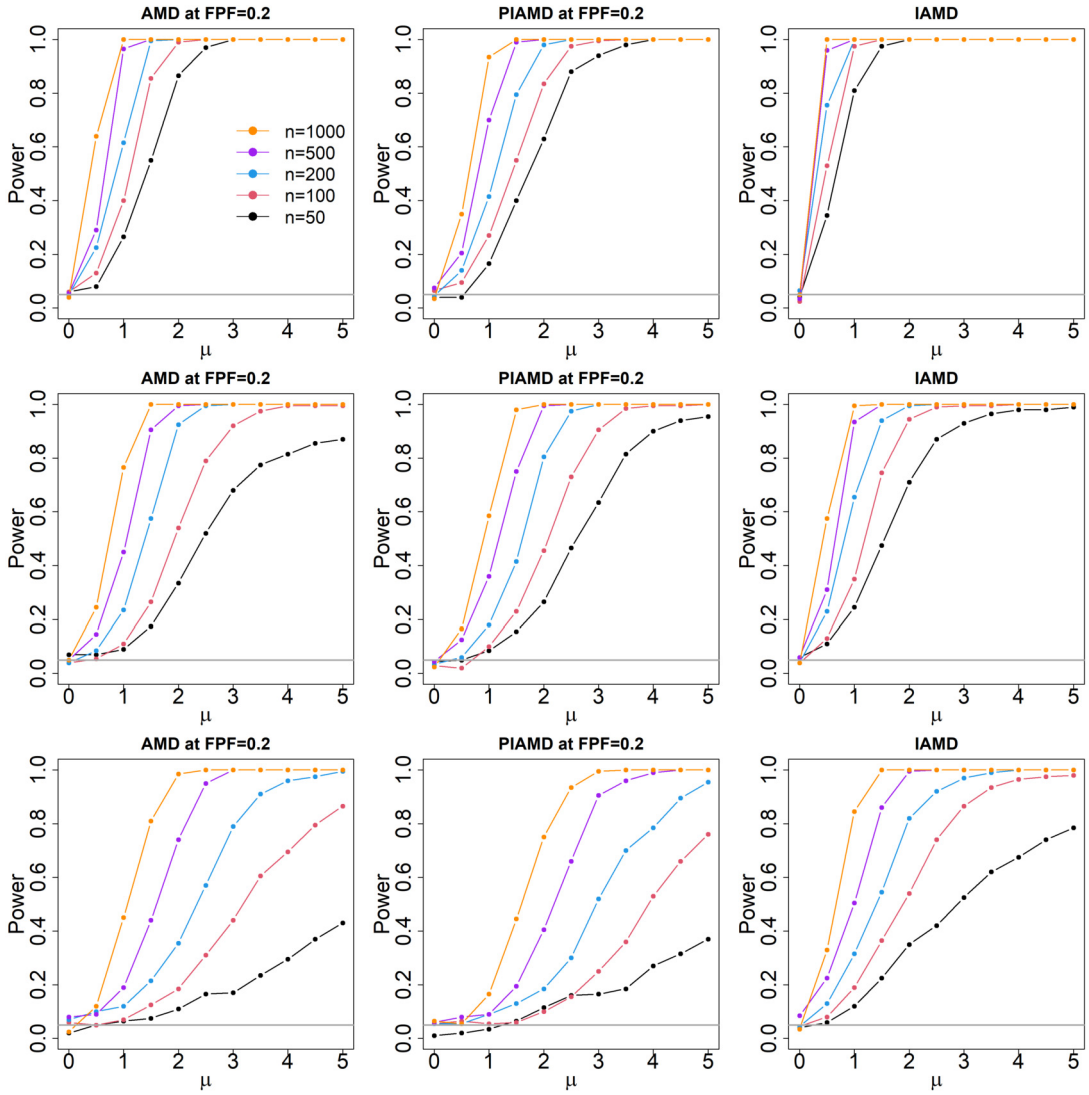

Figure 3: Simulation results. Upper: Scenario 1; lower: Scenario 2; left: TMD at FPF = 0.2; middle: PITMD at FPF = 0.2; right: ITMD. Red, black, blue lines are $n = 50, 100, 200$, respectively. Gray vertical line is 0.05.

5 Ovarian Cancer Study

The ovarian cancer study examined 23 normal and 30 ovarian cancer tissues and quantified their 1523 genes using a microarracy technology (Pepe et al., 2003). Its scientific goal was to identify genes that were overexpressed in ovarian cancer tissues compared to normal tissues. These overexpressed genes would produce specific proteins in blood or urine, which can be used in population cancer screening. As an early phase of the biomarker development, Pepe et al. (2003) discovered the top 100 genes based on the ROC curve-based metrics (Pepe et al., 2001).

We reanalyze the ovarian cancer dataset to demonstrate an advantage of our proposed method over the standard ROC curve analysis. We summarize the discrimination and continuity performances for each of the 1523 genes using the two global measures, ITMD and AUC,

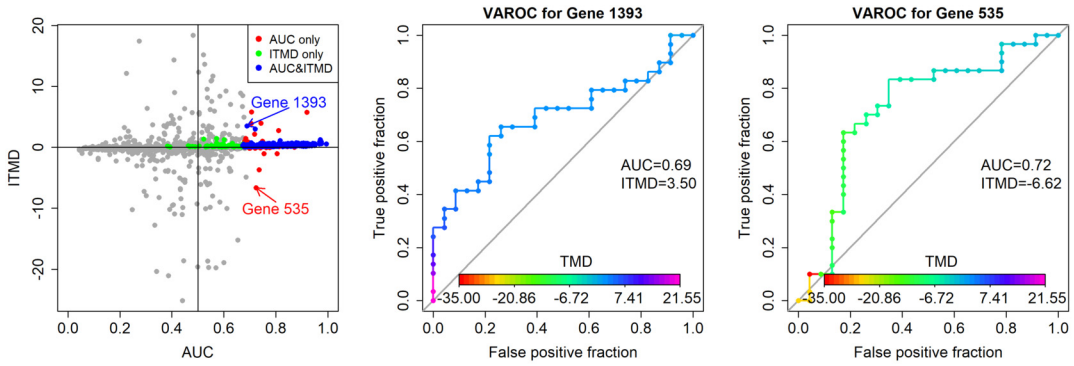

Figure 4: Analysis of ovarian cancer dataset across all thresholds. A two dimensional plot of ITMD versus AUC, where the horizontal and vertical lines are ITMD = 0 and AUC = 0.5, red dots are significant genes based on AUC, not ITMD, green dots are significant genes based on ITMD, not AUC, blue dots are significant genes based on both AUC and ITMD (left), and VAROC curves for gene 1393 (middle) and gene 535 (right).

respectively, as shown in the left panel of Figure 4. The useful genes are expected to lie above (or the right of) the horizontal (or vertical) line, i.e., ITMD $\geqslant 0$ (or AUC $\geqslant 0.5$), which are further classified into three gene sets using one-sided hypothesis tests: (a) 302 significant genes based on ITMD, (b) 314 significant genes based on AUC, and (c) 243 significant genes based on both ITMD and AUC. The two measures are correlated, while some genes with higher AUCs have lower ITMDs, or vice versa. Since AUC is primarily used for evaluating biomarkers, a main advantage of using ITMD is to discover the gene set (c) by excluding the gene set (b). That is, even though the gene set (b) consist of genes with higher AUCs, they are less likely reproducible for biomarker validation studies due to their lower ITMDs, as discussed in Subsection 3.4. For example, we compare the two genes: gene 1393, which has the highest ITMD among the gene set (c), and gene 535, which has the lowest ITMD among the gene set (b), where AUC and ITMD are 0.69 and 3.50 (and 0.72 and -6.62) for gene 1393 (and gene 535), respectively. The ROC curve fails to distinguish between the two genes with similar AUCs, whereas the VAROC curve clearly distinguish them by using different colors, as shown in Figure 5. Thus, given a limited resources, gene 1393, more robust to technical variation than gene 539, is selected for biomarker validation study based on the VAROC curve analysis.

We perform similar analysis based on PITMD and PAUC (and TMD and TPF) with the threshold corresponding to FPF of 0.3. Both of the measures agree in that gene 1483 is ranked best based on the all of the three ROC-based metrics, AUC = 0.99, PAUC = 0.98, and TPF = 1.00, and gene 93 with AUC = 0.93, PAUC = 0.91, and TPF = 0.93 has the highest PITMD (and TMD) among all the other genes that are significant based on both PITMD and PAUC (and both TMD and TPF). As displayed in Figure 5, the two genes with almost perfect discrimination performances have similar ROC curves, but gene 93 with PITMD = 1.26 and TMD = 1.28 has a higher continuity performance than gene 1483 with PITMD = 0.53 and TMD = 0.56. Since the two genes are evaluated at FPF $\leqslant 0.3$, we further use JDP analysis in Figure 5. Both genes are almost not expressed in the normal samples, resulting in similar thresholds, while gene 93 (or gene 1483) is expressed far from (or close to) their thresholds for most of the ovarian cancers. Thus, gene 93 would be more robust to technical variations than gene 1483 and further considered for biomarker validation studies, although gene 93 has a slight lower discrimination

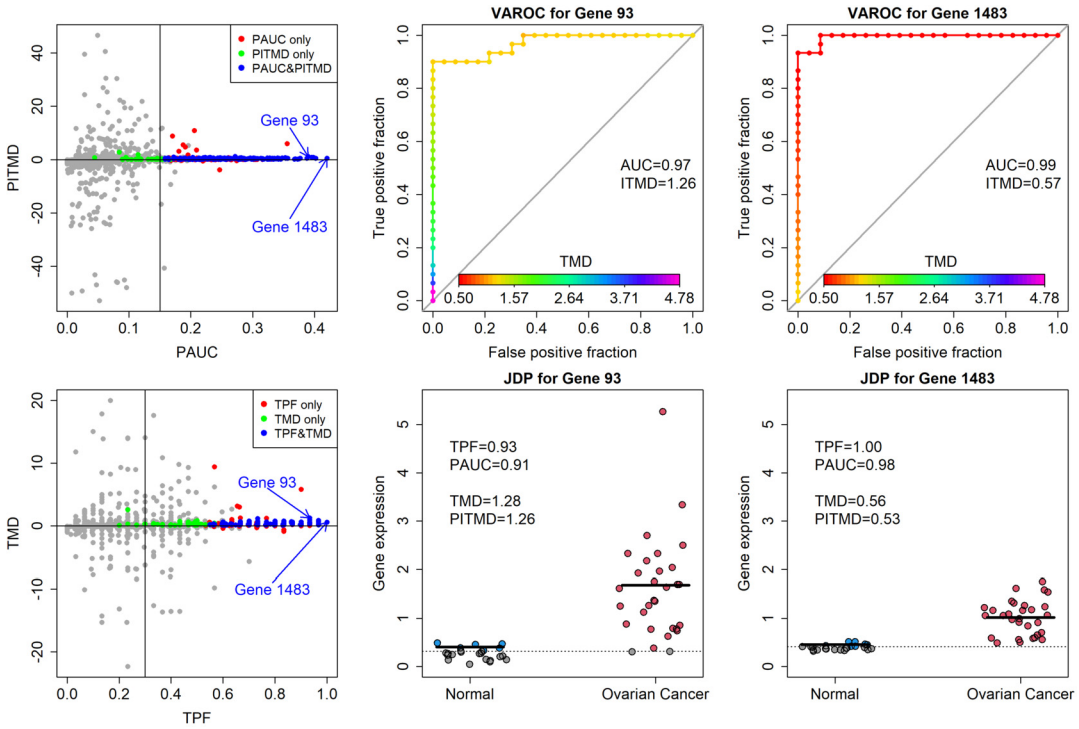

Figure 5: Analysis of ovarian cancer dataset at the threshold corresponding to FPF = 0.3. A two dimensional plot of PITMD versus PAUC, where the horizontal and vertical lines are PITMD = 0 and PAUC = 0.15, red dots are significant genes based on PAUC, not PITMD, green dots are significant genes based on PITMD, not PAUC, blue dots are significant genes based on both PAUC and PITMD (upper left); VAROC curves for gene 93 (upper middle) and gene 1483 (upper right); a two dimensional plot of TMD versus TPF, where the horizontal and vertical lines are TMD = 0 and TPF = 0.3, red dots are significant genes based on TPF, not TMD, green dots are significant genes based on TMD, not TPF, blue dots are significant genes based on both TPF and TMD (bottom left); JDPs for gene 93 (bottom middle) and gene 1483 (bottom right).

performance than gene 1483. We refer to Schummer et al. (1999) and Pepe et al. (2003) for gene labels and their biological interpretations. For these analysis, we used bootstrap t-test with 2000 replications and considered Benjamini and Hochberg adjusted p-value less than 0.05 statistically significant (Benjamini and Hochberg, 1995).

6 Discussion

The ROC curve is primarily used for evaluating biomarkers, but it can be misleading especially when biomarkers are applied to more general populations. From biomarker discovery to validation studies, technical or study-to-study variations are expected. Since the two studies are commonly conducted independently, these sources of variation are neither quantified nor controlled by study design, and are commonly underestimated or ignored in practice. Biomarker studies are therefore less likely to succeed when they focus on biomarkers that are highly sensitive to such variations. For instance, while a biomarker with an AUC of one is statistically ideal,

biologists may consider it unreliable or biologically irrelevant if its values are inconsistent across studies. The proposed VAROC curve is particularly well suited for such situations, providing a robust tool that supports the identification of reproducible biomarkers and contributes to more successful validation studies.

In our data analysis, we use p-values to identify statistically significant biomarkers. However, in practice, many biomarker discovery studies rely solely on ROC-based metrics without accompanying statistical inference. For example, biomarkers with AUC > 0.9 or 0.8 are often regarded as high-performance biomarkers and subsequently tested in validation studies with p-values. Likewise, candidate biomarkers can be evaluated using TMD or ITMD, with or without associated p-values.

Following Pepe's suggestion (Pepe, 2003), we select θ_t by considering an acceptable FPF, such that $\text{FPF}(\theta_t) \equiv t$. In general, evaluating biomarkers at higher θ_t is preferred, but there may be no data point to compute $\text{TPF}(\theta_t)$ or $\text{PAUC}(\theta_t)$ if θ_t is set too high (Pepe et al., 2001). The same limitation applies to $\text{TMD}(\theta_t)$ or $\text{PITMD}(\theta_t)$. Alternatively, the cutoff θ can be determine using other criteria (López-Ratón et al., 2014). For example, Youden's J statistics (Youden, 1950) can be used to estimate the optimal cutoff that maximizes $\text{J}(\theta) = \text{TPF}(\theta) - \text{FPF}(\theta)$ across all possible cutoff values. The continuity performance of the biomarker can then be evaluated by TMD at the corresponding cutoff. Interesting, since TMD is defined as TPM minus FPM, it can also be interpreted as a continuous version of Youden's J statistics.

In some cases, not only higher but also lower biomarker values may indicate a higher likelihood of disease. Several authors (Martínez-Camblor et al., 2017; Martínez-Camblor and Pardo-Fernández, 2019; Martínez-Camblor et al., 2019; Yang et al., 2024) have generalized ROC curves to appropriately evaluate the discrimination performance of biomarkers on the both directions. Similarly, our methods can be extended to accommodate such scenarios when both extremes of biomarker values are meaningful indicators of diseases.

The term "tail" used in this paper focuses on evaluating the performance of Y at a higher cutoff value of θ_t or, equivalently, at a lower FPF value t. More generally, however, θ_t may be set at any point; for example, in some settings it may also be of interest to evaluate Y at a higher FPF value of t. In this broader context, the term "above," which refers to evaluating the performance of Y above θ or θ_t can also be used. Accordingly, the term above MD (AMD) may also be used instead of TMD.

The varoc R package (Chung, 2025) implements the methods developed in this article and is publicly available on CRAN.

Supplementary Material

The R code and data used in Section 5 are available at the Supplementary Material of this paper.

A Appendix

We further explore the relation between ROC and VAROC curves. First, Scenario 1, described in Section 4, is commonly used framework for modeling normally distributed biomarker data. Given a fixed values of μ_0, σ_1, σ_0, the overlap between N_1 and N_0 decreases as μ_1 increases. As a result, both AUC and ITMD are expect to increase with increasing μ_1. We illustrate this relationship through simulation studies using 25 case and 25 control samples, assuming

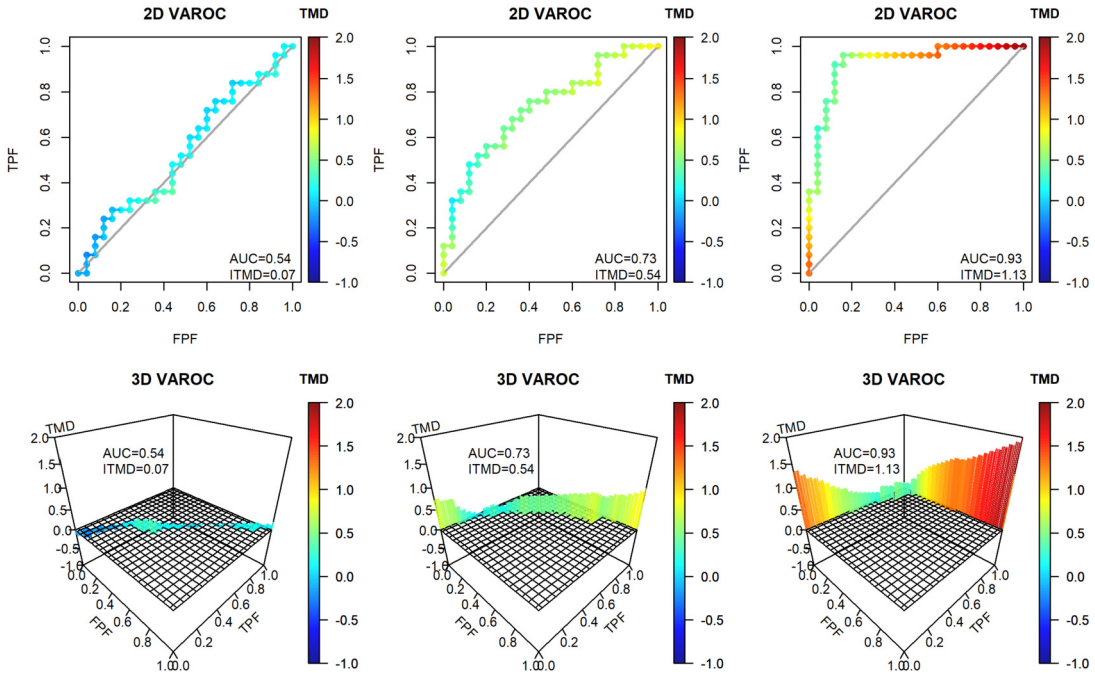

Figure A1: Relations between AUC and ITMD via simulations under the binomial ROC curve models with $\mu_1 = 0, 1$ and 2 (left, middle, and right): two-dimensional and corresponding three-dimensional VAROC curves (upper and bottom).

$\mu_0 = 0$ and $\sigma_0 = \sigma_1 = 1$, and varying μ_1 across 0,1, and 2. When $\mu_1 = 0$, the biomarker X is non-informative, yielding estimated AUC and ITMD values of 0.54 and 0.07, respectively. As μ_1 increases to 1 (and 2), the estimated AUC and ITMD rise to 0.73 and 0.54 (and 0.93 and 1.13), respectively. These results are visualize as two- and three-dimensional VAROC curves in Figure A1.

Next, we explore how ROC curves vary with different values of variances. We consider scenarios (i), (ii), and (iii) described in Subsection 3.3, while additionally considering $(\sigma_1, \sigma_0) = (1, 1), (6, 1), (1, 6)$. Figure A2 presents their theoretical distributions for each scenario. Here, we set a FPF of $t = 0.1$, specifying a corresponding threshold as $\theta_t = \sigma_0 \Phi^{-1}(t) + \mu_0$. For example, if $\sigma_0 = 1$ and $\mu_0 = 0$, $\theta_t = 1.28$. When σ_0 is fixed, θ does not change, and $\text{FPM}(\theta_t)$ remains the same. Thus, the increase of σ_1 solely increases $\text{TPM}(\theta_t)$, leading to an increase in $\text{TMD}(\theta_t)$. On the other hand, if σ_0 increases, θ_t also rises, which causes both $\text{TPM}(\theta_t)$ and $\text{FPM}(\theta_t)$ to increase. However, as displayed in the last row of the panels of Figure A2, the increase in $\text{FPM}(\theta_t)$ is larger than that in $\text{TPM}(\theta_t)$, resulting in a decrease in $\text{TMD}(\theta_t)$. Therefore, the variance shift from σ_1 or σ_0 have different effects on $\text{TMD}(\theta_t)$ and ITMD.

To better understand the effect from these variance shifts, we performed simulation studies under the Gaussian mixture model introduced in Subsection 3.4 under the following two scenarios.

Scenario A: $\mu_1 = 3, 6, 9$, $\sigma_1 = 1, 1.5, 2, \ldots, 6$, $\sigma_0 = 1$,

Scenario B: $\mu_1 = 3, 6, 9$, $\sigma_1 = 1$, $\sigma_0 = 1, 1.5, 2, \ldots, 6$,

where $\mu_0 = 0$ and $\pi = 0.3$ for both scenarios. For each scenario, we generated 500 simulated datasets with $n = 30$ ($n = 15$ for the case and $n = 15$ for the control groups) and computed

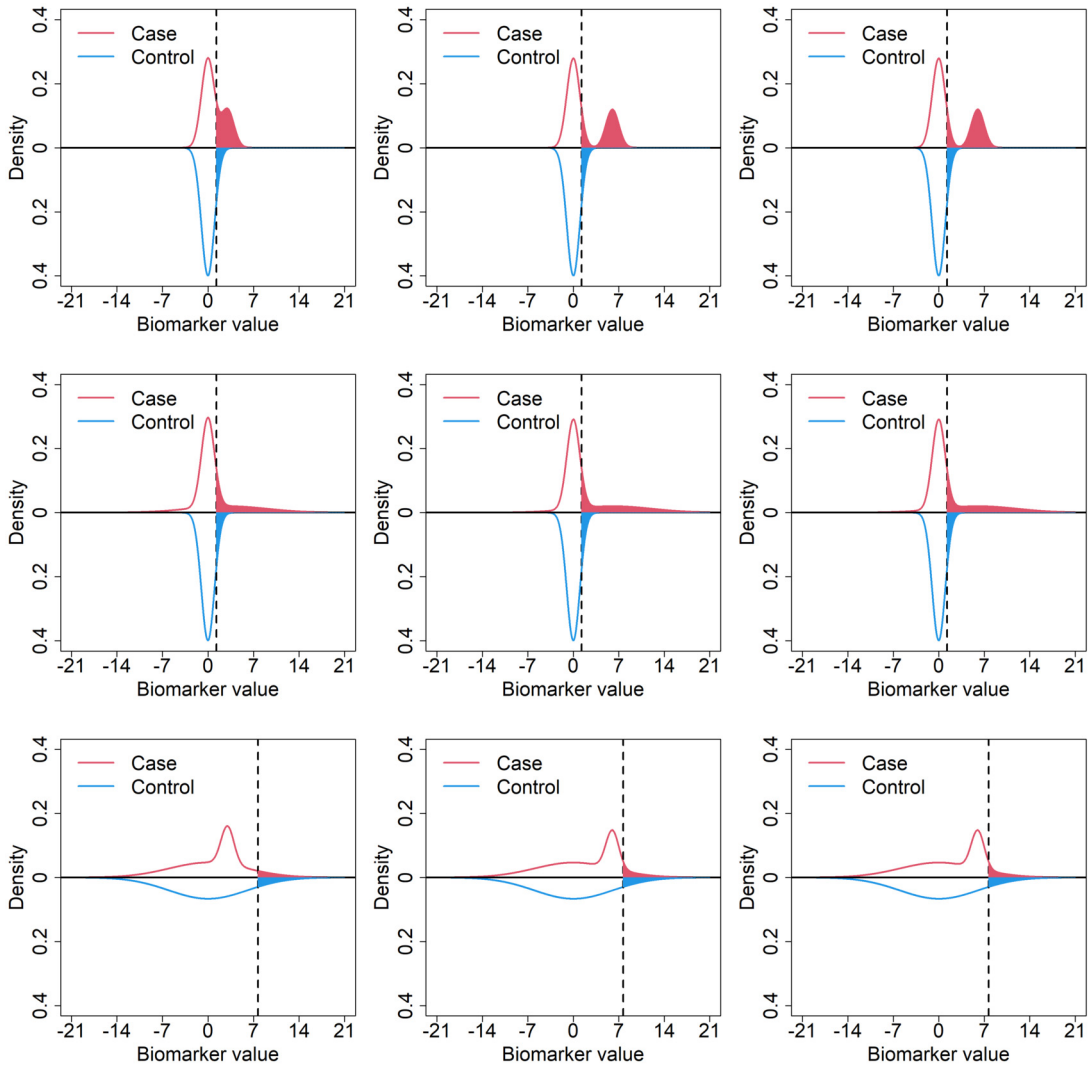

Figure A2: Hypothetical distributions of biomarkers under variation shifts for scenarios (i), (ii), and (iii) (left, middle, and right columns), and (σ_1, σ_0) is (1,1), (6,1), (1,6) (first row), (6,1) (upper, middle, bottom rows). The dashed vertical line represents the threshold value corresponding to the false positive fraction = 10%.

AUC and ITMD values and corresponding p-values. We then averaged these measures and p-values over the 500 iterations. Figure A3 demonstrates the simulation results. Both AUC and ITMD increase as μ_1 increases across all scenarios. On the one hand, as expected from (3.6), AUC decreases as σ_1 or σ_0 increases. On the other hand, ITMD increases as σ_1 increases, but it decreases as σ_0 increases, as explained in the paragraph above. However, the increase in (σ_1, σ_0) leads to a rise in the standard error of ITMD, which results in large (or small) p-values when (σ_1, σ_0) is large (or small). Thus, both AUC and ITMD agree that in a useful biomarkers, $\mu_1 > \mu_0$, and σ_1 and σ_0 are small. Note that directly comparing AUC and ITMD based on their p-values is not straightforward because they evaluate different aspects of biomarkers, i.e., discrimination versus continuity performance.

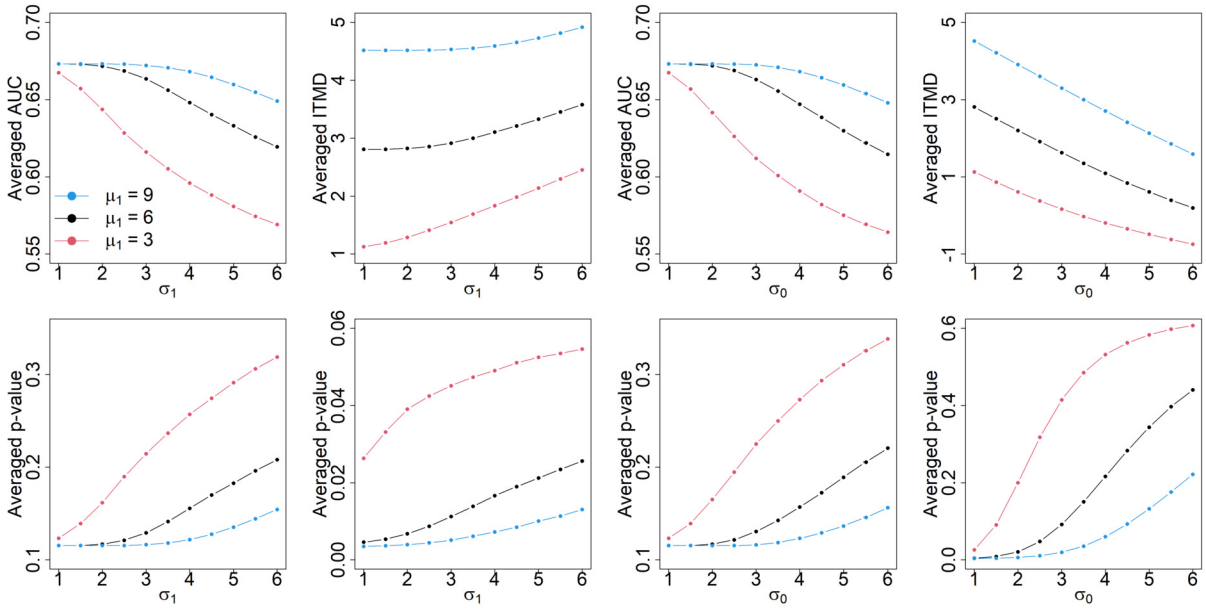

Figure A3: Simulation results: average AUC and ITMD under Scenario A (top first and second columns) and Scenario B (top third and fourth columns), with corresponding average p-values shown in the bottom panels. Blue, black and red lines represents $\mu_1 = 9, 6$, and 3, respectively.

References

Bangdiwala SI, Haedo AS, Natal ML, Villaveces A (2008). The agreement chart as an alternative to the receiver-operating characteristic curve for diagnostic tests. *Journal of Clinical Epidemiology*, 61(9): 866–874. https://doi.org/10.1016/j.jclinepi.2008.04.002

Benjamini Y, Hochberg Y (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B, Statistical Methodology*, 57(1): 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bickel PJ (1965). On some robust estimates of location. *The Annals of Mathematical Statistics*, 36(3): 847–858. https://doi.org/10.1214/aoms/1177700058

Chen LA, Chen DT, Chan W (2010). The distribution-based p-value for the outlier sum in differential gene expression analysis. *Biometrika*, 97(1): 246–253. https://doi.org/10.1093/biomet/asp075

Chung Y (2025). varoc. R package version 1.0.0.

Efron B (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26. https://doi.org/10.1214/aos/1176344552

Feng Z, Pepe MS (2020). Adding rigor to biomarker evaluations—EDRN experience. *Cancer Epidemiology, Biomarkers & Prevention*, 29(12): 2575–2582. https://doi.org/10.1158/1055-9965.EPI-20-0240

Gönen M (2013). Mixtures of receiver operating characteristic curves. *Academic Radiology*, 20(7): 831–837. https://doi.org/10.1016/j.acra.2013.03.003

Huang Y, Sullivan Pepe M, Feng Z (2007). Evaluating the predictiveness of a continuous marker. *Biometrics*, 63(4): 1181–1188. https://doi.org/10.1111/j.1541-0420.2007.00814.x

Huber PJ (1972). The 1972 Wald lecture robust statistics: A review. *The Annals of Mathematical*

Statistics, 43(4): 1041–1067. https://doi.org/10.1214/aoms/1177692459

Lee WC (1999). Probabilistic analysis of global performances of diagnostic tests: Interpreting the Lorenz curve-based summary measures. *Statistics in Medicine*, 18(4): 455–471. https://doi.org/10.1002/(SICI)1097-0258(19990228)18:4<455::AID-SIM44>3.0.CO;2-A

Lee WC, Hsiao CK (1996). Alternative summary indices for the receiver operating characteristic curve. *Epidemiology*, 7(6): 605–611. https://doi.org/10.1097/00001648-199611000-00007

López-Ratón M, Rodríguez-Álvarez MX, Cadarso-Suárez C, Gude-Sampedro F (2014). Optimalcutpoints: An R package for selecting optimal cutpoints in diagnostic tests. *Journal of Statistical Software*, 61: 1–36. https://doi.org/10.18637/jss.v061.i08

Lorenz MO (1905). Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, 9(70): 209–219. https://doi.org/10.1080/15225437.1905.10503443

Martínez-Camblor P, Corral N, Rey C, Pascual J, Cernuda-Morollón E (2017). Receiver operating characteristic curve generalization for non-monotone relationships. *Statistical Methods in Medical Research*, 26(1): 113–123. https://doi.org/10.1177/0962280214541095

Martínez-Camblor P, Pardo-Fernández JC (2019). Parametric estimates for the receiver operating characteristic curve generalization for non-monotone relationships. *Statistical Methods in Medical Research*, 28(7): 2032–2048. https://doi.org/10.1177/0962280217747009

Martínez-Camblor P, Pérez-Fernández S, Díaz-Coto S (2019). Improving the biomarker diagnostic capacity via functional transformations. *Journal of Applied Statistics*, 46(9): 1550–1566. https://doi.org/10.1080/02664763.2018.1554628

Parmigiani G (2019). The fuzzy ROC. arXiv preprint: https://arxiv.org/abs/1903.01868v1

Pepe MS (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, Oxford.

Pepe MS, Etzioni R, Feng Z, Potter JD, Thompson ML, Thornquist M, et al. (2001). Phases of biomarker development for early detection of cancer. *Journal of the National Cancer Institute*, 93(14): 1054–1061. https://doi.org/10.1093/jnci/93.14.1054

Pepe MS, Feng Z, Huang Y, Longton G, Prentice R, Thompson IM, et al. (2008a). Integrating the predictiveness of a marker with its performance as a classifier. *American Journal of Epidemiology*, 167(3): 362–368. https://doi.org/10.1093/aje/kwm305

Pepe MS, Feng Z, Janes H, Bossuyt PM, Potter JD (2008b). Pivotal evaluation of the accuracy of a biomarker used for classification or prediction: Standards for study design. *Journal of the National Cancer Institute*, 100(20): 1432–1438. https://doi.org/10.1093/jnci/djn326

Pepe MS, Longton G, Anderson GL, Schummer M (2003). Selecting differentially expressed genes from microarray experiments. *Biometrics*, 59(1): 133–142. https://doi.org/10.1111/1541-0420.00016

Schummer M, Ng WV, Bumgarner RE, Nelson PS, Schummer B, Bednarski DW, et al. (1999). Comparative hybridization of an array of 21500 ovarian cdnas for the discovery of genes overexpressed in ovarian carcinomas. *Gene*, 238(2): 375–385. https://doi.org/10.1016/S0378-1119(99)00342-X

Stigler SM (1973). The asymptotic distribution of the trimmed mean. *The Annals of Statistics*, 1(3): 472–477. https://doi.org/10.1214/aos/1176342412

Tibshirani R, Hastie T (2007). Outlier sums for differential gene expression analysis. *Biostatistics*, 8(1): 2–8. https://doi.org/10.1093/biostatistics/kxl005

Tukey JW, McLaughlin DH (1963). Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/winsorization 1. *Sankhya. Series A*, 25(3): 331–352.

Vickers AJ, Elkin EB (2006). Decision curve analysis: A novel method for

evaluating prediction models. *Medical Decision Making*, 26(6): 565–574. https://doi.org/10.1177/0272989X06295361

Wu B (2007). Cancer outlier differential gene expression detection. *Biostatistics*, 8(3): 566–575. https://doi.org/10.1093/biostatistics/kxl029

Yang J, Kuan PF, Li X, Li J, Zhou XH (2024). Transformed ROC curve for biomarker evaluation. *Statistics in Medicine*, 43(30): 5681–5697. https://doi.org/10.1002/sim.10268

Youden WJ (1950). Index for rating diagnostic tests. *Cancer*, 3(1): 32–35. https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3

Yuen KK (1974). The two-sample trimmed t for unequal population variances. *Biometrika*, 61(1): 165–170. https://doi.org/10.1093/biomet/61.1.165

Yuen KK, Dixon W (1973). The approximate behaviour and performance of the two-sample trimmed t. *Biometrika*, 60(2): 369–374. https://doi.org/10.1093/biomet/60.2.369