SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2017, Vol. 26(3) 1429–1442
© The Author(s) 2015
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0962280215581694
journals.sagepub.com/home/smm
SAGE

Construction of joint confidence regions for the optimal true class fractions of Receiver Operating Characteristic (ROC) surfaces and manifolds

Leonidas E Bantis,[1] Christos T Nakas,[2] Benjamin Reiser,[3] Daniel Myall[4] and John C Dalrymple-Alford[4,5,6]

Abstract

The three-class approach is used for progressive disorders when clinicians and researchers want to diagnose or classify subjects as members of one of three ordered categories based on a continuous diagnostic marker. The decision thresholds or optimal cut-off points required for this classification are often chosen to maximize the generalized Youden index (Nakas et al., *Stat Med* 2013; 32: 995–1003). The effectiveness of these chosen cut-off points can be evaluated by estimating their corresponding true class fractions and their associated confidence regions. Recently, in the two-class case, parametric and non-parametric methods were investigated for the construction of confidence regions for the pair of the Youden-index-based optimal sensitivity and specificity fractions that can take into account the correlation introduced between sensitivity and specificity when the optimal cut-off point is estimated from the data (Bantis et al., *Biomet* 2014; 70: 212–223). A parametric approach based on the Box–Cox transformation to normality often works well while for markers having more complex distributions a non-parametric procedure using logspline density estimation can be used instead. The true class fractions that correspond to the optimal cut-off points estimated by the generalized Youden index are correlated similarly to the two-class case. In this article, we generalize these methods to the three- and to the general *k*-class case which involves the classification of subjects into three or more ordered categories, where ROC surface or ROC manifold methodology, respectively, is typically employed for the evaluation of the discriminatory capacity of a diagnostic marker. We obtain three- and multi-dimensional joint confidence regions for the optimal true class fractions. We illustrate this with an application to the Trail Making Test Part A that has been used to characterize cognitive impairment in patients with Parkinson's disease.

[1]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA
[2]Laboratory of Biometry, University of Thessaly, Volos, Greece
[3]Department of Statistics, University of Haifa, Haifa, Israel
[4]New Zealand Brain Research Institute, Christchurch, New Zealand
[5]Department of Psychology, University of Canterbury, Christchurch, New Zealand
[6]Department of Medicine, University of Otago, Christchurch, New Zealand

Corresponding author:
Christos T Nakas, Laboratory of Biometry, University of Thessaly, Phytokou street, 38446 N Ionia-Volos, Greece.
Email: cnakas@uth.gr

Keywords

ROC analysis, true class fractions, generalized Youden index, delta method, Trail Making Test

1 Introduction

A three-class approach is useful for progressive disorders when clinicians and researchers aim to characterize patients as members of one of three ordered categories. This approach is useful, for example, when discriminating between people with 'normal cognition', with 'mild cognitive impairment' (MCI), and with 'dementia', such as when pathology exists that causes Parkinson's disease or Alzheimer's disease.[1–4] Decision thresholds and the confidence regions generated by this approach may also help discriminate between competing diagnostic criteria that best separate an intermediate disease state such as MCI from both normal cognition and dementia.[5,6]

These discriminations are often based on a diagnostic marker with a score defined on a continuous scale. The Trail Making Test (TMT) is a visual search test that has been extensively used in neuropsychological assessment.[7] Even the relatively simple Part A of the TMT, in which patients are asked to draw a line through consecutively numbered Arabic numerals presented in circles scattered across a page, is sensitive to cognitive impairment.[8,9] The current study determined cut-off points on the TMT Part A to characterize disease states of 'normal cognition', MCI, and 'dementia' in Parkinson's disease patients. Such a tool might provide a simple quick screen to facilitate initial diagnosis and guide disease management decision making.

The ROC curve is the most common methodological procedure for the evaluation of a continuous- or an ordinal-scaled marker used for classification purposes in a two-class diagnostic problem (typically a non-diseased and a diseased group). In diagnostic problems involving three classes, the ROC surface is typically employed.[10] The ROC surface can be used to assess the discriminatory performance of a diagnostic marker simultaneously for three ordered groups. It additionally provides the corresponding pairwise ROC curves that may be assessed in a post hoc fashion for each pair of the three populations.[11]

The ROC curve is the plot of sensitivity versus 1-specificity of a diagnostic test of interest, as the cut-off point c used for the characterization of disease is varied. Suppose that measurements Y_1 from the non-diseased group follow a distribution with distribution function F_1 (i.e. $Y_1 \sim F_1$) and similarly for the diseased group, $Y_2 \sim F_2$. The specificity of the diagnostic test, for a specific cut-off point c, is $spec(c) = Prob(Y_1 \leq c) = F_1(c)$. The specificity is also known as the true negative fraction (TNF). Similarly, the sensitivity, defined as $sens(c) = Prob(Y_2 > c) = 1 - F_2(c)$, is often called the true positive fraction (TPF). The area under the ROC curve (AUC) is widely used as an overall performance index for the diagnostic marker under consideration. It holds that $\text{AUC} = Prob(Y_1 < Y_2)$.[12]

Regarding the general three-class classification problem, we consider three continuous random variables that refer to the marker scores for each group, namely Y_1, Y_2, Y_3. Without loss of generality, we may assume that higher marker measurements are more indicative of disease and that the ordering of interest is $Y_1 < Y_2 < Y_3$. Two ordered decision thresholds, $c_1 < c_2$, are needed for the characterization of disease states resulting in three possible true class fractions (TCFs), namely TCF_i, $i = 1, 2, 3$, that are defined as follows:

$$TCF_1 = P(Y_1 \leq c_1),\ TCF_2 = P(c_1 < Y_2 \leq c_2),\ TCF_3 = P(Y_3 > c_2)$$

The graph of all possible TCF_i, $i = 1, 2, 3$, triplets obtained based on all possible pairs of the decision threshold values $c_1 < c_2$ represents the ROC surface, that is, $ROC(c_1, c_2) = (TCF_1(c_1), TCF_2(c_1, c_2), TCF_3(c_2))$. Suppose that $Y_1 \sim F_1$, $Y_2 \sim F_2$, and $Y_3 \sim F_3$. The functional form of the ROC surface is[11]

$$ROC(TCF_1, TCF_3) = F_2(F_3^{-1}(1 - TCF_3)) - F_2(F_1^{-1}(TCF_1)) \tag{1}$$

The corresponding index of diagnostic performance for the diagnostic marker under study is the volume under the ROC surface (VUS). It holds that $VUS = P(Y_1 < Y_2 < Y_3)$.[13] Further generalization of the ROC context in the general case of k-class classification, where an ROC manifold (or hyper-surface) is defined, has been described in the literature.[11] Specifically, denote with $Y_1, Y_2, \ldots, Y_k$ the marker scores for each of the k classes: k possible TCFs can be defined based on $c_1 < c_2 < \cdots < c_{k-1}$ ordered decision thresholds. The corresponding TCFs define the ROC manifold. The overall discriminatory capacity of the marker under study is summarized by the the hyper-volume under the ROC manifold (HUM) which is equal to the probability $P(Y_1 < Y_2 < \cdots < Y_k)$.

After a diagnostic marker has been shown to be useful for diagnostic purposes, the selection of an optimal cut-off point based on some optimality criterion is needed. In practice, an optimal cut-off point c^* (two-class case), a pair of cut-off points, $c_1^* < c_2^*$ (three-class case), or a set of $(k - 1)$ ordered cut-off points (k-class case), are needed in order for the practitioner to classify each subject in one of the classes considered. The maximum of the Youden index is a very popular criterion used for cut-off point selection in the two-class case.[14] It is defined as $J = \max_c\{TPF(c) + TNF(c) - 1\} = \max_c\{F_1(c) - F_2(c)\}$. It can be estimated parametrically, non-parametrically, or empirically by plugging in the corresponding distribution estimates of each of the two groups.[15]

Although other criteria can be considered,[16] we will focus on the Youden index and its generalization to three or more classes generalizing and developing established two-class methodology.[15] In the two-class case, the effectiveness of the optimal Youden index-based cut-off estimate $\hat{c}^*$ is examined by the evaluation of the sensitivity and specificity pair associated with c^* along with the corresponding joint confidence region for sensitivity and specificity. Since c^* is estimated from data obtained on both non-diseased and diseased subjects, the corresponding sensitivity and specificity estimates are correlated and their variability changes accordingly due to the estimation of c^*. Typically, this correlation has been overlooked in the literature. Recently, joint confidence regions for the pair of sensitivity and specificity associated with c^* that take into account the aforementioned correlation have been proposed.[15,17] In this work we generalize to the general three- and k-class classification problem.

A generalization of the maximum of the Youden index in the three- and the k-class case has been developed for the empirical[1,18] and the parametric case assuming normality.[19] For the three-class case,

$$\begin{aligned} J_3 &= \max_{c_1, c_2; c_1 < c_2} \{TCF_1 + TCF_2 + TCF_3 - 1\} \\ &= \max_{c_1, c_2; c_1 < c_2} \{F_1(c_1) + F_2(c_2) - F_2(c_1) - F_3(c_2)\} \end{aligned} \tag{2}$$

The estimate of J_3 is associated with an estimate of the pair of cut-off points, $\hat{c}_1^*, \hat{c}_2^*$, that can be used in practice for screening purposes. The associated triplet $(TCF_1(\hat{c}_1^*), TCF_2(\hat{c}_1^*, \hat{c}_2^*), TCF_3(\hat{c}_2^*))$ characterizes the marker under study. The construction of confidence regions that correspond simultaneously to TCF_i, $i = 1, 2, 3$, will allow for proper inference. In the two-class case there is available literature that addresses this issue both in cases where the cut-off is fixed and known[12,20] as well as in cases where the cut-off is estimated by the available data.[15,17] Given that the optimal cut-off points are estimated based on the available data, the TCF_i, $i = 1, 2, 3$, estimates are correlated and their variability changes accordingly similarly to the two-class case. We extend existing methodology[15] and construct simultaneous three-dimensional confidence regions for the TCF_i, $i = 1, 2, 3$, when the pair of cut-off points is estimated from the data. Solutions for this issue in the three-class and general k-class case have not appeared in the literature. We generalize and enrich the two-class case approaches that we have recently proposed.[15]

In Section 2.1, we first present a parametric approach for the construction of joint three-dimensional confidence regions for the triplet TCF_1, TCF_2, and TCF_3 associated with the Youden-based (J_3) pair of cut-offs, based on the delta method. We initially discuss the construction of rectangular confidence regions. We then extend our approach to cases where the correlation and variances of the estimates of TCF_1, TCF_2, and TCF_3 associated with the estimated cut-off pair of $(\hat{c}_1^*, \hat{c}_2^*)$ are taken into account, by introducing 'egg-shaped' three-dimensional confidence regions. In Section 2.2, we robustify our approaches by exploring the use of monotone transformations that lead to marginal normality. We focus on the Box–Cox transformation and illustrate how the variability of the extra parameter of the transformation is taken into account. In Section 3, we develop two non-parametric alternatives. The first is based on kernel smoothing along with smooth bootstrapping[17] and the second is based on the logspline technique.[21] In Section 4, we offer a framework for the generalization in the general k-class case. A large simulation study is presented in Section 5, while the application of our methodological results to the Trails A test when screening for dementia is illustrated in Section 6. We end with a discussion.

2 Parametric approaches

2.1 Delta method approach

Here we assume that $Y_1 \sim N(\mu_1, \sigma_1^2)$, $Y_2 \sim N(\mu_2, \sigma_2^2)$, $Y_3 \sim N(\mu_3, \sigma_3^2)$. Then,

$$TCF_1 = \Phi\left(\frac{c_1 - \mu_1}{\sigma_1}\right), \quad TCF_2 = \Phi\left(\frac{c_2 - \mu_2}{\sigma_2}\right) - \Phi\left(\frac{c_1 - \mu_1}{\sigma_1}\right), \quad TCF_3 = 1 - \Phi\left(\frac{c_2 - \mu_2}{\sigma_2}\right) \tag{3}$$

(TCF_1, TCF_2, TCF_3) define the ROC surface described by equation (1). The corresponding parametric generalized Youden index is

$$J_3(c_1, c_2) = \frac{1}{2}\left\{\Phi\left(\frac{c_1 - \mu_1}{\sigma_1}\right) - \Phi\left(\frac{c_1 - \mu_2}{\sigma_2}\right) + \Phi\left(\frac{c_2 - \mu_2}{\sigma_2}\right) - \Phi\left(\frac{c_2 - \mu_3}{\sigma_3}\right)\right\} \tag{4}$$

The optimal cut-off points maximizing J_3 are defined by[19]

$$\begin{aligned} c_1^* &= \frac{(\mu_2\sigma_1^2 - \mu_1\sigma_2^2) - \sigma_1\sigma_2\sqrt{(\mu_1 - \mu_2)^2 + (\sigma_1^2 - \sigma_2^2)\log\left(\frac{\sigma_1^2}{\sigma_2^2}\right)}}{\sigma_1^2 - \sigma_2^2} \\ c_2^* &= \frac{(\mu_3\sigma_2^2 - \mu_2\sigma_3^2) - \sigma_2\sigma_3\sqrt{(\mu_2 - \mu_3)^2 + (\sigma_2^2 - \sigma_3^2)\log\left(\frac{\sigma_2^2}{\sigma_3^2}\right)}}{\sigma_2^2 - \sigma_3^2} \end{aligned} \tag{5}$$

For $\sigma_1 = \sigma_2 = \sigma_3$, maximizing equation (4) results in $c_1^* = \frac{\mu_1+\mu_2}{2}$ and $c_2^* = \frac{\mu_2+\mu_3}{2}$. As a result, the latter formulas for c_1^*, c_2^* do not follow from equation (5). The corresponding estimates $\hat{c}_1^*, \hat{c}_2^*$ are then obtained by substituting in equations (5) the maximum likelihood estimates of $\mu_1, \mu_2, \mu_3, \sigma_1, \sigma_2, \sigma_3$. The associated estimated optimal triplet of TCF on the ROC surface is defined by $(\widehat{TCF}_1(\hat{c}_1^*), \widehat{TCF}_2(\hat{c}_1^*, \hat{c}_2^*),\ \widehat{TCF}_3(\hat{c}_2^*))$ where $\widehat{TCF}_1(\cdot),\ \widehat{TCF}_2(\cdot),\ \widehat{TCF}_3(\cdot)$ are the maximum likelihood estimates of $TCF_1(\cdot),\ TCF_2(\cdot),\ TCF_3(\cdot)$.

The TCF_i, $i = 1, 2, 3$, are bounded, being probabilities, and thus the triplet $(\widehat{TCF}_1, \widehat{TCF}_2, \widehat{TCF}_3)$ lies in the unit cube. Consequently, use of a normal approximation directly for the construction of confidence regions may be inappropriate, especially for small sample sizes. Accordingly, we use the transformation $\Phi^{-1}(\cdot)$ in order to project the TCF_i, $i = 1, 2, 3$, proportions onto the real line. We define

$$\delta_1 = \Phi^{-1}(TCF_1(c_1^*)), \quad \delta_2 = \Phi^{-1}(TCF_2(c_1^*, c_2^*)), \quad \delta_3 = \Phi^{-1}(TCF_3(c_2^*)) \tag{6}$$

where for the parametric model in equation (3) we obtain

$$\delta_1 = \frac{c_1^* - \mu_1}{\sigma_1}, \quad \delta_2 = \Phi^{-1}\left\{\Phi\left(\frac{c_2^* - \mu_2}{\sigma_2}\right) - \Phi\left(\frac{c_1^* - \mu_1}{\sigma_1}\right)\right\}, \delta_3 = \frac{\mu_2 - c_2^*}{\sigma_2}$$

Denote with $\hat{\delta}_1, \hat{\delta}_2, \hat{\delta}_3$ the corresponding maximum likelihood estimates to which we apply standard normal asymptotic theory. Using the delta method we obtain the variances of the $\hat{\delta}_i$, $i = 1, 2, 3$.[22] Technical details are offered in Section 2.1 of the supplementary material of this article.

To obtain an approximate 95% rectangular parallelepiped region for the optimal TCF we use the conservative Bonferroni adjustment. We thus consider univariate confidence intervals for δ_1, δ_2 and δ_3 of the following form:

$$\hat{\delta}_i \pm 2.3911 \cdot \sqrt{Var(\hat{\delta}_i)}, \quad i = 1, 2, 3 \tag{7}$$

Note that if $(TCF_i^{(l)}), TCF_i^{(u)})$ is a 0.9830% confidence interval for TCF_i then the three-dimensional rectangle $(TCF_1^{(l)}, TCF_1^{(u)}) \times (TCF_2^{(l)}, TCF_2^{(u)}) \times (TCF_3^{(l)}, TCF_3^{(u)})$ is a 95% confidence region for TCF_1, TCF_2, TCF_3. To obtain the desired confidence region in the unit cube, we transform (7) back to the ROC surface space and consider the following form of confidence intervals for the TCFs:

$$\Phi\left(\hat{\delta}_i \pm 2.3911\sqrt{Var(\hat{\delta}_i)}\right), \quad i = 1, 2, 3 \tag{8}$$

Even though such a rectangular confidence region may be easier for practitioners to interpret, ellipsoidal confidence regions are preferable since they can accommodate the correlation between $\hat{\delta}_1, \hat{\delta}_2, \hat{\delta}_3$. The covariance between $\hat{\delta}_i, \hat{\delta}_j$, $(i,j) = 1, 2, 3$, is given in Section 2.1 of the supplementary material of this article. An estimate $\hat{\Sigma}$ of the variance-covariance matrix Σ that corresponds to $(\hat{\delta}_1,\ \hat{\delta}_2,\ \hat{\delta}_3)$ can thus be obtained. The ellipsoid defined by

$$(\mathbf{y} - \mathbf{a})^t \hat{\Sigma}^{-1} (\mathbf{y} - \mathbf{a}) = q_{3;0.95} \tag{9}$$

where $\mathbf{a} = (\hat{\delta}_1, \hat{\delta}_2, \hat{\delta}_3)$ and $q_{3;0.95}$ is the 95th percentile of a χ_3^2 distribution, is an approximate 95% confidence region for the triplet $(\delta_1, \delta_2, \delta_3)$. We transform back to the unit cube using $\Phi(\cdot)$ obtaining an 'egg-shaped' confidence region in the ROC surface space. Simulated examples for both the rectangular and egg-shaped confidence regions are provided in Figure 1 of the supplementary material. This approach will be referred to as 'Delta' in the simulation study of Section 5.

2.1.1 Bootstrap alternative

Since with small sample sizes the delta method approach may not be appropriate for estimating Σ, we may resort to a bootstrap-based approach as follows: we sample with replacement measurements of Y_1, Y_2, and Y_3. We then use the bootstrap samples in order to obtain an estimate of the variance-covariance matrix Σ, namely $\hat{\Sigma}_{(boots)}$. We use the bootstrap-based estimated variances required to obtain the corresponding three-dimensional rectangular confidence region as well as the estimated covariances to obtain the corresponding ellipsoid confidence region from equation (9).

In addition, instead of using χ^2_3 to obtain the confidence region we use the 95th percentile of the bootstrap distribution of

$$\begin{aligned} q^* = (\delta_{\mathbf{1}(boots)} - \hat{\delta}_1, \delta_{\mathbf{2}(boots)} - \hat{\delta}_2, \delta_{\mathbf{3}(boots)} - \hat{\delta}_3)' \hat{\Sigma}^{-1}_{(boots)} (\delta_{\mathbf{1}(boots)} \\ - \hat{\delta}_1, \delta_{\mathbf{2}(boots)} - \hat{\delta}_2, \delta_{\mathbf{3}(boots)} - \hat{\delta}_3) \end{aligned} \tag{10}$$

where $\delta_{i(boots)}$, $i = 1, 2, 3$, is a vector containing the bootstrap values of the corresponding δ_i.[17] We transform back to the unit cube using $\Phi(\cdot)$ in order to obtain an 'egg-shaped' confidence region in the ROC surface space as previously. This approach will be referred to as 'Boots' in the simulation study.

2.2 Box–Cox approach

Assuming that marker measurements are normally distributed is restrictive and can lead to incorrect results when this assumption is violated. Since the ROC curve (and surface) is invariant under monotonic transformations, the Box–Cox transformation to achieve normality is often used in the ROC context.[16,23,24] The Box–Cox transformation is defined by

$$Y_{1(\lambda)} = \begin{cases} \dfrac{Y_1^{\lambda} - 1}{\lambda}, & \lambda \neq 0 \\ \log(Y_1), & \lambda = 0 \end{cases}$$

and similarly for Y_2 and Y_3.

Technical details for the derivation of $\hat{\Sigma}_{(\lambda)}$, the estimate of the variance-covariance matrix of the corresponding $(\hat{\delta}_{1(\lambda)}, \hat{\delta}_{2(\lambda)}, \hat{\delta}_{3(\lambda)})$, are given in Section 2.2 of the supplementary material of this article. The parameter λ affects the information matrix and its variability must be taken into account during the construction of confidence regions when the Box–Cox transformation is employed. This fact has been documented in previous work.[15] Notice that the partial derivatives in the formulas above are analogous to those in the two-class case.[15]

Based on the supplementary material section equations (4) and (5) for the variances and covariances of the $\hat{\delta}_{i(\lambda)}$, $i = 1, 2, 3$, one can straightforwardly construct the desired three-dimensional confidence rectangle and 'egg-shaped' regions from equations (8), (9) after using the Box–Cox transformation. This approach will be referred to as 'Box–Cox' in the simulation study.

2.2.1 Bootstrap alternatives after Box–Cox

The bootstrap-based analogue of the Box–Cox approach involves sampling with replacement from Y_1, Y_2, Y_3 and performing the Box–Cox transformation for each bootstrap sample in order to take into account the variability of the estimate of the parameter λ. In order to construct the ellipsoidal confidence regions one can use χ_3^2 as in equation (9). This approach will be referred to as 'Boots (BC)' in the simulation study. A second option, as follows from equation (10), is to use the 95th percentile of the bootstrap distribution of

$$\begin{aligned} q^{*(\lambda)} = & \left(\delta_{\mathbf{1}^{(\lambda)}(\text{boots})} - \hat{\delta}_{1^{(\lambda)}}, \delta_{\mathbf{2}^{(\lambda)}(\text{boots})} - \hat{\delta}_{2^{(\lambda)}}, \delta_{\mathbf{3}^{(\lambda)}(\text{boots})} - \hat{\delta}_{3^{(\lambda)}}\right)' \\ & \times \hat{\Sigma}_{(\lambda)}^{-1}\left(\delta_{\mathbf{1}^{(\lambda)}(\text{boots})} - \hat{\delta}_{1^{(\lambda)}}, \delta_{\mathbf{2}^{(\lambda)}(\text{boots})} - \hat{\delta}_{2^{(\lambda)}}, \delta_{\mathbf{3}^{(\lambda)}(\text{boots})} - \hat{\delta}_{3^{(\lambda)}}\right) \end{aligned} \tag{11}$$

where $\delta_{\mathbf{i}^{(\lambda)}(\text{boots})}$, $i = 1, 2, 3$, is a vector of the bootstrap values of the $\delta_{i^{(\lambda)}}$ and $\hat{\Sigma}_{(\lambda)}$ is the estimated $\Sigma_{(\lambda)}$ matrix based on the supplementary material section equations (4) and (5). This approach will be referred to as 'Box–Cox-q' in the simulation study.

Alternatively, as in Section 2.1.1, one can replace $\hat{\Sigma}_{(\lambda)}$ in equation (11) with $\hat{\Sigma}_{(\lambda)(\text{boots})}$, the bootstrap-based estimate of $\Sigma_{(\lambda)}$. The latter approach is not assessed in the simulation study presented in Section 5, since in general our simulations did not indicate any improvement when using the bootstrap estimated covariance matrix.

3 Non-parametric approaches

In Section 2, we assumed that the marker scores for all three groups are either normally distributed or can be transformed to normality using a Box–Cox transformation. This transformation to normality may not always be adequate. In such situations non-parametric approaches may be preferable. A simple approach would be to estimate the ROC surface empirically using the standard empirical distribution functions to obtain non-parametric estimates of the TCF_i and then implement a bootstrap procedure in order to obtain confidence regions. Such a procedure would imply sampling m times with replacement from Y_1, Y_2, Y_3 and obtaining the empirical ROC surface, the corresponding estimates of the optimal cut-off points, and the empirical estimates of δ_i, $i = 1, 2, 3$, from equation (6), for each bootstrap sample. Using the m bootstrap triplets of these estimates of δ_i, $i = 1, 2, 3$, one can in turn obtain an estimate of the corresponding variance-covariance matrix and proceed as in the previous section.

However, as pointed out in previous work,[15] for small to moderate sample sizes, a smooth estimate of the ROC surface may be preferable in order to derive valid estimates for Σ, the variance-covariance matrix of the estimated triplet $(\delta_1, \delta_2, \delta_3)$. Here, we investigate both a kernel- and a spline-based approach.

3.1 Kernel-based approach

One may construct a kernel-based estimate of the ROC surface by plugging in the kernel estimate of the underlying distribution of each group and then obtain confidence regions via bootstrapping. However, we found that commonly used fixed bandwidth kernel approaches are not efficient enough to deal with the scenarios presented in our simulations. Here, we explore the use of normal kernels in combination with smooth bootstrapping.[17] Specifically, we consider the normal kernel of the following form for estimating the underlying distribution of the scores of each group:

$$\hat{F}_i^{(k)}(y) = \frac{1}{n_i}\sum_{j=1}^{n_i} \Phi\left(\frac{y - y_{ij}}{h_i}\right) \tag{12}$$

We employ bandwidths equal to $h_i = 0.9\min(sd(y_i), iqr(y_i)/1.34)n_i^{0.2}$ where sd and iqr refer to the standard deviation and interquartile range respectively.[17,25,26] Under this setting, the Youden-index-based cut-off points are obtained in a straightforward manner by plugging in the kernel-based estimates of the three underlying distributions. The variance-covariance matrix of $\hat{\delta}_i$, $i=1$, 2, 3, is obtained by using a smooth bootstrap procedure.[17] Specifically we propose the following algorithm:

- Step 0: Calculate h_1, h_2, h_3 based on Y_1, Y_2, Y_3 respectively. Derive the corresponding delta estimates, $\widehat{\delta_1}$, $\widehat{\delta_2}$ and $\widehat{\delta_3}$, and the normal kernel estimates of F_1, F_2, F_3.
- Step 1: Sample with replacement Y_1, Y_2, and Y_3.
- Step 2: Set $Y_i^{(s)} = Y_i + e_i$, where $e_i \sim N(0, h_i^2)$.
- Step 3: Based on $Y_i^{(s)}$ construct their kernel distribution estimates $(\widehat{F_1}^{(k)}, \widehat{F_2}^{(k)}, \widehat{F_3}^{(k)})$ using h_i of Step 0 and obtain the estimates $\hat{c}_1^{*(k)}, \hat{c}_2^{*(k)}$ of c_1^* and c_2^*.
- Step 4: Obtain the δ_i estimates using the logit transformation of the estimated TCF_i, namely
  - $\hat{\delta}_1^{(b)} = \log it(\widehat{TCF}_1) = \log it\left(\hat{F}_1^{(k)}(\hat{c}_1^{*(k)})\right)$,
  - $\hat{\delta}_2^{(b)} = \log it(\widehat{TCF}_2) = \log it\left(\hat{F}_2^{(k)}(\hat{c}_2^{*(k)}) - \hat{F}_2^{(k)}(\hat{c}_1^{*(k)})\right)$, and
  - $\hat{\delta}_3^{(b)} = \log it(\widehat{TCF}_3) = \log it\left(1 - \hat{F}_3^{(k)}(\hat{c}_2^{*(k)})\right)\cdot$
- Step 5: Repeat steps 1 to 4 m times to obtain m bootstrapped estimates of each delta.
- Step 6: Based on the previous step derive an estimate of Σ (which we denote by $\hat{\Sigma}_{(b)}$), and proceed to the construction of the ellipsoid: $(\mathbf{y}-\mathbf{a})^t\hat{\Sigma}_{(b)}^{-1}(\mathbf{y}-\mathbf{a}) = q_{3;0.95}$ where $\mathbf{a} = (\hat{\delta}_{1(b)}, \hat{\delta}_{2(b)}, \hat{\delta}_{3(b)})$ and $q_{3;0.95}$ is the 95th percentile of the χ_3^2 distribution. Alternatively, one may use the 95th percentile of q^* as given by equation (10).
- Step 7: Transform back to the ROC space by using the inverse logit function and obtain a three-dimensional egg-shaped 95% confidence region for the triplet (TCF_1, TCF_2, TCF_3).

We assess the kernel-based method that involves the smooth bootstrap and is based on q^* for the construction of the confidence region in the simulation study (Section 5) and denote this approach by 'Kernels(SB)-q'.

3.2 Logspline approach

Another convenient choice is the logspline approach.[15,21] We estimate F_1, F_2 and F_3 using the logspline approach as outlined in Section 2.3 of the supplementary material and then, based on these logspline estimates, obtain the associated cut-off points and corresponding TCFs. This approach has been shown to perform satisfactorily in the two-class case.[15] After using the probit (or logit) transformation to project the TCFs onto the real line, we obtain the corresponding $\hat{\delta}_i$, $i=1$, 2, 3. Using the bootstrap, where for each bootstrap sample we repeat the estimation with the logspline approach, we obtain the estimated covariance matrix for $(\hat{\delta}_1, \hat{\delta}_2, \hat{\delta}_3)$. Finally, we proceed as in the previous section in order to construct the corresponding confidence regions for (TCF_1, TCF_2, TCF_3). We note that the knot selection is an automated procedure in the logspline approach and involves using stepwise addition or deletion. The package polspline can be used for logspline estimation in R. The logspline approach is assessed in conjunction with equation (9) for the construction of confidence regions for the triplet (TCF_1, TCF_2, TCF_3) in the simulation study we present in Section 5. This approach will be referred to as 'Logspline' in the simulation study. We also initially examined an alternative version using a bootstrap-based percentile point instead of the 95th percentile of the χ_3^2 distribution. This performed poorly and therefore was excluded from the simulation study.

4 The *k*-class case

There are situations where classification to more than three classes is of interest.[1] Our approaches can be straightforwardly extended to k-class classification problems. The corresponding marker measurements are denoted by $Y_1, Y_2, \ldots, Y_k$ and the underlying distributions by $F_1, F_2, \ldots, F_k$. The ordering of interest is $Y_1 < Y_2 < \cdots < Y_k$. In this case we have $(k-1)$ cut-offs, $c_1 < c_2 < \cdots < c_{k-1}$. By defining a fine grid in the support of the possible cut-offs we obtain, but cannot visualize, the corresponding ROC manifold based on $TCF_1(c_1, \ldots, c_{k-1}), \ldots, TCF_k(c_1, \ldots, c_{k-1})$. The hypervolume of such an ROC manifold equals $1/k!$ for an uninformative marker. Subtleties for the construction of the ROC manifold are given in the literature.[11] The generalized Youden index for the k-class case is defined as follows:[1]

$$\begin{aligned} J_k &= \max_{c_1,\ldots,c_{k-1};c_1 < c_2 < \cdots < c_{k-1}} \{TCF_1 + \cdots + TCF_k - 1\} \\ &= \max_{c_1,\ldots,c_{k-1};c_1 < c_2 < \cdots < c_{k-1}} \big\{F_1(c_1) - F_2(c_1) + \cdots + F_{k-1}(c_{k-1}) - F_k(c_{k-1})\big\} \end{aligned} \tag{13}$$

If the normality assumption is justified for each one of the k groups then as in the three-class case we can obtain

$$c_j^* = \frac{\left(\mu_{j+1}\sigma_j^2 - \mu_j\sigma_{j+1}^2\right) - \sigma_j\sigma_{j+1}\sqrt{(\mu_j - \mu_{j+1})^2 + (\sigma_j^2 - \sigma_{j+1}^2)\log\left(\frac{\sigma_j^2}{\sigma_{j+1}^2}\right)}}{\sigma_j^2 - \sigma_{j+1}^2} \tag{14}$$

and for the case of equal variances $c_j^* = \frac{\mu_j + \mu_{j+1}}{2}, j = 1, \ldots, k-1$. In order to construct k-dimensional confidence regions for the associated optimal $(TCF_1, \ldots, TCF_k)$ point we define $\delta_j = \Phi^{-1}(TCF_j)$ and its corresponding maximum likelihood estimate is denoted by $\hat{\delta}_j$. We can then apply all proposed methods as previously discussed. Both a hyper-ellipsoid and a hyper-rectangular confidence region can be obtained. The hyper-ellipsoid is of the form $(\mathbf{x} - \mathbf{a})^t \hat{\Sigma}^{-1}(\mathbf{x} - \mathbf{a}) = q_{k;0.95}$, where $\mathbf{a} = (\hat{\delta}_1, \ldots, \hat{\delta}_k)$ and $q_{k;0.95}$ is the 95th percentile of a χ_k^2 or can be obtained via bootstrapping similarly to the three-class case. Here, the Bonferroni adjustment implies that the corresponding univariate intervals must have a $0.95^{\frac{1}{k}}$ theoretical coverage in order to obtain an approximate 95% confidence region in the k-dimensional space.

5 Simulation study

We conducted a large simulation study to evaluate our methods in terms of coverage and mean confidence region volumes in the three-class case. We considered different scenarios (detailed in Table 1 of the supplementary material) based on the normal, log-normal, and gamma distributions. The gamma distribution scenarios were included although outside the Box–Cox family of transformations in order to check the robustness of the Box–Cox-based methods. The parameters of these distributions were set in order to achieve J_3 theoretical values of 0.4, 0.6, or 0.8. The sample size scenarios we explore are (50, 50, 50), (100, 100, 100), (200, 200, 200), and (50, 100, 200) for (n_1, n_2, n_3). The nominal coverage was 0.95.

Table 1. Trails A assessment and optimal cut-off points. Cut-off points are transformed back to the original scale in the Box–Cox case for convenience. Also, 95% bootstrap confidence intervals based on estimated percentiles of bootstrap distributions (1000 replications used) are given.

Methods	VUS	J_3	c_1	c_2
'Box–Cox'	0.745 (0.669, 0.821)	0.588 (0.507, 0.670)	46.12 (43.28, 48.96)	81.92 (73.56, 90.29)
'Box–Cox-q'	0.745 (0.669, 0.821)	0.588 (0.507, 0.670)	46.12 (43.28, 48.96)	81.92 (73.56, 90.29)
'Kernels(SB)-q'	0.689 (0.650, 0.729)	0.557 (0.463, 0.650)	48.06 (43.26, 52.86)	81.38 (70.41, 92.35)
'Logspline'	0.752 (0.658, 0.846)	0.600 (0.501, 0.727)	44.91 (39.21, 50.62)	71.43 (57.29, 85.57)

Table 2. TCFs for the three disease states corresponding to Trails A followed by 95% marginal confidence intervals. The class order is $U < MCI < D$.

Methods	TCF_U	TCF_{MCI}	TCF_D
'Box–Cox'	0.789 (0.716, 0.849)	0.624 (0.505, 0.732)	0.764 (0.622, 0.870)
'Box–Cox-q'	0.789 (0.722, 0.845)	0.624 (0.502, 0.734)	0.764 (0.655, 0.851)
'Kernels(SB)-q'	0.813 (0.703, 0.889)	0.574 (0.422, 0.714)	0.726 (0.576, 0.837)
'Logspline'	0.784 (0.642, 0.881)	0.581 (0.403, 0.741)	0.862 (0.716, 0.939)

For the scenarios involving three normal distributions we considered the seven methods described in Sections 2 and 3, namely: 'Delta', 'Boots', 'Box–Cox', 'Boots (BC)', 'Box–Cox-q', 'Kernels(SB)-q', and 'Logspline'. Results are shown in Table 2 of the supplementary material. We only present results for the 'egg-shaped' regions since the rectangular regions result in consistently larger volumes and poorer coverage given that they do not take into account the correlations of the TCFs.

For the normal case, results are satisfactory for almost all J_3 values and sample sizes. For $J_3 = 0.4$ and $(n_1, n_2, n_3) = (50,\ 50,\ 50)$ the coverage is somewhat lower than the nominal one except for the 'Kernels(SB)-q' method. The 'Boots' method provides somewhat better results than the 'Delta' method in terms of both coverage and confidence region volumes for the case $J_3 = 0.4$ and $(n_1, n_2, n_3) = (50,\ 50,\ 50)$. As the sample size increases the methods yield almost identical results, as expected.

The 'Box–Cox' approach provides approximately the same coverage as 'Boots' and 'Delta', with the cost of a larger volume. This is expected since the variability of the extra parameter λ is taken into account. This is the price to pay for being more robust and not assuming that the data are intrinsically normally distributed.

The 'Boots (BC)' method yields the least satisfactory results as compared to the possible parametric alternatives. The non-parametric approaches result in nice coverage in all cases but with substantially larger volumes, as expected. This is the price to pay for not making any parametric assumptions about the underlying data.

The non-normal case scenarios are presented in Table 3 of the supplementary material of the article. We considered 'Box–Cox', 'Boots (BC)', 'Box–Cox-q', 'Kernels(SB)-q', and 'Logspline' given that the use of 'Delta' and 'Boots' is not recommended a priori in these cases. Note that even though the gamma distribution is not in the Box–Cox transformation family the 'Box–Cox' and 'Box–Cox-q' procedures still provide reasonable results. The usefulness of the Box–Cox transformation approach in the ROC context even for distributions which are not strictly in this family has been pointed out previously.[14,15,24]

We observe that most methods provide satisfactory coverage results in almost all cases. Specifically, the 'Box–Cox' approach performs nicely although for small sample sizes and

$J_3 = 0.4$ its coverage is somewhat reduced. This is corrected with the method 'Box–Cox-q' which provides better coverage for small sample sizes with a cost of a somewhat larger volume. The 'Box–Cox-q' approach seems to improve 'Box–Cox' in terms of coverage in almost all cases both in Tables 2 and 3 of the supplementary material of the article. As sample sizes increase these two methods yield similar results as expected. The method 'Boots (BC)' performs poorly compared to its competitors. The non-parametric approaches, that is, 'Logspline' and 'Kernels(SB)-q', yield coverage close to the nominal level in most cases, but, as in the normal case, with substantially larger volumes. The 'Kernels(SB)-q' approach results in smaller volumes than 'Logspline', a preferable property that will result in tighter marginal confidence intervals for the TCFs. Comparing 'Logspline' and 'Kernels(SB)-q' no clear winner in terms of coverage exists. Regarding the parametric approaches, the 'Box–Cox-q' method provides the most satisfactory coverage although with somewhat larger volumes than 'Box–Cox'.

6 Application

A total of 245 patients with Parkinson's disease underwent the TMT Part A (also referred to as 'Trails A' in the sequel) as a routine examination. Based on a battery of cognitive tests used for the characterization of cognitive impairment, 170 patients were classified as unimpaired (U), scoring within the normal range for the test battery. Fifty-two patients were classified as having MCI, while 23 patients were classified as having dementia (D). In terms of the TMT Part A, the latter group is expected to have a slow mean completion time, while patients with MCI are expected to have completion times that are intermediate relative to Parkinson's disease patients who show normal cognition (U). This dataset is available in the online supplementary material of this article.

Unimpaired patients (U) had a mean completion time of 37.71 (±11.39) seconds, similar to that expected of an older population of otherwise healthy controls. Patients with MCI had an average of 59.73 (±19.57) seconds, while patients with dementia (D) had an average of 122.83 (±57.09) seconds. The Anderson–Darling test for normality was used. Data did not support the normality assumption in general ($A = 1.43$, $p = 0.001$ for U, $A = 0.69$, $p = 0.069$ for MCI, $A = 0.87$, $p = 0.022$ for D).

The Box–Cox transformation was used, which resulted in $p > 0.1$ for the Anderson–Darling test for all three groups. As a result, the delta method approach and a bootstrap alternative were employed for the transformed measurements ('Box–Cox' and 'Box–Cox-q'), while the kernel and logspline methods were employed for the untransformed measurements ('Kernels(SB)-q', 'Logspline').

VUS, J_3 and the corresponding optimal cut-off points along with 95% confidence intervals are shown in Table 1. The 'Kernels(SB)-q' method seems to underestimate the VUS and J_3, however, this does not affect the optimal cut-off points and TCFs. We mention that the empirical VUS is 0.754 (0.674, 0.834).

Resulting TCFs are given in Table 2, while the corresponding ROC surfaces along with their 'egg-shaped' confidence regions for the TCFs are illustrated in Figure 1. For this particular example 'Boots (BC)' (not shown) and 'Box–Cox-q' provide virtually identical results, which was shown not to hold in general. As expected from the simulation study, the non-parametric confidence regions are much larger than those based on the Box–Cox transformation. All four methods result in cut-off points in the vicinity of 45 for c_1 and in the vicinity of 80 for c_2. These would result in TCF_U around 80%, TCF_{MCI} around 60% and TCF_D around 75%. We conclude that a quick screen based on Trails A would classify patients with a completion time of less than 45 s as normal, patients in the range 45–80 s as MCI, and patients above 80 s as D. The Trails A test is a quickly administered diagnostic marker for cognitive impairment in Parkinson's disease and the estimated cut-off points could be taken into consideration in clinical practice.

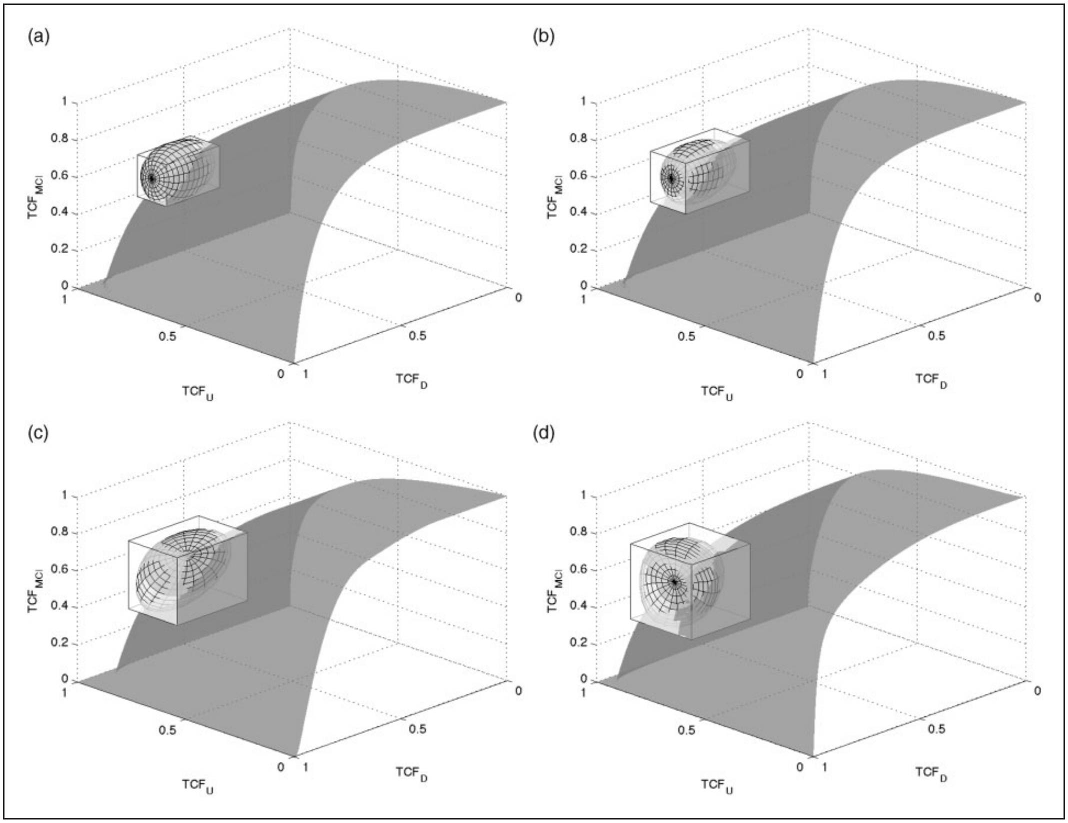

Figure 1. ROC surfaces and corresponding proposed 95% confidence regions for the Trails A test. (a) 'Box–Cox' approach. (b) 'Box–Cox-q' approach. (c) 'Kernels(SB)-q' approach. (d) 'Logspline' approach.

7 Discussion

We generalized and expanded on our two-class methods for the construction of confidence regions for the optimal TCF in the ROC context to the three- and general k-class cases. These methods can be useful in practice when interpreting the results of the assessment of a diagnostic marker that results in three or more classes. Several parametric and non-parametric approaches have been proposed and discussed generalizing previous results.[15] Also, a kernel-based approach has been introduced here. The simulation study of Section 5 has highlighted strengths and weaknesses of the proposed approaches according to the distributional properties underlying the markers measurements.

Parametric approaches may not work as expected in practice when extreme departures from normality are present and the Box–Cox transformation does not adequately address the problem. The proposed non-parametric approaches, logspline and kernel-based, provide suitable alternatives which cover a very wide range of applications in practice.

In contrast to other authors who used the Box–Cox transformation approach for ROC curve analyses[23] we found it necessary to take into account the variability due to estimating the transformation parameter lambda. This may be due to the fact that previous authors considered one-dimensional problems such as confidence intervals for the Youden index while our construction of confidence regions in the ROC space is a multi-dimensional problem.

Our work revolves around the use of the generalized Youden index for cut-off point selection in the three- and k-class classification problems. The generalized Youden index apart from its simplicity has a useful clinical interpretation as the accuracy of the diagnostic marker under consideration. Future research may concern the use of our approaches when other indices are preferred for cut-off point selection in the three-class case. Such indices have recently been proposed.[27]

We have illustrated our approach from a dataset of Parkinson's disease patients based on the TMT Part A as a screen for cognitive impairment. We have generated cut-off values with suitable confidence intervals using this simple test for three cognitive states associated with this neurodegenerative condition. The three-class approach has been shown to be convenient in clinical practice in applications that cannot be accommodated by the two-class case.[1,2,9,11,19,27] As a result our methods may find a wide range of applicability in such situations, given that there are no alternative approaches that will result in correct coverage when the optimal cut-off points are estimated from the data.

A `Matlab` package (named `egg3d`) for the implementation of 'Delta', 'Boots', 'Box–Cox', 'Boots (BC)', 'Box–Cox-q', and 'Kernels(SB)-q' is offered with the supplementary material of the article, along with the dataset used for the application in Section 6. Information on the use of the package is provided in Section 4 of the supplementary material. R and Matlab code for the implementation of the 'Logspline' method is available from the authors upon request.

Supplementary material

This paper is accompanied by supplementary material which includes technical details for the delta method approach, the Box–Cox approach, and the logspline approach for the construction of confidence regions for TCF triplets, as well as the simulation scenarios and results (Tables 1 to 3 therein), and information on the use of the `Matlab` package `egg3d` which was produced by the authors and can be used for the implementation of the proposed methodologies.

Acknowledgment

The authors gratefully acknowledge the helpful comments from the editor and referees, which have improved the content and presentation of this paper.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

1. Nakas CT, Dalrymple-Alford JC, Anderson TJ, et al. Generalization of Youden index for multiple-class classification problems applied to the assessment of externally validated cognition in Parkinson disease screening. *Stat Med* 2013; **32**: 995–1003.
2. Dalrymple-Alford JC, MacAskill MR, Nakas CT, et al. The MoCA: Well-suited screen for cognitive impairment in Parkinson disease. *Neurol* 2010; **75**: 1717–1725.
3. Yarnall AJ, Rochester L and Burn DJ. Mild cognitive impairment in Parkinson's disease. *Age Ageing* 2013; **42**: 567–576.

4. Gerstenecker A and Mast B. Mild cognitive impairment: A history and the state of current diagnostic criteria. *Int Psychoger* 2014; **27**: 1–13. DOI: 10.1017/S1041610214002270.
5. Stephan BCM, Minett T, Pagett E, et al. Diagnosing Mild Cognitive Impairment (MCI) in clinical trials: A systematic review. *BMJ Open* 2013; **3**: e001909.
6. Barker RA and Williams-Gray CH. Mild cognitive impairment and Parkinson's disease –Something to remember. *J Parkin Dis* 2014; **4**: 651–656.
7. Lezak MD, Howieson DB, Bigler ED, et al. *Neuropsychological assessment*. 5th edn. Oxford: Oxford University Press, 2012, p.1200.
8. Galvin JE, Powlishta KK, Wilkins K, et al. Predictors of preclinical Alzheimer disease and dementia: A clinicopathologic study. *Arch Neurol* 2005; **62**: 758–765.
9. Dalrymple-Alford JC, Livingston L, Macaskill MR, et al. Characterizing mild cognitive impairment in Parkinson's disease. *Mov Disord* 2011; **26**: 629–636.
10. Krzanowski WJ and Hand DJ. *ROC curves for continuous data*. Boca Raton, FL: Chapman and Hall/CRC, 2009, p.232.
11. Nakas CT and Yiannoutsos CT. Ordered multiple-class ROC analysis with continuous measurements. *Stat Med* 2004; **23**: 3437–3449.
12. Pepe MS. *The statistical evaluation of medical tests for classification and prediction*. Oxford: Oxford University Press, 2004, p.318.
13. Scurfield BK. Multiple-event forced-choice tasks in the theory of signal detectability. *J Math Psychol* 1996; **40**: 253–269.
14. Fluss R, Faraggi D and Reiser B. Estimation of the Youden Index and its associated cutoff point. *Biomet J* 2005; **47**: 458–472.
15. Bantis LE, Nakas CT and Reiser B. Construction of confidence regions in the ROC space after the estimation of the optimal Youden index-based cut-off point. *Biomet* 2014; **70**: 212–223.
16. Zou KH, Liu A, Bandos AI, et al. *Statistical evaluation of diagnostic performance: Topics in ROC analysis*. Boca Raton: Chapman and Hall/CRC, 2011, p.245.
17. Yin J and Tian L. Joint inference about sensitivity and specificity at the optimal cut-off point associated with Youden index. *Comput Stat Dat Anal* 2014; **77**: 1–13.
18. Nakas CT, Alonzo TA and Yiannoutsos CT. Accuracy and cut-off point selection in three-class classification problems using a generalization of the Youden index. *Stat Med* 2010; **29**: 2946–2955.
19. Luo J and Xiong C. Youden index and associated cut-points for three ordinal diagnostic groups. *Commun Stat Simul Comput* 2013; **42**: 1213–1234.
20. Kosinski AS, Chen Y and Lyles RH. Sample size calculations for evaluating a diagnostic test when the gold standard is missing at random. *Stat Med* 2010; **29**: 1572–1579.
21. Kooperberg C and Stone CJ. A study of logspline density estimation. *Comput Stat Dat Anal* 1992; **12**: 327–348.
22. Schisterman EF and Perkins N. Confidence intervals for the Youden index and corresponding optimal cut-point. *Commun Stat Simul Comput* 2007; **36**: 549–563.
23. Molanes-López EM and Letón E. Inference of the Youden index and associated threshold using empirical likelihood for quantiles. *Stat Med* 2011; **30**: 2467–2480.
24. Schisterman EF, Reiser B and Faraggi D. ROC analysis for markers with mass at zero. *Stat Med* 2006; **25**: 623–638.
25. Silverman BW. *Density estimation for statistics and data analysis*. Boca Raton, FL: Chapman and Hall, 1986, p.176.
26. Faraggi D and Reiser B. Estimation of the area under the ROC curve. *Stat Med* 2002; **21**: 3093–3106.
27. Attwood K, Tian L and Xiong C. Diagnostic thresholds with three ordinal groups. *J Biopharm Stat* 2014; **24**: 608–633.