Journal Title
XX(X):1–16

DOI: 10.1177/ToBeAssigned
www.sagepub.com/

Affinity-based measures of biomarker performance evaluation

Miguel de Carvalho[1], Bradley J. Barney[2] and Garritt L. Page[3]

Abstract

We propose new summary measures of biomarker accuracy which can be used as companions to existing diagnostic accuracy measures. Conceptually, our summary measures are tantamount to the so-called Hellinger affinity and we show that they can be regarded as measures of agreement constructed from similar geometrical principles as Pearson correlation. We develop a covariate-specific version of our summary index, which practitioners can use to assess the discrimination performance of a biomarker, conditionally on the value of a predictor. We devise nonparametric Bayes estimators for the proposed indexes, derive theoretical properties of the corresponding priors, and assess the performance of our methods through a simulation study. The proposed methods are illustrated using data from a prostate cancer diagnosis study.

Keywords

Biomarker, covariate-specific diagnostic, Hellinger affinity, summary measure

1 Introduction

Accurate diagnosis is a key target of diagnostic decision-making. Before a biomarker is routinely applied in practice, it is important to evaluate its performance in discriminating between diseased and non-diseased subjects. The most well-known summary accuracy measures are the AUC (Area Under the receiver operating characteristic Curve) and the Youden index;[1] other summary indexes can be found in Pepe[2] (Section 4.3.3). These well-known summary measures at times gloss over important differences between diseased and non-diseased subjects. Particularly, in genomic studies it is known that gene expression data can differ substantially between diseased and non-diseased subjects while having a similar mean.[3–5] Gene expression data may present bimodal and multimodal patterns (e.g. Figure 1 in Wang and Tian[5]) and it is well known that the AUC is not tailored for this type of setting.

To shed some light on why the AUC may fail, let Y_D and $Y_{\bar{D}}$ be random variables representing the test results for diseased and non-diseased subjects, and let F_D and $F_{\bar{D}}$ be the corresponding distribution functions. Formally, the AUC consists of $P(Y_D > Y_{\bar{D}})$ and it is typically argued that AUC $= 0.5$ for a test that does no better than chance in discriminating between diseased and non-diseased individuals, while AUC $= 1$ for a test that perfectly distinguishes between diseased and non-diseased subjects. While AUC is widely used in practice, Figure 1 illustrates a setting under which the AUC is known to perform poorly. Regarding this setting, Lee and Hsiao[6] (p. 606) make the following comment:

> "For the two populations of the diseased and the non-diseased [...] the marker perfectly separates the two. Therefore, any clinician (or epidemiologist) will have no trouble in choosing a decision rule for the marker, that is, high and low

[1]School of Mathematics, University of Edinburgh, Scotland, UK
[2]Department of Pediatrics, University of Utah, Salt Lake City, Utah, USA
[3]Department of Statistics, Brigham Young University, Provo, Utah, USA

Corresponding author:
Miguel de Carvalho
Email: miguel.decarvalho@ed.ac.uk

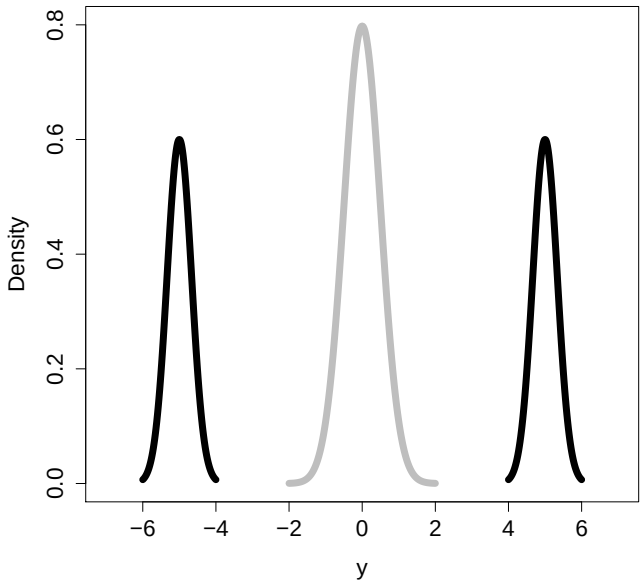


Figure 1. The separation trap: Perfect discrimination but AUC $= 0.5$. Details on the truncated normal densities used to construct this instance can be found in Example 2; the black and grey lines respectively denote the densities of the biomarkers of the diseased and non-diseased subjects.

> cutoff points. Nevertheless, adopting the AUC as the measure of overall performance leads one to conclude that the marker is not better than flipping a fair coin (its AUC is 0.5)"

Throughout, we will refer to the situation in Figure 1 as the 'separation trap,' since one has perfect discrimination but AUC $= 0.5$. As can be seen from Figure 1, even though the populations of diseased and non-diseased subjects are perfectly separated, half of the diseased-subjects are predicted to have a test result higher than the non-diseased subjects, and thus AUC $= 0.5$. Another shortcoming of the AUC is that it is defined for the situation where larger values of the biomarker are more indicative of disease, and this is not always the case as illustrated on the prostate cancer study data example from Section 4. While one can always consider $1 -$ AUC when lower values of the biomarker are more indicative of disease, it is desirable in practice to have methods that can be readily computed and interpreted without assuming anything about the biomarker threshold(s) that demarcate positive and negative test diagnoses.[3–5]

But beyond the AUC, the Youden index (YI) also falls into the separation trap. To see this recall that YI $= \max_{y\in\mathbb{R}} \{F_{\bar{D}}(y) - F_D(y)\}$, with YI $= 0$ corresponding to complete overlap ($F_{\bar{D}}(y) = F_D(y)$), and it is often argued that YI $= 1$ when the distributions are 'completely separated'. It is straightforward to show that in the example in Figure 1, it holds that YI $= 1/2$, while the distributions of the markers for diseased and non-diseased subjects are completely separated—thus confirming that the Youden index would fall into the separation trap. The optimal cutoff region yielded through the Youden index is

$$\arg\max_{y\in\mathbb{R}} \{F_{\bar{D}}(y) - F_D(y)\} = [2, 4].$$

Interestingly, however, the more sensible cutoff region $[-4, -2] \cup [2, 4]$ could have been obtained by adjusting the definition of Youden index to consider the absolute value of the difference between distribution functions, but even this modified index would be equal to 1/2.

A main goal of this article is to propose new diagnostic accuracy measures that: i) accommodate the separation trap; ii) do not require knowing in advance if larger values of the biomarker are more indicative of disease; and iii) can be used as companions, or possibly as alternatives, to existing diagnostic accuracy measures. Conceptually, our summary measures can be motivated by first considering a geometric interpretation of covariance and Pearson correlation. By recalling the well-known fact that for zero-mean

finite-variance random variables X and Y, the covariance can be interpreted as an inner product between random variables,[7] it follows that Pearson correlation

$$\rho = \frac{\text{cov}(X,Y)}{\text{sd}(X)\text{sd}(Y)} = \cos(X \angle Y), \tag{1}$$

can be interpreted as a cosine of the angle between X and Y. This simple geometric interpretation is handy for understanding some basic properties of the Pearson correlation, including the fact that just like a cosine, ρ is between -1 and 1, and that orthogonality in such context corresponds to the case where there is no agreement between X and Y, that is, $\text{cov}(X,Y) = 0$. The summary measures we develop, introduced in Section 2, are constructed along the same lines, but are based on evaluating the level of agreement between densities (of the biomarkers of diseased and non-diseased subjects) instead of focusing on random variables—as Pearson correlation does.

Recognizing that biomarker densities may depend on one or more covariates such as age, we consider the impact that covariates have on the discriminatory ability per the metric we develop. This follows current trends in research associated with AUC (e.g., Inácio de Carvalho et al.[8]); the fact that our diagnostic accuracy measures are designed without assuming anything about the biomarker threshold(s) that demarcate positive and negative test diagnoses is particularly advantageous for the covariate-dependent situation. We devise a covariate-specific version of our main summary index, which practitioners can use to assess the discrimination performance of a biomarker, conditionally on the value of a covariate. We develop nonparametric Bayesian estimators for all proposed indexes and evaluate the numerical performance of a specific implementation in detail through a simulation study. Using Bayesian nonparametric and semiparametric inference for evaluating the performance of a biomarker is not unprecedented,[8–17] and doing so provides a great deal of flexibility particularly in the dependent case (i.e., a covariate is present). An additional computational advantage of our covariate-specific summary measure with respect to that of Inácio de Carvalho et al.[8] is that it avoids the need of computing conditional quantiles over a grid of covariates—a task which requires a substantial computational investment. More importantly, as we elaborate below, our summary measures do not fall into the separation trap depicted in Figure 1.

The article is organized as follows. In the next section we introduce the proposed measures along with the corresponding inference tools. In Section 3 we conduct a simulation study. Section 4 offers an illustration of our methods in a prostate cancer diagnosis case study. Proofs are included in the online supplementary materials.

2 Affinity measures of biomarker accuracy

2.1 Angle-based summary measures of biomarker accuracy

Our summary measures are built on similar construction principles as Pearson correlation, but instead of looking at the angle between random variables as in (1), we work directly with the densities of the biomarker outcomes for diseased and non-diseased subjects, that is $f_D = \text{d}F_D/\text{d}y$ and $f_{\bar{D}} = \text{d}F_{\bar{D}}/\text{d}y$, respectively. Thus, in place of the covariance inner product we use $\langle f_D, f_{\bar{D}} \rangle = \int_{-\infty}^{\infty} f_D(y) f_{\bar{D}}(y)\,\text{d}y$, and in place of the standard deviation (sd) norms, we use $\|f_D\| = \{\int_{-\infty}^{\infty} f_D^2(y)\,\text{d}y\}^{1/2} < \infty$, and $\|f_{\bar{D}}\| = \{\int_{-\infty}^{\infty} f_{\bar{D}}^2(y)\,\text{d}y\}^{1/2} < \infty$. The starting point for the construction of our measure is given by a standardized inner product defined as:

$$\bar{\kappa} = \frac{\langle f_D, f_{\bar{D}} \rangle}{\|f_D\| \|f_{\bar{D}}\|}. \tag{2}$$

For a biomarker with perfect discriminatory ability we would have $\bar{\kappa} = 0$, as f_D would be orthogonal to $f_{\bar{D}}$. The higher the value of $\bar{\kappa}$, the lower the discriminatory ability of the corresponding biomarker. Indeed, similar to the Pearson correlation, our measure can be interpreted as an angle between f_D and $f_{\bar{D}}$. However, since $f_D \geqslant 0$ and $f_{\bar{D}} \geqslant 0$ it follows that $\langle f_D, f_{\bar{D}} \rangle \geqslant 0$, and thus the angle between f_D and $f_{\bar{D}}$ can only be acute or right, that is $f_D \angle f_{\bar{D}}$ is in $[0, \pi/2]$, and thus $\bar{\kappa}$ is in $[0, 1]$. Orthogonality between the biomarker outcome for diseased and non-diseased subjects corresponds to a biomarker that perfectly discriminates between diseased and non-diseased subjects.

Following the terminology in de Carvalho et al.[18] (Definition 2), we refer to $\bar{\kappa}$ in (2) as a measure of *compatibility*. Roughly speaking, *compatibility* is defined as a standardized inner product. However natural the $\bar{\kappa}$ in (2) may appear, in practice it would constrain us to work with square-integrable densities. To retain the main ingredients of the construction above and to avoid the issue of being constrained to square-integrable densities, we resort to a seminal square-root characterization.[19] Since f_D and $f_{\bar{D}}$ are valid densities, it follows that $\|\sqrt{f_D}\| = \|\sqrt{f_{\bar{D}}}\| = 1$, and thus we define our summary measure as

$$\kappa = \frac{\langle \sqrt{f_D}, \sqrt{f_{\bar{D}}} \rangle}{\|\sqrt{f_D}\| \|\sqrt{f_{\bar{D}}}\|} = \langle \sqrt{f_D}, \sqrt{f_{\bar{D}}} \rangle = \int_{-\infty}^{+\infty} \sqrt{f_D(y)} \sqrt{f_{\bar{D}}(y)}\,\text{d}y. \tag{3}$$

Some comments are in order. Similar to (2), orthogonality between the biomarker outcome for diseased and non-diseased subjects corresponds to the case where the biomarker is perfect, that is $\kappa = 0$ for a perfect test—which perfectly discriminates diseased subjects from non-diseased subjects—and $\kappa = 1$ for a useless test—for which $f_D = f_{\bar{D}}$.

Interestingly, the measure κ in (3) is known in mathematical statistics under the name of Hellinger affinity,[20] but we are unaware of applications of the concept in the statistical evaluation of biomarkers. In context, κ can be interpreted as a measure of the level of agreement between the densities of the biomarker outcomes for diseased and non-diseased subjects, or equivalently, as a measure of the highest possible biomarker accuracy.

Example 1. (Binormal affinity). To fix ideas, consider the case when both diseased and non-diseased populations have normal biomarker densities. Thus, $f_D(y) = \phi(y \mid \mu_D, \sigma_D^2)$ and $f_{\bar{D}}(y) = \phi(y \mid \mu_{\bar{D}}, \sigma_{\bar{D}}^2)$. As stated in Table 1:

$$\kappa = \sqrt{\frac{2\sigma_D\sigma_{\bar{D}}}{\sigma_D^2 + \sigma_{\bar{D}}^2}} \exp\left\{ -\frac{1}{4}\frac{(\mu_D - \mu_{\bar{D}})^2}{\sigma_D^2 + \sigma_{\bar{D}}^2} \right\}. \tag{4}$$

As expected, for a useless test—that is $\mu_D = \mu_{\bar{D}}$ and $\sigma_D = \sigma_{\bar{D}}$—it holds that $\kappa = 1$. For fixed $\sigma_D > 0$ and $\sigma_{\bar{D}} > 0$ it follows that as $\mu_D \to \infty$ and $\mu_{\bar{D}} \to -\infty$, that is as populations become more separated, then $\kappa \to 0$. Indeed, as it can be seen from Figure 2 the more separated the populations—that is the more orthogonal they are—the closer κ gets to zero. Notice also that, in this setting, the larger the AUC the lower κ.

We further explore the relationship between AUC and κ in the *proper binormal* setting, which is an important special case of the binormal setting in Example 1. To ensure the ROC curve is truly *proper* (i.e., everywhere concave and thus amenable to summarizing by AUC), the binormal model must constrain σ_D to equal $\sigma_{\bar{D}}$.[21] As the population means move apart, the (AUC, κ) pairs trace the curve shown in Figure 3. Because AUC is not misleading in the proper binormal model, the relationship between AUC and κ in this straightforward setting provides some direction in interpreting κ. For instance, in the proper binormal model, if AUC > 0.95 would be deemed excellent diagnostic accuracy, then $\kappa < 0.51$ would likewise be considered excellent. Understanding the matching between κ and AUC helps to provide guidance on how much κ is expected to be for previously-studied biomarkers, such as those reported in Table 2.9 in Zhou et al.[22] by relying on what is known about their AUC.

It need not always be the case that larger values of AUC pair with smaller values of κ, and there are actually situations for which AUC and κ may recommend different decisions, as will be seen in Examples 2 and 3.

Example 2. (Separation trap). Let's revisit the setting from Figure 1. The exact setup is

$$\begin{cases} f_D(y) = 1/2\phi_{\text{T}}(y \mid -6, -4, -5, 1/3^2) + 1/2\phi_{\text{T}}(y \mid 4, 6, 5, 1/3^2), \\ f_{\bar{D}}(y) = \phi_{\text{T}}(y \mid -2, 2, 0, 1/4^2). \end{cases}$$

Here $\phi_{\text{T}}(y \mid a, b, \mu, \sigma^2)$ is the density of a truncated normal with lower bound a and upper bound b. In this case it holds that

$$\begin{aligned} \kappa &= \int_{-\infty}^{\infty} \sqrt{f_D(y)}\sqrt{f_{\bar{D}}(y)}\,\mathrm{d}y \\ &= \int_{-6}^{-4} \sqrt{f_D(y)}\sqrt{f_{\bar{D}}(y)}\,\mathrm{d}y + \\ &\quad \int_{-2}^{2} \sqrt{f_D(y)}\sqrt{f_{\bar{D}}(y)}\,\mathrm{d}y + \int_{4}^{6} \sqrt{f_D(y)}\sqrt{f_{\bar{D}}(y)}\,\mathrm{d}y \\ &= 0. \end{aligned}$$

Thus, κ claims that both populations are perfectly separated—and so it would not fall into the separation trap. We briefly note that the separation trap shortcomings of AUC and the Youden Index would be eliminated by applying the (nonmonotonic) absolute value transformation to the biomarker. However, it would require knowledge of the underlying densities in order to determine this is an advantageous transformation for this instance. The summary measures we advocate require no such knowledge, which we count as a decided advantage. Dependence on such information becomes increasingly restrictive if the densities are covariate-dependent—such as in the setting to be discussed in Section 2.2—because a suitable transformation could also be covariate-dependent.

Table 1 contains the affinity for the bibeta and (potentially improper) bigamma models. Note that Dorfman et al.[23] consider the proper bigamma model with constant shape parameters. For completeness we include derivations of these expressions in the supplementary materials.

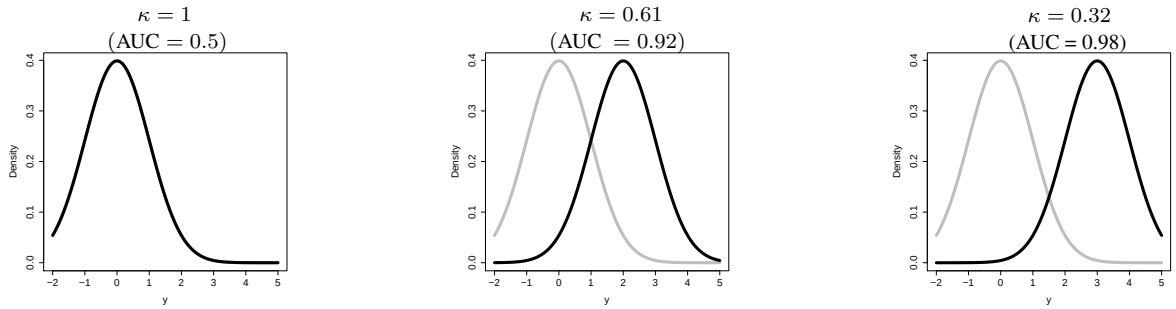


Figure 2. Affinity for binormal model from Example 1; the black and grey lines respectively denote the densities of the biomarkers of the diseased and non-diseased subjects; the configurations of parameters are as follows: a) $(\mu_D, \sigma_D) = (\mu_{\bar{D}}, \sigma_{\bar{D}}) = (0, 1)$; b) $(\mu_D, \sigma_D) = (2, 1)$ and $(\mu_{\bar{D}}, \sigma_{\bar{D}}) = (0, 1)$; c) $(\mu_D, \sigma_D) = (3, 1)$ and $(\mu_{\bar{D}}, \sigma_{\bar{D}}) = (0, 1)$.

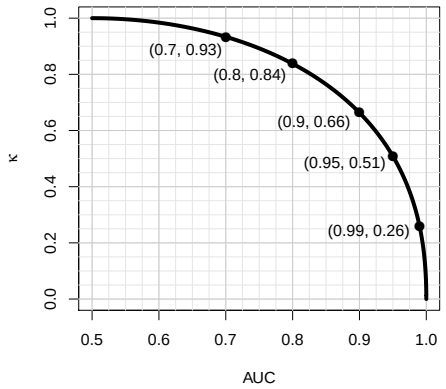


Figure 3. Correspondence between AUC and κ in the proper binormal model. Because AUC is a reasonable summary of biomarker accuracy in this model (and we argue κ always is), this figure can be used to build intuition on the interpretation of the magnitude of κ for those familiar with interpreting AUC.

Table 1. Affinity (κ) for bibeta, bigamma, and binormal models; here, α_D and $\alpha_{\bar{D}}$ are the shape parameters and β_D and $\beta_{\bar{D}}$ are the rate parameters of the corresponding gamma distributions

Model	Affinity
Bibeta	$\frac{B((a_D+a_{\bar{D}})/2,(b_D+b_{\bar{D}})/2)}{\{B(a_D,b_D)B(a_{\bar{D}},b_{\bar{D}})\}^{1/2}}$
Bigamma	$\left[\frac{{\beta_D}^{\alpha_D}{\beta_{\bar{D}}}^{\alpha_{\bar{D}}}}{\Gamma(\alpha_D)\Gamma(\alpha_D)}\right]^{1/2} \frac{\Gamma((\alpha_D+\alpha_{\bar{D}})/2)}{((\beta_D+\beta_{\bar{D}})/2)^{(\alpha_D+\alpha_{\bar{D}})/2}}$
Binormal	$\sqrt{\frac{2\sigma_D\sigma_{\bar{D}}}{\sigma_D^2+\sigma_{\bar{D}}^2}} \exp\left\{-\frac{1}{4}\frac{(\mu_D-\mu_{\bar{D}})^2}{\sigma_D^2+\sigma_{\bar{D}}^2}\right\}$

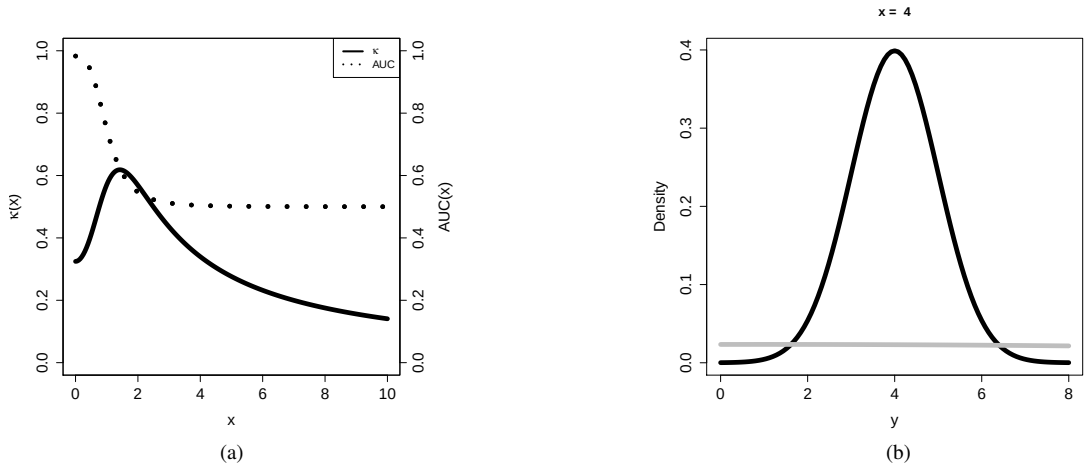
2.2 Properties and covariate-specific affinity

The following proposition documents two elementary properties associated with our measure of biomarker accuracy.

Proposition 1. *Affinity, as defined in* (3)*, obeys the following properties:*

1) $\kappa \in [0, 1]$.

2) κ *is invariant to monotone increasing data transformations.*



(a) (b)

Figure 4. a) Covariate-specific affinity (solid line) for binormal model from Example 3, and corresponding covariate-specific AUC (dotted line); b) Density of the diseased (black line) and non-diseased (grey line) subjects, for $x = 4$.

A proof of Proposition 1 can be found in the online supplementary materials. Interestingly, just like affinity, the AUC is also invariant to monotone increasing data transformations.[2] Affinity is also invariant to whether we work with a test for which larger values of the biomarker are more indicative of disease, or the other way around; this is an obvious consequence of the fact that $\langle\sqrt{f_D}, \sqrt{f_{\bar{D}}}\rangle = \langle\sqrt{f_{\bar{D}}}, \sqrt{f_D}\rangle$. Thus, for instance, binormal affinity in (4) is the same, regardless of whether $\mu_D > \mu_{\bar{D}}$ or vice versa. For the lack of better terminology, below we refer to an *upper-tailed biomarker* as one for which larger values of the biomarker are more indicative of disease, and to a *lower-tailed biomarker* as to one where larger values of the biomarker are less indicative of disease. Another parallel to the AUC is the fact that κ can be regarded as an area under a curve, with the curve of interest being

$$c(y) = \sqrt{f_D(y) f_{\bar{D}}(y)}.$$

Another interesting aspect is that κ can also be regarded as an average of the square root of a likelihood ratio, in the sense that

$$\kappa = \int_{-\infty}^{\infty} \sqrt{\frac{f_D(y)}{f_{\bar{D}}(y)}} f_{\bar{D}}(y)\,\mathrm{d}y = E_{\bar{D}}\left(\sqrt{\frac{f_D(Y_{\bar{D}})}{f_{\bar{D}}(Y_{\bar{D}})}}\right).$$

If covariates are available the question arises of how to conduct a covariate-specific analysis for measuring biomarker accuracy using affinity. A natural extension of (3) to the conditional setting is

$$\kappa(x) = \langle\sqrt{f_{D|x}}, \sqrt{f_{\bar{D}|x}}\rangle = \int_{-\infty}^{+\infty} \sqrt{f_D(y \mid x)}\sqrt{f_{\bar{D}}(y \mid x)}\,\mathrm{d}y, \tag{5}$$

where $x \in \mathcal{X} \subseteq \mathbb{R}^p$ is a covariate, $f_{D|x} = f_D(\cdot \mid x)$, and $f_{\bar{D}|x} = f_{\bar{D}}(\cdot \mid x)$. Below we refer to $\kappa(x)$ as the covariate-specific affinity. As with κ, it holds that $\kappa(x) \in [0, 1]$, and that $\kappa(x)$ is invariant to monotone increasing data transformations.

Example 3. (Binormal covariate-specific affinity). Extending Example 1 to allow covariate dependent densities, suppose that $f_D(y \mid x) = \phi(y \mid \mu_D(x), \sigma_D^2(x))$ and $f_{\bar{D}}(y \mid x) = \phi(y \mid \mu_{\bar{D}}(x), \sigma_{\bar{D}}^2(x))$. It then follows that

$$\kappa(x) = \sqrt{\frac{2\sigma_D(x)\sigma_{\bar{D}}(x)}{\sigma_D^2(x) + \sigma_{\bar{D}}^2(x)}} \exp\left\{-\frac{1}{4}\frac{\{\mu_D(x) - \mu_{\bar{D}}(x)\}^2}{\sigma_D^2(x) + \sigma_{\bar{D}}^2(x)}\right\}.$$

In particular, for $\mu_D(x) = x$ and $\mu_{\bar{D}}(x) = x - 3$, and $\sigma_D(x) = 1$ and $\sigma_{\bar{D}}(x) = 1 + x^2$, we obtain the covariate-specific affinity plotted in Figure 4(a). As it can be observed from Figure 4, for values of the predictor between 0 and approximately 1.2, both κ and AUC agree that the quality of the test deteriorates as x increases (κ increases and AUC decreases). As x increases beyond 1.2, each measure suggests a different conclusion as to how the test accuracy changes with x. To understand the reason for this, we

analyze in further detail the case of $x = 4$, whose corresponding densities are plotted in Figure 4(b). In the case $x = 4$ we have an AUC $= 0.504$ whereas $\kappa = 0.34$. Thus, on the one hand the AUC $= 0.504$ suggests that the test is quite poor, whereas the value of $\kappa = 0.34$ suggests that it could be satisfactory, though far from excellent. The intuition underlying this lack of agreement is as follows: κ is taking into account that around 95% of the mass for the test values for diseased subjects will be on the $[0, 8]$ interval, whereas around 95% of the mass for the test values of non-diseased subjects will be on the $[-30, 38]$ interval.

2.3 Nonparametric Bayesian inference for affinity and covariate-specific affinity

In this section we discuss Bayesian nonparametric estimators for affinity, as defined in (3), and covariate-specific affinity, as defined in (5). Let $\{Y_{D,i}\}_{i=1}^{n_D}$ and $\{Y_{\bar{D},i}\}_{i=1}^{n_{\bar{D}}}$ be random samples from F_D and $F_{\bar{D}}$. We propose to estimate κ in (3) by modeling each conditional density f_D and $f_{\bar{D}}$ as an infinite mixture model of the type

$$f(y) = \int_{\Theta} K(y \mid \theta)\, G(\mathrm{d}\theta), \tag{6}$$

where K is a kernel and G is a random mixing measure. The corresponding induced prior is

$$\kappa = \int_{-\infty}^{+\infty} \left\{ \int_{\Theta} K(y \mid \theta)\, G_D(\mathrm{d}\theta) \right\}^{1/2} \left\{ \int_{\Theta} K(y \mid \theta)\, G_{\bar{D}}(\mathrm{d}\theta) \right\}^{1/2} \mathrm{d}y. \tag{7}$$

A natural approach is to consider each G as a Dirichlet process,[24] and to rely on normal kernels, in which case (6) becomes a so-called Dirichlet process mixture of normal kernels,

$$\begin{aligned} f(y) &= \int_{\mathbb{R}\times(0,\infty)} \phi(y \mid \mu, \sigma^2)\, G(\mathrm{d}\mu, \mathrm{d}\sigma^2), \qquad G \sim \mathrm{DP}(\alpha, G_0), \\ &= \sum_{h=1}^{\infty} \pi_h \phi(y \mid \mu_h, \sigma_h^2). \end{aligned} \tag{8}$$

Here $\alpha > 0$ is the so-called precision parameter, G_0 is the centering distribution function, or baseline measure, and we use the notation $G \sim \mathrm{DP}(\alpha, G_0)$ to represent that G follows a Dirichlet process (DP). A celebrated representation of the DP is the so-called stick-breaking construction.[25] According to this representation a random distribution function G follows a DP if it admits a representation of the type

$$G = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h}, \quad \theta_h \overset{\text{iid}}{\sim} G_0,$$

where $\pi_1 = V_1$, and $\pi_h = V_h \prod_{k<h}(1 - V_k)$, with $V_h \overset{\text{iid}}{\sim} \mathrm{Beta}(1, \alpha)$, for $h = 2, \ldots$. The θ_h are known as atoms, the π_h as mixing weights, and the V_h are the so-called stick-breaking weights.

For regression data, $\{(x_{D,i}, Y_{D,i})\}_{i=1}^{n_D}$ and $\{(x_{\bar{D},i}, Y_{\bar{D},i})\}_{i=1}^{n_{\bar{D}}}$, we propose to estimate $\kappa(x)$ in (3) by modeling each density f_D and $f_{\bar{D}}$ as an infinite mixture model of regressions

$$f(y \mid x) = \int_{\Theta} K(y \mid \theta)\, G_x(\mathrm{d}\theta), \tag{9}$$

where K is a kernel and G_x is a covariate-specific random mixing measure. The corresponding induced prior is

$$\kappa(x) = \int_{-\infty}^{+\infty} \left\{ \int_{\Theta} K(y \mid \theta)\, G_{D,x}(\mathrm{d}\theta) \right\}^{1/2} \left\{ \int_{\Theta} K(y \mid \theta)\, G_{\bar{D},x}(\mathrm{d}\theta) \right\}^{1/2} \mathrm{d}y. \tag{10}$$

A natural approach is to consider each G_x as a dependent Dirichlet process (DDP),[26] and to rely on normal kernels in which case (9) becomes an infinite mixture of regression models,

$$f(y \mid x) = \int_{\mathbb{R}\times(0,\infty)} \phi(y \mid \mu, \sigma^2)\, G_x(\mathrm{d}\mu, \mathrm{d}\sigma^2). \tag{11}$$

Because of the support properties in Theorem 4 of Barrientos et al.,[27] we consider a 'single-weights' DDP[28,29]

$$G_x = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_{x,h}}.$$

The random support locations $\theta_{x,h}$ are, for $h = 1, 2, \ldots$ independent and identically distributed realizations from a stochastic process over the covariate space $\mathcal{X}$ and the weights $\{\pi_h\}_{h=1}^{\infty}$ match those from a standard DP; in this specific version of (2.3) we obtain

$$f(y \mid x) = \sum_{h=1}^{\infty} \pi_h \phi(y \mid \mu_h(x), \sigma_h^2). \tag{12}$$

To achieve a reasonable tradeoff between flexibility and parsimony, in practice we choose to model $\mu(x)$ as a linear model, that is, $\mu_h(x) = B_x^{\mathrm{T}}\beta_h$, where $B_x^{\mathrm{T}} \equiv B^{\mathrm{T}}$ corresponds to the cubic B-spline basis evaluated at the predictor. Finally, to facilitate prior specification we suggest *standardizing* the biomarkers (i.e., $Z_{Di} = (Y_{Di} - \bar{Y}_D)/s_D$ and $Z_{\bar{D}j} = (Y_{\bar{D}j} - \bar{Y}_{\bar{D}})/s_{\bar{D}}$) and rescaling the covariate (i.e., $\min\{x_{\bar{D}}, x_D\} = -1$ and $\max\{x_{\bar{D}}, x_D\} = 1$). Having estimated the densities on the standardized data, the location-scale adjustment may be applied to easily convert to densities for the untransformed data.

We now present a specific embodiment of our model. Let $B_{\bar{D}i}^{\mathrm{T}}$ represent a q-vector with the cubic B-spline representation of $x_{\bar{D}i}$, with $x_{\bar{D}i}$ having been rescaled to lie in $[-1, 1]$. The assumptions for the non-diseased population in the conditional case are that

$$\begin{aligned}
f_{\bar{D}}(Z_{\bar{D}i} \mid x_{\bar{D}i}) &= \int \phi(Z_{\bar{D}i} \mid B_{\bar{D}i}^{\mathrm{T}}\beta, \sigma^2)\, \mathrm{d}G_{\bar{D}}(\beta, \sigma^2) \\
G_{\bar{D}}(\beta, \sigma^2) \mid G_{\bar{D}0}(\beta, \sigma^2) &\sim \mathrm{DP}(1, G_{\bar{D}0}(\beta, \sigma^2)) \\
G_{\bar{D}0}(\beta, \sigma^2) &\equiv \mathrm{N}(\beta_{\bar{D}0}, \Sigma_{\bar{D}0}) \times \mathrm{IG}(\text{shape} = 1, \text{rate} = 50) \\
\beta_{\bar{D}0} &\sim \mathrm{N}(0, I) \\
\Sigma_{\bar{D}0} &\sim \mathrm{IWish}(\nu = q, \text{scale} = qI),
\end{aligned}$$

where IG and IWish respectively denote the inverse Gamma and inverse Wishart distributions. Two aspects of this specification are particularly noteworthy. First, it is assumed that the number and locations of all knots are known, although this could be relaxed. Second, the prior on the within-cluster variance (i.e., σ^2) was chosen to favor variances much less than one. The justification for this is immediate when recognizing that the likelihood is on standardized data with a marginal sample variance of one; the within-cluster variance ought to be substantially smaller than the marginal variance. The assumptions are analogous for the diseased population; the only difference is the substitution of D for $\bar{D}$. To apply the model specification without conditioning on any covariate, we can simply set $B_{\bar{D}i}^{\mathrm{T}} = 1$. While the discussion above focuses on the single covariate setup, the B-spline specification for $\mu_h(x)$ can be easily extended to the p covariate setting via a generalized additive model.[8]

2.4 Theoretical properties on induced priors

This section includes theoretical properties on the induced priors for the summary measures introduced in Section 2.1. We recall that the support of a random probability measure consists of the set of all elements for which every open neighborhood has positive probability;[30] depending on the metric defining the neighborhood, the support has a different name (e.g. one refers to Hellinger support if neighborhoods are yielded via the Hellinger distance). Priors with a large support are desirable as they are not too concentrated in a certain specific region of the parameter space.

Although in practice we model the densities from which κ is estimated with a Dirichlet process mixture, as in (8) and (12), below we document theoretical results which apply more generally to (6) and (9) and only require that the mixing distribution has a full weak support, which includes the Dirichlet process as a particular case. In what follows, we assume the same setting as in Lijoi et al.,[31] namely:

A_1) The random mixing distribution(s) has (have) full weak support.

A_2) $\int_{-\infty}^{+\infty} K(y \mid \theta)\, \mathrm{d}y = 1$, for $\theta \in \Theta$.

A_3) $\theta \mapsto K(y \mid \theta)$ is bounded, continuous, and $\mathbb{B}_\Theta$-measurable for $y \in \mathbb{R}$.

A_4) The family of mappings $\{\theta \mapsto K(y \mid \theta) : y \in C\}$, is uniformly equicontinuous, for every compact $C \subset \mathbb{R}$.

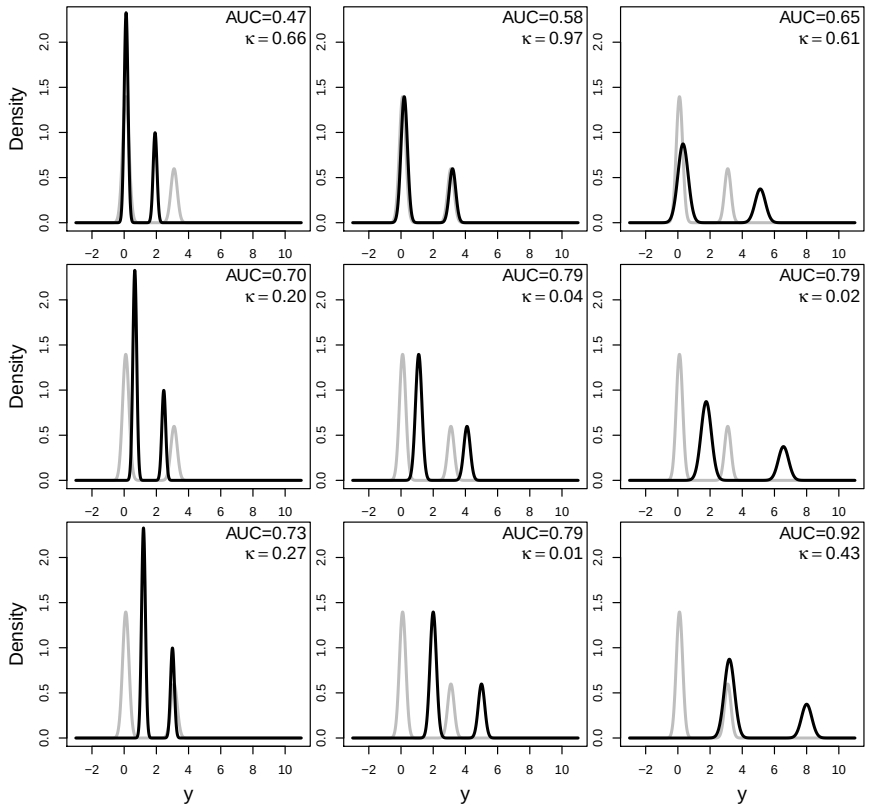


Figure 5. Densities for the second unconditional simulation study setting in Table 2; the black and grey lines respectively denote the densities of the biomarkers of the diseased and non-diseased subjects.

Here, A_1 is a condition on the support of the mixing, whereas A_2–A_4 are regularity conditions on the kernel. Under these conditions, it can be shown that $f(y)$ in (6) has full Hellinger support.[31] As a consequence, the following result holds.

Theorem 1. *Suppose A_1–A_4 and let $(\Omega, \mathcal{A}, P)$ be the probability space associated with the infinite mixture model in (6), which induces $\kappa = \int \sqrt{f_D(y)}\sqrt{f_{\bar{D}}(y)}\,\mathrm{d}y$. Let κ^ω be a realization of the κ index under (6). Then, for every $\varepsilon > 0$, it holds that $P\{\omega \in \Omega : |\kappa^\omega - \kappa| < \varepsilon\} > 0$.*

Under the same conditions as above, it can be shown that $f(y \mid x)$ in (9) has full Hellinger support.[27] Thus, the following analogous result to Theorem 1 holds for the covariate-specific version of our summary measure as defined in (5).

Theorem 2. *Suppose A_1–A_4 and let $(\Omega, \mathcal{A}, P)$ be the probability space associated with the infinite mixture of regression models in (9), which induces $\kappa(x) = \int \sqrt{f_D(y \mid x)}\sqrt{f_{\bar{D}}(y \mid x)}\,\mathrm{d}y$. Let $\kappa^\omega(x)$ be a trajectory of covariate-specific affinity $\kappa(x)$ under (9). Then, for $x_1, \ldots, x_n \in \mathcal{X}$, for every positive integer n and $\varepsilon > 0$, it holds that $P\{\omega \in \Omega : |\kappa^\omega(x_i) - \kappa(x_i)| < \varepsilon,\ i = 1, \ldots, n\} > 0$.*

Proofs of Theorems 1 and 2 can be found in the online supplementary materials.

3 Simulation study

3.1 Data generating processes

The simulation settings are summarized in Table 2. The simulation employed pairs of biomarker distributions that were either conditional on a single uniformly distributed covariate or were unconditional. In the unconditional settings, each distribution was either normal or a mixture of normals, and the means and standard deviations were systematically altered so that a range of κ and AUC values were considered. In terms of the conditional setting, we consider the same scenarios as in Inácio de Carvalho

Table 2. Simulation study settings

Scenario	Non-Diseased ($f_{\bar{D}}$)	Diseased (f_D)	Notes†
Unconditional #1	$\phi(.4, .8)$	$\phi(\mu_D, \sigma_D)$	1)
Unconditional #2	$.7\phi(.1, .2) + .3\phi(3.1, .2)$	$.7\phi(\mu_{1D}, \sigma_D) + .3\phi(\mu_{2D}, \sigma_D)$	2)
Conditional #1	$\phi(.5 + x_{\bar{D}}, 1.5)$	$\phi(2 + 4x_D, 2)$	3)
Conditional #2	$\phi(\sin(\pi(x_{\bar{D}} + 1)), .5)$	$\phi(.5 + x_D^2, 1)$	3)
Conditional #3	$\phi(\sin(\pi x_{\bar{D}}), \sqrt{.2 + .5\exp(x_{\bar{D}})})$	$(1 + \exp(-x))^{-1}\phi(x_D, .5) + (1 + \exp(x))^{-1}\phi(x_D^3, 1)$	3)

† Each unconditional setting has nine distinct pairs of means and standard deviations where for 1): μ_D in $\{.8, 1.6, 3.2\}$ and σ_D in $\{.8, 1.2, 1.6\}$ and 2): $\sigma_D = .2c$ and (μ_{1D}, μ_{2D}) in $\{(.2c, 3.2c), (1.1c, 4.1c), (2c, 5c)\}$, with c in $\{.6, 1, 1.6\}$. For the conditional setting we have 3): $x_{\bar{D}}, x_D \overset{\text{iid}}{\sim} \text{Unif}(-1, 1)$.

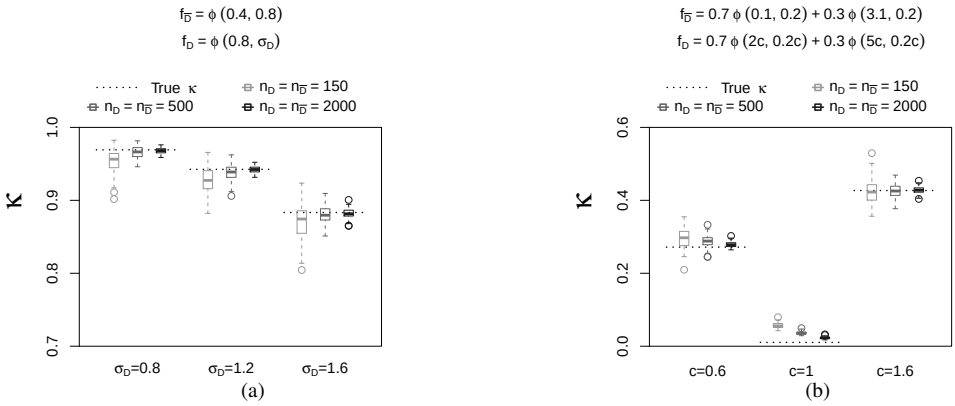


Figure 6. κ estimates (average across 100 simulations) along with true values in the unconditional scenarios of the simulation study (Table 2): a) the first unconditional setting (normals); b) the second unconditional setting (mixtures of normals).

et al.[8] because these scenarios were originally constructed for an investigation on a related topic: the performance of a Bayesian nonparametric model for ROC estimation with covariate-dependent biomarker distributions. The conditional distributions were normal or a mixture of normals, and the covariate's effect on the mean and standard deviation were modeled according to varying levels of complexity. We note that our simulation study includes cases leading where the true ROC curve is improper (e.g., Unconditional # 1, say $\sigma_D = 1.2$). We consider such settings for full generality but note that while they may be meaningful from a conceptual perspective, they may not necessarily be sensible for many practical settings—as they would correspond to locally worse than chance performance.[21,22,32–34]

Figure 5 depicts the density pairs from the second unconditional setting; the plots for the remaining scenarios are included in the supplementary materials. Of particular note is the pattern of possibilities for κ and AUC when the biomarker densities are mixtures of normals. In particular, the middle plot in Figure 5 displays a situation where κ is particularly adept at identifying the distinctiveness of the diseased and non-diseased populations as can be seen by the very small κ value. However, our convention that AUC be computed assuming the biomarker will be one-sided forces the AUC to be lower than might be expected given the distinctiveness of the populations. (By "one-sided" we refer to the situation when a positive test region is comprised either of all values greater than some threshold c—an "upper-tailed" test—or of all values less than c—a "lower-tailed" test—thus prohibiting noncontiguous positive test regions.) While it is certainly possible to entertain more flexible regions at which the biomarker would be considered to have a positive result, this would require another nontrivial step before AUC could even be calculated, whereas such a step is not needed to calculate κ.

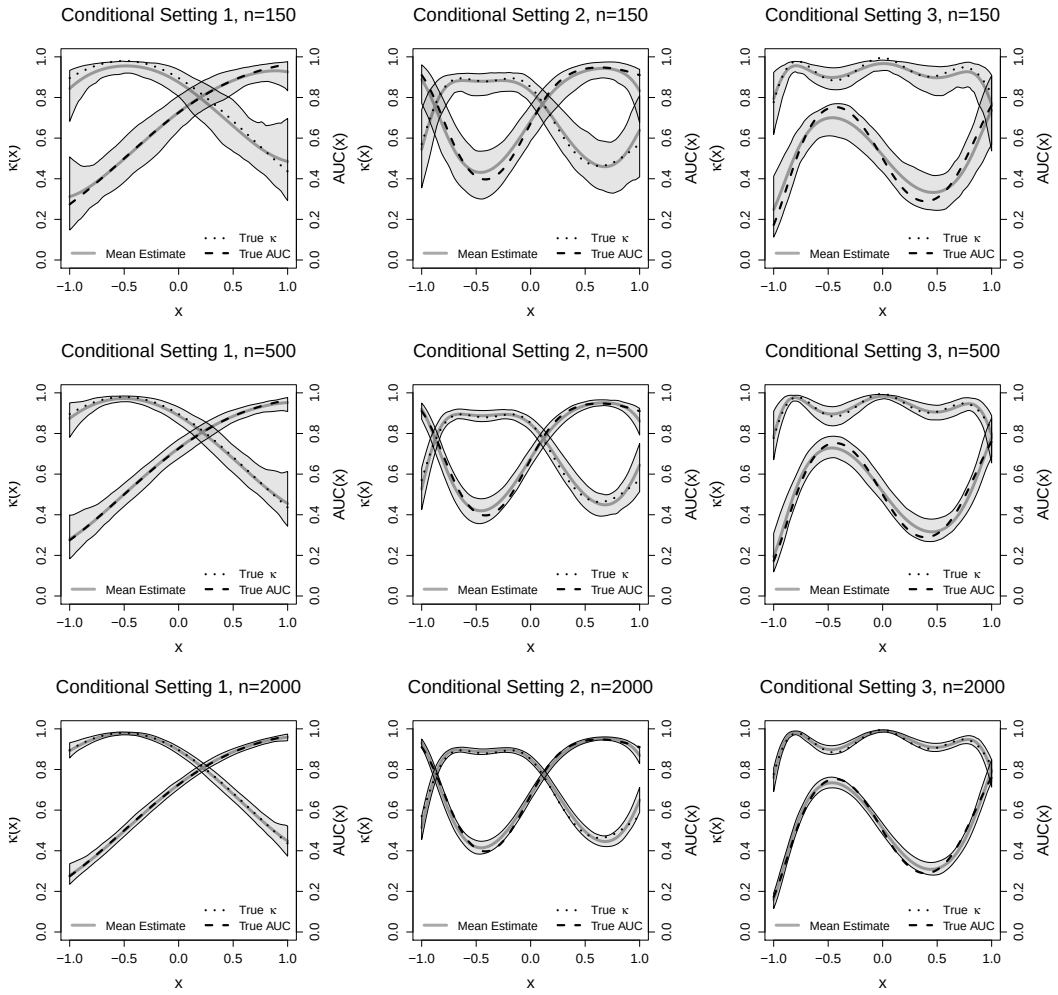


Figure 7. Estimated covariate-specific affinity, $\kappa(x)$, and covariate-specific AUC, AUC(x), across the 100 simulated data sets for the conditional settings described in Table 2. The bands represent the pointwise empirical 2.5th and 97.5th percentiles of the 100 point estimates, while the dark grey lines represent the average of the 100 estimates.

3.2 Monte Carlo simulation study

For each setting in Table 2, we generated 100 data sets from f_D and from $f_{\bar{D}}$. The sample sizes were varied at $n_D = n_{\bar{D}} = 150$, 500, or 2000 to provide some sense of how reliably κ and AUC were estimated in moderate to large samples. In implementing the model, no additional knots for the cubic B-splines were included; this lets us ascertain the covariate-dependent model's flexibility in the absence of extra knots. Additionally, because the covariate values were simulated from the Unif(-1,1) distribution, we did not rescale the covariate prior to computing the B-spline representation. For each synthetic data set κ was estimated by collecting 300 MCMC iterates after a burn in of 2000 and thinning of 40. The Blocked Gibbs sampler[35,36] was employed setting the upper bound of mixture components to 20. All computation was carried out using the ROCstudio package that can be executed in the statistical software R.[37]

Selected results from the unconditional settings are summarized in Figure 6, which depicts the Monte Carlo average (across 100 simulations) of the estimated values for κ, along with the actual values. In part a), which was characterized by each population having a normal distribution, κ is estimated with little bias. Not surprisingly, the bias is reduced by having larger sample sizes. In part b), which was characterized by each population having a mixture of normals distribution, the same pattern was exhibited, but as the true κ approaches 0 more bias tends to be induced. The supplementary material contains the remainder of the simulation results where this is further illustrated.

The results from the conditional settings are summarized in Figure 7. For each of the 100 simulated data sets, the conditional means for κ and AUC were estimated at values of x ranging from -1 to 1. The pointwise averages of the 100 estimated means are plotted in this figure, as well as the 2.5th and 97.5th empirical percentiles of these estimated means. This gives some sense for how variable the estimates are (primarily attributable to differences between the 100 simulated data sets). Point estimates of AUC(x) and $\kappa(x)$ were quite successful in estimating the corresponding true values. Predictably, the estimates exhibited less variability as more data were available. Recall that a strength of $\kappa(x)$ is that it is not susceptible to the separation trap, nor does it require us to distinguish between upper- and lower-tailed biomarkers. This distinction for AUC(x) explains why the AUC is sometimes estimated to be well below 0.5. Given these advantages of $\kappa(x)$ over AUC(x), it is even more notable that $\kappa(x)$ can be reliably estimated. An important collateral suggestion of the simulation is that the model is quite flexible even if the cubic B-spline basis does not include additional knots, though of course knots may be added if desired.

4 Revisiting a prostate cancer diagnosis study

We now turn our attention to an application that has been regularly employed to demonstrate biomarker accuracy that is covariate-dependent.

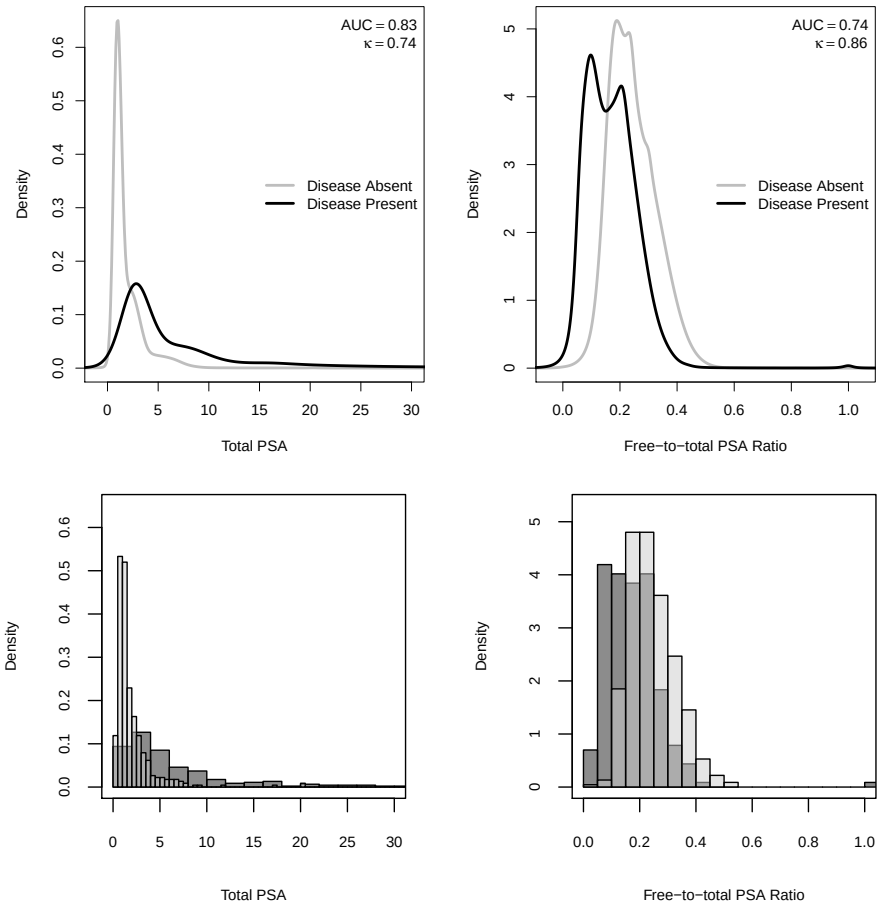


Figure 8. Top: DPM-based estimated densities along with AUC and κ values when age is not considered. The black and grey lines respectively denote the densities of the biomarkers of the diseased and non-diseased subjects. Bottom: Overlapping histograms. For display purposes, the largest total PSA values ($< 3\%$ of the $n = 683$ observations) are not shown in the leftmost plots. All total PSA values were below 100.

4.1 Study data and preliminary considerations

The data were gathered from the Beta-Carotene and Retinol Efficacy Trial (CARET)—a lung cancer prevention trial, conducted at the Fred Hutchinson Cancer Research Center. During this study longitudinal measurements of two Prostate Specific Antigen (PSA)-based biomarkers were collected for 71 prostate cancer cases and 70 controls. The biomarker measurements were taken on males between 46 and 80 years old. The number of repeated measures per subject ranged from one to nine, with $n = 683$ total observations. Further details on this study can be found, for instance, in Etzioni et al.[38] and Pepe.[2] To make our inferences directly comparable with those of Rodriguez and Martinez[12]—who consider a Gaussian process prior-based model for AUC(x)—we follow the latter authors and ignore the longitudinal nature of the data; however, for reference, we also include in the supplementary materials the results from restricting analysis to each subject's last available observation. A test based on total PSA concentration (Biomarker 1, ng/ml) was assumed to have a positive test result if the measurement was sufficiently large. Conversely, a test based on the free-to-total PSA ratio (Biomarker 2, f/t) was assumed to have a positive test result if the measurement was sufficiently small.

The direction of the tendency is of no consequence in estimating κ which provides an intrinsic advantage relative to AUC. In estimating AUC, we must consider the direction of the biomarker, that is whether larger values of the biomarker are more indicative of disease or the other way around. A main goal below will be on illustrating how the proposed methods can be used to assess which biomarker might screen better for prostate cancer.

4.2 PSA-based analysis

We first fit the unconditional model (i.e., sans covariate so that $x_i^{\mathrm{T}} = 1$) by collecting 1 800 MCMC iterates after a burn in of 20 000 and thinning of 100. To visualize differences between the biomarkers we provide Figure 8. For each biomarker, the estimated density among cases and controls are superimposed. It is readily apparent that there are differences between cases and controls, and that the direction of the differences depends on which biomarker we consider. Both univariate summaries, κ and AUC, signal a preference for the first biomarker as a screening mechanism. The 95% credible interval of κ associated with total PSA concentration is (0.69, 0.78) and for AUC is (0.80, 0.87), while for free-to-total PSA ratio the interval for κ is (0.82, 0.90) and for AUC (0.70, 0.78) respectively.

4.3 PSA-based analysis with age-adjustment

It is well known that PSA levels may be age-dependent—for both diseased and non-diseased subjects—since both benign prostate conditions and prostate cancer become more common with age. With this in mind, we obtained conditional density estimates for each biomarker in each population to estimate $\kappa(\text{age})$ and AUC(age) by fitting the conditional model and collecting 1 800 MCMC iterates after a burn in of 20 000 and thinning of 100 and using the same specifications as before. In model fitting, the patients' ages were first rescaled from the interval $[46.75, 80.83]$ to the interval $[-1, 1]$, and, following numerical evidence from Inácio de Carvalho et al.[39] (Section 3), we elected to not to include any additional knots in the cubic B-splines.

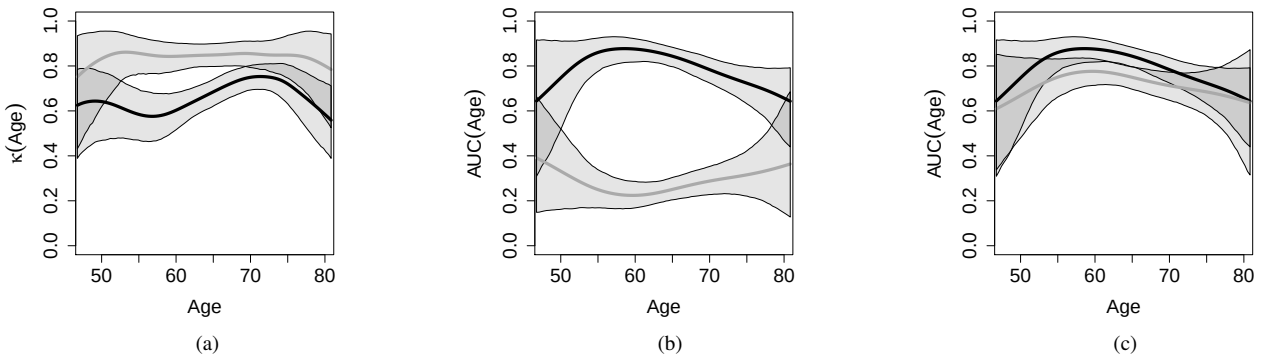


Figure 9. Means and 95% pointwise credible intervals for the age-adjusted affinity and AUC of two biomarkers in cases and controls. a) is the age-adjusted affinity; b) is the age-adjusted AUC if both biomarkers have upper-tailed biomarkers; c) is the age-adjusted AUC if the second biomarker biomarker is lower-tailed. In each panel, the black and grey lines respectively denote the first and second biomarkers.

Figure 9 displays the posterior mean and pointwise 95% credible intervals for κ and AUC as a function of age. Notice first that for total PSA concentration our estimated AUC(age) is very similar to that found in Figure 4 of Rodriguez and Martinez,[12] with the largest discriminatory power occurring when an individual is in their late 50s. Regarding comparisons with κ, generally speaking total PSA concentration exhibits less affinity than free-to-total PSA ratio between the distributions of those with and without a prostate cancer diagnosis. This suggests that a biomarker based on total PSA concentration would be preferred to a test based on free-to-total PSA ratio. The first biomarker's affinity appears to be sensitive to the subject's age. The AUC seems to indicate that total PSA concentration is a reasonably good diagnostic biomarker, while κ seems to be even more optimistic regarding the test's ability ($\kappa \approx 0.6$); a similar conclusion holds for free-to-total PSA ratio. In addition, κ more clearly identifies the difference in screening ability of the two biomarkers for males aged 55 to 70. Furthermore, it is invariant to whether the biomarker is assumed to be lower- or upper-tailed.

Finally, both analyses suggest total PSA concentration is a better alternative than free-to-total PSA ratio in screening older males with lung cancer for prostate cancer. This latter conclusion is supported even more emphatically by $\kappa(\text{age})$ than by $\text{AUC}(\text{age})$, as seen in Figure 9.

5 Discussion

In this paper we show how Hellinger affinity can be used as a natural summary measure for assessing the performance of a biomarker. The summary measure has several desirable properties that motivate its use as a supplement, if not competitor, to other existing summaries such as AUC and the Youden index. Affinity shares some of the properties of the AUC—such as invariance to monotone increasing transformations—, but it does not fall into the separation trap, whereas both the AUC and the Youden index would. Indeed, a principal advantage of κ is that it is readily calculated and interpreted without assuming anything about the biomarker threshold(s) that demarcate positive and negative test diagnoses. This can be especially beneficial if, for instance, a biomarker's distribution when the disease is present favors both atypically low and atypically high values. Affinity-based measures can be framed into the same geometrical principles as Pearson correlation, and they focus on the overlap between f_D and $f_{\bar{D}}$ rather than on always presuming that larger values of a biomarker are more indicative of disease. Nonparametric Bayes estimators for affinity and covariate-specific affinity are discussed, and theoretical properties of the corresponding priors have been derived. While it could be natural to fit parametric models such as the ones in Table 1, the added flexibility of the proposed inferences allows us to model biomarker accuracy in a way that offers flexibility and robustness against misspecification. The proposed methods can be readily applied to the case of data with ties, including situations where data are discrete or ordinal. Indeed, the proposed construct can be readily applied to discrete or ordinal data defined over a set $\mathcal{Y}$ by considering the discrete affinity as

$$\kappa = \sum_{y \in \mathcal{Y}} \sqrt{f_D(y)} \sqrt{f_{\bar{D}}(y)}.$$

For discrete data, one would need however to consider in Equations (7) and (10) kernels with a discrete support.

While not explored here, our summary measure has the potential to be applied to the more general setting where $p > 1$ biomarkers per subject are available. Indeed, if $f_D(y)$ and $f_{\bar{D}}(y)$ denote the joint distributions of the p biomarkers, for diseased and non-diseased subjects, similarly to (3) one can define

$$\kappa = \langle \sqrt{f_D}, \sqrt{f_{\bar{D}}} \rangle = \int_{\mathbb{R}^p} \sqrt{f_D(y)} \sqrt{f_{\bar{D}}(y)} \, \mathrm{d}y.$$

However, estimation of κ would become more challenging in the multivariate case than in the univariate case presented in this article. Future work could also entail nonparametric Bayesian inference for a covariate-specific version of the so-called overlap coefficient[5] which can be defined as

$$\text{OVL}(x) = \int_{-\infty}^{\infty} \min\{f_D(y \mid x), f_{\bar{D}}(y \mid x)\} \, \mathrm{d}y.$$

The index in (5) would quantify the proportion of overlap area between $f_D(y \mid x)$ and $f_{\bar{D}}(y \mid x)$, and thus it could be used as a companion to $\text{AUC}(x)$ and $\kappa(x)$. Finally, another direction which we may revisit in future work rests on the study of $\kappa(x)$ and $\text{OVL}(x)$ on settings where a gold standard test is unavailable.

Funding

This work was partially supported by FCT (Fundação para a Ciência e a Tecnologia, Portugal), through the projects PTDC/MAT-STA/28649/2017 and UID/MAT/00006/2019. Part of the research was conducted while the second author was at Brigham Young University.

Supplemental material

Supplementary materials are available and include proofs of theoretical results, derivations for entries in Table 1, simulation setting figures, and a supplement to the data analysis.

References

1. Youden W. Index for rating diagnostic tests. *Cancer* 1950; 3(1): 32–35.
2. Pepe M. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. New York: Oxford University Press, 2003.
3. Parodi S, Pistoia V and Muselli M. Not proper ROC curves as new tool for the analysis of differentially expressed genes in microarray experiments. *BMC bioinformatics* 2008; 9(1): 410.
4. Silva-Fortes C, Turkman MAA and Sousa L. Arrow plot: A new graphical tool for selecting up and down regulated genes and genes differentially expressed on sample subgroups. *BMC Bioinformatics* 2012; 13(1): 147.
5. Wang D and Tian L. Parametric methods for confidence interval estimation of overlap coefficients. *Computational Statistics and Data Analysis* 2017; 106: 12–26. DOI:10.1016/j.csda.2016.08.013.
6. Lee WC and Hsiao C. Alternative summary indices for the receiver operating characteristic curve. *Epidemiology* 1996; 7(6): 605–611.
7. Williams D. *Probability with Martingales*. Cambridge, UK: Cambridge University Press, 1991.
8. Inácio de Carvalho V, Jara A, Hanson T et al. Bayesian nonparametric ROC regression modeling. *Bayesian Analysis* 2013; 8(3): 623–646.
9. Erkanli A, Sung M, Jane Costello E et al. Bayesian semi-parametric ROC analysis. *Statistics in Medicine* 2006; 25(22): 3905–3928. DOI: 10.1002/sim.2496.
10. Gu J, Ghosal S and Roy A. Bayesian bootstrap estimation of ROC curve. *Statistics in Medicine* 2008; 27(26): 5407–5420. DOI: 10.1002/sim.3366.
11. Branscum AJ, Johnson WO and Baron AT. Robust medical test evaluation using flexible bayesian semiparametric regression models. *Epidemiology Research International* 2013; Article ID 131232.
12. Rodriguez A and Martinez J. Bayesian semiparametric estimation of covariate-dependent ROC curves. *Biostatistics* 2014; 15(2): 353–369. DOI:10.1093/biostatistics/kxt044.
13. Inácio de Carvalho V, Jara A and de Carvalho M. Bayesian nonparametric approaches for ROC curve inference. In Mitra R and Müller P (eds.) *Nonparametric Bayesian Inference in Biostatistics*. Cham: Springer. ISBN 978-3-319-19517-9 978-3-319-19518-6, 2015. pp. 327–344. DOI:10.1007/978-3-319-19518-6_16.
14. Branscum A, Johnson W, Hanson T et al. Flexible regression models for ROC and risk analysis, with or without a gold standard. *Statistics in Medicine* 2015; 34(30): 3997–4015.
15. Johnson W and de Carvalho M. Bayesian nonparametric biostatistics. In Mitra R and Müller P (eds.) *Nonparametric Bayesian Inference in Biostatistics*. Cham: Springer, 2015. pp. 15–54.
16. Inácio de Carvalho V, de Carvalho M, Alonzo T et al. Functional covariate-adjusted partial area under the specificity-ROC curve with an application to metabolic syndrome diagnosis. *Annals of Applied Statistics* 2016; 10(3): 1472–1495.
17. Inácio de Carvalho V and Branscum AJ. Bayesian nonparametric inference for the three-class Youden index and its associated optimal cutoff points. *Statistical Methods in Medical Research* 2018; 27: 0962280217742538.
18. de Carvalho M, Page GL and Barney BJ. On the geometry of Bayesian inference. *Bayesian Analysis* 2018, in press; .
19. Bhattacharyya A. On a measure of divergence between two multinomial populations. *Sankhyā: the indian journal of statistics* 1946; : 401–406.
20. van der Vaart A. *Asymptotic Statistics*. Cambridge, UK: Cambridge University Press, 1998. ISBN 978-0-521-49603-2.
21. Hillis SL and Berbaum KS. Using the mean-to-sigma ratio as a measure of the improperness of binormal ROC curves. *Academic radiology* 2011; 18(2): 143–154.
22. Zhou XH, McClish DK and Obuchowski NA. *Statistical Methods in Diagnostic Medicine*. 2nd ed. ed. New York: Wiley, 2011.
23. Dorfman DD, Berbaum KS, Metz CE et al. Proper receiver operating characteristic analysis: The bigamma model. *Academic radiology* 1997; 4(2): 138–149.
24. Ferguson T. A Bayesian analysis of some nonparametric problems. *Annals of Statistics* 1973; 1: 209–230.
25. Sethuraman J. A constructive definition of Dirichlet priors. *Statistica Sinica* 1994; 4: 639–650.
26. MacEachern S. Dependent Dirichlet processes. *Unpublished manuscript, Department of Statistics, The Ohio State University* 2000; .
27. Barrientos A, Jara A and Quintana F. On the support of MacEachern's dependent Dirichlet processes and extensions. *Bayesian Analysis* 2012; 7: 277–310. DOI:10.1214/12-BA709.
28. De Iorio M, Müller P, Rosner G et al. An ANOVA model for dependent random measures. *Journal of the American Statistical Association* 2004; 99(465): 205–215. DOI:10.1198/016214504000000205.

29. De Iorio M, Johnson W, Müller P et al. Bayesian nonparametric nonproportional hazards survival modeling. *Biometrics* 2009; 65(3): 762–771. DOI:10.1111/j.1541-0420.2008.01166.x.
30. Ghosal S and Van der Vaart A. *Fundamentals of Nonparametric Bayesian Inference*, volume 44. Cambridge, UK: Cambridge University Press, 2017.
31. Lijoi A, Prünster I and Walker S. Extending Doob's consistency theorem to nonparametric densities. *Bernoulli* 2004; 10(4): 651–663. DOI:10.3150/bj/1093265634.
32. Pan X and Metz CE. The "proper" binormal model: parametric receiver operating characteristic curve estimation with degenerate data. *Academic Radiology* 1997; 4(5): 380–389.
33. Wagner RF, Metz CE and Campbell G. Assessment of medical imaging systems and computer aids: a tutorial review. *Academic Radiology* 2007; 14(6): 723–748.
34. Bandos AI, Guo B and Gur D. Estimating the area under roc curve when the fitted binormal curves demonstrate improper shape. *Academic Radiology* 2017; 24(2): 209–219.
35. Ishwaran H and James LF. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* 2001; : 161–173.
36. Ishwaran H and James LF. Approximate Dirichlet Process Computing in Finite Normal Mixtures: Smoothing and Prior Information. *Journal of Computational and Graphical Statistics* 2002; 11(3): 508–532. DOI:10.1198/106186002411.
37. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL https://www.R-project.org/.
38. Etzioni R, Pepe M, Longton G et al. Incorporating the time dimension in receiver operating characteristic curves: A case study of prostate cancer. *Medical Decision Making* 1999; 19(3): 242–251. DOI:10.1177/0272989X9901900303.
39. Inácio de Carvalho V, de Carvalho M and Branscum A. Nonparametric Bayesian covariate-adjusted estimation of the Youden index. *Biometrics* 2017; 73: 1279–1288. DOI:10.1111/biom.12686.