Received: 22 July 2019 | Accepted: 24 August 2019
DOI: 10.1002/sta4.249

Check for updates

ORIGINAL ARTICLE

WILEY

Inference in receiver operating characteristic surface analysis via a trinormal model-based testing approach

Samuel Noll[1] | Reinhard Furrer[1,2] | Benjamin Reiser[3] | Christos T. Nakas[4,5]

[1] Department of Mathematics, University of Zurich, Zurich, Switzerland

[2] Department of Computational Science, University of Zurich, Zurich, Switzerland

[3] Department of Statistics, University of Haifa, Haifa, Israel

[4] Department of Agriculture, Crop Production and Rural Environment, University of Thessaly, Volos 38446, Greece

[5] Department of Clinical Chemistry, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland

Correspondence
Christos T. Nakas, School of Agriculture, Laboratory of Biometry, University of Thessaly, Volos 38446, Greece.
Email: cnakas@uth.gr

Funding information
Israel Science Foundation, Grant/Award Number: 387/15

Receiver operating characteristic (ROC) analysis is the methodological framework of choice for the assessment of diagnostic markers and classification procedures in general, in both two-class and multiple-class classification problems. We focus on the three-class problem for which inference usually involves formal hypothesis testing using a proxy metric such as the volume under the ROC surface (VUS). In this article, we develop an existing approach from the two-class ROC framework. We define a hypothesis-testing procedure that directly compares two ROC surfaces under the assumption of the trinormal model. In the case of the assessment of a single marker, the corresponding ROC surface is compared with the chance plane, that is, to an uninformative marker. A simulation study investigating the proposed tests with existing ones on the basis of the VUS metric follows. Finally, the proposed methodology is applied to a dataset of a panel of pancreatic cancer diagnostic markers. The described testing procedures along with related graphical tools are supported in the corresponding R-package `trinROC`, which we have developed for this purpose.

KEYWORDS

Box–Cox transformation, Delta method, pancreatic cancer biomarkers, ROC analysis, trinormal ROC model, volume under the ROC surface (VUS)

1 | INTRODUCTION

In the general two-class diagnostic problem, consider a set of individuals that belongs to one of two diagnostic classes, for example, either to the nondiseased group, denoted by D^-, or to the diseased group, D^+. A (bio)marker or, in general, a classifier results in a measurement for each individual on a single univariate continuum $X=x$. By convention, without loss of generality, higher marker values are considered to be associated with higher disease risk. On the basis of a specific cut-off point c (also referred to as a threshold value in the literature), on the continuum of marker measurements, an individual with $x \leq c$ is allocated to the nondiseased group; otherwise, (when $x>c$) to the diseased group. We assume that a reference standard is available that provides accurate information on the true status (i.e., diseased and nondiseased) of each individual in the study.

The receiver operating characteristic (ROC) curve depicts and summarizes the overlap of the distributions of the individuals from the two diagnostic classes. Based on the reference standard, the ROC curve is defined in the unit square as the curve connecting the points of the misclassification probability of a nondiseased individual (1 − specificity) on the x-axis against the probability of a diseased individual being correctly classified (sensitivity) on the y-axis, for each cut-off point c, as c varies in the diagnostic marker's measurements support. For an uninformative marker, that is, distributions of the nondiseased and diseased individuals with quasicomplete overlap, we obtain a curve that is close to the main

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2020 The Authors. Stat published by John Wiley & Sons Ltd

source: https://doi.org/10.7892/boris.148631 | downloaded: 27.12.2020

diagonal of the unit square (also called the chance diagonal). If a marker results in complete separation between the two distributions, we obtain an ROC curve that passes through the point with coordinates (0,1) in the ROC space (i.e., the unit square). The area under the ROC curve (AUC) is an overall index of the diagnostic accuracy of the marker under study and can be obtained by integrating the function of the ROC curve, $\mathrm{AUC} = \int_0^1 G_+(F_-^{-1}(1-t))\,dt$, where F_ℓ is the cumulative distribution function (cdf) of class D^ℓ, where ℓ=−,+ and G_ℓ=1− F_ℓ is the corresponding survival function. It can be shown that the AUC is equal to the probability that a randomly chosen diseased individual attains a higher value than a randomly chosen nondiseased individual (Krzanowski & Hand, 2009). Statistical inference on the discriminatory power of a marker can be based on the ROC or the ROC curve.

Parametric and nonparametric approaches have been proposed in the literature for the assessment of the diagnostic accuracy of a marker or for the comparison of the diagnostic accuracy of competing markers in two-class problems. Specifically, widely used nonparametric approaches (DeLong et al., 1988), where two or more markers are compared through their corresponding empirical AUC. A test that involves the empirical ROC curve itself using a permutation testing procedure also exists (Venkatraman, 2000; Venkatraman & Begg, 1996). Notice that the empirical AUC is equivalent to the Mann–Whitney *U* statistic (Bamber, 1975).

The binormal model is widely used in the parametric setting. It assumes that the i=1, …, n_ℓ, are independent and identically distributed measurements from class D^ℓ and are either normally distributed, that is, $X_{\ell i} \sim N(\mu_\ell, \sigma_\ell^2)$ for ℓ=−,+, or they can be transformed to normality through a common, latent transformation procedure (Box & Cox, 1964) adapted to the ROC curve context (Molodianovitch et al., 2006). Under the binormal model, the ROC curve has the form $\mathrm{ROC}(t) = \Phi\left(\frac{\mu_+ - \mu_-}{\sigma_+} + \frac{\sigma_-}{\sigma_+}\Phi^{-1}(t)\right)$, t∈[0,1], with Φ being the cdf of the standard normal distribution. McClish (1989) proposed an estimate for the variance of the AUC on the basis of the binormal model. Wieand et al. (1989) introduced a statistical test for comparing two classifiers under the binormal model on the basis of the respective expression of the AUC. An approach for the comparison of two classifiers using the binormal model assumptions and the corresponding ROC curve directly, without making use of the AUC, exists in the literature (Metz & Kronman, 1980; Metz et al., 1984). The corresponding parameters of the ROC curve are compared, resulting in an approximately chi-squared distributed test that makes use of the shape of the ROC curve directly, rather than its corresponding AUC, with the latter being a proxy/summarizing function.

Whereas classification procedures to one of two classes have been investigated for over half a century, multiple-class classification problems have only recently acquired more attention (Nakas & Reiser, 2018; Liu et al., 2018). A detailed overview of the three-class diagnostic setting is provided in Nakas (2014). For the latter, an intermediate class D^0 is considered, which can be a transitional or early-stage disease class in practice. Clinical examples involve dementia disease states in Alzheimer's disease, Parkinson's disease, or HIV, with patients belonging to classes with normal cognition, mild cognitive impairment, or dementia. The ROC surface was introduced as a generalization of the ROC curve, and the summarizing index of the volume under the ROC surface (VUS) for the assessment of the diagnostic accuracy of a marker was used in a three-class classification task (Scurfield, 1996).

The three-class diagnostic setting is briefly described as follows. A diagnostic marker yields measurements X=x on a continuous scale for all three groups. Without loss of generality, we assume that higher values tend to be associated with increased severity of disease. We assume that the true class membership for each individual is known (i.e., the reference standard is independently available). The i=1, …, n_ℓ, independent and identically distributed measurements from class D^ℓ are denoted by $X_{\ell i}$ for ℓ=−,0,+. We write f_ℓ as the probability density function in class D^ℓ and F_ℓ as the corresponding cdf in group D^ℓ. G_ℓ=1− F_ℓ is the survival function. An intuitive decision rule regarding the application of the diagnostic marker for assigning subjects into three ordinal diagnostic groups is based on a pair of cut-off points c_- and c_+, where c_-<c_+. Then, assign individuals with $x \le c_-$ to the healthy class D^-, individuals with c_-<$x \le c_+$ to D^0, and x>c_+ to the diseased group D^+. We set t_-= $F_-(c_-)$ and t_+=$G_+(c_+)$. The ROC surface is defined as the probability that a randomly selected subject from group D^0 has a test result between c_- and c_+, which can be written as

$$z = \mathrm{ROCs}(t_-, t_+) = F_0(c_+) - F_0(c_-) = F_0(G_+^{-1}(t_+)) - F_0(F_-^{-1}(t_-)), \tag{1}$$

where G_+^{-1} and F_-^{-1} are the inverse functions of G_+ and F_-, respectively (Nakas & Yiannoutsos, 2004). The function z=ROCs(t_-,t_+) defines a surface in the unit cube, which means (t_-, t_+, z) ∈ [0,1]×[0,1]×[0,1]. The assumption c_-<c_+ implies that the ROC surface is only defined on the domain $t_+ < G_+(F_-^{-1}(t_-))$. It follows that perfect discrimination is present if there is complete separation between the distributions of the three classes; that is, if ROCs(1,1)=1. On the other hand, if the three distributions are congruent, that is, $P(X_{-i} \le x) = P(X_{0i} \le x) = P(X_{+i} \le x)$ for each measurement x, then the ROC surface is equivalent to the chance plane defined by the equation $t_- + t_+ + z = 1$. Such a classifier will have no better discriminatory power than a random allocation function and is deemed to be an uninformative classifier.

The VUS is defined as

$$\mathrm{VUS} = \int_0^1 \int_0^{G_+(F_-^{-1}(t_-))} \left(F_0(G_+^{-1}(t_+)) - F_0(F_-^{-1}(t_-))\right) dt_-\, dt_+, \tag{2}$$

which summarizes the global diagnostic accuracy for trichotomous tests. It holds that (Mossman, 1999; Xiong et al., 2006)

$$\text{VUS} = P(X_- < X_0 < X_+),$$

that is, the probability that the results of the diagnostic test from a randomly selected triple with one individual from each diagnostic group will be ranked in the correct order. From the definition of the ROC surface and the convention of the ordered classes according to the disease state, it follows that the boundaries of the VUS are given by 1/6 for an uninformative marker and 1 in the case of perfect separation between the corresponding distributions of the three classes under study.

In what follows, we review the trinormal ROC surface model in Section 2 and existing VUS-based testing approaches relevant to hypothesis testing in the ROC surface analytical framework in Section 3. We then introduce in Section 4 a trinormal model-based ROC testing framework extending the ROC curve setting (Metz & Kronman, 1980; Metz et al., 1984). Section 5 presents a simulation study examining the proposed approaches and comparing them with existing inferential procedures in the ROC surface framework. Section 6 offers an illustration using data from a published study on pancreatic cancer diagnostic markers (Leichtle et al., 2013). We end with a discussion and refer to the use of the accompanying R-package `trinROC`, which we have specifically built for ROC surface analysis applications.

2 | THE TRINORMAL ROC MODEL

Considering that the data from the three classes D^ℓ, ℓ=−, 0, + are normally distributed, we write the i=1, …, n_ℓ, independent and identically distributed measurements from class D^ℓ as $X_{\ell i} \sim N(\mu_\ell, \sigma_\ell^2)$. Otherwise, transformation functions such as the Box–Cox approach may be applied in order to obtain data that can reasonably be considered to be normally distributed. Bantis et al. (2015) have described a detailed implementation of the Box–Cox approach in the ROC surface context.

We reformulate the functional form in Equation (1) as follows:

$$\text{ROCs}(t_-, t_+) = \Phi\left(\frac{\Phi^{-1}(1-t_+)+d}{c}\right) - \Phi\left(\frac{\Phi^{-1}(t_-)+b}{a}\right), \tag{3}$$

where the parameters a, b, c, and d are given by

$$a = \frac{\sigma_0}{\sigma_-}, \qquad b = \frac{\mu_- - \mu_0}{\sigma_-}, \qquad c = \frac{\sigma_0}{\sigma_+}, \qquad d = \frac{\mu_+ - \mu_0}{\sigma_+}. \tag{4}$$

Estimation of an ROC surface using the trinormal model can be performed by estimating the parameters a, b, c, and d with maximum likelihood estimators of the means $\widehat{\mu}_\ell = \sum_{i=1}^{n_\ell} X_{\ell i}/n_\ell$ and variances $\widehat{\sigma}_\ell^2 = \sum_{i=1}^{n_\ell}(X_{\ell i} - \widehat{\mu}_i)^2/n_\ell$ for ℓ=−,0,+.

The corresponding VUS is then obtained by reformulating Equation (2), as follows (Xiong et al., 2006):

$$\widehat{\text{VUS}} = \iint_{\mathcal{A}} \Phi\left(\frac{\Phi^{-1}(1-t_+)+\widehat{d}}{\widehat{c}}\right) - \Phi\left(\frac{\Phi^{-1}(t_-)+\widehat{b}}{\widehat{a}}\right) \mathrm{d}t_- \,\mathrm{d}t_+,$$

where $\mathcal{A} = \{(t_-, t_+) | 0 \le t_- \le 1,\ 0 \le t_+ \le \Phi((\widehat{\mu}_+ - \widehat{\mu}_- + \widehat{\sigma}_- \Phi^{-1}(t_-))/\widehat{\sigma}_+)\}$. A shorter expression for the VUS is given by

$$\widehat{\text{VUS}} = \int_{-\infty}^{\infty} \Phi(\widehat{a}s - \widehat{b})\Phi(-\widehat{c}s + \widehat{d})\varphi(s)\,\mathrm{d}s,$$

where φ is the density of the standard normal distribution.

3 | COMMON METHODS: HYPOTHESIS TESTING WITH THE VUS

The existing literature only involves testing procedures that use overall summary indices of the ROC surface for the assessment of the diagnostic accuracy of markers in the three-class setting (Krzanowski & Hand, 2009). As a result, the comparison of Classifier 1 with Classifier 2 will most often involve testing H_0:VUS_1=VUS_2 against H_1:$\text{VUS}_1 \neq \text{VUS}_2$. When diagnostic markers that have been tested/applied to the same set of patients are compared, there is an inherent correlation between marker measurements. This correlation is passed down to the corresponding estimated VUSs and has to be taken into account in the respective hypothesis-testing procedures. In the simple case where we investigate the performance of a single marker, the null hypothesis of interest is H_0:VUS_1=1/6, where 1/6 is the VUS of the chance plane. Representative tests involving VUS are presented in the sequel.

3.1 | Trinormal VUS testing approaches

In the trinormal ROC model framework, one can use the following testing procedures for the comparison of paired diagnostic markers (statistic Z_{paired}) and unpaired diagnostic markers (statistic Z_{unpaired}) and for the assessment of a single marker, respectively (statistic Z_{single}), assuming that marker values are normally distributed in each class. The unpaired case arises when the two markers being compared have been tested independently on different sets of individuals. The paired case is more frequent in practice when researchers compare a panel of markers whose data arise from the same experiment on the same set of individuals.

$$Z_{\text{paired}} = \frac{\widehat{\text{VUS}}_1 - \widehat{\text{VUS}}_2}{\sqrt{\widehat{\text{Var}}(\widehat{\text{VUS}}_1) + \widehat{\text{Var}}(\widehat{\text{VUS}}_2) - 2\widehat{\text{Cov}}(\widehat{\text{VUS}}_1, \widehat{\text{VUS}}_2)}}, \tag{5}$$

$$Z_{\text{unpaired}} = \frac{\widehat{\text{VUS}}_1 - \widehat{\text{VUS}}_2}{\sqrt{\widehat{\text{Var}}(\widehat{\text{VUS}}_1) + \widehat{\text{Var}}(\widehat{\text{VUS}}_2)}}, \tag{6}$$

$$Z_{\text{single}} = \frac{\widehat{\text{VUS}}_1 - \frac{1}{6}}{\sqrt{\widehat{\text{Var}}(\widehat{\text{VUS}}_1)}}. \tag{7}$$

These statistics are considered to follow a standard normal distribution under the null hypothesis. Detailed formulae for the computation of the variance and covariance estimates have been proposed (Xiong et al., 2006, 2007). We use "VUS test" to denote such an approach in the simulation section.

3.2 | An empirical VUS test

Although the trinormal model assumes normality, the empirical nonparametric approach poses no parametric distributional assumptions on the data but is based on the empirical cdf for each class. For each triplet (X_{-i}, X_{0j}, X_{+k}) of measurements from the three classes, the function of correct orderings (ratings) is as follows:

$$\text{cr}(X_{-i}, X_{0j}, X_{+k}) = \begin{cases} 1, & \text{if } X_{-i} < X_{0j} < X_{+k}, \\ \frac{1}{2}, & \text{if } X_{-i} = X_{0j} < X_{+k} \text{ or } X_{-i} < X_{0j} = X_{+k}, \\ \frac{1}{6}, & \text{if } X_{-i} = X_{0j} = X_{+k}, \\ 0, & \text{else.} \end{cases}$$

Then the empirical VUS can be estimated by

$$\widetilde{\text{VUS}} = \frac{1}{n_- n_0 n_+} \sum_{i=1}^{n_-} \sum_{j=1}^{n_0} \sum_{k=1}^{n_+} \text{cr}(X_{-i}, X_{0j}, X_{+k}). \tag{8}$$

Nonparametric methods involving $\widetilde{\text{VUS}}$ for the comparison of competing diagnostic markers and for the assessment of the diagnostic accuracy of a single marker in discriminating between three diagnostic groups exist in the literature (Dreiseitl et al., 2000; Nakas & Yiannoutsos, 2004). Hypothesis testing is based on the asymptotic normality of U statistics (Hoeffding, 1948) and utilize Equations (5), (6), and (7) by substituting $\widehat{\text{VUS}}$ with $\widetilde{\text{VUS}}$. Bootstrap-based testing in this setting, where the variances in Equations (5), (6), and (7) are estimated by bootstrapping, is feasible (Nakas & Yiannoutsos, 2004). The bootstrap approach is computationally faster and to be preferred in practice instead of the fully empirical approach given their similar performance (Nakas & Yiannoutsos, 2004). We use "Boot test" to denote the bootstrap approach.

4 | HYPOTHESIS TESTING USING THE ROC SURFACE DIRECTLY: A TRINORMAL MODEL-BASED ROC TEST

4.1 | Comparing two markers

In the two-class case, that is, under the ROC curve framework, Metz and Kronman (1980) and Metz et al. (1984) obtained testing procedures, for the unpaired and paired cases, to compare two ROC curves in a binormal setting using only the parameters of the binormal model. We adapted this concept in the three-class setting (under the ROC surface framework) and propose a test statistic on the basis of the parameters of the trinormal model-based ROC surface given in Equation (3). Denote the two markers by Classifier k, for k=1,2, with their corresponding parameters indexed accordingly. Estimates $(\widehat{a}_k, \widehat{b}_k, \widehat{c}_k, \widehat{d}_k)^T$ obtained from Equation (4) are asymptotically multivariate normally distributed (Dorfman & Alf, 1968). This result holds for the binormal model-based ROC curve, but it is trivial to extend for the ROC surface. Consequently, the null hypothesis of interest is H_0: $a_1=a_2$, $b_1=b_2$, $c_1=c_2$, $d_1=d_2$ against H_1: $a_1\neq a_2$ or $b_1\neq b_2$ or $c_1\neq c_2$ or $d_1\neq d_2$. In order to assess the evidence against the hypothesis of equality of two ROC surfaces, we obtain the following test statistic, which is approximately chi-squared distributed with four degrees of freedom:

$$X^2 = \left(\widehat{a}_1 - \widehat{a}_2 \;\; \widehat{b}_1 - \widehat{b}_2 \;\; \widehat{c}_1 - \widehat{c}_2 \;\; \widehat{d}_1 - \widehat{d}_2\right) \widehat{W}^{-1} \begin{pmatrix} \widehat{a}_1 - \widehat{a}_2 \\ \widehat{b}_1 - \widehat{b}_2 \\ \widehat{c}_1 - \widehat{c}_2 \\ \widehat{d}_1 - \widehat{d}_2 \end{pmatrix}, \tag{9}$$

where for unpaired data, $\widehat{W} = \widehat{W}_1 + \widehat{W}_2$ is the sum of the covariance matrices of the trinormal model parameters estimates $\widehat{a}_k$, $\widehat{b}_k$, $\widehat{c}_k$, and $\widehat{d}_k$. The single covariance matrices are

$$\widehat{W}_k = \begin{pmatrix} \widehat{\sigma}^2_{a_k} & \widehat{\sigma}_{a_k b_k} & \widehat{\sigma}_{a_k c_k} & \widehat{\sigma}_{a_k d_k} \\ \widehat{\sigma}_{b_k a_k} & \widehat{\sigma}^2_{b_k} & \widehat{\sigma}_{b_k c_k} & \widehat{\sigma}_{b_k d_k} \\ \widehat{\sigma}_{c_k a_k} & \widehat{\sigma}_{c_k b_k} & \widehat{\sigma}^2_{c_k} & \widehat{\sigma}_{c_k d_k} \\ \widehat{\sigma}_{d_k a_k} & \widehat{\sigma}_{d_k b_k} & \widehat{\sigma}_{d_k c_k} & \widehat{\sigma}^2_{d_k} \end{pmatrix} \text{ with entries given by the Delta method as}$$

$$\widehat{W}_k = \begin{pmatrix} \frac{\widehat{a}_k^2}{2}\left(\frac{1}{n_0}+\frac{1}{n_-}\right) & \frac{\widehat{a}_k\widehat{b}_k}{2n_-} & \frac{\widehat{a}_k\widehat{c}_k}{2n_0} & 0 \\ \frac{\widehat{a}_k\widehat{b}_k}{2n_-} & \frac{\widehat{b}_k^2}{2n_-}+\frac{\widehat{a}_k^2}{n_0}+\frac{1}{n_-} & 0 & \frac{\widehat{a}_k\widehat{c}_k}{n_0} \\ \frac{\widehat{a}_k\widehat{c}_k}{2n_0} & 0 & \frac{\widehat{c}_k^2}{2}\left(\frac{1}{n_0}+\frac{1}{n_+}\right) & \frac{\widehat{c}_k\widehat{d}_k}{2n_+} \\ 0 & \frac{\widehat{a}_k\widehat{c}_k}{n_0} & \frac{\widehat{c}_k\widehat{d}_k}{2n_+} & \frac{\widehat{d}_k^2}{2n_+}+\frac{\widehat{c}_k^2}{n_0}+\frac{1}{n_+} \end{pmatrix}, \quad k = 1, 2. \tag{10}$$

The null hypothesis will be rejected if $X^2 > X^2_\alpha$, that is, if the test statistic exceeds the chi-squared distribution with four degrees of freedom quantile at a predefined significance level α.

When the marker measurements are unpaired, the estimated parameters $\widehat{a}_1, \widehat{b}_1, \widehat{c}_1, \widehat{d}_1$ are independent from $\widehat{a}_2, \widehat{b}_2, \widehat{c}_2, \widehat{d}_2$, and hence all covariances for estimates between the two ROC surfaces are zero. As a consequence, $\widehat{W}$ can be written as the sum of the covariances of the two sets of parameters. However, when the marker measurements are paired, in order to account for the correlation introduced, we define $\widehat{W}^* = \widehat{W}_1 + \widehat{W}_2 - \widehat{C}$. The trinormal model-based ROC test may then be written as

$$X^2 = \left(\widehat{a}_1 - \widehat{a}_2 \;\; \widehat{b}_1 - \widehat{b}_2 \;\; \widehat{c}_1 - \widehat{c}_2 \;\; \widehat{d}_1 - \widehat{d}_2\right) \widehat{W}^{*^{-1}} \begin{pmatrix} \widehat{a}_1 - \widehat{a}_2 \\ \widehat{b}_1 - \widehat{b}_2 \\ \widehat{c}_1 - \widehat{c}_2 \\ \widehat{d}_1 - \widehat{d}_2 \end{pmatrix}, \tag{11}$$

which follows approximately a chi-squared distribution with four degrees of freedom.

The entries of $\widehat{C}$ are given by

$$\widehat{C} = \begin{pmatrix} 2\widehat{\sigma}_{a_1a_2} & \widehat{\sigma}_{a_1b_2} + \widehat{\sigma}_{a_2b_1} & \widehat{\sigma}_{a_1c_2} + \widehat{\sigma}_{a_2c_1} & \widehat{\sigma}_{a_1d_2} + \widehat{\sigma}_{a_2d_1} \\ \widehat{\sigma}_{a_1b_2} + \widehat{\sigma}_{a_2b_1} & 2\widehat{\sigma}_{b_1b_2} & \widehat{\sigma}_{b_1c_2} + \widehat{\sigma}_{b_2c_1} & \widehat{\sigma}_{b_1d_2} + \widehat{\sigma}_{b_2d_1} \\ \widehat{\sigma}_{a_1c_2} + \widehat{\sigma}_{a_2c_1} & \widehat{\sigma}_{b_1c_2} + \widehat{\sigma}_{b_2c_1} & 2\widehat{\sigma}_{c_1c_2} & \widehat{\sigma}_{c_1d_2} + \widehat{\sigma}_{c_2d_1} \\ \widehat{\sigma}_{a_1d_2} + \widehat{\sigma}_{a_2d_1} & \widehat{\sigma}_{b_1d_2} + \widehat{\sigma}_{b_2d_1} & \widehat{\sigma}_{c_1d_2} + \widehat{\sigma}_{c_2d_1} & 2\widehat{\sigma}_{d_1d_2} \end{pmatrix}$$

and are estimated similarly to the two-class case. Repeated applications of the Delta method are used to show that the entries of the symmetric $\widehat{W}^*$ entries are given by

$$\widehat{w}^*_{11} = \frac{\widehat{a}_1^2}{2}\left(\frac{1}{n_0} + \frac{1}{n_-}\right) + \frac{\widehat{a}_2^2}{2}\left(\frac{1}{n_0} + \frac{1}{n_-}\right) - 2\left(\frac{\widehat{\rho}_0^2\widehat{a}_1\widehat{a}_2}{2n_0} + \frac{\widehat{\rho}_-^2\widehat{a}_1\widehat{a}_2}{2n_-}\right),$$

$$\widehat{w}^*_{12} = \frac{\widehat{a}_1\widehat{b}_1}{2n_-} + \frac{\widehat{a}_2\widehat{b}_2}{2n_-} - \frac{\widehat{a}_1\widehat{b}_2\widehat{\rho}_-}{2n_-} - \frac{\widehat{a}_2\widehat{b}_1\widehat{\rho}_-}{2n_-},$$

$$\widehat{w}^*_{13} = \frac{\widehat{a}_1\widehat{c}_1}{2n_0} + \frac{\widehat{a}_2\widehat{c}_2}{2n_0} - \frac{\widehat{a}_1\widehat{c}_2\widehat{\rho}_0}{2n_0} - \frac{\widehat{a}_2\widehat{c}_1\widehat{\rho}_0}{2n_0},$$

$$\widehat{w}^*_{22} = \frac{\widehat{b}_1^2}{2n_-} + \frac{\widehat{a}_1^2}{n_0} + \frac{1}{n_-} + \frac{\widehat{b}_2^2}{2n_-} + \frac{\widehat{a}_2^2}{n_0} + \frac{1}{n_-} - 2\left(\frac{\widehat{\rho}_-^2\widehat{b}_1\widehat{b}_2}{2n_0} + \frac{\widehat{\rho}_0^2\widehat{a}_1\widehat{a}_2}{n_0} + \frac{\widehat{\rho}_-}{n_-}\right),$$

$$\widehat{w}^*_{24} = \frac{\widehat{a}_1\widehat{c}_1}{n_0} + \frac{\widehat{a}_2\widehat{c}_2}{n_0} - \frac{\widehat{a}_1\widehat{c}_2\widehat{\rho}_0}{n_0} - \frac{\widehat{a}_2\widehat{c}_1}{n_0},$$

$$\widehat{w}^*_{33} = \frac{\widehat{c}_1^2}{2}\left(\frac{1}{n_0} + \frac{1}{n_+}\right) + \frac{\widehat{c}_2^2}{2}\left(\frac{1}{n_0} + \frac{1}{n_+}\right) - 2\left(\frac{\widehat{\rho}_0^2\widehat{c}_1\widehat{c}_2}{2n_0} + \frac{\widehat{\rho}_+^2\widehat{c}_1\widehat{c}_2}{2n_+}\right),$$

$$\widehat{w}^*_{34} = \frac{\widehat{c}_1\widehat{d}_1}{2n_+} + \frac{\widehat{c}_2\widehat{d}_2}{2n_+} - \frac{\widehat{c}_1\widehat{d}_2\widehat{\rho}_+}{2n_+} - \frac{\widehat{c}_2\widehat{d}_1\widehat{\rho}_+}{2n_+},$$

$$\widehat{w}^*_{44} = \frac{\widehat{d}_1^2}{2n_+} + \frac{\widehat{c}_1^2}{n_0} + \frac{1}{n_+} + \frac{\widehat{d}_2^2}{2n_+} + \frac{\widehat{c}_2^2}{n_0} + \frac{1}{n_+} - 2\left(\frac{\widehat{\rho}_+^2\widehat{d}_1\widehat{d}_2}{2n_+} + \frac{\widehat{\rho}_0^2\widehat{c}_1\widehat{c}_2}{n_0} + \frac{\widehat{\rho}_+}{n_+}\right),$$

$$\widehat{w}^*_{14} = \widehat{w}^*_{23} = 0,$$

where $\widehat{\rho}_\ell$ are the corresponding pairwise Pearson correlation coefficients. The above elements are sufficient to define this symmetric matrix. We reject H_0 if $X^2 > X^2_\alpha$ just as we did in the unpaired case. We use "ROC test" to denote such an approach in Section 5.

4.2 | Assessment of a single marker

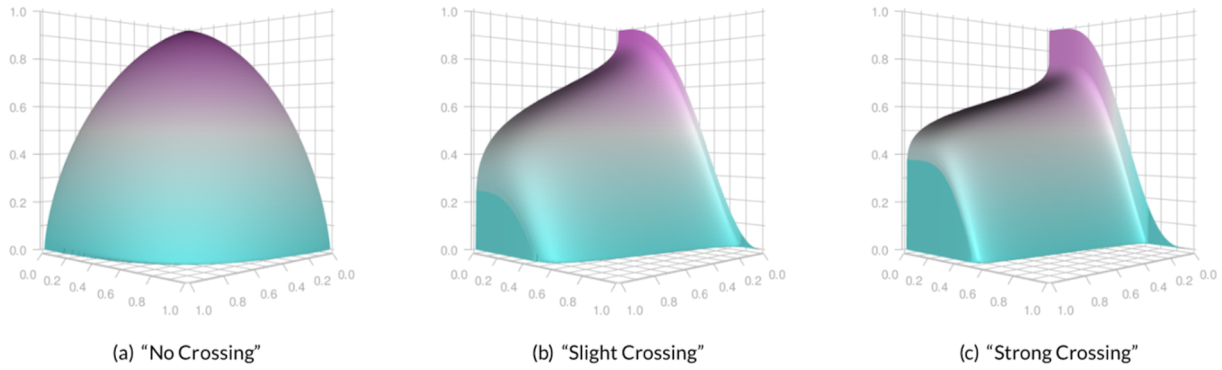
In the previous section, we have seen how two ROC surfaces can be compared. It is also possible to asses a single marker using a similar strategy, namely, by comparing an ROC surface with the chance plane (i.e., an uninformative ROC surface) leading to the null hypothesis H_0: a_1=0, b_1=0, c_1=1, and d_1=0. The corresponding test statistic is

$$X^2 = \left(\widehat{a}_1 - 1 \;\; \widehat{b}_1 \;\; \widehat{c}_1 - 1 \;\; \widehat{d}_1\right)\widehat{W}_1^{-1}\begin{pmatrix} \widehat{a}_1 - 1 \\ \widehat{b}_1 \\ \widehat{c}_1 - 1 \\ \widehat{d}_1 \end{pmatrix}, \tag{12}$$

with $\widehat{W}_1$ defined as in Equation (10). Under the null hypothesis, X^2 follows approximately a chi-squared distribution with four degrees of freedom.

5 | SIMULATION STUDY

The simulation study consists of three parts, where in the first two parts we investigate the performance of the proposed testing procedures given in Sections 3 and 4. In the first part, we investigate tests that assess single markers in their deviation from the chance plane in order to evaluate whether a classifier performs significantly better than a random allocation procedure. In the second part, we assess the tests relevant to the comparison of two diagnostic markers. In these first two simulation parts, we sampled the data from underlying normal distributions. The third part of the simulation study evaluates the performance of the proposed procedures after applying the Box–Cox transformation for data normalization. Log-normal and gamma distributions were considered for the sampling scenarios. The simulation was carried out using `R` (RStudio Team, 2016) and the R-package `trinROC` (Noll, 2019) specifically developed for the implementation of the proposed and widely used competing methodologies in a ROC surface analysis.



(a) "No Crossing" (b) "Slight Crossing" (c) "Strong Crossing"

FIGURE 1 ROC surfaces representing the general shape of the surface for each level of "crossing" in the single-marker simulation study (based on samples with a theoretical VUS=0.4 and n_e=100). The effect of differences between variances of D^-, D^0, and D^+ on the ROC surface is apparent. The ROC surface crosses the chance plane in the "slight crossing" and "strong crossing" scenarios. ROC, receiver operating characteristic; VUS, volume under the ROC surface

The following scenarios were considered (for all three simulation parts): Three different sample size scenarios n_e = 25, 50, 100, for each class (i.e., $n_-=n_0=n_+$).

The factor stated as "crossing" flags the existence/importance of differences between variances of the three classes. Three "crossing" scenarios were considered: "no crossing" (equal variance between the three classes), "slight crossing" (an intermediate situation), and "strong crossing" (significant differences in variances between the three classes). Figure 1 displays the effect on the ROC surface for the three different scenarios of differences in variances between the three groups. When differences between variances of the distributions of the three classes exist, the ROC surface crosses the chance plane.

For the simulation on the comparison between markers, we also distinguish between unpaired and paired data by sampling from bivariate normal distributions using ρ=0 and ρ=0.5 for each of D^-, D^0, and D^+. Typically, paired data arise when marker measurements are obtained from markers applied on the same set of individuals, whereas unpaired data are less frequent in practice considering cases where measurements arise from independent studies for the markers under comparison. We considered the trinormal model-based ROC test (ROC test), the trinormal VUS test (VUS test), and the bootstrap-based test (Boot test, with 500 bootstrap replications). Each result is based on 1000 iterations and an α level of 5%.

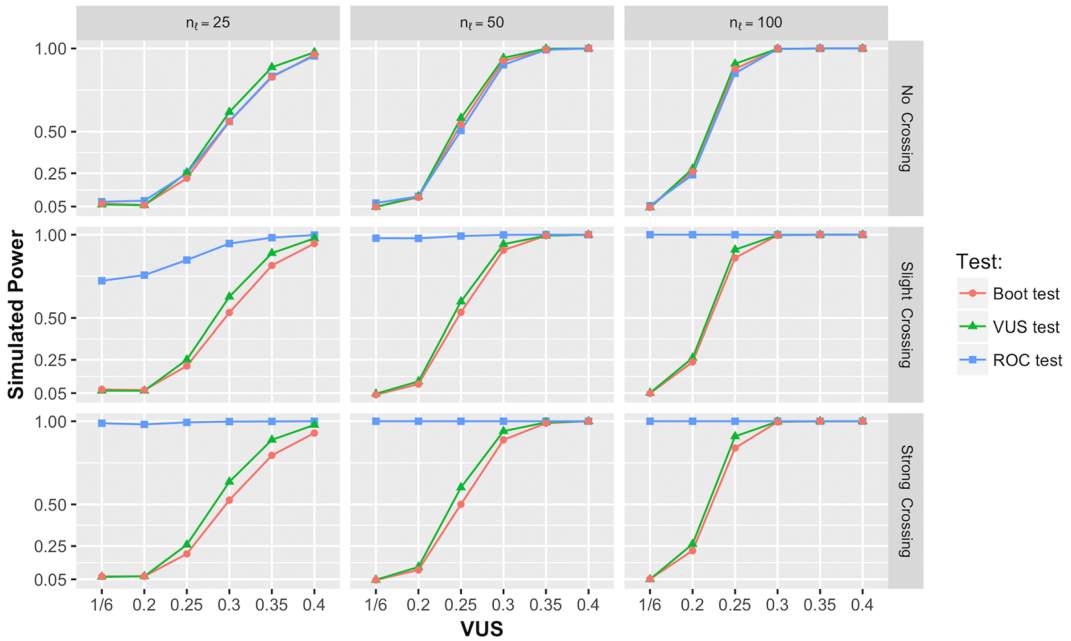


FIGURE 2 Empirical power per 1,000 iterations and an α significance level of 5%, based on the simulation results of the tests that assess single markers. The columns indicate the sample size, and the rows represent the three different assumptions of variability in D^-, D^0, and D^+. Each dot corresponds to a different scenario with increasing VUS. ROC, receiver operating characteristic; VUS, volume under the ROC surface

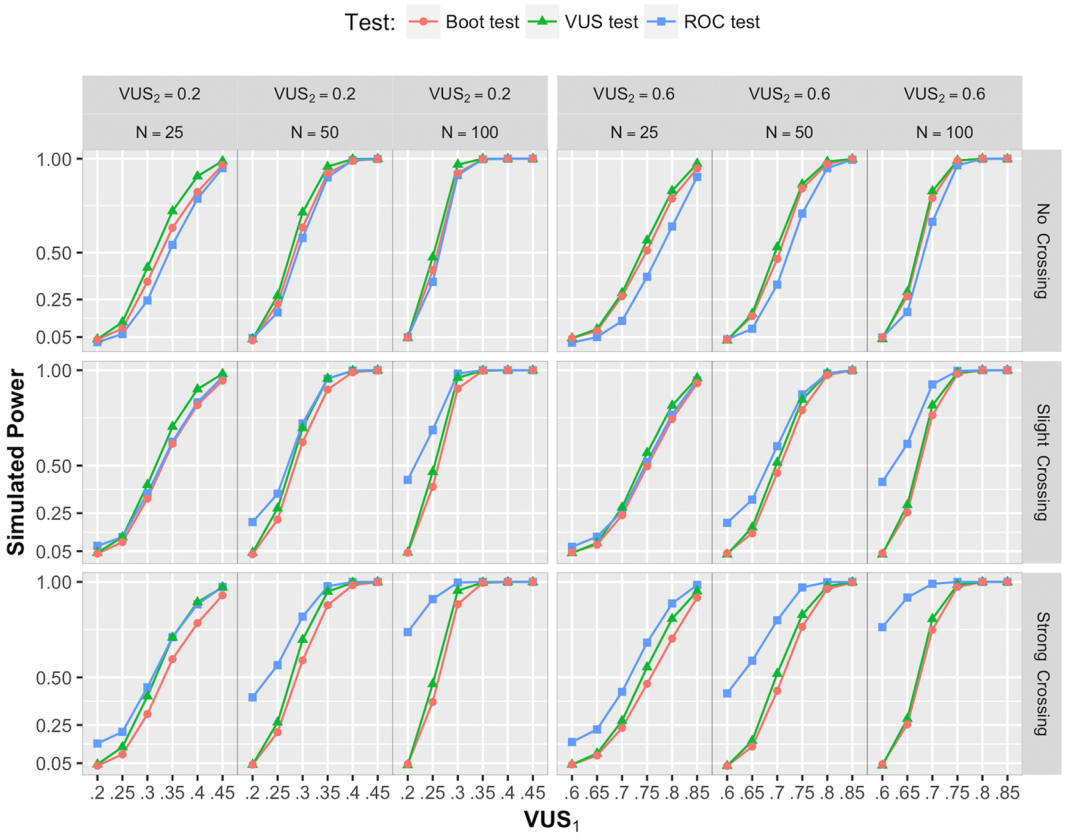


FIGURE 3 Empirical power per 1,000 iterations and an α significance level of 5%, based on the simulation results of the tests that assess comparisons of two markers. The left panel represents the results of VUS_2=1/6, whereas the right panel represents those for VUS_2=0.3. The paired data scenarios are shown. Rows represent the three different assumptions of variability for D^-, D^0, and D^+. ROC, receiver operating characteristic; VUS, volume under the ROC surface

5.1 | Single-marker assessment tests

Means and standard deviations for the distributions of D^-, D^0, and D^+ were chosen so that the true VUS equals 1/6, 0.2, 0.25, 0.3, 0.35, and 0.4 for each one of the six scenarios studied in this part of the simulation. These VUS scenarios cover cases with markers of increasing accuracy starting from the case of a marker with uninformative corresponding VUS (equal to 1/6). For each one of the six different VUS scenarios, the factor "crossing" was set at the following: "no crossing" implying $(\sigma_-, \sigma_0, \sigma_+) = (1, 1, 1)$, "slight crossing" with $(\sigma_-, \sigma_0, \sigma_+) = (1, 1.5, 2)$, and "strong crossing" with $(\sigma_-,$

TABLE 1 Parameters used in the Box–Cox simulation

VUS	Crossing	Gamma									Log-normal					
		D^-			D^0			D^+			D^-		D^+		D^+	
		α_-	β_-	s_-	α_0	β_0	s_0	α_+	β_+	s_+	μ_-	σ_-	μ_0	σ_0	μ_+	σ_+
1/6	No crossing	3	4	0.00	3	4	0.00	3	4	0.00	0.00	1.00	0.00	1.00	0.00	1.00
	Slight crossing	4	5	0.00	5	6	−10.21	6	7	−20.42	0.00	1.00	0.00	1.50	0.00	2.00
	Strong crossing	5	6	0.00	8	9	−46.87	11	12	−93.75	0.00	1.00	0.00	2.00	0.00	3.00
0.3	No crossing	3	4	0.00	3	4	2.55	3	4	5.11	0.00	1.00	0.40	1.00	0.81	1.00
	Slight crossing	4	5	0.00	5	6	−4.75	6	7	−9.51	0.00	1.00	0.60	1.50	1.19	2.00
	Strong crossing	5	6	0.00	8	9	−36.06	11	12	−72.11	0.00	1.00	0.79	2.00	1.57	3.00

Note. The shift parameter s_ℓ indicates how the samples from the classes were shifted in order to attain the desired VUS.

Abbreviation: VUS, volume under the receiver operating characteristic surface.

TABLE 2 Results of the Box–Cox simulation

	Distribution	Crossing	n_e	Original data			Box–Cox transformed data		
				$\widehat{VUS}$	ROC test	VUS test	$\widehat{VUS}$	ROC test	VUS test
VUS_1									
1/6	Log-normal	No crossing	25	0.162	0.602	0.198	0.166	0.081	0.079
			50	0.168	0.694	0.208	0.167	0.052	0.060
			100	0.166	0.747	0.246	0.167	0.052	0.065
		Slight crossing	25	0.252	0.957	0.511	0.164	0.730	0.066
			50	0.262	0.998	0.762	0.166	0.974	0.069
			100	0.268	1.000	0.883	0.168	1.000	0.042
		Strong crossing	25	0.303	0.998	0.822	0.168	0.985	0.062
			50	0.309	1.000	0.916	0.166	1.000	0.054
			100	0.309	1.000	0.956	0.167	1.000	0.051
	Gamma	No crossing	25	0.167	0.173	0.071	0.167	0.076	0.063
			50	0.167	0.174	0.070	0.167	0.069	0.057
			100	0.166	0.180	0.073	0.166	0.055	0.057
		Slight crossing	25	0.180	0.518	0.061	0.160	0.621	0.062
			50	0.179	0.808	0.050	0.159	0.926	0.057
			100	0.182	0.985	0.088	0.161	1.000	0.072
		Strong crossing	25	0.184	0.983	0.036	0.157	0.995	0.086
			50	0.185	1.000	0.065	0.159	1.000	0.072
			100	0.187	1.000	0.108	0.161	1.000	0.062
VUS_1,VUS_2									
0.3,0.3	Log-normal	No crossing	25	0.267	0.572	0.157	0.293	0.036	0.046
			50	0.274	0.692	0.185	0.297	0.037	0.049
			100	0.271	0.780	0.195	0.299	0.052	0.053
		Slight crossing	25	0.344	0.943	0.212	0.289	0.351	0.041
			50	0.344	0.994	0.276	0.296	0.766	0.061
			100	0.340	1.000	0.429	0.300	0.982	0.051
		Strong crossing	25	0.363	0.998	0.231	0.295	0.831	0.046
			50	0.354	1.000	0.320	0.296	0.996	0.059
			100	0.343	1.000	0.427	0.299	1.000	0.043
	Gamma	No crossing	25	0.283	0.119	0.056	0.313	0.033	0.046
			50	0.286	0.138	0.065	0.316	0.034	0.046
			100	0.287	0.153	0.060	0.317	0.035	0.042
		Slight crossing	25	0.314	0.277	0.068	0.302	0.317	0.039
			50	0.316	0.531	0.084	0.303	0.716	0.043
			100	0.320	0.836	0.133	0.306	0.981	0.070
		Strong crossing	25	0.328	0.820	0.077	0.293	0.958	0.056
			50	0.333	0.982	0.118	0.299	0.999	0.057
			100	0.336	1.000	0.206	0.303	1.000	0.061

Note. The upper half depicts a single-marker assessment simulation with a theoretical VUS_1=1/6. The lower half displays a simulation study of comparison of two markers under H_0: VUS_1=VUS_2=0.3.

Abbreviations: ROC, receiver operating characteristic; VUS, volume under the ROC surface.

$\sigma_0, \sigma_+) = (1, 2, 3)$. Equidistant means between the three classes were used for fulfilling the VUS scenario assumptions. Figure 2 summarizes the first part of the simulation.

For the "no crossing" scenarios, we find that the three tests have a similar empirical power, with the trinormal model-based VUS test only very slightly dominating the others. As the sample size n_ℓ increases, the empirical power also increases for VUS≠1/6 in all the tests. The cases of different variances between D^-, D^0, and D^+ are reflected on the second and third rows of Figure 2. The proposed trinormal model-based ROC test has consistently higher power than its competitors. As it is designed to detect differences in the parameters of the ROC surface, it rejects the null hypothesis much faster than the VUS-based tests. It correctly rejects when VUS=1/6 given that the shape of the ROC surface deviates from the chance plane. For a standard deviation ratio of 1:2:3 between D^-, D^0, and D^+ and sample size of 50 or bigger for each class, the trinormal model-based ROC test will virtually always reject the null hypothesis, whereas VUS-based tests remain invariant for such differences relative to the VUS=1/6 scenario. This result was expected, given that VUS is just a proxy metric that does not capture the shape of the ROC surface per se.

5.2 | Comparison of two classifiers

In the second part of the simulation study, two marker comparisons were performed. We chose to investigate two main scenarios: in the first, a set of increasing $VUS_1 = 0.2, 0.25, 0.3, 0.35, 0.4, 0.45$ is compared with an uninformative marker with VUS_2=0.2. In the second, VUS_1=0.6, 0.65, 0.7, 0.75, 0.8, 0.85, whereas $VUS_2 = 0.6$. As in the first part of the simulation, samples were drawn from normal distributions with equidistant means and standard deviations depending on the factor "crossing," such that the scenarios' VUS assumptions were fulfilled. For the two marker comparison simulations, we defined "no crossing" as $(\sigma_{-1}, \sigma_{01}, \sigma_{+1}, \sigma_{-2}, \sigma_{02}, \sigma_{+2}) = (1, 1, 1, 1, 1, 1)$, "slight crossing" as (1, 1.25, 1.5, 1, 1.5, 2), and "strong crossing" as (1, 1.5, 2, 1, 2, 3), in analogy to the single-marker assessment simulation.

Figure 3 depicts the simulation results for the paired data case where, for each class, measurements were drawn from bivariate normal distributions with ρ=0.5. As the simulation results for the unpaired data scenarios turned out to be very similar to those obtained from the paired setting, they are not presented here but can be found in the Supporting Information. The following two scenarios can be found also in the Supporting Information: $VUS_1 = 1/6, 0.2, 0.25, 0.3, 0.35, 0.4$, which is compared with an uninformative marker $VUS_2 = 1/6$, and $VUS_1 = 0.3, 0.35, 0.4, 0.45, 0.5, 0.6$, which is compared with VUS_2=0.3.

Similarly to the first part of the simulation, we find that for an increasing sample size n_ℓ, all the tests result in higher power. In the case of equal variances, the rejection ratios of the trinormal model-based ROC test are slightly weaker than the rejection ratios of the bootstrap test and the trinormal VUS test. The VUS test slightly dominates the bootstrap test. When the variances of the classes differ, the trinormal model-based ROC test starts to dominate the other tests at the latest from a sample size n_ℓ=50.

5.3 | A simulation study involving the Box–Cox transformation

In the third part of the simulation, we sample from distributions other than normal. We calculated the performance of the VUS test and the ROC test introduced above before and after the application of the Box–Cox transformation (Bantis et al., 2015) with the nonnormally distributed data for single-marker assessment and for the comparison of two markers. As the results of the boot test are invariant under monotone transformations, we excluded this test from this simulation part.

We sampled from log-normal and gamma distributions. We set n_ℓ=25, 50, 100. For the control of the "crossing" factor, as there exists no closed formula to compute the theoretical VUS of nonnormal data, we had to numerically calculate suitable parameters of the distributions in order to obtain the desired VUS and variability among the three classes. The parameters considered are given in Table 1. Each result is based on 1,000 iterations and an α significance level of 5%.

TABLE 3 Overview of the 12 pancreatic cancer diagnostic markers

	Markers											
	WBC	**TSH**	**CEAE**	**HCT**	**ALBP**	**ALB**	**CRP**	**HGB**	**A1GL**	**ALBU**	**A1GLP**	**CA-199**
Emp. VUS	0.18	0.18	0.37	0.37	0.44	0.49	0.43	0.4	0.47	0.46	0.53	0.58
Trin. VUS	0.18	0.21	0.37	0.39	0.4	0.41	0.41	0.41	0.44	0.44	0.49	0.6
p values												
Trin. ROC test	.12	5.1e−08	4.2e−07	6.3e−10	4.7e−07	6.7e−10	3.6e−05	3.9e−10	2.7e−10	1.2e−11	2.0e−13	0
Trin. VUS test	.63	.31	8.8e−05	2.1e−05	1.5e−05	9.4e−06	6.1e−05	6.8e−06	5.1e−07	5.6e−07	1.2e−08	5.6e−13
Boot test	.7	.75	9.7e−04	6.8e−04	8.3e−06	3.6e−07	6.6e−05	1.0e−04	7.3e−07	3.1e−06	2.8e−08	3.3e−10

Abbreviations: ROC, receiver operating characteristic; VUS, volume under the ROC surface.

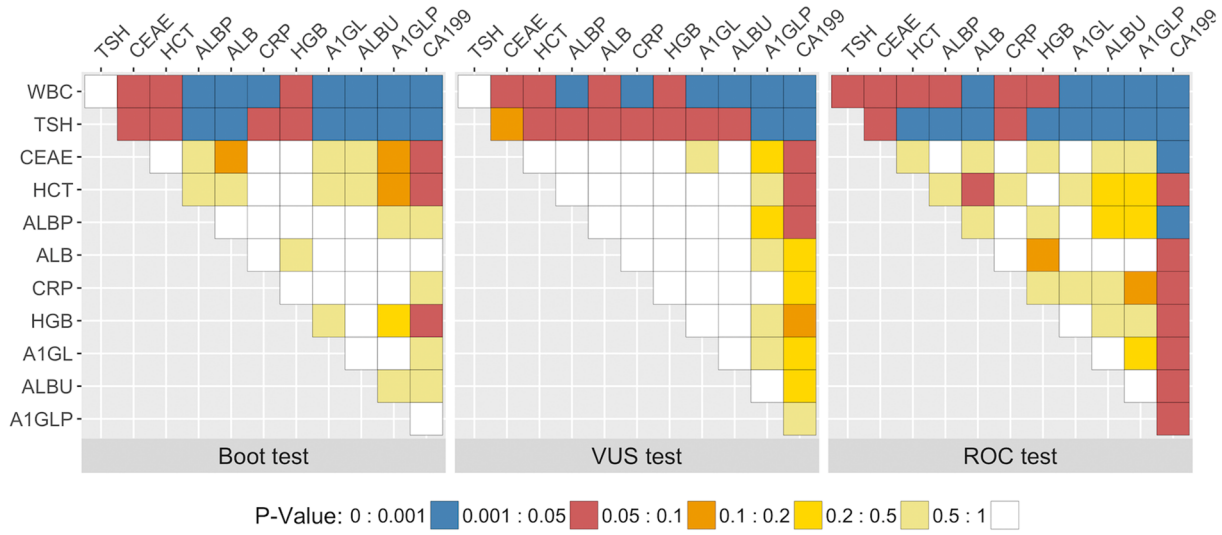


FIGURE 4 Pairwise comparisons of the 12 markers displayed in Table 3. The three heat matrices display the FDR-adjusted *p* values obtained by the empirical VUS test, the trinormal model VUS-based test, and the trinormal model ROC-based test after a Box–Cox transformation. FDR, false discovery rate; ROC, receiver operating characteristic; VUS, volume under the ROC surface

Table 2 shows the results of the Box–Cox simulation in the single marker and comparison of two marker cases. The results are consistent with those of the first two simulations. In the single-marker assessment part, we observe a strong increase in the simulated power of the trinormal-based ROC test due to the variability among the three classes. The results of the VUS test stay invariant around the significance level. The same behaviour is observable in the case of comparing VUS_1=VUS_2=0.3.

6 | APPLICATION TO A PANCREATIC CANCER DATASET

We investigate a panel of 12 diagnostic markers on the basis of measurements from 106 individuals who underwent a complete assessment for pancreatic cancer. The dataset used was first described in (Leichtle et al., 2013). The three classes under study consisted of pancreatic cancer patients (D^+), pancreatitis patients (D^0), and healthy controls (D^-).

Table 3 shows the empirical VUS and the corresponding trinormal VUS of the Box–Cox transformed data as well as the *p* values of the single-marker assessment of the three tests introduced above.

A single-marker investigation yielded highly significant *p* values (<.001) for most markers for all three tests. Only marker WBC was not significant for any test, whereas marker TSH was only significant for the proposed test but not for the VUS-based ones. The VUS for the markers deemed significant through the VUS-based tests varies between 0.18 and 0.6. False discovery rate-adjusted *p* values for pairwise comparisons are shown as heat maps in Figure 4.

Marker CA-199 is consistently the best among the markers under consideration for differentiating between the three classes. However, this result is more pronounced using the trinormal model ROC-based test, because the distributions of marker measurements are highly variable, negatively affecting VUS-based tests only. The utility of marker CA-199 in pancreatic cancer assessment is well documented in the literature (Leichtle et al., 2013).

7 | DISCUSSION

In this article, we have introduced a trinormal model ROC-based test that can be used to assess diagnostic markers in a three-class setting, both for the assessment of a single marker and for the comparison of two competing markers arising from the same set of data or from independent measurements. Application of the Box–Cox transformation in the ROC surface framework (Bantis et al., 2015), prior to the use of the proposed test, can be considered when significant departures from normality assumptions exist for marker measurements.

Typically, in omics applications, thousands of biomarkers may be assessed simultaneously through computational procedures that involve ROC analysis techniques. A metric such as the VUS may fail to recognize significant biomarkers in cases where the corresponding VUS has a low value; however, the corresponding biomarker is informative. In these cases, the ROC surface crosses the chance plane leading to a low VUS. Modelling the surface per se for hypothesis testing may reveal valuable diagnostic markers in cases where the VUS would fail.

The R-package `trinROC` that accompanies this article is available on CRAN, involves testing and plotting options, and can be used as a general-purpose package for three-class diagnostic testing in the ROC framework. Description of its use is offered in the relevant vignette and reference manual.

The research of Prof. Benjamin Reiser was supported by the Israel Science Foundation (Grant 387/15).

DATA AVAILABILITY STATEMENT

Data used in Section 6 are available at a public repository (ResearchGate page of corresponding author https://doi.org/10.13140/RG.2.2.20621.08160).

ORCID

Reinhard Furrer https://orcid.org/0000-0002-6319-2332
Benjamin Reiser https://orcid.org/0000-0002-9797-9029
Christos T. Nakas https://orcid.org/0000-0003-4155-722X

REFERENCES

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, *12*(4), 387–415. https://doi.org/10.1016/0022-2496(75)90001-2

Bantis, L. E., Nakas, C. T., Reiser, B., Myall, D., & Dalrymple-Alford, J. C. (2015). Construction of joint confidence regions for the optimal true class fractions of receiver operating characteristic (ROC) surfaces and manifolds. *Statistical Methods in Medical Research*, *26*(3), 1429–1442. https://doi.org/10.1177/0962280215581694

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 211–252.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 837–845.

Dorfman, D. D., & Alf, E. (1968). Maximum likelihood estimation of parameters of signal detection theory a direct solution. *Psychometrika*, *33*(1), 117–124.

Dreiseitl, S., Ohno-Machado, L., & Binder, M. (2000). Comparing three-class diagnostic tests by three-way ROC analysis. *Medical Decision Making*, *20*(3), 323–331.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, *19*(3), 293–325. https://doi.org/10.1214/aoms/1177730196

Krzanowski, W. J., & Hand, D. J. (2009). *ROC curves for continuous data*. London: CRC Press.

Leichtle, A. B., Ceglarek, U., Weinert, P., Nakas, C. T., Nuoffer, J. M., Kase, J., … Fiedler, G. M.(2013). Pancreatic carcinoma, pancreatitis, and healthy controls: Metabolite models in a three-class diagnostic dilemma. *Metabolomics*, *9*(3), 677–687.

Liu, S., Sun, X., Xu, W., Zhang, Y., & Dai, J. (2018). Null distribution of volume under ordered three-class ROC surface (VUS) with continuous measurements. *IEEE Signal Processing Letters*, *25*(12), 1855–1859.

McClish, D. K. (1989). Analyzing a portion of the ROC curve. *Medical decision making*, *9*(3). PMID: 2668680. https://doi.org/10.1177/0272989X8900900307

Metz, C. E., & Kronman, H. B. (1980). Statistical significance tests for binormal ROC curves. *Journal of Mathematical Psychology*, *22*(3), 218–243. https://doi.org/10.1016/0022-2496(80)90020-6

Metz, C. E., Wang, P. L., & Kronman, H. B. (1984). A new approach for testing the significance of differences between ROC curves measured from correlated data. *Information Processing in Medical Imaging, 8th conference, Brussels, 29 August - 2 September 1983*, 432–445.

Molodianovitch, K., Faraggi, D., & Reiser, B. (2006). Comparing the areas under two correlated ROC curves: Parametric and non-parametric approaches. *Biometrical Journal*, *48*(5), 745–757. https://doi.org/10.1002/bimj.200610223

Mossman, D. (1999). Three-way ROCs. *Medical Decision Making*, *19*(1), 78–89.

Nakas, C. T. (2014). Developments in ROC surface analysis and assessment of diagnostic markers in three-class classification problems. *REVSTAT–Stat J*, *12*(1), 43–65.

Nakas, C. T., & Reiser, B. (2018). Editorial for the special issue of "statistical methods in medical research" on "advanced ROC analysis". *Statistical Methods in Medical Research*, *27*(3), 649–650. https://doi.org/10.1177/0962280217742536

Nakas, C. T., & Yiannoutsos, C. T. (2004). Ordered multiple-class ROC analysis with continuous measurements. *Statistics in Medicine*, *23*(22), 3437–3449. https://doi.org/10.1002/sim.1917

Noll, S (2019). trinROC: Statistical Tests for Assessing Trinormal ROC Data. https://CRAN.R-project.org/package=trinROC

RStudio Team (2016). Rstudio: Integrated development environment for R [computer software manual]. Boston, MA. Retrieved from http://www.rstudio.com/

Scurfield, B. K. (1996). Multiple-event forced-choice tasks in the theory of signal detectability. *Journal of Mathematical Psychology*, *40*(3), 253–269. https://doi.org/10.1006/jmps.1996.0024

Venkatraman, E. S. (2000). A permutation test to compare receiver operating characteristic curves. *Biometrics*, *56*(4), 1134–1138. https://doi.org/10.1111/j.0006-341X.2000.01134.x

Venkatraman, E. S., & Begg, C. B. (1996). A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment. *Biometrika*, *83*(4). https://doi.org/10.1093/biomet/83.4.835

Wieand, S., Gail, M. H., James, B. R., & James, K. L. (1989). A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika*, *76*(3), 585. https://doi.org/10.1093/biomet/76.3.585

Xiong, C., van Belle, G., Miller, J. P., & Morris, J. C. (2006). Measuring and estimating diagnostic accuracy when there are three ordinal diagnostic groups. *Statistics in Medicine*, *25*(7), 1251–1273. https://doi.org/10.1002/sim.2433

Xiong, C., van Belle, G., Miller, J. P., Yan, Y., Gao, F., Yu, K., & Morris, J. C. (2007). A parametric comparison of diagnostic accuracy with three ordinal diagnostic groups. *Biometrical Journal*, *49*(5), 682–693. https://doi.org/10.1002/bimj.200610359

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Noll S, Furrer R, Reiser B, Nakas C. Inference in receiver operating characteristic surface analysis via a trinormal model-based testing approach. *Stat*. 2019;8:e249. https://doi.org/10.1002/sta4.249