Likelihood-type confidence regions for optimal sensitivity and specificity of a diagnostic test

Likelihood-type confidence regions for optimal sensitivity and specificity of a diagnostic test

Gianfranco Adimari[a], Duc Khanh To[a,b,c,*], Monica Chiogna[d], Francesca Scatozza[e] and Antonio Facchiano[e]

[a]Department of Statistical Sciences, University of Padova, Via C. Battisti, 241, Padova, 35121, Italy
[b]Department of Information and Engineering, University of Padova, Via Gradenigo, 6/b, Padova, 35131, Italy
[c]Faculty of Mathematics and Computer Science, University of Science - Vietnam National University, Nguyen Van Cu street, 227, Ho Chi Minh city, 70000, Vietnam
[d]Department of Statistical Sciences "Paolo Fortunati", University of Bologna, Via Belle Arti, 41, Bologna, 40126, Italy
[e]Laboratorio di Oncologia Molecolare, Istituto Dermopatico dell'Immacolata, IDI-IRCCS, Via Monti di Creta, 104, Rome, 00167, Italy

ARTICLE INFO

Keywords:
Empirical likelihood
Box-Cox transformation
Youden index
ROC analysis

ABSTRACT

New methods are proposed that provide approximate joint confidence regions for the optimal sensitivity and specificity of a diagnostic test, i.e., sensitivity and specificity corresponding to the optimal cutpoint as defined by the Youden index criterion. Such methods are semi-parametric or non-parametric and attempt to overcome the limitations of alternative approaches. The proposed methods are based on empirical likelihood pivots, giving rise to likelihood-type regions with no predetermined constraints on the shape and automatically range-respecting. The proposal covers three situations: the binormal model, the binormal model after the use of Box-Cox transformations and the fully non-parametric model. In the second case, it is also shown how to use two different transformations, for the healthy and the diseased subjects. The finite sample behaviour of our methods is investigated using simulation experiments. The simulation results also show the advantages offered by our methods when compared with existing competitors. Illustrative examples, involving three real datasets, are also provided.

1. Introduction

In medical research, diagnostic tests (or biomarkers) with continuous values are widely employed to attempt to distinguish between diseased and non-diseased subjects. The diagnostic accuracy of a test can be assessed by its receiver operating characteristic (ROC) curve.

The test result can be dichotomized at a specified cutpoint. Given the cutpoint, the sensitivity is the probability of a true positive, i.e., the probability that the test correctly identifies a diseased subject. The specificity is the probability of a true negative, i.e., the probability that the test correctly identifies a non-diseased subject. When one varies the cutpoint throughout the entire real line, the resulting pairs (1 − specificity, sensitivity) form the ROC curve (see Pepe, 2003, as general reference).

Let X denote the result of a continuous diagnostic test for a non-diseased subject and Y the result of the test for a diseased subject. We assume, without loss of generality, that high test values indicate a high likelihood of disease (if not, we could just change the signs of X and Y). Then, for a given cutpoint τ, the sensitivity and the specificity of the test are

$$\begin{aligned} \theta(\tau) &= \Pr(Y > \tau) = 1 - F_Y(\tau), \\ \eta(\tau) &= \Pr(X \leq \tau) = F_X(\tau), \end{aligned}$$

respectively, where $F_X(\cdot)$ denotes the cumulative distribution function of X and $F_Y(\cdot)$ the cumulative distribution function of Y. The sensitivity depends only on the diseased population; the specificity depends only on the non-diseased population.

*Corresponding author
gianfranco.adimari@unipd.it (G. Adimari); duckhanh.to@unipd.it (D.K. To); monica.chiogna2@unibo.it (M. Chiogna); f.scatozza@idi.it (F. Scatozza); a.facchiano@idi.it (A. Facchiano)
ORCID(s): 0000-0002-7811-912X (G. Adimari); 0000-0002-4641-0764 (D.K. To); 0000-0002-7238-3739 (M. Chiogna); 0000-0002-4243-2392 (A. Facchiano)

In practice, to make a diagnosis, that is, classifying a subject as either diseased or healthy, a diagnostic cutpoint is required. Choosing a high cutpoint level produces a high specificity (i.e. low likelihood of a false positive result), and a low sensitivity (i.e. high likelihood of a false negative result); choosing a low cutpoint level gives opposite results.

To select an "optimal" diagnostic cutpoint, there exists a variety of approaches. Among them, the one based on the Youden index (Youden, 1950), J, is certainly the most popular. The Youden index is the maximum of the sum of sensitivity and specificity minus one, i.e., $J = \max_\tau J(\tau)$, with

$$J(\tau) = \theta(\tau) + \eta(\tau) - 1.$$

The corresponding optimal cutpoint, τ^+ say, is the value that maximizes $J(\tau)$, i.e., the cutpoint that has the largest value in the associated sum of sensitivity and specificity. This approach, hence, maximizes the sum of the two correct classification probabilities. From a geometric perspective, the point on the ROC curve corresponding to this criterion maximizes the vertical distance from the bisector line (the no-discrimination line) in the unit square. Sensitivity and specificity at the optimal cutpoint τ^+ are relevant measures for the diagnostic ability of the test.

In practice, the ROC curve of a new diagnostic test or biomarker is unknown, being unknown (at least partially) the distributions of the test results, for both diseased and non-diseased populations. Hence, the statistical evaluation of the discriminatory ability of the test is obtained by making inferences about its ROC curve and other quantities of interest, such as optimal cutpoints and associated sensitivities and specificities. Generally, the inference is based on data from some suitable sample of patients for which the disease status can be exactly assessed through a so-called gold standard test (GS).

When an optimal cutpoint is estimated from data, from both diseased and healthy samples, the corresponding estimated sensitivity and specificity are correlated, and joint inference is necessary to take into account such a correlation. Methods to build joint confidence regions for sensitivity and specificity at the optimal cutpoint based on the Youden index are proposed by Bantis et al. (2014) and Yin and Tian (2014). Both works deal with parametric (or semi-parametric) and non-parametric methods. More precisely, Bantis et al. (2014) first consider a parametric approach for the classical binormal model, and use asymptotic normality of the maximum likelihood estimator (MLE) $\hat{\Delta} = \left((\hat{\mu}_y - \hat{\tau}^+)/\hat{\sigma}_y, (\hat{\mu}_x - \hat{\tau}^+)/\hat{\sigma}_x)\right)^T$, where $\hat{\mu}_x$, $\hat{\mu}_y$, $\hat{\sigma}_x^2$ and $\hat{\sigma}_y^2$ are MLEs for population's means and variances, of which $\hat{\tau}^+$, the estimator for τ, is a (smooth) function. Once the variance-covariance matrix of $\hat{\Delta}$ has been estimated, confidence regions for the optimal sensitivity and specificity are obtained by mapping, in the ROC space, elliptical regions for the parameter, say Δ, estimated through $\hat{\Delta}$. Estimation of the variance-covariance matrix can be obtained in a (cumbersome) closed form or by using a standard non-parametric bootstrap approach. In this last case, the proposed method is somewhat definable as semi-parametric. When the hypothesis of normality is not adequate, the authors suggest resorting, when possible, to a single Box-Cox transformation or using a fully non-parametric approach. The proposed non-parametric method is based on log spline models to get smooth ROC estimation, the associated optimal cutpoint, as well as the corresponding pair of estimated optimal sensitivity and specificity. The use of the asymptotic normality of the estimators (after a logit transformation) and a bootstrap estimate for the variance-covariance matrix complete the proposal. Also, Yin and Tian (2014) first consider the binormal model and propose a semi-parametric method, based on a suitable generalized pivot for the pair of optimal sensitivity and specificity together with bootstrap calibration. When normality is not satisfied, again the use of a single Box-Cox transformation, if effective, is suggested. Alternatively, a fully non-parametric approach is proposed, based on kernel estimators of F_X and F_Y, associated estimators of the optimal cutpoint and optimal sensitivity and specificity, and bootstrap calibration. To improve the coverage probability of the proposed non-parametric (elliptical) regions, the use of suitable transformations (such as the logit transformation) is also considered. Recently, Schaible and Yin (2021) propose the same schemes used in Yin and Tian (2014) to construct confidence regions for optimal predictive values when the disease prevalence is known.

The main limits of the methodologies proposed in the literature and indicated above are: (i) the related confidence regions have (or derive from regions that have) elliptical shape because they are based on the asymptotic normality of appropriate estimators or pivots; (ii) when the normality of the biomarker does not meet, application of Box-Cox transformations is considered, but limited to a single transformation for both populations. In particular, the elliptical shape is ultimately a predetermined constraint, often not justified in the light of the observed data: the resulting regions might indeed include parameter values less plausible than some of the values left outside.

In this paper, we propose new methods that provide approximate joint confidence regions for the sensitivity and specificity of a biomarker, corresponding to the optimal cutpoint as defined by the Youden index criterion. Such methods are semi-parametric or non-parametric, and attempt to overcome the above-mentioned limitations.

Our proposal is based on empirical likelihood pivots, giving rise to likelihood-type regions with no predetermined constraints on the shape and automatically range-respecting. The proposal covers three situations: the binormal model, the binormal model after the use of Box-Cox transformations and the fully non-parametric model. Importantly, in the second case, we show how to use two different transformations, for the healthy and the diseased subjects.

The paper is organized as follows. In Section 2, we briefly describe an idea in Adimari and Chiogna (2010), which represents our starting point here. In Section 3, we present the proposed methodology, whose accuracy has been assessed by an extensive simulation study described in Section 4. In the study we also compare, in terms of coverage levels, our methods with those present in the literature, mentioned above. In Section 5 we illustrate the use of the techniques presented through the application to three sets of real data. Finally, we close with some remarks left in Section 6.

2. Background

Let $x_1, \dots, x_m$ be a random sample from X, i.e., the test results from m non-diseased patients, and $y_1, \dots, y_n$ a random sample from Y, i.e., the test results from n diseased patients. The true disease status of each patient is assumed to be ascertained by a GS test. Moreover, let $\hat{F}_X$ denote the empirical distribution function based on $x_1, \dots, x_m$ and $\hat{F}_Y$ the empirical distribution function based on $y_1, \dots, y_n$. In particular, hence, for a fixed value t,

$$\hat{F}_X(t) = \frac{1}{m} \sum_{i=1}^{m} I(x_i \le t),$$

where $I(\cdot)$ denotes the indicator function. Recall that we consider a continuous diagnostic test and that, for a fixed cutpoint τ, $\theta = \theta(\tau) = 1 - F_Y(\tau)$ and $\eta = \eta(\tau) = F_X(\tau)$. Let F_X^{-1} denote the inverse function of F_X. Starting with the relation $1 - \theta = F_Y(F_X^{-1}(\eta))$ (i.e., the η-quantile τ of X equals the $(1-\theta)$-quantile of Y), Adimari and Chiogna (2010) define the empirical likelihood (Owen, 2001) quantity

$$\begin{aligned} \ell(\theta, \eta, \tau) \quad = \quad & 2m \left\{ \hat{F}_X(\tau) \log \frac{\hat{F}_X(\tau)}{\eta} + \left[1 - \hat{F}_X(\tau)\right] \log \frac{1 - \hat{F}_X(\tau)}{1 - \eta} \right\} \\ & + 2n \left\{ \hat{F}_Y(\tau) \log \frac{\hat{F}_Y(\tau)}{1 - \theta} + \left[1 - \hat{F}_Y(\tau)\right] \log \frac{1 - \hat{F}_Y(\tau)}{\theta} \right\}, \end{aligned}$$

for $\theta \in (0,1)$, $\eta \in (0,1)$, $\tau \in \mathcal{T}$, where $\mathcal{T} = \left[\max\{x_{(1)}, y_{(1)}\}, \min\{x_{(m)}, y_{(n)}\}\right)$, and $x_{(i)}$, $i = 1, \dots, m$, and $y_{(j)}$, $j = 1, \dots, n$, denote the order statistics from the samples. When $\tau \notin \mathcal{T}$, then $\ell(\theta, \eta, \tau) = +\infty$. For readers convenience, some details are given in the Appendix. The Authors prove that, under very weak conditions, the quantity $\ell(\theta, \eta, \tau)$ is asymptotically pivotal. More precisely, for each triplet τ_0, $\theta_0 = 1 - F_Y(\tau_0)$ and $\eta_0 = F_X(\tau_0)$ of true parameter values, when $\min\{m, n\} \to +\infty$ and $\lim m/n$ is finite and not zero,

$$\ell(\theta_0, \eta_0, \tau_0) \xrightarrow{d} \chi^2_2,$$

where χ^2_2 indicates the chi-squared distribution with 2 degrees of freedom. Then, the authors propose to use such a result to construct non-parametric confidence regions for a pair of parameters, for example, (θ_0, η_0), when the third remaining parameter is fixed. Therefore, in the example just considered, the set

$$\mathcal{R}_\alpha = \left\{(\theta, \eta) : \ell(\theta, \eta, \tau_0) \le c_\alpha\right\},$$

where $\alpha \in (0,1)$ and c_α is such that $\Pr(\chi^2_2 > c_\alpha) = \alpha$, is a confidence region with nominal coverage probability $1 - \alpha$ for the pair (sensitivity, specificity), at a fixed cutpoint τ_0. At nominal confidence level $1 - \alpha$, $\mathcal{R}_\alpha$ gives the pairs of values for the sensitivity and specificity, associated with τ_0, which are compatible with the observed data. Regions obtained by $\ell(\theta, \eta, \tau_0)$ retain all good features of empirical likelihood confidence regions, whose shape and orientation are determined only by the data, and are range-respecting.

3. The proposed method

Let τ^+ be the true optimal cutpoint based on the Youden index approach, and let θ_0^+ and η_0^+ be the corresponding values of sensitivity and specificity. If τ^+ was known, the pivot $\ell(\theta^+, \eta^+, \tau^+)$ could be used directly to build approximate confidence regions for (θ_0^+, η_0^+). In practice, τ^+ is unknown and must be estimated from the available data.

Let $\hat{\tau}^+$ be a suitable estimator for τ^+. By resorting to the plug-in method, we can consider the quantity $\ell(\theta^+, \eta^+, \hat{\tau}^+)$, where an estimate replaces an unknown value. Unfortunately, such a replacement is not untroublesome. Indeed, the standard χ^2 approximation no longer holds in such a case. However, in the following, we resort to a simple and fast (parametric) bootstrap calibration and apply it to the distribution of $\ell_*(\theta_0^+, \eta_0^+) \equiv \ell(\theta_0^+, \eta_0^+, \hat{\tau}^+)$ when a parametric (or semi-parametric) model can be assumed for the data. This allows us to obtain estimates of the quantiles of the distribution of $\ell_*(\theta^+, \eta^+)$, which we will use to define the desired confidence regions. We will illustrate such a procedure, starting from the binormal case and then extending the proposed approach to situations where the biomarker distribution is far from the normal one but can reach the normal shape after proper transformation. We will propose also a completely non-parametric approach.

3.1. Binormal model

In some practical situations, it is reasonable to assume that the diagnostic test has a normal distribution, in both populations of healthy and diseased subjects. Let $X \sim N(\mu_x, \sigma_x^2)$ and $Y \sim N(\mu_y, \sigma_y^2)$. To formalize the assumption that the values of the test are positively associated with the disease status, we impose hereafter that $\mu_y > \mu_x$.

For the binormal model, the optimal cutpoint provided by the criterion based on the Youden index is obtained analytically as

$$\tau^+ = \mu_x - \frac{(\mu_y - \mu_x) - (\sigma_y/\sigma_x)\sqrt{(\mu_y - \mu_x) + ((\sigma_y/\sigma_x)^2 - 1)\sigma_x^2 \log((\sigma_y/\sigma_x)^2)}}{(\sigma_y/\sigma_x)^2 - 1}.$$

When variances are equal, i.e. $\sigma_x^2 = \sigma_y^2$, then $\tau^+ = (\mu_x + \mu_y)/2$. Then, by the plug-in principle, a consistent estimator of τ^+ can be obtained by substituting in the above expressions the empirical counterparts $\bar{X} = \frac{1}{m}\sum_{i=1}^m X_i$, $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$, $S_x = \sqrt{\frac{1}{m-1}\sum_{i=1}^m (X_i - \bar{X})^2}$ and $S_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y})^2}$ of μ_x, μ_y, σ_x and σ_y, respectively.

Let $\Phi(\cdot)$ denote the cumulative distribution function of the standard normal. Given the observed data $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$, we compute the estimate $\hat{\tau}^+$ and the corresponding optimal estimated sensitivity $\hat{\theta}^+ = 1 - \Phi((\hat{\tau}^+ - \bar{y})/S_y)$ and specificity $\hat{\eta}^+ = \Phi((\hat{\tau}^+ - \bar{x})/S_x)$. To conveniently calibrate $\ell_*(\theta^+, \eta^+)$ we resort to an adjusted bootstrap procedure:

1. use parametric bootstrap to get B bootstrap samples $\{x\}_b$ and $\{y\}_b$, for $b = 1, \ldots, B$, of sizes m and n, respectively;
2. add to each bootstrap sample the extremes (min and max) of the corresponding original sample, to obtain "enlarged" bootstrap samples of size $m + 2$ and $n + 2$, respectively;
3. compute $(\bar{x}_b, \bar{y}_b, S_{xb}, S_{yb})$, $\hat{\tau}_b^+$ and $\ell(\hat{\theta}^+, \hat{\eta}^+, \hat{\tau}_b^+)$ from the b-th pair of (enlarged) bootstrap samples, where $\hat{\theta}^+$ and $\hat{\eta}^+$ are the estimates obtained from the original data;
4. get the estimate $\hat{c}_\alpha$ as the sample quantile of order $1 - \alpha$ from the values $\ell(\hat{\theta}^+, \hat{\eta}^+, \hat{\tau}_b^+)$, $b = 1, \ldots, B$.

Then, the set

$$\mathcal{R}_\alpha = \left\{(\theta^+, \eta^+) : \ell_*(\theta^+, \eta^+) \leq \hat{c}_\alpha\right\},$$

is a confidence region, with nominal coverage $1 - \alpha$, for the optimal pair (sensitivity, specificity), corresponding to the Youden index criterion. Step 2 in the above-given procedure controls for possible contractions of the convex hulls of bootstrap samples, to reduce the effects of the so-called convex hull problem, i.e., to reduce the probability that $\ell(\hat{\theta}^+, \hat{\eta}^+, \hat{\tau}_b^+)$ may be not finite. Moreover, we process only (enlarged) bootstrap samples whose averages respect the fixed order $(\mu_y > \mu_x)$.

3.2. Normal models after Box-Cox transformations

Suppose now that distributions of X and Y cannot reasonably be considered normal. Moreover, assume that $X > 0$ and $Y > 0$, with probability 1. In such a situation, Box-Cox transformations can help to achieve normality and are frequently used in ROC studies (Bantis et al., 2014; Fluss et al., 2005). However, to the best of our knowledge, papers present in the literature discuss, in this regard, the same transformation for the test results, both for healthy and diseased subjects. Although this approach has the advantage of leaving the ROC curve unchanged (due to its invariance property with respect to increasing monotonic transformations), it can be inappropriate in cases where the two distributions of X and Y are very different from each other. In the following, we show how a more general approach involving two different transformations may be used, at least for the aim of this paper.

Let

$$X^{(\lambda_1)} = \begin{cases} \frac{X^{\lambda_1}-1}{\lambda_1} & \lambda_1 \neq 0 \\ \log(X) & \lambda_1 = 0 \end{cases} \qquad Y^{(\lambda_2)} = \begin{cases} \frac{Y^{\lambda_2}-1}{\lambda_2} & \lambda_2 \neq 0 \\ \log(Y) & \lambda_2 = 0, \end{cases}$$

the transformed test results. The parameters defining the transformations are denoted by λ_1 and λ_2. Our proposal starts from the following observation. Since the optimal cutpoint τ^+ meets the relation

$$\begin{aligned} \tau^+ &= \arg\max_\tau \left[F_X(\tau) - F_Y(\tau)\right] = \arg\max_\tau \left[\Pr(X \leq \tau) - \Pr(Y \leq \tau)\right] \\ &= \arg\max_\tau \left[\Pr\left(X^{(\lambda_1)} \leq \tau^{(\lambda_1)}\right) - \Pr\left(Y^{(\lambda_2)} \leq \tau^{(\lambda_2)}\right)\right], \end{aligned}$$

where $\tau^{(\lambda_1)}$ and $\tau^{(\lambda_2)}$ are the transformed values of the same generic cutpoint τ, the double transformation leaves unchanged the optimal values of sensitivity and specificity. Moreover,

$$\theta_0^+ = 1 - \Pr\left(Y^{(\lambda_2)} \leq \tau^{+(\lambda_2)}\right) \quad \text{and} \quad \eta_0^+ = \Pr\left(X^{(\lambda_1)} \leq \tau^{+(\lambda_1)}\right).$$

Clearly, in the transformed scales, such optimal values match with two different cutpoint values.

In practice, λ_1 and λ_2 are estimated by using the maximum likelihood approach (Box and Cox, 1964). Let $\hat{\lambda}_1$ and $\hat{\lambda}_2$ be the estimates, and set $\hat{\tau}_1^+ = \hat{\tau}^{+(\hat{\lambda}_1)}$ and $\hat{\tau}_2^+ = \hat{\tau}^{+(\hat{\lambda}_2)}$, where $\hat{\tau}^+$ is some estimate of τ^+ . Then, denoting by $\hat{F}_X^{(1)}$ and $\hat{F}_Y^{(2)}$ the empirical distribution functions obtained by the sample from $X^{(\hat{\lambda}_1)}$ and $Y^{(\hat{\lambda}_2)}$, respectively, the empirical likelihood pivot that we will use is

$$\begin{aligned} \ell(\theta^+, \eta^+, \hat{\tau}^+) &= 2m \left\{ \hat{F}_X^{(1)}(\hat{\tau}_1^+) \log \frac{\hat{F}_X^{(1)}(\hat{\tau}_1^+)}{\eta^+} + \left[1 - \hat{F}_X^{(1)}(\hat{\tau}_1^+)\right] \log \frac{1 - \hat{F}_X^{(1)}(\hat{\tau}_1^+)}{1 - \eta^+} \right\} \\ &\quad + 2n \left\{ \hat{F}_Y^{(2)}(\hat{\tau}_2^+) \log \frac{\hat{F}_Y^{(2)}(\hat{\tau}_2^+)}{1 - \theta^+} + \left[1 - \hat{F}_Y^{(2)}(\hat{\tau}_2^+)\right] \log \frac{1 - \hat{F}_Y^{(2)}(\hat{\tau}_2^+)}{\theta^+} \right\}, \end{aligned} \tag{1}$$

which is infinite when $\hat{\tau}_1^+$ is out of the range of the sample from $X^{(\hat{\lambda}_1)}$, or $\hat{\tau}_2^+$ is out of the range of the sample from $Y^{(\hat{\lambda}_2)}$. Observe that now the pivot in (1) depends on three estimated nuisance parameters: $\hat{\lambda}_1$, $\hat{\lambda}_2$ and $\hat{\tau}^+$.

To obtain confidence regions for the pair (θ_0^+, η_0^+), we propose the following algorithm. Given the observed data $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$, we firstly compute the estimates $\hat{\lambda}_1$ and $\hat{\lambda}_2$. Then, using transformed data $x'_1, \ldots, x'_m$ and $y'_1, \ldots, y'_n$, say, we obtain sample means and standard deviations $\bar{x}'$, $\bar{y}'$, S'_x, S'_y, and maximize, with respect to τ,

$$\Phi\left(\frac{\tau^{(\hat{\lambda}_1)} - \bar{x}'}{S'_x}\right) - \Phi\left(\frac{\tau^{(\hat{\lambda}_2)} - \bar{y}'}{S'_y}\right)$$

to get $\hat{\tau}^+$. Next, we can get the estimates $\hat{\theta}^+ = 1 - \Phi\left(\frac{\hat{\tau}_2^+ - \bar{y}'}{S'_y}\right)$ and $\hat{\eta}^+ = \Phi\left(\frac{\hat{\tau}_1^+ - \bar{x}'}{S'_x}\right)$. Finally, we use bootstrap calibration:

1. get B parametric bootstrap samples $\{x'\}_b$ and $\{y'\}_b$, for $b = 1, \ldots, B$, of sizes m and n, respectively;
2. add to each bootstrap sample the extremes (min and max) of the corresponding original (transformed) sample, to obtain "enlarged" bootstrap samples of size $m + 2$ and $n + 2$, respectively;

3. from the b-th pair of (enlarged) bootstrap samples, compute $(\bar{x}'_b, \bar{y}'_b, S'_{xb}, S'_{yb})$, $\hat{\tau}^+_b$, $\hat{\tau}^+_{1b}$, $\hat{\tau}^+_{2b}$ and $\ell(\hat{\theta}^+, \hat{\eta}^+, \hat{\tau}^+_b)$, using (1) where $\hat{\theta}^+$ and $\hat{\eta}^+$ are the estimates obtained from the original (transformed) data;
4. get the estimate $\hat{c}_\alpha$ as the sample quantile of order $1-\alpha$ from the values $\ell(\hat{\theta}^+, \hat{\eta}^+, \hat{\tau}^+_b)$, $b = 1, \dots, B$.

The set

$$\mathcal{R}_\alpha = \left\{(\theta^+, \eta^+) : \ell_*(\theta^+, \eta^+) \le \hat{c}_\alpha\right\},$$

is a confidence region with nominal coverage probability $1-\alpha$, for the optimal pair (sensitivity, specificity), corresponding to the Youden index criterion. In the bootstrap procedure, we process only (enlarged) bootstrap samples whose averages respect the order between the averages of the original (transformed) samples.

It is finally worth remarking that the use of the two transformations represents only a tool to obtain a consistent inferential result, in situations where this is necessary. To estimate the ROC curve, for instance, practitioners would use the original data and compute the empirical ROC curve. The double transformation enters into play only to obtain the confidence regions for the pair (θ^+_0, η^+_0) of optimal sensitivity and specificity, corresponding to the cutpoint τ^+. Observe moreover that the estimate $\hat{\tau}^+$ is provided on the original scale.

3.3. Fully non-parametric approach

There are cases in which the use of Box-Cox transformations is not entirely satisfactory, and a completely non-parametric approach is more effective. Let $\tilde{J}(\tau) = \tilde{F}_X(\tau) - \tilde{F}_Y(\tau)$ be an estimated version of the Youden index, based on suitable kernel estimators of F_X and F_Y. In particular, here we consider the well-known Gaussian kernel, so that, for instance,

$$\tilde{F}_X(\tau) = \frac{1}{m}\sum_{i=1}^{m} \Phi\left(\frac{\tau - x_i}{h_m}\right),$$

where h_m denotes an appropriate bandwidth sequence. Maximization of $\tilde{J}(\tau)$, gives a consistent estimator (Fluss et al., 2005) $\tilde{\tau}^+$ for the optimal cutpoint τ^+. Therefore, we can use the pivot $\ell_*(\theta^+, \eta^+) = \ell(\theta^+, \eta^+, \tilde{\tau}^+)$ to build approximate confidence regions for (θ^+_0, η^+_0).

In order to calibrate the distribution of $\ell_*(\theta^+_0, \eta^+_0)$, we can resort to a standard non-parametric bootstrap procedure:

1. given the observed data $x_1, \dots, x_m$ and $y_1, \dots, y_n$, obtain the estimates $\tilde{\tau}^+$, $\tilde{\theta}^+ = 1 - \tilde{F}_Y(\tilde{\tau}^+)$ and $\tilde{\eta}^+ = \tilde{F}_X(\tilde{\tau}^+)$;
2. get B non-parametric bootstrap samples $\{x\}_b$ and $\{y\}_b$, for $b = 1, \dots, B$, of sizes m and n, respectively;
3. add to each bootstrap sample the extremes (min and max) of the original sample, to obtain "enlarged" bootstrap samples of size $m+2$ and $n+2$, respectively;
4. from the b-th pair of (enlarged) bootstrap samples, compute the estimate $\tilde{\tau}^+_b$, and $\ell(\tilde{\theta}^+, \tilde{\eta}^+, \tilde{\tau}^+_b)$, where $\tilde{\theta}^+$ and $\tilde{\eta}^+$ are the estimate obtained from the original data;
5. get the estimate $\hat{c}_\alpha$ as the sample quantile of order $1-\alpha$ from the values $\ell(\tilde{\theta}^+, \tilde{\eta}^+, \tilde{\tau}^+_b)$, $b = 1, \dots, B$.

The set

$$\mathcal{R}_\alpha = \left\{(\theta^+, \eta^+) : \ell_*(\theta^+, \eta^+) \le \hat{c}_\alpha\right\},$$

is a confidence region with nominal coverage $1-\alpha$, for the optimal pair (sensitivity, specificity), corresponding to the Youden index criterion. In the bootstrap procedure, we process only (enlarged) bootstrap samples whose averages respect the fixed order ($\mu_y > \mu_x$). Enlarged samples again serve to reduce the possible effects of the convex hull problem.

As for the choice of the bandwidth sequences, h_m and h_n, we suggest resorting to results in Azzalini (1981). More precisely, we set $h_m = K_x S_x m^{-1/3}$ and $h_n = K_y S_y n^{-1/3}$, and use such sequences also in the bootstrap process. An empirical study, preliminary to our simulations discussed in the next section, suggests us, as a good compromise, to set $K_x = K_y = 1.1$.

Observe that, in all cases covered in sections 3.1-3.3, the point estimates, say $\hat{\theta}^+_\ell$ and $\hat{\eta}^+_\ell$, obtainable by minimizing the functions $\ell(\cdot, \cdot, \hat{\tau}^+)$ and $\ell(\cdot, \cdot, \tilde{\tau}^+)$, coincide with the empirical estimates, i.e. $\hat{\theta}^+_\ell = 1 - \hat{F}_Y(\hat{\tau}^+)$ (or $\hat{\theta}^+_\ell = 1 - \hat{F}_Y(\hat{\tau}^+_2)$) and $\hat{\eta}^+_\ell = \hat{F}_X(\hat{\tau}^+)$ (or $\hat{\eta}^+_\ell = \hat{F}_X(\hat{\tau}^+_1)$), in the semi-parametric case, and $\hat{\theta}^+_\ell = 1 - \hat{F}_Y(\tilde{\tau}^+)$ and $\hat{\eta}^+_\ell = \hat{F}_X(\tilde{\tau}^+)$, in the non-parametric case.

4. Simulation study

From a practical point of view, it is essential to investigate the accuracy of confidence regions, built according to the techniques above presented, in samples of finite size. In this section, we conduct some simulation experiments to achieve this goal. We focus on sixteen scenarios, three of which are directly related to the binormal model, that allow us to evaluate the technique proposed in Section 3.1. The remaining thirteen scenarios allow us to evaluate both the technique presented in Section 3.2, involving Box-Cox transformation (Emp BC) and the fully non-parametric technique (Emp NP), discussed in Section 3.3. The considered scenarios are listed below:

(I) $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{N}(2.321, 3.925)$; the true optimal cutpoint based on the Youden index approach is $\tau^+ = 1.282$; the corresponding specificity and sensitivity are $\eta_0^+ = 0.9$ and $\theta_0^+ = 0.7$.
(II) $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{N}(1.645, 0.393)$, $\tau^+ = 0.842$, $\eta_0^+ = 0.8$ and $\theta_0^+ = 0.9$.
(III) $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{N}(3.988, 5.111)$, $\tau^+ = 1.645$, $\eta_0^+ = 0.95$ and $\theta_0^+ = 0.85$.
(IV) $X \sim \mathcal{G}a(5,1)$, $Y \sim \mathcal{G}a(2.474, 0.185)$, $\tau^+ = 7.992$, $\eta_0^+ = 0.9$ and $\theta_0^+ = 0.7$.
(V) $X \sim \mathcal{G}a(5,1)$, $Y \sim \mathcal{G}a(3.894, 0.277)$, $\tau^+ = 7.999$, $\eta_0^+ = 0.9$ and $\theta_0^+ = 0.8$.
(VI) $X \sim \mathcal{G}a(2, 0.5)$, $Y \sim \mathcal{G}a(2, 0.072)$, $\tau^+ = 9.056$, $\eta_0^+ = 0.94$ and $\theta_0^+ = 0.86$.
(VII) $\log(X) \sim \mathcal{N}(2, 0.1)$, $\log(Y) \sim \mathcal{N}(2.6, 0.25)$, $\tau^+ = 10.418$, $\eta_0^+ = 0.861$ and $\theta_0^+ = 0.696$.
(VIII) $\log(X) \sim \mathcal{N}(2, 0.1)$, $\log(Y) \sim \mathcal{N}(2.8, 0.25)$, $\tau^+ = 10.980$, $\eta_0^+ = 0.895$ and $\theta_0^+ = 0.790$.
(IX) $\log(X) \sim \mathcal{N}(2, 0.1)$, $\log(Y) \sim \mathcal{N}(3.1, 0.25)$, $\tau^+ = 12.065$, $\eta_0^+ = 0.940$ and $\theta_0^+ = 0.889$.
(X) $\log(X) \sim \mathcal{N}(0, 0.25)$, $Y \sim \mathcal{G}a(3,1)$, $\tau^+ = 1.706$, $\eta_0^+ = 0.857$ and $\theta_0^+ = 0.756$.
(XI) $\log(X) \sim \mathcal{N}(0, 0.25)$, $Y \sim \mathcal{G}a(3.6,1)$, $\tau^+ = 1.835$, $\eta_0^+ = 0.888$ and $\theta_0^+ = 0.833$.
(XII) $\log(X) \sim \mathcal{N}(0, 0.563)$, $Y \sim \mathcal{G}a(4.6,1)$, $\tau^+ = 2.190$, $\eta_0^+ = 0.852$ and $\theta_0^+ = 0.895$.
(XIII) $X \sim \mathcal{B}e(1, 1.8)$, $Y \sim \mathcal{B}e(2, 3/4)$, $\tau^+ = 0.568$, $\eta_0^+ = 0.779$ and $\theta_0^+ = 0.760$.
(XIV) $X \sim \mathcal{B}e(1, 1.5)$, $Y \sim \mathcal{B}e(2, 1/3)$, $\tau^+ = 0.731$, $\eta_0^+ = 0.860$ and $\theta_0^+ = 0.803$.
(XV) $X \sim \mathcal{N}(10,1)$, $Y \sim 0.5\mathcal{N}(12.16, 1) + 0.5\mathcal{N}(16.16, 5)$, $\tau^+ = 11.37$, $\eta_0^+ = 0.915$ and $\theta_0^+ = 0.844$.
(XVI) $X \sim 0.5\mathcal{N}(-1,1) + 0.5\mathcal{N}(2,1)$, $Y \sim 0.5\mathcal{N}(2.5, 1.5) + 0.5\mathcal{N}(4,1)$, $\tau^+ = 2.194$, $\eta_0^+ = 0.788$ and $\theta_0^+ = 0.782$.

For each scenario, we generated 10,000 pairs of samples, with different sample sizes. A generated pair is processed only if the sample averages respect the fixed order ($\mu_y > \mu_x$). As for the parametric and non-parametric bootstrap calibration, we use $B = 1000$.

Moreover, we compare our proposed methods to the existing approaches: the delta-based method with or without common Box-Cox transformation discussed in Bantis et al. (2014), and the generalized pivot quantile (GPQ) method and smoothed bootstrap with arcsin-square-root transformation method (BTATII) proposed by Yin and Tian (2014). In particular, we apply the delta-based (Delta) and the GPQ methods in scenarios (I)-(III), where two distributions are normal. For the non-normal cases, the Delta method with common Box-Cox transformation (Delta BC) and the BTATII methods are considered. For the BTATII method, we use the well-known Gaussian kernel with the rule-of-thumb bandwidth $h_j = 0.9 \min\{\mathrm{sd}(x_j), \mathrm{IQR}(x_j)/1.34\} n_j^{-1/5}$, where $\mathrm{sd}(x_j)$ and $\mathrm{IQR}(x_j)$ are the standard deviation and the inner quartile range of two samples, for $j = 1, 2$ (non-diseased and diseased groups), and we consider 500 smoothed bootstrap samples, as in Yin and Tian (2014).

For three levels of nominal coverage $1 - \alpha$, that is, 0.90, 0.95 and 0.99, Tables 1-5 give the estimated coverage probabilities of the confidence regions obtained by using the proposed methods and the existing approaches (Delta, Delta BC, GPQ and BTATII). Each table refers to a particular chosen parametric setting for data generation. Simulation results show that confidence regions based on $\ell_*(\theta^+, \eta^+)$ are sufficiently accurate in almost all considered cases for which their use is appropriate. In particular, confidence regions seem to be effective for binormal models or binormal models after transformation (scenarios I-XII), even at the smallest considered sample sizes, and when the model is relatively close to the Box-Cox transformation family (Gamma distribution). Confidence regions for the non-parametric case generally require a higher sample size, as expected. Its usefulness is proven, in particular, by the results in Table 5. Clearly, in all cases, accuracy increases with sample sizes, and the closer the values of θ_0^+ and η_0^+ to 1, the larger the required sample sizes.

Under scenarios with normal distributions (see Table 1), our proposed empirical likelihood method has a similar or better performance compared to the Delta or the GPQ methods. Under scenarios with non-normal distributions, the Delta BC method performs well only when the two distributions are both gamma or log-normal, see Tables 2 and

Table 1
Simulation results for scenarios I - III: binormal settings. Emp is the proposed empirical likelihood method, Delta is the delta-based approach, and GPQ is the generalized pivot quantile method.

Sample size	Nominal coverage: 0.90			Nominal coverage: 0.95			Nominal coverage: 0.99		
	Emp	Delta	GPQ	Emp	Delta	GPQ	Emp	Delta	GPQ
	Scenario I: $\theta_0^+ = 0.7$, $\eta_0^+ = 0.9$								
(30, 30)	0.899	0.883	0.874	0.945	0.938	0.930	0.981	0.983	0.984
(50, 50)	0.901	0.896	0.882	0.950	0.945	0.935	0.992	0.988	0.987
(50, 75)	0.894	0.894	0.891	0.949	0.943	0.944	0.991	0.986	0.988
(75, 50)	0.895	0.887	0.887	0.950	0.939	0.940	0.989	0.987	0.987
(75, 75)	0.897	0.896	0.886	0.949	0.943	0.941	0.991	0.986	0.985
(100, 100)	0.901	0.893	0.893	0.950	0.946	0.944	0.990	0.987	0.988
	Scenario II: $\theta_0^+ = 0.8$, $\eta_0^+ = 0.9$								
(30, 30)	0.906	0.887	0.876	0.955	0.934	0.939	0.980	0.981	0.987
(50, 50)	0.894	0.897	0.883	0.948	0.943	0.938	0.989	0.986	0.986
(50, 75)	0.894	0.895	0.890	0.948	0.943	0.944	0.988	0.987	0.990
(75, 50)	0.906	0.892	0.890	0.947	0.945	0.941	0.991	0.987	0.986
(75, 75)	0.902	0.893	0.893	0.952	0.944	0.943	0.989	0.986	0.989
(100, 100)	0.905	0.897	0.890	0.952	0.948	0.942	0.991	0.989	0.986
	Scenario III: $\theta_0^+ = 0.85$, $\eta_0^+ = 0.95$								
(50, 50)	0.884	0.890	0.886	0.927	0.941	0.942	0.953	0.984	0.990
(50, 75)	0.897	0.895	0.891	0.934	0.946	0.946	0.959	0.986	0.987
(75, 50)	0.889	0.893	0.883	0.947	0.942	0.938	0.985	0.984	0.987
(75, 75)	0.902	0.894	0.891	0.949	0.947	0.946	0.987	0.987	0.988
(100, 100)	0.903	0.897	0.893	0.948	0.946	0.944	0.989	0.988	0.988

3. This method shows unsatisfactory coverage probability compared to our proposed Emp BC and Emp NP methods when two distributions have different forms as in scenarios (X)-(XII), see Table 4, or under the beta distributions (see Table 5). The performance of the BTATII method appears generally inconsistent; in fact, this method shows low coverage probabilities in several settings, regardless of increasing the sample sizes (see Tables 2-5). Under scenarios (XII)-(XIV), with two beta distributions, only our Emp NP method seems to be effective, and this also occurs in mixture settings (XV)-(XVI) (see Table 6). Note that, in the mixture settings, we can reasonably apply our Emp BC method and the Delta DC method only in scenario (XV), where the probability of having negative values from X and Y is very low.

5. Applications

In this section, we apply our proposed methods to three real datasets concerning Alzheimer's disease, breast cancer and melanoma. The three datasets are chosen to cover the different situations discussed in Section 3. In the applications, we set $B = 1000$.

5.1. Alzheimer's disease

We consider data from Alzheimer's Disease Neuroimaging Initiative (ADNI, `adni.loni.usc.edu`). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see `www.adni-info.org`.

Here, we focus on T-tau (total-tau) protein, a well-known biological marker for AD. We consider subjects who are registered in three ADNI projects: ADNI 1, ADNI 2 and ADNI GO. The true disease status of subjects was diagnosed employing neuropsychological tests and consisted of three categories: cognitively normal (CN), mild cognitive impairment (MCI), and Alzheimer's disease (AD). We want to investigate some aspects related to the

Table 2
Simulation results for scenarios IV-VI: bi-gamma settings. Emp BC is the proposed empirical likelihood method with two different Box-Cox transformations, Emp NP is the proposed empirical likelihood method with kernel estimation, Delta BC is the delta-based approach with common Box-Cox transformation, BTATII is the smoothed bootstrap with arcsin-square-root transformation.

Sample size	Nominal coverage: 0.90				Nominal coverage: 0.95				Nominal coverage: 0.99			
	Emp BC	Emp NP	Delta BC	BTATII	Emp BC	Emp NP	Delta BC	BTATII	Emp BC	Emp NP	Delta BC	BTATII
	Scenario IV: $\theta_0^+ = 0.7$, $\eta_0^+ = 0.9$											
(30, 30)	0.873	0.844	0.895	0.908	0.927	0.854	0.944	0.959	0.975	0.860	0.985	0.967
(50, 50)	0.870	0.921	0.903	0.889	0.927	0.953	0.954	0.950	0.984	0.971	0.991	0.982
(50, 75)	0.876	0.914	0.905	0.873	0.935	0.948	0.950	0.935	0.985	0.970	0.989	0.974
(75, 50)	0.855	0.919	0.900	0.873	0.922	0.967	0.948	0.938	0.982	0.993	0.988	0.985
(75, 75)	0.873	0.916	0.901	0.867	0.933	0.959	0.950	0.932	0.985	0.992	0.990	0.982
(100, 100)	0.875	0.915	0.904	0.862	0.935	0.940	0.964	0.925	0.985	0.995	0.990	0.980
(150, 150)		0.919		0.841		0.965		0.912		0.995		0.976
	Scenario V: $\theta_0^+ = 0.8$, $\eta_0^+ = 0.9$											
(30, 30)	0.874	0.860	0.890	0.902	0.935	0.873	0.943	0.949	0.972	0.882	0.984	0.949
(50, 50)	0.874	0.919	0.895	0.888	0.934	0.954	0.943	0.940	0.987	0.978	0.986	0.970
(50, 75)	0.886	0.908	0.903	0.874	0.936	0.951	0.951	0.931	0.987	0.976	0.988	0.963
(75, 50)	0.862	0.906	0.898	0.879	0.927	0.952	0.947	0.937	0.983	0.990	0.988	0.979
(75, 75)	0.875	0.908	0.900	0.876	0.939	0.954	0.949	0.928	0.986	0.993	0.989	0.974
(100, 100)	0.883	0.909	0.905	0.866	0.941	0.961	0.954	0.923	0.987	0.993	0.989	0.974
(150, 150)		0.905		0.855		0.957		0.912		0.995		0.973
	Scenario VI: $\theta_0^+ = 0.86$, $\eta_0^+ = 0.94$											
(50, 50)	0.897	0.828	0.896	0.875	0.942	0.838	0.946	0.938	0.977	0.844	0.987	0.921
(50, 75)	0.889	0.848	0.898	0.861	0.945	0.862	0.949	0.924	0.977	0.867	0.989	0.923
(75, 50)	0.880	0.910	0.897	0.864	0.942	0.938	0.945	0.923	0.987	0.948	0.985	0.954
(75, 75)	0.885	0.908	0.901	0.864	0.948	0.941	0.949	0.924	0.988	0.952	0.988	0.957
(100, 100)	0.886	0.926	0.898	0.859	0.946	0.964	0.948	0.920	0.988	0.983	0.989	0.967
(150, 150)		0.922		0.848		0.968		0.909		0.996		0.967

ability of T-tau protein to distinguish healthy people from Alzheimer's patients. More precisely, our goal is to obtain confidence regions for their optimal sensitivity and specificity. For the analysis, we treat data only for subjects who are younger than 75 years and for two groups: CN and AD. The total number of units is 321, with 214 CN and 107 AD.

As often happens in practice, we consider the log-transformation of T-tau protein values, in our analysis. Since (see Figure 1) the empirical distributions of the transformed values support the assumption of normality in both groups (CN and AD), we apply the method proposed in Section 3.1 to construct confidence regions for the optimal sensitivity and specificity of such biomarker. Figure 2 shows the regions, of nominal levels 0.90, 0.95 and 0.99, produced by our method. The (empirical) point estimates for the optimal sensitivity and specificity are $\hat{\theta}_\ell^+ = 0.748$ and $\hat{\eta}_\ell^+ = 0.790$, respectively. A look at Figure 2, allows us to conclude that, when used with the optimal cutpoint provided by the Youden index criterion, the T-tau protein has a quite good ability to discriminate between healthy people and Alzheimer's patients.

5.2. Breast cancer

We use data previously analyzed in Patrício et al. (2018). Data refer to 64 women with breast cancer, recruited from the Gynaecology Department of the University Hospital Centre of Coimbra (CHUC) between 2009 and 2013, and 52 healthy volunteers. Each diseased patient's diagnosis came from positive mammography and was histologically confirmed, and data are collected before surgery and treatment. Blood samples together with clinical, demographic and anthropometric data were collected for all participants, during the first consultation. The final dataset includes 9 variables: Age, BMI, Glucose, Insulin, HOMA (Homeostasis Model Assessment index), Leptin, Adiponectin, Resistin and MCP-1 (Chemokine Monocyte Chemoattractant Protein 1). By using Support Vector Machines (SVM) models, Patrício et al. (2018) find that the best combination (in terms of sensitivity and specificity) of such variables involves Age, BMI, Glucose and Resistin as predictors.

Table 3
Simulation results for scenarios VII-IX: bi-log normal settings. Emp BC is the proposed empirical likelihood method with two different Box-Cox transformations, Emp NP is the proposed empirical likelihood method with kernel estimation, Delta BC is the delta-based approach with common Box-Cox transformation, BTATII is the smoothed bootstrap with arcsin-square-root transformation.

Sample size	Nominal coverage: 0.90				Nominal coverage: 0.95				Nominal coverage: 0.99			
	Emp BC	Emp NP	Delta BC	BTATII	Emp BC	Emp NP	Delta BC	BTATII	Emp BC	Emp NP	Delta BC	BTATII
	Scenario VII: $\theta_0^+ = 0.696$, $\eta_0^+ = 0.861$											
(30, 30)	0.865	0.883	0.878	0.854	0.936	0.922	0.931	0.930	0.985	0.935	0.980	0.978
(50, 50)	0.869	0.908	0.895	0.833	0.929	0.958	0.942	0.908	0.983	0.988	0.984	0.975
(50, 75)	0.866	0.908	0.895	0.819	0.927	0.957	0.946	0.892	0.984	0.990	0.987	0.967
(75, 50)	0.862	0.879	0.893	0.822	0.932	0.936	0.947	0.896	0.983	0.989	0.986	0.969
(75, 75)	0.865	0.896	0.897	0.818	0.932	0.950	0.946	0.893	0.984	0.991	0.988	0.966
(100, 100)	0.866	0.901	0.896	0.793	0.927	0.953	0.947	0.873	0.983	0.992	0.989	0.957
(150, 150)		0.910		0.785		0.956		0.863		0.994		0.952
	Scenario VIII: $\theta_0^+ = 0.790$, $\eta_0^+ = 0.895$											
(30, 30)	0.887	0.849	0.880	0.871	0.940	0.864	0.933	0.946	0.979	0.871	0.983	0.954
(50, 50)	0.874	0.907	0.897	0.857	0.936	0.948	0.946	0.924	0.986	0.973	0.988	0.970
(50, 75)	0.877	0.912	0.900	0.849	0.936	0.955	0.948	0.917	0.988	0.979	0.988	0.965
(75, 50)	0.870	0.888	0.887	0.847	0.935	0.946	0.942	0.912	0.986	0.987	0.987	0.972
(75, 75)	0.877	0.899	0.897	0.850	0.934	0.949	0.947	0.911	0.986	0.991	0.987	0.969
(100, 100)	0.867	0.895	0.894	0.841	0.935	0.953	0.948	0.907	0.987	0.990	0.990	0.971
(150, 150)		0.896		0.817		0.949		0.887		0.992		0.963
	Scenario IX: $\theta_0^+ = 0.888$, $\eta_0^+ = 0.940$											
(50, 50)	0.903	0.863	0.893	0.904	0.944	0.877	0.945	0.934	0.974	0.882	0.986	0.906
(50, 75)	0.900	0.866	0.895	0.901	0.946	0.882	0.943	0.927	0.977	0.887	0.987	0.916
(75, 50)	0.887	0.914	0.893	0.889	0.948	0.946	0.944	0.931	0.991	0.963	0.984	0.933
(75, 75)	0.891	0.917	0.890	0.890	0.950	0.949	0.942	0.933	0.991	0.963	0.987	0.948
(100, 100)	0.886	0.921	0.903	0.894	0.947	0.965	0.949	0.936	0.989	0.991	0.988	0.962
(150, 150)		0.926		0.893		0.968		0.933		0.995		0.968

As an illustration of our methods, we consider as biomarker the following linear combination of the three predictors: Glucose + Resistin - 1.5×BMI. The coefficients of combination and the exclusion of Age are justified by a preliminary analysis based on logistic regression. The estimated AUC (area under the ROC curve) for the new biomarker (say, T) is roughly 0.84. We aim to construct confidence regions for optimal sensitivity and specificity of T.

Empirical distributions of T, for healthy and diseased subjects, appear far from normality. Therefore, taking into account also the relatively small sample sizes, we use the method proposed in Section 3.2 involving Box-Cox transformations. The estimated Box-Cox parameters are $\hat{\lambda}_1 = -0.197$ and $\hat{\lambda}_2 = -1.070$, for healthy and diseased groups, respectively. Figure 3 shows Q-Q normal plots for T, before and after transformations, whereas Figure 4 shows the confidence regions of nominal levels 0.90, 0.95 and 0.99 produced by our method. The (empirical) point estimates for the optimal sensitivity and specificity are $\hat{\theta}_\ell^+ = 0.797$ and $\hat{\eta}_\ell^+ = 0.75$, respectively. A look at Figure 4, allows us to conclude that, when used with the optimal cutpoint provided by the Youden index criterion, the biomarker T has a satisfactory ability to discriminate between healthy and diseased women, especially in terms of sensitivity.

5.3. Melanoma

In a recent paper, Cesati et al. (2020) analyzed the gene expression of 27 cytokines/chemokines, at the tissue level, in a sample of patients with primary melanoma, metastatic melanoma and healthy. The used data comes from the GENT2 database and the study aimed to identify effective biomarkers to predict the presence of the tumour. By using SVM, the authors find, among other things, that the combination of 4 cytokines, namely, IL-1Ra, IL-7, MIP-1a, and MIP-1b, is the most effective signature to discriminate melanoma patients.

We use the same dataset, and by an inspection of the boxplots in Figure 5, representing the empirical distributions for the expressions of such genes for healthy and diseased patients, we consider as biomarker the linear combination MIP1-a − IL1Ra − IL-7 + MIP-1b. The boxplots giving the empirical distributions of the investigated biomarker say W, for healthy and diseased patients are presented in Figure 6. Figure 7 shows the related Q-Q normal plots.

Table 4
Simulation results for scenarios X - XII: log-normal and gamma distributions. Emp BC is the proposed empirical likelihood method with two different Box-Cox transformations, Emp NP is the proposed empirical likelihood method with kernel estimation, Delta BC is the delta-based approach with common Box-Cox transformation, BTATII is the smoothed bootstrap with arcsin-square-root transformation.

Sample size	Nominal coverage: 0.90				Nominal coverage: 0.95				Nominal coverage: 0.99			
	Emp BC	Emp NP	Delta BC	BTATII	Emp BC	Emp NP	Delta BC	BTATII	Emp BC	Emp NP	Delta BC	BTATII
	Scenario X: $\theta_0^+ = 0.756$, $\eta_0^+ = 0.857$											
(30, 30)	0.861	0.884	0.837	0.854	0.940	0.928	0.902	0.917	0.988	0.948	0.971	0.974
(50, 50)	0.882	0.903	0.836	0.833	0.936	0.950	0.900	0.904	0.988	0.988	0.969	0.968
(50, 75)	0.877	0.899	0.815	0.816	0.938	0.948	0.885	0.889	0.985	0.990	0.963	0.961
(75, 50)	0.872	0.891	0.833	0.834	0.930	0.942	0.905	0.904	0.985	0.990	0.971	0.969
(75, 75)	0.876	0.892	0.830	0.824	0.938	0.947	0.902	0.894	0.983	0.990	0.972	0.961
(100, 100)	0.877	0.897	0.824	0.814	0.933	0.953	0.895	0.889	0.986	0.991	0.969	0.961
(150, 150)		0.903		0.795		0.956		0.874		0.993		0.951
	Scenario XI: $\theta_0^+ = 0.833$, $\eta_0^+ = 0.888$											
(30, 30)	0.900	0.867	0.827	0.854	0.945	0.895	0.893	0.925	0.982	0.911	0.967	0.935
(50, 50)	0.890	0.901	0.836	0.859	0.945	0.947	0.899	0.913	0.991	0.980	0.968	0.959
(50, 75)	0.885	0.906	0.817	0.851	0.943	0.954	0.888	0.904	0.990	0.985	0.962	0.954
(75, 50)	0.876	0.896	0.829	0.858	0.931	0.943	0.896	0.912	0.985	0.987	0.967	0.961
(75, 75)	0.887	0.905	0.827	0.856	0.941	0.953	0.904	0.910	0.989	0.991	0.969	0.960
(100, 100)	0.881	0.900	0.829	0.858	0.941	0.953	0.898	0.910	0.988	0.991	0.969	0.964
(150, 150)		0.907		0.843		0.953		0.903		0.991		0.959
	Scenario XII: $\theta_0^+ = 0.895$, $\eta_0^+ = 0.852$											
(30, 30)	0.891	0.880	0.812	0.845	0.949	0.918	0.885	0.873	0.973	0.939	0.962	0.880
(50, 50)	0.884	0.896	0.805	0.855	0.940	0.944	0.878	0.890	0.991	0.983	0.962	0.922
(50, 75)	0.883	0.896	0.810	0.848	0.946	0.942	0.882	0.886	0.989	0.985	0.962	0.927
(75, 50)	0.890	0.897	0.792	0.869	0.945	0.947	0.869	0.904	0.991	0.986	0.952	0.933
(75, 75)	0.884	0.906	0.796	0.859	0.941	0.954	0.871	0.899	0.988	0.991	0.956	0.937
(100, 100)	0.887	0.906	0.802	0.862	0.947	0.952	0.877	0.902	0.991	0.990	0.960	0.946
(150, 150)		0.905		0.859		0.953		0.904		0.992		0.949

A look at these figures suggests that the distributions of W may be not normal and that Box-Cox transformations cannot be applied. Therefore, to construct confidence regions for optimal sensitivity and specificity, we use the fully non-parametric approach described in Section 3.3.

The sample sizes are $m = 210$ and $n = 350$. The desired confidence regions, of nominal level 0.90, 0.95 and 0.99, are shown in Figure 8. As one can see, the combination W has a good ability to discriminate between healthy and melanoma patients since optimal sensitivity and specificity always are above 0.89 with a confidence of 0.95. As discussed in Cesati et al. (2020), the use of cytokines/chemokines at the tissue level may have applications for diagnostic purposes and to predict response to therapy or prognosis. The results of our analysis seem to confirm those obtained by Cesati and colleagues: measuring the expression of selected genes may represent, soon, a quantitative approach for an automatic process useful to identify suspect samples.

6. Final remarks

In this article, we present a new approach for the construction of semi-parametric and non-parametric confidence regions for optimal sensitivity and specificity of a diagnostic test, according to the Youden index criterion. The proposed approach is based on empirical likelihood pivots and applies in the case of the binormal model, as well as in the case of the binormal model after Box-Cox transformations or in a completely non-parametric setting. The approach is new in itself and shows how to use, eventually, two different Box-Cox transformations for populations of healthy and diseased subjects. Unlike those obtained from other techniques proposed in the literature, the regions built with our method have no constraints on the shape (which depends only on the data) and are automatically range-respecting. The accuracy of the new confidence regions is evaluated through simulation experiments. The simulation results show that our methods perform at least as well as competitors present in the literature, showing themselves more accurate in situations where

Table 5
Simulation results for scenarios XIII and XIV: bi-beta settings. Emp BC is the proposed empirical likelihood method with two different Box-Cox transformations, Emp NP is the proposed empirical likelihood method with kernel estimation, Delta BC is the delta-based approach with common Box-Cox transformation, BTATII is the smoothed bootstrap with arcsin-square-root transformation.

Sample size	Nominal coverage: 0.90				Nominal coverage: 0.95				Nominal coverage: 0.99			
	Emp BC	Emp NP	Delta BC	BTATII	Emp BC	Emp NP	Delta BC	BTATII	Emp BC	Emp NP	Delta BC	BTATII
	Scenario XIII: $\theta_0^+ = 0.779$, $\eta_0^+ = 0.760$											
(30, 30)	0.832	0.896	0.614	0.830	0.914	0.947	0.702	0.892	0.983	0.973	0.824	0.953
(50, 50)	0.800	0.897	0.526	0.810	0.887	0.948	0.627	0.874	0.972	0.991	0.781	0.946
(50, 75)	0.772	0.897	0.586	0.790	0.872	0.949	0.684	0.858	0.972	0.988	0.829	0.933
(75, 50)	0.767	0.897	0.361	0.781	0.868	0.955	0.469	0.851	0.963	0.993	0.649	0.928
(75, 75)	0.736	0.900	0.389	0.783	0.849	0.953	0.501	0.857	0.961	0.992	0.696	0.939
(100, 100)	0.685	0.900	0.308	0.769	0.805	0.956	0.412	0.842	0.945	0.993	0.607	0.922
(150, 150)		0.903		0.744		0.958		0.821		0.992		0.907
	Scenario XIV: $\theta_0^+ = 0.803$, $\eta_0^+ = 0.806$											
(50, 50)	0.412	0.901	0.071	0.821	0.566	0.952	0.113	0.879	0.826	0.993	0.244	0.931
(50, 75)	0.325	0.902	0.094	0.828	0.474	0.952	0.151	0.880	0.759	0.992	0.291	0.931
(75, 50)	0.296	0.902	0.010	0.805	0.443	0.957	0.026	0.862	0.735	0.992	0.090	0.922
(75, 75)	0.212	0.907	0.015	0.816	0.342	0.960	0.031	0.872	0.644	0.994	0.097	0.930
(100, 100)	0.105	0.905	0.002	0.804	0.197	0.961	0.006	0.869	0.468	0.993	0.028	0.931
(150, 150)		0.905		0.777		0.958		0.849		0.994		0.921

Table 6
Simulation results for scenarios XV and XVI: mixture settings. Emp BC is the proposed empirical likelihood method with two different Box-Cox transformations, Emp NP is the proposed empirical likelihood method with kernel estimation, Delta BC is the delta-based approach with common Box-Cox transformation, BTATII is the smoothed bootstrap with arcsin-square-root transformation.

Sample size	Nominal coverage: 0.90				Nominal coverage: 0.95				Nominal coverage: 0.99			
	Emp BC	Emp NP	Delta BC	BTATII	Emp BC	Emp NP	Delta BC	BTATII	Emp BC	Emp NP	Delta BC	BTATII
	Scenario XV: $\theta_0^+ = 0.915$, $\eta_0^+ = 0.884$											
(50, 50)	0.823	0.913	0.870	0.897	0.921	0.948	0.931	0.928	0.977	0.966	0.981	0.931
(50, 75)	0.810	0.926	0.876	0.898	0.908	0.960	0.934	0.923	0.973	0.976	0.986	0.937
(75, 50)	0.801	0.906	0.808	0.895	0.898	0.953	0.883	0.927	0.973	0.986	0.968	0.951
(75, 75)	0.791	0.908	0.811	0.893	0.884	0.955	0.896	0.933	0.973	0.989	0.975	0.955
(100, 100)	0.764	0.918	0.764	0.895	0.850	0.961	0.863	0.933	0.963	0.993	0.964	0.964
(150, 150)		0.916		0.896		0.956		0.930		0.993		0.965
	Scenario XVI: $\theta_0^+ = 0.788$, $\eta_0^+ = 0.782$											
(30, 30)		0.879		0.696		0.929		0.772		0.960		0.886
(50, 50)		0.855		0.646		0.925		0.724		0.983		0.838
(50, 75)		0.828		0.613		0.909		0.690		0.978		0.803
(75, 50)		0.874		0.635		0.930		0.711		0.980		0.818
(75, 75)		0.874		0.635		0.930		0.711		0.980		0.818
(100, 100)		0.851		0.595		0.926		0.669		0.986		0.778
(150, 150)		0.864		0.560		0.929		0.639		0.985		0.747

the marker distributions have different shapes in the populations of healthy and diseased subjects (see Table 4). In the beta model, where the use of Box-Cox transformation is not appropriate, our non-parametric Emp NP approach appears to be the only one effective (see Table 5). The same happens in the case of mixture models (see Table 6).

Our approach involving a (double) Box-Cox transformation, is a semi-parametric approach. In principle, when properly applied, a semi-parametric technique is more efficient than a fully non-parametric one. Typically, the increased efficiency is relevant for inference with small samples. This is also reflected in our simulation results, where, when $n = m = 30$, confidence regions based on Emp BC are generally more accurate than those obtained by non-parametric

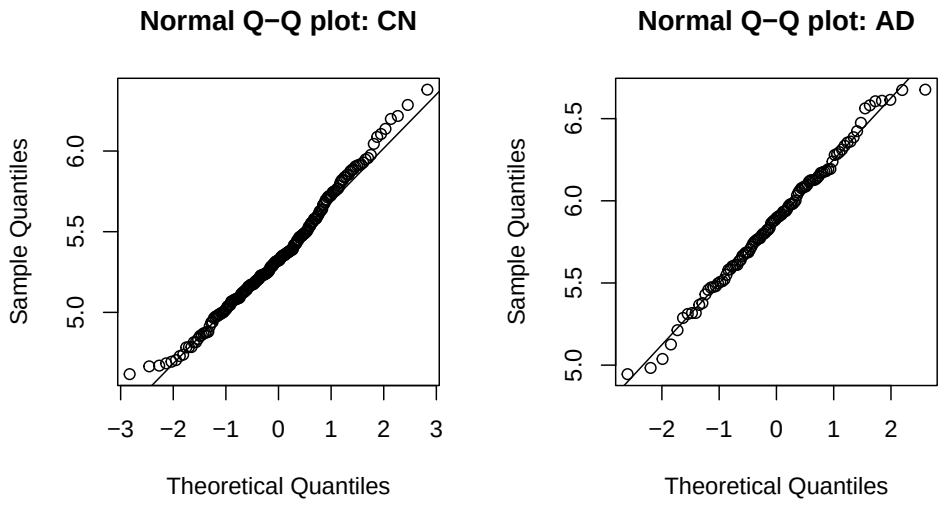


Figure 1: Q-Q nomral plots for log(Tau-protein).

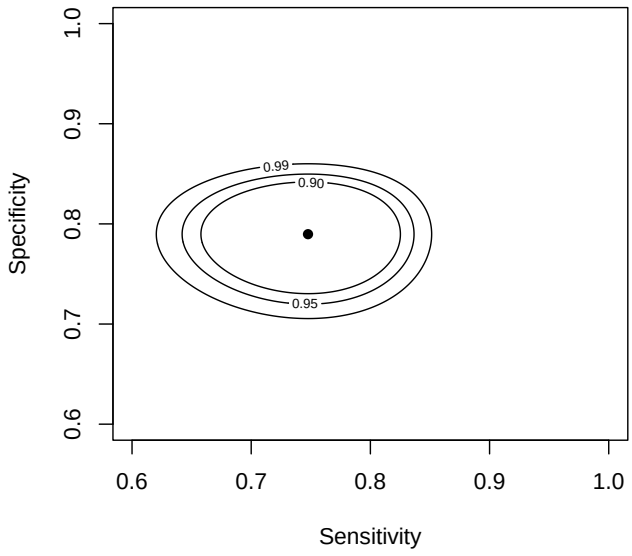
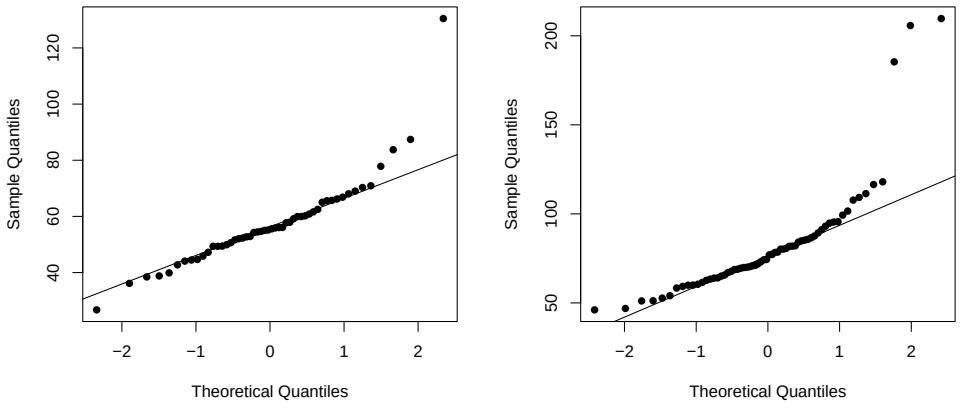


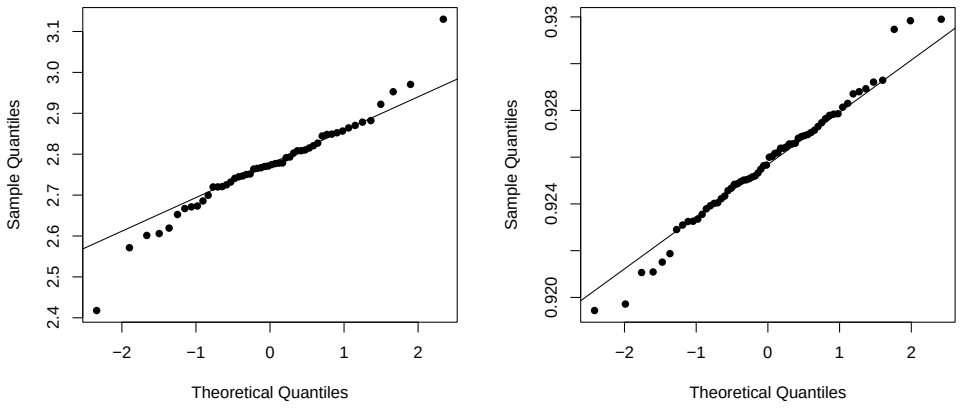
Figure 2: Confidence regions for the optimal sensitivity and specificity for the log of T-tau protein, with nominal coverage: 0.90, 0.95 and 0.99.

techniques; of course, when the double transformation is adequate (scenarios (X)-(XII)). To further highlight this aspect, we report below (Table 7) some simulation results concerning Emp BC and Emp NP (the best performing alternative in large samples), for even smaller sample sizes, which refer to scenario XI.

Therefore, although the double transformation violates the invariance property of the ROC curve (as pointed out also by a reviewer), for the achievement of the main goal of the paper, i.e., to build confidence regions for optimal sensitivity and specificity, when correctly applied, the Emp BC approach appears useful. As already mentioned, the double transformation leaves unchanged the optimal values of sensitivity and specificity, and the estimated optimal



(a) Healthy group: before Box-Cox (b) Cancer group: before Box-Cox



(c) Healthy group: after Box-Cox (d) Cancer group: after Box-Cox

Figure 3: Q-Q normal plot for the new biomarker T, for healthy and diseased patients before and after Box-Cox transformation.

Table 7
Scenario XI, small sample sizes: comparison between Emp BC and the best performing alternative in large samples.

Sample size	Nominal coverage: 0.90		Nominal coverage: 0.95		Nominal coverage: 0.99	
	Emp BC	Emp NP	Emp BC	Emp NP	Emp BC	Emp NP
(20, 20)	0.884	0.738	0.908	0.748	0.925	0.753
(25, 25)	0.884	0.816	0.934	0.837	0.961	0.848

cutpoint $\hat{\tau}^+$ is given in the original scale. Clearly, the user must take into account the "limit" of the procedure and use it only for the purpose for which it was designed.

In Section 1, we recall that Schaible and Yin (2021) recently proposed a technique for the construction of confidence regions for optimal predictive values of a biomarker, when the prevalence of the disease can be given as known. Our approach can be easily extended to achieve the same goal. Let pp and np be the positive and the negative predictive values for a biomarker, that is, the probability of disease under a positive test result, and the probability of no disease under a negative test result, respectively. Then, given the disease prevalence, say π, we have that

$$pp(\tau) = \frac{\theta(\tau)\pi}{\theta(\tau)\pi + (1-\eta(\tau))(1-\pi)} \quad \text{and} \quad np(\tau) = \frac{\eta(\tau)(1-\pi)}{(1-\theta(\tau))\pi + \eta(\tau)(1-\pi)}, \tag{2}$$

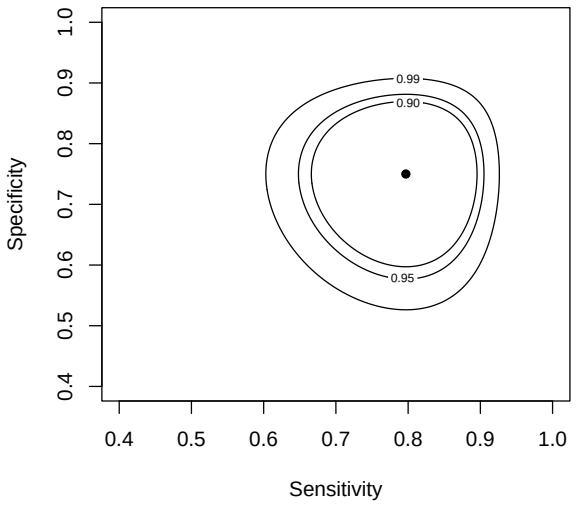


Figure 4: Confidence regions for the optimal sensitivity and specificity for the combination T = Glucose + Resistin - 1.5×BMI, with nominal coverage: 0.90, 0.95 and 0.99.

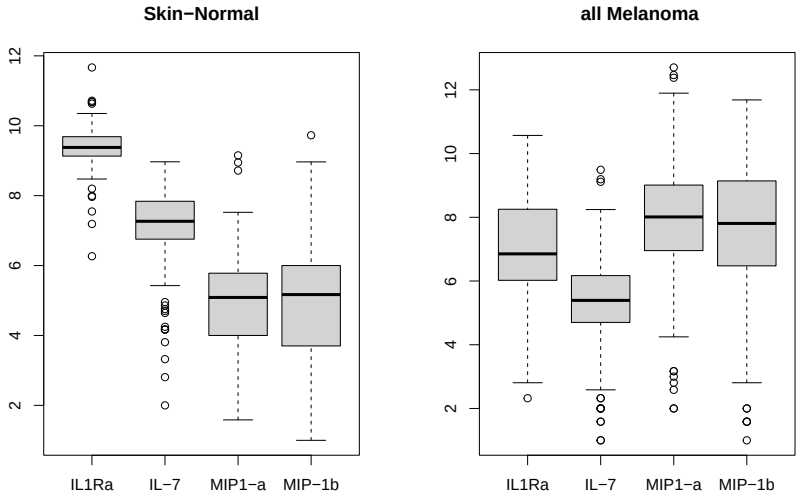


Figure 5: Boxplot for the distributions of the four cytokines in melanoma dataset.

at a fixed cutpoint τ. This means that the pair $(pp(\tau), np(\tau))$ is a one-to-one function of $(\theta(\tau), \eta(\tau))$. As a consequence, a confidence region for the optimal pair of predictive values, say (pp_0^+, np_0^+), can be obtained by simply mapping a region for (θ_0^+, η_0^+) with desired nominal coverage, constructed with one of our three proposed techniques. In other words, a confidence region for (pp_0^+, np_0^+), with nominal coverage $1-\alpha$, consists of all the pairs (pp^+, np^+) whose anti-image (θ^+, η^+), according to (2), belong to the set $\mathcal{R}_\alpha = \left\{(\theta^+, \eta^+) : \ell_*(\theta^+, \eta^+) \leq \hat{c}_\alpha\right\}$.

However, predictive values strictly depend on the disease prevalence π, which may significantly affect inference. To avoid problems, positive and negative likelihood ratios can be considered. The positive likelihood ratio of a diagnostic test is defined as the ratio between the probability of correctly classifying a diseased patient and the probability of incorrectly classifying a non-diseased patient when the result of the test is positive. The negative likelihood ratio is the

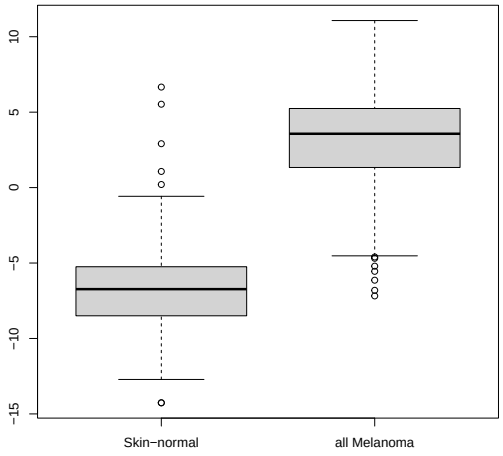


Figure 6: Boxplot for the distributions of the combination MIP1-a − IL1Ra − IL-7 + MIP-1b

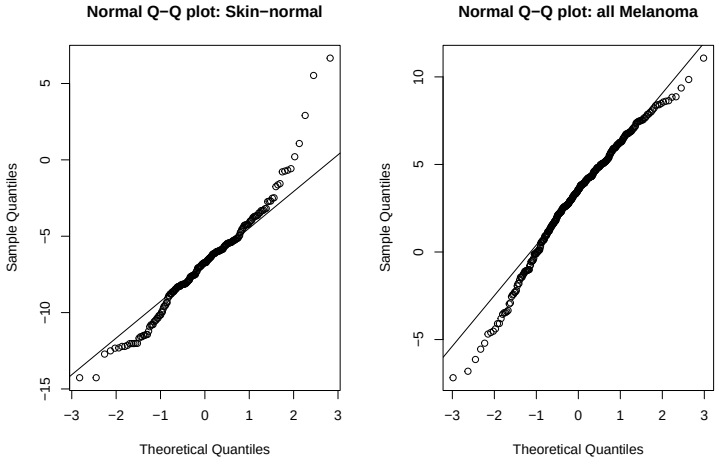


Figure 7: Q-Q normal plots for the combination MIP1-a − IL1Ra − IL-7 + MIP-1b.

ratio between the probability of incorrectly classifying a diseased patient and the probability of correctly classifying a non-diseased patient when the test result is negative. Then, likelihood ratios are functions of the test's sensitivity and specificity. More precisely, for the positive likelihood ratio (pl) and the negative likelihood ratio (nl), we have

$$pl(\tau) = \frac{\theta(\tau)}{1-\eta(\tau)}, \quad \text{and} \quad nl(\tau) = \frac{1-\theta(\tau)}{\eta(\tau)},$$

at a fixed cutpoint τ. Since, the pair $(pl(\tau), nl(\tau))$ is a one-to-one function of $(\theta(\tau), \eta(\tau))$, again a confidence region for the optimal pair of likelihood ratios, say (pl_0^+, nl_0^+), can be obtained by simply mapping a region for (θ_0^+, η_0^+) with desired nominal coverage, constructed with one of our three proposed techniques. Just as an example, Figure 9 shows confidence regions for the optimal pair of likelihood ratios, for the test W and data considered in Section 5.3.

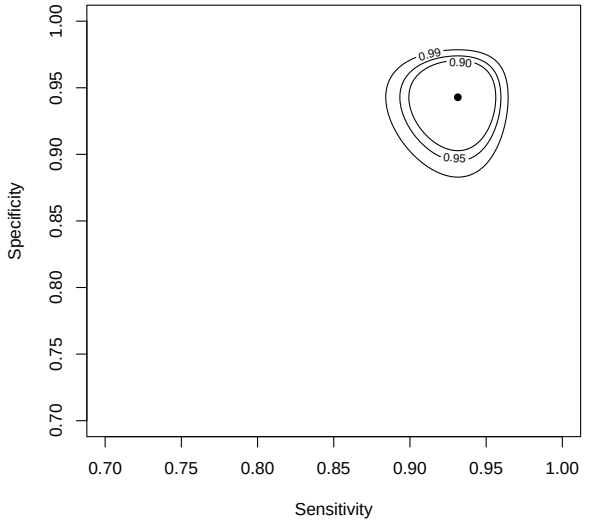


Figure 8: Confidence regions for the optimal sensitivity and specificity for the combination W = MIP1-a − IL1Ra − IL-7 + MIP-1b, with nominal coverage: 0.90, 0.95 and 0.99.

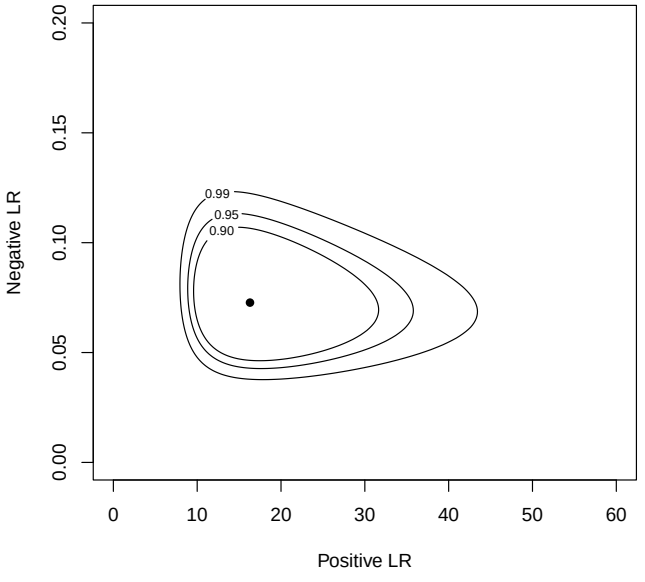


Figure 9: Confidence regions for the optimal likelihood ratios for the combination W = MIP1-a − IL1Ra − IL-7 + MIP-1b, with nominal coverage: 0.90, 0.95 and 0.99.

Acknowledgments

Data collection and sharing for the application in Section 5.1 was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers

Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (`www.fnih.org`). The grantee organisation is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the Ministero dell'Istruzione, dell'Università e della Ricerca-Italy (grant number DIFO_ECCELLENZA18_01).

The authors acknowledge an Associated Editor and two anonymous Reviewers whose valuable suggestions contributed to improve presentation of the contents.

Supplementary data

Supplemental material including R codes used in applications can be found online.

References

Adimari, G., Chiogna, M., 2010. Simple nonparametric confidence regions for the evaluation of continuous-scale diagnostic tests. Int J Biostat 6.

Azzalini, A., 1981. A note on the estimation of a distribution function and quantiles by a kernel method. Biometrika 68, 326–328.

Bantis, L.E., Nakas, C.T., Reiser, B., 2014. Construction of confidence regions in the ROC space after the estimation of the optimal youden index-based cut-off point. Biometrics 70, 212–223.

Box, G.E., Cox, D.R., 1964. An analysis of transformations. J R Stat Soc B 26, 211–243.

Cesati, M., Scatozza, F., D'arcangelo, D., Antonini-Cappellini, G.C., Rossi, S., Tabolacci, C., Nudo, M., Palese, E., Lembo, L., Di Lella, G., et al., 2020. Investigating serum and tissue expression identified a cytokine/chemokine signature as a highly effective melanoma marker. Cancers 12, 3680.

Fluss, R., Faraggi, D., Reiser, B., 2005. Estimation of the Youden Index and its associated cutoff point. Biometrical J 47, 458–472.

Owen, A.B., 2001. Empirical likelihood. Chapman and Hall/CRC.

Patrício, M., Pereira, J., Crisóstomo, J., Matafome, P., Gomes, M., Seiça, R., Caramelo, F., 2018. Using resistin, glucose, age and BMI to predict the presence of breast cancer. BMC Cancer 18, 1–8.

Pepe, M.S., 2003. The statistical evaluation of medical tests for classification and prediction. Oxford University Press.

Schaible, B.J., Yin, J., 2021. Joint confidence region estimation on predictive values. Pharm Stat 20, 1147–1167.

Yin, J., Tian, L., 2014. Joint inference about sensitivity and specificity at the optimal cut-off point associated with Youden index. Comput Stat Data An 77, 1–13.

Youden, W.J., 1950. Index for rating diagnostic tests. Cancer 3, 32–35.

Appendix

Let $\boldsymbol{p} = (p_1, \dots, p_m)$ and $\boldsymbol{q} = (q_1, \dots, q_n)$ denote probability vectors, with $p_i > 0$ and $q_j > 0$, representing multinomial distributions on the data points $\{x_1, \dots, x_m\}$ and $\{y_1, \dots, y_n\}$, respectively. The empirical likelihood for (θ, η), at fixed τ, is defined as

$$L(\theta, \eta, \tau) = \sup_{\boldsymbol{p},\boldsymbol{q}} \prod_{i=1}^{m} p_i \prod_{j=1}^{n} q_j \tag{3}$$

subject to the following constraints: $p_i > 0$, $q_j > 0$, $\sum_{i=1}^{m} p_i = 1$, $\sum_{j=1}^{n} q_j = 1$, $\sum_{i=1}^{m} p_i \mathrm{I}(x_i \le \tau) = \eta$ and $\sum_{j=1}^{n} q_j \mathrm{I}(\tau \le y_j) = 1 - \theta$, Under the constraints $\sum_{i=1}^{m} p_i = 1$ and $\sum_{j=1}^{n} q_j = 1$, the empirical likelihood $L(\theta, \eta, \tau)$ in (3) reaches its maximum $m^m n^n$, at $p_i = 1/m$ and $q_j = 1/n$ for all i, j. Thus, the empirical log-likelihood ratio for (θ, η)

is defined as

$$\ell(\theta,\eta,\tau) = \sup_{\boldsymbol{p},\boldsymbol{q}} \left\{ -2\sum_{i=1}^{m} \log\left(mp_i\right) - 2\sum_{j=1}^{n} \log\left(nq_j\right) \right\}, \tag{4}$$

subject to the constraints mentioned above. Since we have two independent samples, from (4), we can write:

$$\ell(\theta,\eta,\tau) = \ell_1(\eta,\tau) + \ell_2(\theta,\tau),$$

where

$$\begin{aligned}
\ell_1(\eta,\tau) &= \sup_{\boldsymbol{p}} \left\{ -2\sum_{i=1}^{m} \log\left(mp_i\right) \,:\, p_i > 0, \sum_{i=1}^{m} p_i = 1, \sum_{i=1}^{m} p_i \mathrm{I}(x_i \le \tau) = \eta \right\}, \\
\ell_2(\theta,\tau) &= \sup_{\boldsymbol{q}} \left\{ -2\sum_{j=1}^{n} \log\left(nq_j\right) \,:\, q_j > 0, \sum_{j=1}^{n} q_j = 1, \sum_{j=1}^{n} q_j \mathrm{I}(y_j \le \tau) = 1-\theta \right\}.
\end{aligned}$$

Thus, we can find the maximum of each component separately: $\ell_1(\eta,\tau)$ is the empirical log-likelihood ratio for the η-quantile of X, and $\ell_2(\theta,\tau)$ is the empirical log-likelihood ratio for the $(1-\theta)$-quantile of Y. By using the Lagrange multiplier method, we have

$$\ell_1(\eta,\tau) = 2m \left\{ \widehat{F}_X(\tau) \log \frac{\widehat{F}_X(\tau)}{\eta} + \left[1-\widehat{F}_X(\tau)\right] \log \frac{1-\widehat{F}_X(\tau)}{1-\eta} \right\}$$

if $\tau \in \left[x_{(1)}, x_{(m)}\right)$, otherwise $\ell_1(\eta,\tau) = +\infty$; and

$$\ell_2(\theta,\tau) = 2n \left\{ \widehat{F}_Y(\tau) \log \frac{\widehat{F}_Y(\tau)}{1-\theta} + \left[1-\widehat{F}_Y(\tau)\right] \log \frac{1-\widehat{F}_Y(\tau)}{\theta} \right\}$$

if $\tau \in \left[y_{(1)}, y_{(n)}\right)$, otherwise $\ell_2(\theta,\tau) = +\infty$. Here, $x_{(i)}, i = 1,\ldots,m$ and $y_{(j)}, j = 1,\ldots,n$, denote the order statistics from the samples. Therefore,

$$\begin{aligned}
\ell(\theta,\eta,\tau) &= 2m \left\{ \widehat{F}_X(\tau) \log \frac{\widehat{F}_X(\tau)}{\eta} + \left[1-\widehat{F}_X(\tau)\right] \log \frac{1-\widehat{F}_X(\tau)}{1-\eta} \right\} \\
&\quad + 2n \left\{ \widehat{F}_Y(\tau) \log \frac{\widehat{F}_Y(\tau)}{1-\theta} + \left[1-\widehat{F}_Y(\tau)\right] \log \frac{1-\widehat{F}_Y(\tau)}{\theta} \right\},
\end{aligned}$$

for $\theta \in (0,1)$, $\eta \in (0,1)$, $\tau \in \mathcal{T}$, where $\mathcal{T} = \left[\max\{x_{(1)}, y_{(1)}\}, \min\{x_{(m)}, y_{(n)}\}\right)$. When $\tau \notin \mathcal{T}$, then $\ell(\theta,\eta,\tau) = +\infty$.