REVSTAT – Statistical Journal
Volume 12, Number 1, March 2014, 43–65

DEVELOPMENTS IN ROC SURFACE ANALYSIS AND ASSESSMENT OF DIAGNOSTIC MARKERS IN THREE-CLASS CLASSIFICATION PROBLEMS

Author: Christos T. Nakas
– Laboratory of Biometry, University of Thessaly,
School of Agricultural Sciences, Phytokou street, 38446 Volos, Greece
cnakas@uth.gr

Abstract:

- This article reviews current state of the art of ROC surface analysis and illustrates its use through an application on a pancreatic cancer diagnostic marker. Receiver Operating Characteristic (ROC) surfaces have been studied in the literature essentially only during the last decade and are considered as a natural generalization of ROC curves in three-class diagnostic problems. This article presents the definition, construction, modelling, and utility of the ROC surface while trying to provide an extensive reference list in the topic. It describes methodology for inference based on the Volume Under the ROC surface (VUS) and methodology for decision making through the selection of optimal cut-off points using the notion of the generalized Youden index as the optimality criterion of choice. It ends with a discussion regarding future directions for research in this field of knowledge.

Key-Words:

- *generalized Youden index; receiver operating characteristic (ROC) surface; true class fraction (TCF); volume under the ROC surface (VUS).*

AMS Subject Classification:

- 62C99, 62P10.

1. INTRODUCTION

Receiver Operating Characteristic (ROC) curve analysis has been an active area of research since the early 1950s. The ROC curve depicts the quality of a diagnostic marker in a two-class classification problem. It illustrates the trade-off between sensitivity and specificity as the cut-off point for decision making varies through possible values of the diagnostic marker. Put more formally, suppose that, in a two-class classification problem, a diagnostic marker results in measurements $X_1 \sim F_1$ from the first class under study and $X_2 \sim F_2$ from the second class under study. Suppose that, in general, values from X_2 are larger than values from X_1 but X_1 and X_2 are not perfectly separated, *i.e.* there is an amount of overlap between measurements from the two-classes.[1] A cut-off point c is selected for decision making which will result in the fractions of specificity, defined as $\text{spec}(c) = P(X_1 \leq c)$, and sensitivity, defined as $\text{sens}(c) = P(X_2 > c)$. The fractions of sensitivity (or else True Positive Fraction, TPF) and specificity (True Negative Fraction, TNF) vary as the cut-off point c varies. The ROC curve is defined as the graph depicting $(1 - P(X_1 \leq c), P(X_2 > c)) = (1 - \text{spec}(c), \text{sens}(c))$ in the unit square $[0, 1] \times [0, 1]$, as c varies. Equivalently, the ROC curve is the graph of the function $\text{ROC}(t) = 1 - F_2(F_1^{-1}(1 - t))$, where $t \in [0, 1]$. The Area Under the ROC Curve (AUC) is equivalent to $P(X_1 < X_2)$ and it is the most widely used index for the quantification of the performance of a diagnostic marker in the two-class setting. A useful diagnostic marker will result in an ROC curve with AUC close to 1. A diagnostic marker with AUC close to 0.5 will, in general, be considered as uninformative. The AUC takes on values in $[0.5, 1]$ if the condition that measurements from X_1 are in general smaller than those from X_2 actually holds. The non-parametric estimate of the AUC is equivalent to the Wilcoxon–Mann–Whitney statistic (Pepe, 2003). Formal assessment of the quality of a diagnostic marker based on the AUC consists of testing the null hypothesis, $H_0 : \text{AUC} = 0.5$ versus the alternative of interest through the statistic $z = \{(\text{AUC} - 0.5)/\text{se}(AUC)\} \sim N(0, 1)$, where se(AUC) is the standard error of AUC, estimated, *e.g.*, using the bootstrap. If H_0 is rejected, the diagnostic marker under study is considered to be useful and a cut-off point c must be chosen for decision-making purposes. Use of the maximum of the Youden index (J) is a widely adopted approach for cut-off point selection. The Youden index is defined as $J = \max_c\{\text{sens}(c) + \text{spec}(c) - 1\} = \max_c\{F_1(c) - F_2(c)\}$, as a result, the value of c that maximizes J is chosen. ROC curve analysis is presented in detail in a number of well-written books, such as Pepe (2003) and Zhou *et al.* (2011).

Notions of ROC curve analysis have been extended to accommodate problems of three-class and multiple-class classification. The ROC surface has been proposed as a natural generalization of the ROC curve for the assessment of diagnostic markers in three-class classification problems. The ROC surface was

[1] However, we do not impose any type of stochastic ordering by $X_1 < X_2$.

introduced by Scurfield (1996). The Volume Under the ROC Surface (VUS) was proposed as an index for the assessment of the diagnostic accuracy of the marker under consideration. Unfortunately, the latter article received very little attention probably because it only described the theoretical construction of the ROC surface and did not provide any related application. A similar construction was proposed independently, a few years later though, by Mossman (1999) which was implemented in Mathematica by Heckerling (2001). Inference regarding the VUS, based on Mossman's construction, using non-parametric statistics, was studied by Dreiseitl *et al.* (2000). The ROC surface construction, and the generalization of this construction in multiple-class classification problems, in a non-parametric context, was proposed in Nakas and Yiannoutsos (2004). Interestingly, the latter construction unifies the approaches of Mossman and Scurfield in a natural way and thus offered the framework and the theoretical basis for extending ROC curve analysis concepts in multiple-class classification problems. This construction has been reinvented at least a couple of times later on (*e.g.* Xiong *et al.*, 2006; Li and Fine, 2008), however, in Xiong *et al.* (2006) the parametric framework is studied extensively supplementing the work in Nakas and Yiannoutsos (2004). Given the theoretical basis for the ROC surface, several articles appeared in the literature during the last 10 years generalizing notions from ROC curve analysis. ROC surfaces are overviewed in the textbook on ROC analysis by Krzanowski and Hand (2009).

In the following sections the ROC surface analysis literature will be reviewed and unified, and an illustration offering insight on the use of ROC surfaces will be described. The Discussion in Section 5 will constitute an effort to provide guidance for future research to the interested reader.

2. ROC SURFACE ANALYSIS

2.1. Description of the problem

To define formally the general three-class classification problem, suppose that n_1 measurements from Class 1, denoted by X_1, follow a distribution with cumulative distribution function F_1 (*i.e.* $X_1 \sim F_1$), and similarly for n_2 measurements from Class 2, $X_2 \sim F_2$, and for n_3 measurements from Class 3, $X_3 \sim F_3$. A decision rule that classifies subjects in one of these classes can be defined using two ordered threshold points $c_1 < c_2$. Specifically, suppose that the ordering of interest is $X_1 < X_2 < X_3$. The researcher's goal is the assessment of the quality of a diagnostic marker in classifying correctly subjects from the three ordered classes.

2.2. Definition

The construction of the ROC surface is based on the following algorithm: Decide for Class 1 when a measurement is less than c_1, for Class 2 when it is between c_1 and c_2, for Class 3 otherwise. This decision rule will result in three True Class Fractions (TCFs) and six False Class Fractions (FCFs). Then, $\mathrm{TCF}_1 = P(X_1 \leq c_1)$, $\mathrm{TCF}_2 = P(c_1 < X_2 \leq c_2)$, and $\mathrm{TCF}_3 = P(X_3 > c_2)$. Also, $\mathrm{FCF}_{12} = P(c_1 \leq X_1 \leq c_2)$ and the remaining five possible FCF_{ij}, $i, j = 1, 2, 3$, $i \neq j$ are defined accordingly. Varying c_1, c_2 in the union of the supports of F_1, F_2, and F_3, $(\mathrm{TCF}_1, \mathrm{TCF}_2, \mathrm{TCF}_3)$ can be plotted in a three-dimensional coordinate system to produce the ROC surface in the unit cube. The True Class Fractions take values in $[0, 1]$ with corner coordinates $\{(1, 0, 0),\ (0, 1, 0),\ (0, 0, 1)\}$. Thus, the ROC surface is the 3-dimensional plot in the unit cube depicting $(F_1(c_1), F_2(c_2) - F_2(c_1), 1 - F_3(c_2))$, for all cut-off points (c_1, c_2), with $c_1 < c_2$. The functional form of the ROC surface is $\mathrm{ROC}_s(\mathrm{TCF}_1, \mathrm{TCF}_3) = F_2(F_3^{-1}(1 - \mathrm{TCF}_3)) - F_2(F_1^{-1}(\mathrm{TCF}_1))$ (Nakas and Yiannoutsos, 2004). It can be seen that this is a generalization of the ROC curve in three dimensions since projecting the ROC surface to the plane defined by TCF_2 versus TCF_1, *i.e.* setting $\mathrm{TCF}_3 = 0$, the ROC curve between Classes 1 and 2 is produced, *i.e.* $\mathrm{ROC}(\mathrm{TCF}_1) = 1 - F_2(F_1^{-1}(\mathrm{TCF}_1))$. The latter is the equivalent construction of the ROC curve depicting $(\mathrm{TCF}_1(c_1), \mathrm{TCF}_2(c_1))$ instead of $(\mathrm{FCF}_{12}(c_1), \mathrm{TCF}_2(c_1))$. Similarly, the projection of the ROC surface to the plane defined by the axes TCF_2, TCF_3, yields the ROC curve between Classes 2 and 3, *i.e.* $\mathrm{ROC}(\mathrm{TCF}_3) = F_2(F_3^{-1}(1 - \mathrm{TCF}_3))$, the latter being the functional form of TCF_2 versus TCF_3 analogous to specificity versus sensitivity rather than the other way around. For reasons of brevity, a pictorial representation will be provided in Section 4.

2.3. The Volume Under the ROC Surface (VUS)

The Volume Under the ROC Surface (VUS) is equal to $P(X_1 < X_2 < X_3)$. An unbiased non-parametric estimator of VUS is given by

$$\widehat{\mathrm{VUS}} = \frac{1}{n_1 n_2 n_3} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} I(X_{1i}, X_{2j}, X_{3k}) \ ,$$

where $I(X_1, X_2, X_3)$ equals one if X_1, X_2, X_3 are in the correct order and zero otherwise (Dreiseitl *et al.*, 2000). The definition of $I(X_1, X_2, X_3)$ can be adapted to adjust for the presence of ties. Specifically, when ties are present, define: $I(X_1, X_2, X_3) = 1/2$ if $X_1 = X_2 < X_3$ or if $X_1 < X_2 = X_3$ and $I(X_1, X_2, X_3) = 1/6$ if $X_1 = X_2 = X_3$, and $I(X_1, X_2, X_3) = 0$ (or 1 if perfectly ordered) otherwise.

The expected value of VUS will be then

$$P(X_1 < X_2 < X_3) + \frac{1}{2}\,P(X_1 < X_2 = X_3) \\ + \frac{1}{2}\,P(X_1 = X_2 < X_3) + \frac{1}{6}\,P(X_1 = X_2 = X_3)\ .$$

The VUS takes the value $1/3! = 1/6$ when the three distributions completely overlap and the value one when the three classes are perfectly discriminated in the correct order. Parametric approaches for the estimation of VUS have been discussed in Xiong *et al.* (2006). Kang and Tian (2013) offer an extensive study comparing possible parametric and non-parametric approaches for the estimation of VUS in terms of bias and root mean square error.

In several situations in practice researchers may wish to limit the study of the ROC surface to a clinically relevant range of measurement values. In such cases the partial VUS has been defined in Xiong *et al.* (2006). The partial VUS generalizes the notion of the partial AUC in the two-class problem (see *e.g.* Zhou *et al.*, 2011).

2.4. ROC surface modelling

Restate the functional form of the ROC surface, by writing $\mathrm{TCF}_1 = p_1$ and $\mathrm{TCF}_3 = p_3$, as follows:

(2.1)

$$\mathrm{ROC_s}(p_1, p_3) = \begin{cases} F_2\big(F_3^{-1}(1-p_3)\big) - F_2\big(F_1^{-1}(p_1)\big)\,, & \text{if } \ F_1^{-1}(p_1) \le F_3^{-1}(1-p_3)\ , \\ 0\,, & \text{otherwise}\,. \end{cases}$$

Then, VUS is defined as

$$\mathrm{VUS} \ = \int_0^1 \int_0^{1-F_3(F_1^{-1}(p_1))} \mathrm{ROC_s}(p_1, p_3)\ \mathrm{d}p_3\,\mathrm{d}p_1\ .$$

2.4.1. Empirical and non-parametric estimation

The empirical estimator of the ROC surface can be obtained by replacing the distribution functions in the definition of the ROC surface with their empirical counterparts. The empirical, non-parametric estimator of the ROC surface is

$$\widehat{\mathrm{ROC_s}}(p_1, p_3) = \begin{cases} \widehat{F}_2\big(\widehat{F}_3^{-1}(1-p_3)\big) - \widehat{F}_2\big(\widehat{F}_1^{-1}(p_1)\big)\,, & \text{if } \ \widehat{F}_1^{-1}(p_1) \le \widehat{F}_3^{-1}(1-p_3)\ , \\ 0\,, & \text{otherwise}\,, \end{cases}$$

where $\widehat{F}_1, \widehat{F}_2$ and $\widehat{F}_3$ are the empirical distribution functions for the measurements from the three classes.

Most recently, kernel approaches for the estimation of the ROC surface have been studied (Kang and Tian, 2013). Specifically, F_1, F_2, and F_3, can be modeled through Gaussian kernel estimators of the form $F_i(t) = 1/n_i \Sigma_{j=1}^{n_i} \Phi\{(t - X_{ij})/h_i\}$, for $i = 1, 2, 3$. For the bandwidth h_i, which controls the amount of smoothing, Kang and Tian (2013) have considered $h_i = \{4/(3n_i)\}^{1/5} \min(\mathrm{SD}_i, \mathrm{IQR}_i/1.349)$; here, SD_i and IQR_i are the standard deviation and interquartile range, respectively, for the X_i measurements.

Bayesian non-parametric estimation of the ROC surface based on Finite Polya Tree (FPT) prior distributions for the three-classes was studied by Inácio *et al.* (2011). The model is specified hierarchically and involves the specification of independent FPT prior distributions for F_i, for $i = 1, 2, 3$, conditional on a set of hyperparameters, *i.e.*

$$F_i \mid c_i, \theta_i \sim \mathrm{FPT}_{J_i}(F_{\theta_i}, c_i) \ , \qquad i = 1, 2, 3 \ .$$

Suppose, that the F_i are centered at $F_{\boldsymbol{\theta}_i} = N(\mu_i, \sigma_i)$, where $\boldsymbol{\theta}_i = (\mu_i, \sigma_i)$. The mixing parameters μ_i have independent normal prior distributions $N(a_{\mu_i}, b_{\mu_i})$, whereas σ_i have independent gamma prior distributions $\Gamma(a_{\sigma_i}, b_{\sigma_i})$. Hyperparameters are considered fixed. The levels of the finite Polya trees are set equal to J_i, and are used to determine the level of detail that is accommodated by the model; mathematical subtleties on the model can be found in Inácio *et al.* (2011).

2.4.2. Parametric estimation

Under the assumption of normality for F_1, F_2, and F_3 (*i.e.* $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$, $X_3 \sim N(\mu_3, \sigma_3^2)$), Xiong *et al.* (2006) used the model in (2.1) to describe the general framework of the ROC surface and the VUS. The parametric form of the ROC surface is

$$\begin{aligned} \mathrm{ROC_s}(p_1, p_3) \ = \ & \Big\{ \Phi\big(\beta_1 + \beta_2 \Phi^{-1}(1 - p_3)\big) - \Phi\big(\beta_3 + \beta_4 \Phi^{-1}(p_1)\big) \Big\} \\ & \times \ \mathbb{1}_{\{\beta_3 + \beta_4 \Phi^{-1}(p_1) \leq \beta_1 + \beta_2 \Phi^{-1}(1 - p_3)\}}(p_1, p_3) \ , \end{aligned}$$

where $\mathbb{1}$ denotes the indicator function, Φ is the distribution function of the standard normal, and $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)^{\mathrm{T}}$ specifies the parameters of the ROC surface. If the normality assumption is valid, the components of $\boldsymbol{\beta}$ may be expressed as functions of the means and variances of the three normal distributions which model F_1, F_2, and F_3, as follows:

$$\beta_1 = \frac{\mu_3 - \mu_2}{\sigma_2} \, , \qquad \beta_2 = \frac{\sigma_3}{\sigma_2} \, , \qquad \beta_3 = \frac{\mu_1 - \mu_2}{\sigma_2} \, , \qquad \beta_4 = \frac{\sigma_1}{\sigma_2} \ .$$

Kang and Tian (2013) have considered the use of the Box–Cox transformation for non-normally distributed data prior to the use of the parametric normal model and have compared with the kernel approach they proposed in terms of the bias and accuracy of the estimation of the VUS (see §2.4.1).

Under the Bayesian parametric paradigm, in order to find estimates for the beta parameters, a Markov Chain Monte Carlo approach is needed. A Metropolis–Hastings algorithm or a Gibbs sampler can be employed. The use of the Metropolis–Hastings algorithm with uninformative normal priors for the means and uninformative gamma prior distributions for the standard deviations is recommended in Inácio *et al.* (2011). However, studies focusing on Bayesian parametric approaches for the ROC surface have not appeared in the literature yet.

2.4.3. Semi-parametric estimation

Semi-parametric estimation of the ROC surface was studied by Li and Zhou (2009) generalizing the results of the two-class case in Hsieh and Turnbull (1996) and by Nze Ossima *et al.* (2013), generalizing the results of the two-class case in Gönen and Heller (2010). The estimation of the ROC surface of a diagnostic marker with continuous measurements given covariate information has been considered in Li *et al.* (2012). Specifically, suppose that the measurements of the diagnostic marker under study can be modeled through the following general regression model for a set of p covariates, $\boldsymbol{Z} = (Z_1, ..., Z_p)^{\mathrm{T}}$,

$$g(X_i) = \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\beta}_i + \sigma_i\,\varepsilon\ , \qquad i = 1, 2, 3\ , \tag{2.2}$$

where g is a strictly monotone increasing function, $\boldsymbol{\beta}_i = (\beta_{i1}, ..., \beta_{ip})^{\mathrm{T}}$ are the regression coefficients for Class i, σ_i is a class-specific scale parameter, and ε is the error following a common distribution function G with support $(-\infty, \infty)$ for all three classes. Then, the construction of the ROC surface is based on the rule: Decide for Class 1 when the diagnostic marker's measurement estimate from (2.2) is less than c_1, for Class 2 when it is between c_1 and c_2, for Class 3 otherwise.

2.5. Inference based on the VUS

Formal assessment of the diagnostic accuracy of a marker in a three-class classification problem via its VUS can be based on testing the null hypothesis $H_0 : \mathrm{VUS} = 1/6$ versus the alternative of interest. The test statistic is

$$Z_1 = \frac{\widehat{\mathrm{VUS}} - 1/6}{\sqrt{\mathrm{var}(\widehat{\mathrm{VUS}})}} \sim N(0, 1)\ . \tag{2.3}$$

The $\widehat{\text{VUS}}$ is the non-parametric estimate of VUS. Then, Z_1 is normally distributed based on results from U-statistics theory (Pepe, 2003). Variance of $\widehat{\text{VUS}}$ can be estimated by using U-statistics methodology or the bootstrap (Nakas and Yiannoutsos, 2004). The bootstrap approach consists of sampling with replacement n_1, n_2, n_3 subjects from the initial samples from X_1, X_2, X_3 respectively, and calculating the VUS for each of the b replications of this procedure. The bootstrap estimate of the variance of VUS is the sample variance of the b bootstrap VUSs (Nakas and Yiannoutsos, 2004). Properties of non-parametric estimators of the variance of $\widehat{\text{VUS}}$ have been studied by Guangming *et al.* (2013). Based on Z_1, 95% confidence intervals for VUS can be constructed in a straightforward fashion. Wan (2012) proposed an empirical likelihood confidence interval for the non-parametric estimate of VUS.

The parametric approach for confidence interval construction for VUS is studied in Xiong *et al.* (2006). Confidence intervals are constructed based on the Delta method, otherwise the bootstrap can be used, where for each bootstrap replication the parametric VUS is calculated. Non-parametric predictive inference for the ROC surface and the VUS is developed in Coolen-Maturi *et al.* (2013).

Regarding the comparison of VUSs, consider the case where two markers (A and B) are measured on the same $n = n_1 + n_2 + n_3$ specimens which are classified by a gold standard procedure into three ordered disease classes. Let $(X_1^{\text{A}}, X_2^{\text{A}}, X_3^{\text{A}})$ and $(X_1^{\text{B}}, X_2^{\text{B}}, X_3^{\text{B}})$ be the values for markers A and B, respectively.

To compare VUS^{A} and VUS^{B} via their non-parametric, empirical estimates, Dreiseitl *et al.* (2000) proposed a U-statistics approach. Specifically, the null hypothesis $H_0 : \text{VUS}^{\text{A}} = \text{VUS}^{\text{B}}$ is tested by calculating

$$Z_2 = \frac{\widehat{\text{VUS}}^{\text{A}} - \widehat{\text{VUS}}^{\text{B}}}{\sqrt{\text{var}\big(\widehat{\text{VUS}}^{\text{A}}\big) + \text{var}\big(\widehat{\text{VUS}}^{\text{B}}\big) - 2\,\text{cov}\big(\widehat{\text{VUS}}^{\text{A}}, \widehat{\text{VUS}}^{\text{B}}\big)}},$$

and then comparing this value to a standard normal distribution. The variance and covariance of $\widehat{\text{VUS}}$ can be estimated using the estimators provided in Dreiseitl *et al.* (2000). Alternatively, the bootstrap can be used to test H_0 as in Nakas and Yiannoutsos (2004). Xiong *et al.* (2007) have studied the parametric analogue for the comparison of VUSs based on the results in Xiong *et al.* (2006), while Tian *et al.* (2011) consider the parametric approach using notions of generalized pivots. Inference for specific TCFs is studied in Dong *et al.* (2011, 2013).

2.6. The ROC umbrella

The notion of the ROC surface has been generalized to accommodate cases with umbrella or tree orderings (*i.e.* $X_1 < X_3 > X_2$ or $X_2 > X_1 < X_3$, respec-

tively) between the three classes under study by Nakas and Alonzo (2007). The ROC surface and VUS reviewed in the previous sections are not applicable when such orderings are of interest. Specifically, these approaches do not allow one to assess the ability of a marker to differentiate two disease classes from a third disease class without requiring a specific monotone order for the three disease classes under study. The derivation of an ROC surface for the ordering $X_2 > X_1 < X_3$ is reviewed here, however, the derivation is analogous for the other ordering.

Using the fact that $(X_2 > X_1 < X_3) = (X_1 < X_2 < X_3) \cup (X_1 < X_3 < X_2)$, or equivalently $P(X_2 > X_1 < X_3) = P(X_1 < X_2 < X_3) + P(X_1 < X_3 < X_2)$, the construction of two ROC surfaces (say A and B) corresponding to the orderings $X_1 < X_2 < X_3$ and $X_1 < X_3 < X_2$, respectively, is possible. These are the plots of the points: $\left(\mathrm{TCF}_1^{\mathrm{A}}(c_1, c_2), \mathrm{TCF}_2^{\mathrm{A}}(c_1, c_2), \mathrm{TCF}_3^{\mathrm{A}}(c_1, c_2)\right)$ and $\left(\mathrm{TCF}_1^{\mathrm{B}}(c_1, c_2), \mathrm{TCF}_2^{\mathrm{B}}(c_1, c_2), \mathrm{TCF}_3^{\mathrm{B}}(c_1, c_2)\right)$, respectively, with $(c_1, c_2) \in \mathbb{R}^2$ and $c_1 < c_2$.

The umbrella ordering can be viewed however on a single graph in the unit cube by plotting on the same axes defined by $x = \mathrm{TCF}_1^{\mathrm{A}}$, $y = \mathrm{TCF}_2^{\mathrm{A}}$, $z = \mathrm{TCF}_3^{\mathrm{A}}$, in turn

$$\left(\mathrm{TCF}_1^{\mathrm{A}}(c_1, c_2),\ \mathrm{TCF}_2^{\mathrm{A}}(c_1, c_2),\ \mathrm{TCF}_3^{\mathrm{A}}(c_1, c_2)\right)$$

and

$$\left(1 - \mathrm{TCF}_1^{\mathrm{B}}(c_1, c_2),\ 1 - \mathrm{TCF}_2^{\mathrm{B}}(c_1, c_2),\ 1 - \mathrm{TCF}_3^{\mathrm{B}}(c_1, c_2)\right),$$

with $(c_1, c_2) \in \mathbb{R}^2$ and $c_1 < c_2$. It can be shown that surfaces A, B thus constructed on a single graph, are disjoint.

The resulting umbrella ROC graph is a diagnostic plot for the visual assessment of the degree of separation in the given ordering of the three populations based on the samples. The volume under surface A plus the volume over surface B can be used for inference. We refer to this summary measure as the umbrella volume (UV). UV is equivalently the sum of the volumes under the ROC surfaces A and B corresponding to the monotone orderings $X_1 < X_2 < X_3$ and $X_1 < X_3 < X_2$, respectively. The umbrella ROC graph contains both ordered ROC surfaces.

The non-parametric unbiased estimator of the volume of the umbrella ROC graph $P(X_2 > X_1 < X_3)$ is:

$$\widehat{UV} = \frac{1}{n_1 n_2 n_3} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} I_U(X_{1i}, X_{2j}, X_{3k}) \ ,$$

where $I_U(X_1, X_2, X_3)$ equals one if $X_2 > X_1 < X_3$ and zero otherwise; the UV varies from zero to one and is equal to $P(X_1 < X_2 < X_3) + P(X_1 < X_3 < X_2) = 1/6 + 1/6 = 1/3$ when the three distributions completely overlap and equals one when the three classes are perfectly discriminated in the given ordering.

In practice, ties may occur between measurements in the three disease classes, in which case $I_U(X_1, X_2, X_3)$=1 if $X_1 < X_2 = X_3$, $I_U(X_1, X_2, X_3)$=1/2 if $X_1 = X_2 < X_3$ or if $X_1 = X_3 < X_2$, and $I_U(X_1, X_2, X_3)$=1/6 if $X_1 = X_2 = X_3$. The expected value of UV will then be

$$P(X_1 < X_2 < X_3) + P(X_1 < X_3 < X_2) + P(X_1 < X_2 = X_3) + \frac{1}{2} P(X_1 = X_2 < X_3) + \frac{1}{2} P(X_1 = X_3 < X_2) + \frac{1}{6} P(X_1 = X_2 = X_3) \ .$$

Comparison of umbrella ROC volumes in a non-parametric framework has been studied in Alonzo and Nakas (2007), while the umbrella ROC has not been studied in the parametric framework yet. Alonzo *et al.* (2009) provide a comparison of tests for restricted orderings in the three-class case, illustrating the usefulness of ROC surfaces and ROC umbrellas in different applied contexts.

2.7. The ROC manifold

For the k-class problem, with $k > 3$, based on a single diagnostic marker, an ROC manifold may be constructed as described in Nakas and Yiannoutsos (2004). Using $k-1$ ordered decision thresholds c_j, $j = 1, ..., k-1$, with $c_1 < \cdots < c_{k-1}$, define a decision rule as in the three-class case given above. Then k TCFs are defined in a k-dimensional space. The ROC manifold is produced by varying the $k-1$ ordered decision thresholds. The Hypervolume Under the ROC Manifold (HUM) is

$$\text{HUM} \ = \ P\Big\{(X_1 < X_2) \cap \cdots \cap (X_{k-1} < X_k)\Big\} .$$

The HUM will vary from $1/k!$ to 1, taking the value $1/k!$ for a completely uninformative marker and the value 1 when the k populations are perfectly separated.

A non-parametric unbiased estimate of HUM is

$$\widehat{\text{HUM}} \ = \ \frac{1}{n_1 \cdots n_k} \sum_{i_1=1}^{n_1} \cdots \sum_{i_k=1}^{n_k} I(X_{1i_1}, ..., X_{ki_k}) \ ,$$

where the n_i, for $i = 1, ..., k$, are the sample sizes from the k populations and the function $I(X_1, ..., X_k)$ is defined in analogy to the three-class case. The ROC manifold and HUM have not been studied in a parametric framework yet. Theoretical extensions relating to the general k-class problem are studied in Davidov and Herman (2012).

2.8. Other topics in three- and k-class ROC methodology

Computational aspects regarding the calculation of the VUS or HUM when computational complexity is an issue have also appeared in the literature (Waegeman *et al.*, 2008a,b; Clémençon *et al.*, 2013). Alternative approaches for the generalization of the ROC curve in three- and multiple-class classification problems have been proposed by Yang and Carlin (2000), Hand and Till (2001), Wan and Zhang (2009) and Yang and Zhao (2010). These approaches, however, address specific research questions in the sense that they do not offer a complete theoretical framework for the generalization of ROC curve analysis and will not be studied further in this review. Generalizations of ROC analysis notions when the gold standard is continuous-scale rather than categorical has been studied by Obuchowski (2006) and by Shiu and Gatsonis (2012).

In the two-class case, considerable amount of research has been conducted to address issues where no gold standard is available for the characterization of the true status of the subjects in the study, or when the gold standard information is available for a fraction of the subjects in the study, *i.e.*, in the presence of verification bias (see *e.g.* Pepe, 2003; Zhou *et al.*, 2011). Only a few papers have appeared that introduce these notions in ROC surface analysis (Chi and Zhou, 2009; Wang *et al.*, 2011; Kang *et al.*, 2013b). Bantis *et al.* (2013) have used a cubic spline smoothing approach to model the ROC surface when measurements are subject to a limit of detection.

Theoretical properties of the ROC surface and ROC manifold that span beyond the scopes of the current article have been studied in Scurfield *et al.* (1998), He and Frey (2006), He *et al.* (2006), Everson and Fieldsend (2006), Edwards and Metz (2007), Sahiner *et al.* (2008), He and Frey (2008), He and Frey (2009), He *et al.* (2010), Schubert *et al.* (2011), Edwards and Metz (2012), and Edwards (2013).

3. THE GENERALIZED YOUDEN INDEX

3.1. Definition

A three-class Youden index has been recently proposed for the assessment of accuracy and cut-off point selection in the three-class setting (Nakas *et al.*, 2010, 2013). Specifically, define:

$$\begin{aligned} J_3 &= \max_{c_1,c_2}\Big\{\mathrm{TCF}_1 + \mathrm{TCF}_2 + \mathrm{TCF}_3 - 1\Big\} \\ &= \max_{c_1,c_2}\Big\{F_1(c_1) + F_2(c_2) - F_2(c_1) - F_3(c_2)\Big\}. \end{aligned} \tag{3.1}$$

This is a constrained optimization problem with $c_1 < c_2$. This latter condition will always be true if a usual stochastic order of the form $P(X_1 > x) \leq P(X_2 > x) \leq P(X_3 > x)$ holds. The pair of cut-off points c_1, c_2 that corresponds to J_3 is considered optimal and can be used in practice for decision making in the three-class case. As in the two-class setting, weights can be added to the definition of J_3 to reflect the relative importance of the three TCFs.

3.2. Properties

The generalized Youden index lends itself to a natural unification of the two- and three-class analysis approaches. Denote by $J_{3;(1,2,3)}$ the J_3 index corresponding to the ordering $X_1 < X_2 < X_3$ and by $J_{2;(i,j)}$, the ordinary Youden index corresponding to the ordering $X_i < X_j$, for $i, j = 1, 2, 3$. Then, by the definitions of J_2 and J_3 above, it follows that

$$\begin{aligned} J_{3;(1,2,3)} &= \max_{c_1,c_2}\Big\{F_1(c_1) - F_2(c_1) + F_2(c_2) - F_3(c_2)\Big\} \\ &= \max_{c_1}\Big\{F_1(c_1) - F_2(c_1)\Big\} + \max_{c_2}\Big\{F_2(c_2) - F_3(c_2)\Big\} \\ &= J_{2;(1,2)} + J_{2;(2,3)} \ . \end{aligned}$$

Thus, J_3 is the sum of the Youden index for the two-class analysis of classes 1 and 2 and the Youden index for the two-class analysis of classes 2 and 3. This result holds if weights are introduced in the definition of J_3 since λ can be set to one and $\nu^* = \nu/\lambda$, $\mu^* = \mu/\lambda$ can be used instead of ν, μ in the definition of J_3^+. Then, $J^+_{3;(1,2,3)} = \max_{c_1,c_2}\{\nu^* \cdot \mathrm{TCF}_1 + \mu^* \cdot \mathrm{TCF}_2 + \mathrm{TCF}_3 - 1\} = J^+_{2;(1,2)} + J^+_{2;(2,3)}$. This result also holds whenever the ordering $X_1 < X_2 < X_3$ is true, thus $c_1 < c_2$. A counterexample, where the ordering is not true and, as a result, the property does not hold, can easily be constructed. As a rule of thumb, pairwise AUCs for adjacent classes can reveal the correct order, which in turn can be used for the three-class analysis. From the property above it follows that J_3 takes on values in $[0, 2]$. To define J_3 in $[0, 1]$, Luo and Xiong (2013) proposed using $J_3/2$.

3.3. Estimation

Note that J_3 can be estimated non-parametrically by using the empirical distribution functions in the definition in (3.1), *i.e.* $\widehat{J}_3 = \max_{c_1,c_2}\{\widehat{F}_1(c_1) + \widehat{F}_2(c_2) - \widehat{F}_2(c_1) - \widehat{F}_3(c_2)\}$, or parametrically based on distributional assumptions for the data. Empirical non-parametric estimation of the generalized Youden index has been considered in Nakas *et al.* (2010, 2013), while parametric estimation based on normality assumptions has been described in Luo and Xiong (2012, 2013). Luo and Xiong created an R-package (`DiagTest3Grp`) for the estimation of the VUS, generalized Youden index and respective optimal cut-off

points under the parametric normal model, of which further details can be found in Luo and Xiong (2012). Estimation and use of the generalized Youden index for non-parametric predictive inference is studied in Coolen-Maturi *et al.* (2013).

3.4. Other measures of discrimination ability

The generalized Youden index can serve as an index of the discrimination ability of a diagnostic marker for the purpose of selecting the cut-off points that may be used for decision making, while the VUS is the measure of choice for the evaluation of the discrimination ability of the marker under study *per se*. The reason for the selective use of different measures of discrimination ability is the interpretation of the measure itself. Other measures for the evaluation of the discrimination ability of a marker rising from the definition of the ROC surface has also been proposed in the literature (*e.g.* Van Calster *et al.*, 2012a,b) but have not received much attention from the research community. Use of a general cost function for the selection of cut-off points in multiple-class diagnostic testing has been studied in Skaltsa *et al.* (2012).

4. ILLUSTRATION OF ROC SURFACE ANALYSIS

CA19-9 is a standard pancreatic cancer diagnostic marker. Measurements on 40 pancreatic cancer patients, 23 pancreatitis patients, and 40 healthy controls were available. The dataset that is used here for illustrative purposes is part of the dataset in Leichtle *et al.* (2013). Evaluation of CA19-9 in terms of its diagnostic ability to differentiate between the three classes in the order

$$\texttt{Controls} < \texttt{Pancreatitis} < \texttt{Cancer}$$

is illustrated. Descriptive statistics are given in Table 1, while respective boxplots are depicted in Figure 1.

Table 1: Descriptive statistics for CA 19-9 marker measurements for the three classes under study.

	`Controls`	`Pancreatits`	`Cancer`
mean	6.94	22.50	200.46
sd	4.74	30.88	237.85
median	6.60	8.51	111.60
min	0.6	2.5	0.6
max	20.67	121.80	971.50
N	40	23	40

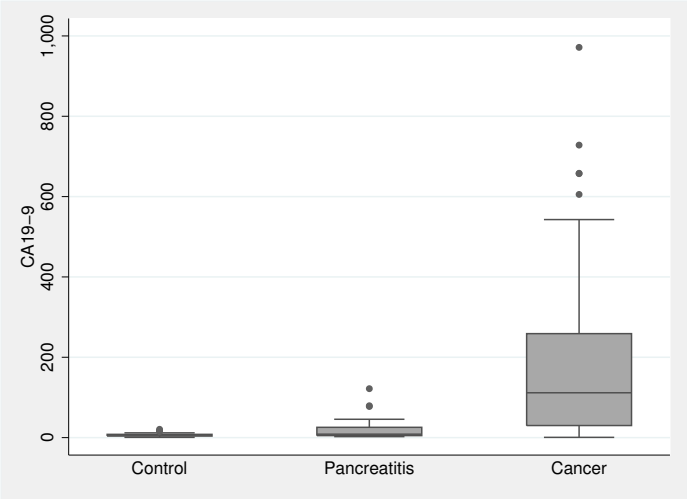

Figure 1: Boxplots of CA 19-9 marker measurements for the three classes under study.

Frequentist and Bayesian non-parametric ROC surfaces are depicted in Figure 2. The empirical non-parametric VUS is equal to 0.528 (95% CI: $0.403, 0.654$; $p < 0.001$), while the VUS based on the Bayesian non-parametric approach is equal to 0.550 (95% CI: $0.455, 0.652$; $p < 0.001$). The generalized Youden index J_3 is 0.929, resulting in $c_1 = 8.40$ and $c_2 = 25.60$. The cut-off point c_1 corresponds to the diagnosis between pancreatitis patients and healthy controls, while c_2 discriminates between pancreatic cancer patients and pancreatitis patients.

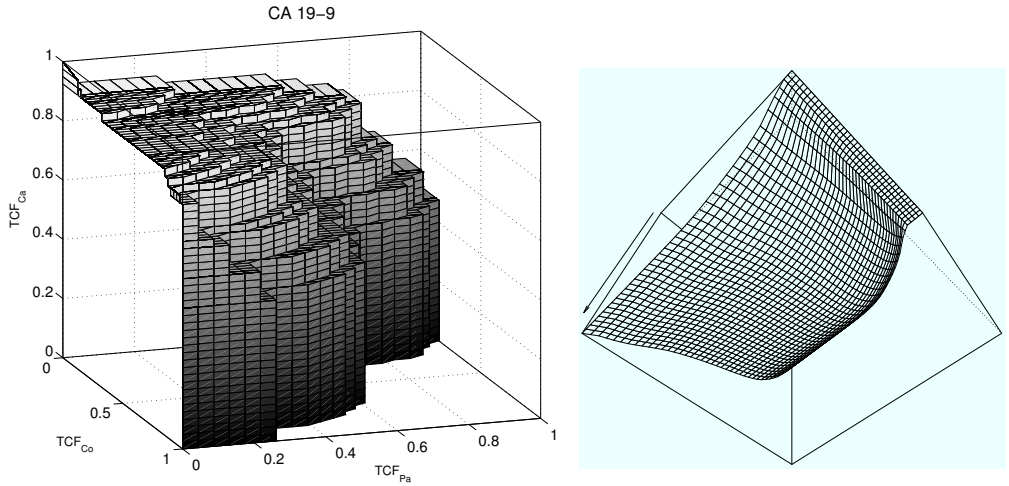

Figure 2: Non-parametric ROC surface for the CA 19-9 data (left panel) and Bayesian non-parametric model from a different viewpoint (right panel).

The corresponding TCF for healthy controls is equal to 80.00%. Regarding pancreatitis patients TCF is just 30.40%, while for pancreatic cancer patients TCF is 82.50%. Compare with the parametric approach that the `DiagTest3Grp` R-package employs: VUS $= 0.519$ (95% CI: $0.385, 0.653$; $p < 0.001$), $J_3 = 1.22 = 0.61 \times 2$, with $c_1 = 16.17$ and $c_2 = 86.62$, corresponding to the TCF triplet (0.974, $0.562, 0.684$) respectively. Unfortunately, the aforementioned R-package does not offer a graph for the ROC surface. However, the Shapiro–Wilk test rejects the normality assumption for all three groups in the study (with $p < 0.001$). Non-parametric approaches are thus considered as more reliable in our example.

Data analysis was conducted using `R version 3.0.0` (R Foundation for Statistical Computing, `http://www.R-project.org`), `Matlab R2013a` (MathWorks Inc., Natick, MA), and `Stata 11.2` (StataCorp LP, College Station, TX).

5. DISCUSSION

ROC surface analysis is a valuable tool for three-class classification problems as it generalizes ROC curve analysis in a natural way within the ROC framework. The utility of ROC surface analysis is demonstrated by the numerous applications that have already appeared in diverse scientific fields (*e.g.* Yu, 2012; Ratnasamy *et al.*, 2008; Yiannoutsos *et al.*, 2008; Abraham *et al.*, 2009; Wandishin and Mullen, 2009; Dalrymple-Alford *et al.*, 2010; Tremont *et al.*, 2011; Dunngalvin *et al.*, 2011; Bruña *et al.*, 2012; Cianferoni *et al.*, 2012; Coleman *et al.*, 2013; Migliaretti *et al.*, 2013; Leichtle *et al.*, 2013).

Until now, researchers have mainly dealt with geometric properties of the ROC surface itself and with generalizations of theoretical findings from the two-class case. Many issues remain to be resolved. Multiple-class classification within the ROC framework and the notion of the ROC umbrella have only scantly been dealt with. Based on the probabilistic properties of the VUS and UV, the claim that the ROC surface and VUS can also be used for three-class analysis when the classes are nominal instead of ordinal (*e.g.* Li and Fine, 2008) seems to be flawed. As a result, theoretical developments for the robustification of the framework of ROC surface analysis are still needed. Other topics of future research include further generalizations from the two-class case. Specifically, issues of future research include time-varying ROC surfaces and generalized linear modelling approaches for the ROC surface along the lines presented in Pepe (2003). The study of predictive values in the three-class and multiple-class case is also of interest. An initial attempt is presented in Yiannoutsos *et al.* (2008). Reclassification issues have just started attracting the interest of researchers in the field. Li *et al.* (2013) have extended the notions in Pencina *et al.* (2012) regarding the net reclassification improvement and integrated discrimination improvement for the k-class

case. Pepe and Thompson (2000) have studied the issue of combination of diagnostic markers in the two-class case via maximizing the area under the ROC curve and have compared this approach with the combination of the diagnostic markers measurements using logistic regression and linear discriminant analysis. Zhang and Li (2011) and Kang *et al.* (2013a) have generalized these results for ROC surface analysis by considering the combinations that maximize the VUS. Currently there is ongoing research on this topic regarding different approaches for VUS maximization using combinations of diagnostic markers.

The generalized Youden index is a simple, useful loss-function for the selection of the optimal cut-off points that can be used for decision-making based on a diagnostic marker of interest in the three-class case. Modelling approaches summarized here and in Kang and Tian (2013), could be employed to develop further practices for the choice of cut-off points after the construction of the ROC surface. Non-parametric predictive inference methods also offer a valuable framework for decision-making in three-class ROC analysis (Coolen *et al.*, 2013; Coolen-Maturi *et al.*, 2013).

R-packages for the implementation of ROC surface analysis tools are of great importance. Researchers interested in using ROC surface methodology should be able to use the Comprehensive R Archive Network repository for their research needs.

ACKNOWLEDGMENTS

The author wishes to thank Dr. Alexander Leichtle for providing the CA19-9 data and Dr. Vanda Inácio de Carvalho for providing the R-code for the implementation of the Bayesian non-parametric approach.

REFERENCES

Abraham, A. G.; Duncan, D. D.; Gange, S. J. and West, S. (2009). Computer-aided assessment of diagnostic images for epidemiological research, *BMC Medical Research Methodology*, **9**, art. no. 74.

Alonzo, T. A. and Nakas, C. T. (2007). Comparison of ROC umbrella volumes with an application to the assessment of lung cancer diagnostic markers, *Biometrical Journal*, **49**, 654–664.

Alonzo, T. A.; Nakas, C. T.; Yiannoutsos, C. T. and Bucher, S. (2009). A comparison of tests for restricted orderings in the three-class case, *Statistics in Medicine*, **28**, 1144–1158.

Bantis, L. E.; Tsimikas, J. V. and Georgiou, S. D. (2013). Smooth ROC curves and surfaces for markers subject to a limit of detection using monotone natural cubic splines, *Biometrical Journal*, **55**, 719–740.

Bruña, R.; Poza, J.; Gòmez, C.; García, M.; Fernández, A. and Hornero, R. (2012). Analysis of spontaneous MEG activity in mild cognitive impairment and Alzheimer's disease using spectral entropies and statistical complexity measures, *Journal of Neural Engineering*, **9**, art. no. 036007.

Chi, Y. Y. and Zhou, X.-H. (2009). Receiver operating characteristic surfaces in the presence of verification bias, *Journal of the Royal Statistical Society. Ser.* C, **57**, 1–23.

Cianferoni, A.; Garrett, J. P.; Naimi, D. R.; Khullar, K. and Spergel, J. M. (2012). Predictive values for food challenge-induced severe reactions: Development of a simple food challenge score, *Israel Medical Association Journal*, **14**, 24–28.

Clémençon, S.; Robbiano, S. and Vayatis, N. (2013). Ranking data with ordinal labels: Optimality and pairwise aggregation, *Machine Learning*, **91**, 67–104.

Coleman, D. J.; Silverman, R. H.; Rondeau, M. J.; Lloyd, H. O.; Khanifar, A. A. and Chan, R. V. P. (2013). Age-related macular degeneration: Choroidal ischaemia?, *British Journal of Opthalmology*, **97**, 1020–1023.

Coolen, F. P. A.; Coolen-Schrijner, P.; Coolen-Maturi, T. and Elkhafifi, F. F. (2013). Nonparametric Predictive Inference for Ordinal Data, *Communications in Statistics—Theory and Methods*, **42**, 3478–3496.

Coolen-Maturi, T.; Elkhafifi, F. F. and Coolen, F. P. A. (2013). Nonparametric predictive inference for three-group ROC analysis, *Technical Report No 1307; Department of Mathematical Sciences, University of Durham, UK*, `http://www.npi-statistics.com/NPI-3ROC-report-1307.pdf`.

Davidov, O. and Herman, A. (2012). Ordinal dominance curve based inference for stochastically ordered distributions, *Journal of the Royal Statistical Society, Ser.* B:, **74**, 825–847.

Dalrymple-Alford, J.C.; MacAskill, M.R.; Nakas, C.T.; Livingston, L.; Graham, C.; Crucian, G. P.; Melzer, T. R.; Kirwan, J.; Keenan, R.; Wells, S.; Porter, R. J.; Watts, R. and Anderson, T. J. (2010). The MoCA: Well suited screen for cognitive impairment in Parkinson disease, *Neurology*, **75**, 1717–1725.

Dong, T.; Tian, L.; Hutson, A. and Xiong, C. (2011). Parametric and non-parametric confidence intervals of the probability of identifying early disease stage given sensitivity to full disease and specificity with three ordinal diagnostic groups, *Statistics in Medicine*, **30**, 3532–3545.

Dong, T.; Kang, L.; Hutson, A.; Xiong, C. and Tian, L. (2013). Confidence interval estimation of the difference between two sensitivities to the early disease stage, *Biometrical Journal*, DOI: 10.1002/bimj.201200012.

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

Dunngalvin, A.; Daly, D.; Cullinane, C.; Stenke, E.; Keeton, D.; Erlewyn-Lajeunesse, M.; Roberts, G. C.; Lucas, J. and Hourihane, J. O. (2011). Highly accurate prediction of food challenge outcome using routinely available clinical data, *Journal of Allergy and Clinical Immunology*, **127**, 633–639.

Edwards, D. C. (2013). Validation of monte carlo estimates of three-class ideal observer operating points for normal data, *Academic Radiology*, **20**, 908–914.

Edwards, D. C. and Metz, C. E. (2007). Optimization of restricted ROC surfaces in three-class classification tasks, *IEEE Transactions on Medical Imaging*, **26**, 1345–1356.

Edwards, D. C. and Metz, C. E. (2012). The three-class ideal observer for univariate normal data: Decision variable and ROC surface properties, *Journal of Mathematical Psychology*, **56**, 256–273.

Everson, R. M. and Fieldsend, J. E. (2006). Multi-class ROC analysis from a multi-objective optimisation perspective, *Pattern Recognition Letters*, **27**, 918–927.

Gönen, M. and Heller, G. (2010). Lehmann family of ROC curves, *Medical Decision Making*, **30**, 509–517.

Guangming, P.; Xiping, W. and Wang, Z. (2013). Nonparametric statistical inference for $P(X < Y < Z)$, *Sankhya* A, **75**, 118–138.

Hand, D. J. and Till, R. J. (2001). A simple generalisation of the area under the ROC curve for multiple-class classification problems, *Machine Learning*, **45**, 171–186.

He, X. and Frey, E. C. (2006). Three-class ROC analysis—the equal error utility assumption and the optimality of three-class ROC surface using the ideal observer, *IEEE Transactions on Medical Imaging*, **25**, 979–986.

He, X. and Frey, E. C. (2008). The meaning and use of the volume under a three-class ROC surface (VUS), *IEEE Transactions on Medical Imaging*, **27**, 577–588.

He, X. and Frey, E. C. (2009). The validity of three-class Hotelling trace (3-HT) in describing three-class task performance: Comparison of three-class volume under ROC surface (VUS) and 3-HT, *IEEE Transactions on Medical Imaging*, **28**, 185–193.

He, X.; Gallas, B. D. and Frey, E. C. (2010). Three-class ROC analysis: Toward a general decision theoretic solution, *IEEE Transactions on Medical Imaging*, **29**, 206–215.

He, X.; Metz, C.E.; Tsui, B. M. W.; Links, J. M. and Frey, E. C. (2006). Three-class ROC analysis - A decision theoretic approach under the ideal observer framework, *IEEE Transactions on Medical Imaging*, **25**, 571–581.

Heckerling, P. S. (2001). Parametric three-way Receiver Operating Characteristic surface analysis using Mathematica, *Medical Decision Making*, **21**, 409–417.

Hsieh, F. and Turnbull, B. W. (1996). Nonparametric and semiparametric estimation of the receiver operating characteristic curve, *Annals of Statistics*, **24**, 25–40.

Inácio, V.; Turkman, A. A.; Nakas, C. T. and Alonzo, T. A. (2011). Nonparametric Bayesian estimation of the three-way receiver operating characteristic surface, *Biometrical Journal*, **53**, 1011–1024.

Kang, L. and Tian, L. (2013). Estimation of the volume under the ROC surface with three ordinal diagnostic categories, *Computational Statistics and Data Analysis*, **62**, 39–51.

Kang, L.; Xiong, C.; Crane, P. and Tian, L. (2013a). Linear combinations of biomarkers to improve diagnostic accuracy with three ordinal diagnostic categories, *Statistics in Medicine*, **32**, 631–643.

Kang, L.; Xiong, C. and Tian, L. (2013b). Estimating confidence intervals for the difference in diagnostic accuracy with three ordinal diagnostic categories without a gold standard, *Computational Statistics and Data Analysis*, **68**, 326–338.

Krzanowski, W. J. and Hand, D. J. (2009). *ROC Curves for Continuous Data*, Chapman & Hall/CRC, Boca Raton.

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

Li, J. and Fine, J. (2008). ROC analysis with multiple classes and multiple tests: methodology and its application in microarray studies, *Biostatistics*, **9**, 566–576.

Li, J. and Zhou, X.-H. (2009). Nonparametric and semiparametric estimation of the three-way receiver operating characteristic surface, *Journal of Statistical Planning and Inference*, **139**, 4133–4142.

Li, J. L.; Zhou, X.-H. and Fine, J. P. (2012). A regression approach to ROC surface, with applications to Alzheimer's disease, *Science China Mathematics*, **55**, 1583–1595.

Li, J. L.; Jiang, B. and Fine, J. P. (2013). Multicategory reclassification statistics for assessing improvements in diagnostic accuracy, *Biostatistics*, **14**, 382–394.

Luo, J. and Xiong, C. (2012). `DiagTest3Grp`: An R package for analyzing diagnostic tests with three ordinal groups, *Journal of Statistical Software*, **51**, 1–24.

Luo, J. and Xiong, C. (2013). Youden index and associated cut-points for three ordinal diagnostic groups, *Communications in Statistics—Simulation and Computation*, **42**, 1213–1234.

Migliaretti, G.; Ciaramitaro, P.; Berchialla, P.; Scarinzi, C.; Andrini, R.; Orlando, A. Faccani, G. (2013). Teleconsulting for minor head injury: The piedmont experience, *Journal of Telemedicine and Telecare*, **19**, 33–35.

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

Nakas, C. T. and Alonzo, T. A. (2007). ROC graphs for assessing the ability of a diagnostic marker to detect three disease classes with an umbrella ordering, *Biometrics*, **63**, 603–609.

Nakas, C. T.; Alonzo, T. A. and Yiannoutsos, C. T. (2010). Accuracy and cut-off point selection in three-class classification problems using a generalization of the Youden index, *Statistics in Medicine*, **29**, 2946–2955.

Nakas, C. T.; Dalrymple-Alford, J. C.; Anderson, T. and Alonzo, T. A. (2013). Generalization of Youden index for multiple-class classification problems applied to the assessment of externally validated cognition in Parkinson disease screening, *Statistics in Medicine*, **32**, 995–1003.

Nakas, C. T. and Yiannoutsos, C. T. (2004). Ordered multiple-class ROC analysis with continuous measurements, *Statistics in Medicine*, **23**, 3437–3449.

Nze Ossima, A. D.; Daurès, J. P.; Bessaoud, F. and Trétarre, B. (2013). The generalized Lehmann ROC curves: Lehmann family of ROC surfaces, *Journal of Statistical Computation and Simulation*, DOI: 10.1080/00949655.2013.831863.

Obuchowski, N. A. (2006). An ROC-type measure of diagnostic accuracy when the gold standard is continuous-scale, *Statistics in Medicine*, **25**, 481–493.

Pencina, M. J.; D'Agostino Sr, R. B. and Demler, O. V. (2012). Novel metrics for evaluating improvement in discrimination: net reclassification and integrated discrimination improvements for normal variables and nested models, *Statistics in Medicine*, **31**, 101–113.

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

Pepe, M. S. and Thompson, M. L. (2000). Combining diagnostic test results to increase accuracy, *Biostatistics*, **1**, 123–140.

Ratnasamy, C.; Kinnamon, D. D.; Lipshultz, S. E. and Rusconi, P. (2008). Associations between neurohormonal and inflammatory activation and heart failure in children, *American Heart Journal*, **155**, 527–533.

Sahiner, B.; Chan, H. P.ăand Hadjiiski, L. M. (2008). Performance analysis of three-class classifiers: Properties of a 3-D ROC surface and the normalized volume under the surface for the ideal observer, *IEEE Transactions on Medical Imaging*, **27**, 215–227.

Schubert, C. M.; Thorsen, S. N. and Oxley, M. E. (2011). The ROC manifold for classification systems, *Pattern Recognition*, **44**, 350–362.

Scurfield, B. K. (1998). Generalization of the theory of signal detectability to n-event m-dimensional forced-choice tasks, *Journal of Mathematical Psychology*, **42**, 5–31.

Scurfield, B. K. (1996). Multiple-event forced-choice tasks in the theory of signal detectability, *Journal of Mathematical Psychology*, **40**, 253–269.

Skaltsa, K.; Jover, L.; Fuster, D. and Carrasco, J. L. (2012). Optimum threshold estimation based on cost function in a multistate diagnostic setting, *Statistics in Medicine*, **31**, 1098–1109.

Shiu, S. Y. and Gatsonis, C. (2012). On ROC analysis with nonbinary reference standard, *Biometrical Journal*, **54**, 457–480.

Tian, L.; Xiong, C.; Lai, C. Y. and Vexler, A. (2011). Exact confidence interval estimation for the difference in diagnostic accuracy with three ordinal diagnostic groups, *Journal of Statistical Planning and Inference*, **141**, 549–558.

Tremont, G.; Papandonatos, G. D.; Springate, B.; Huminski, B.; McQuiggan, M. D.; Grace, J.; Frakey, L. and Ott, B. R. (2011). Use of the telephone-administered Minnesota Cognitive Acuity Screen to detect mild cognitive impairment, *American Journal of Alzheimer's Disease and other Dementias*, **26**, 555–562.

Van Calster, B.; Van Belle, V.; Vergouwe, Y. and Steyerberg, E. W. (2012a). Discrimination ability of prediction models for ordinal outcomes: Relationships between existing measures and a new measure, *Biometrical Journal*, **54**, 674–685.

Van Calster, B.; Vergouwe, Y.; Looman, C. W. N.; Van Belle, V.; Timmerman, D. and Steyerberg, E. W. (2012b). Assessing the discriminative ability of risk models for more than two outcome categories, *European Journal of Epidemiology*, **27**, 761–770.

Waegeman, W.; De Baets, B.ăand Boullart, L. (2008a). Learning layered ranking functions with structured support vector machines, *Neural Networks*, **21**, 1511–1523.

Waegeman, W.; De Baets, B. and Boullart, L. (2008b). On the scalability of ordered multi-class ROC analysis, *Computational Statistics and Data Analysis*, **52**, 3371–3388.

Wan, S. (2012). An empirical likelihood confidence interval for the volume under ROC surface, *Statistics and Probability Letters*, **82**, 1463–1467.

Wan, S. and Zhang, B. (2009). Semiparametric ROC surfaces for continuous diagnostic tests based on two test measurements, *Statistics in Medicine*, **28**, 2370–2383.

Wandishin, M. S. and Mullen, S. J. (2009). Multiclass ROC analysis, *Weather and Forecasting*, **24**, 530–547.

Wang, Z.; Zhou, X.-H. and Wang, M. (2011). Evaluation of diagnostic accuracy in detecting ordered symptom statuses without a gold standard, *Biostatistics*, **12**, 567–581.

Xiong, C.; van Belle, G.; Miller, J. P. and Morris, J. C. (2006). Measuring and estimating diagnostic accuracy when there are three ordinal diagnostic groups, *Statistics in Medicine*, **25**, 1251–1273.

Xiong, C.; van Belle, G.; Miller, J. P.; Yan, Y.; Gao, F.; Yu, K. and Morris, J. C. (2007). A parametric comparison of diagnostic accuracy with three ordinal diagnostic groups, *Biometrical Journal*, **49**, 682–693.

Yang, H. and Carlin, D. (2000). ROC surface: A generalization of ROC curve analysis, *Journal of Biopharmaceutical Statistics*, **10**, 183–196.

Yang, H. and Zhao, L. (2010). A Method of Estimating and Comparing Volumes Under Receiver Operating Characteristic (ROC) Surfaces, *Statistics in Biopharmaceutical Research*, **2**, 279–291.

Yiannoutsos, C. T.; Nakas, C. T. and Navia, B. A. (2008). Assessing multiple-group diagnostic problems with multi-dimensional receiver operating characteristic surfaces: Application to proton MR Spectroscopy (MRS) in HIV-related neurological injury, *Neuroimage*, **40**, 248–255.

Yu, T. (2012). ROCS: Receiver operating characteristic surface for class-skewed high-throughput data, *PLoS ONE*, **7**, at. no. e40598.

Zhang, Y. and Li, J. (2011). Combining multiple markers for multi-category classification: An ROC surface approach, *Australian and New Zealand Journal of Statistics*, **53**, 63–78.

Zhou, X.-H.; Obuchowski, N. A. and McClish, D. K. (2011). *Statistical Methods in Diagnostic Medicine*, Second Edition, Wiley, New York.