Biometrics 70, 212–223
March 2014

DOI: 10.1111/biom.12107

Construction of Confidence Regions in the ROC Space after the Estimation of the Optimal Youden Index-Based Cut-Off Point

Leonidas E. Bantis,[1] Christos T. Nakas,[2,*] and Benjamin Reiser[3]

[1]Department of Statistics and Actuarial-Financial Mathematics, University of the Aegean, 83200 Samos, Greece
[2]Laboratory of Biometry, University of Thessaly, Phytokou Street, 38446 Volos, Greece
[3]Department of Statistics, University of Haifa, Haifa 31905, Israel
[*]*email:* cnakas@uth.gr

Summary. After establishing the utility of a continuous diagnostic marker investigators will typically address the question of determining a cut-off point which will be used for diagnostic purposes in clinical decision making. The most commonly used optimality criterion for cut-off point selection in the context of ROC curve analysis is the maximum of the Youden index. The pair of sensitivity and specificity proportions that correspond to the Youden index-based cut-off point characterize the performance of the diagnostic marker. Confidence intervals for sensitivity and specificity are routinely estimated based on the assumption that sensitivity and specificity are independent binomial proportions as they arise from the independent populations of diseased and healthy subjects, respectively. The Youden index-based cut-off point is estimated from the data and as such the resulting sensitivity and specificity proportions are in fact correlated. This correlation needs to be taken into account in order to calculate confidence intervals that result in the anticipated coverage. In this article we study parametric and non-parametric approaches for the construction of confidence intervals for the pair of sensitivity and specificity proportions that correspond to the Youden index-based optimal cut-off point. These approaches result in the anticipated coverage under different scenarios for the distributions of the healthy and diseased subjects. We find that a parametric approach based on a Box–Cox transformation to normality often works well. For biomarkers following more complex distributions a non-parametric procedure using logspline density estimation can be used.

Key words: Delta method; Logspline; ROC curve; Sensitivity; Specificity; Youden index.

1. Introduction

Diagnostic markers are essential in clinical practice for the assessment of the discrimination between classes in different stages of disease. The utility of a diagnostic marker is routinely checked using receiver operating characteristic (ROC) methodology. In the simple two-class problem the ROC curve is constructed and the area under the ROC curve (AUC) is often used for marker assessment. Suppose that, for a specific diagnostic marker with continuous measurements where the discrimination between non-diseased and diseased subjects is of interest, $X \sim F_X$ are the measurements of non-diseased subjects and $Y \sim F_Y$ are the measurements of diseased subjects. Without loss of generality, we assume that larger values of the marker are more indicative of disease. Using a cut-off point c to decide that a subject is non-diseased when a measurement is less than c and that the subject is diseased otherwise, the specificity of the marker is $spec(c) = P[X \leq c]$ and the sensitivity of the marker is $sens(c) = P[Y > c]$. Then, the ROC curve is constructed by plotting the points $(1 - spec(c), sens(c)), c \in (-\infty, \infty)$ in the ROC curve space. The ROC curve space corresponds to the unit square. A screening test with high discriminative power will produce an ROC curve close to the point with coordinates $(0, 1)$ in the unit square. An equivalent construction of the ROC curve is produced by plotting $(spec(c), sens(c)), c \in (-\infty, \infty)$ in the unit square. The area under the ROC curve (AUC) is equal to $P[X < Y]$ (Pepe, 2003, p.78).

Once the diagnostic accuracy of the marker is established, the selection of an optimal cut-off point is needed which will be used by practitioners for screening purposes. The cut-off point must be selected based on an optimality criterion. The maximum of the Youden index is quite often used in practice (e.g., Rhemrev et al., 2010; Chen et al., 2011) because of its simplicity and its interpretation as the accuracy of a diagnostic marker by clinicians. It is defined as $J = max_c\{sens(c) + spec(c) - 1\} = max_c\{F_X(c) - F_Y(c)\}$ and it is the maximum distance from the ROC curve to the main diagonal ($sens = 1 - spec$). The relative importance of sensitivity and specificity for any given problem can be reflected in the choice of the optimal cut-off point by introducing weights, ν and μ, in the previous definition as follows: $J^* = max_c\{\nu \cdot sens(c) + \mu \cdot spec(c) - 1\}$. These weights are often difficult to assess (Greiner, Pfeiffer, and Smith, 2000, and references cited therein). J^* is commonly referred to as the generalized Youden index. Use of the Youden and generalized Youden index approach has been examined in a number of articles (e.g., Fluss, Faraggi, and Reiser, 2005; Schisterman and Perkins, 2007; Skaltsa, Jover, and Carrasco, 2010). Zou et al. (2013) discuss these two approaches along with other possible metrics for obtaining an optimal cut-off point.

The resulting optimal cut-off point corresponds to a pair of sensitivity and specificity values that characterize the diagnostic marker under study. Confidence intervals for the specificity (or 1-specificity) and sensitivity pair are routinely

calculated based on classic binomial distribution formulas (Pepe, 2003). However, when the optimal cut-off point is estimated from the data, rather than considered fixed, the corresponding specificity and sensitivity proportions are correlated and their variability changes accordingly. This correlation needs to be taken into account in order to construct confidence intervals with the correct coverage. The development of confidence intervals for sensitivity and specificity computed at the Youden Index based optimal cut-off point has not been addressed in the literature. McClish (2012) has recently emphasized the importance of examining the ROC curves of biomarkers in a region around the optimal point.

In this article, we study possible approaches for the construction of confidence regions in the ROC space when the optimal cut-off point is estimated from the data based on the Youden index. We illustrate that the standard approach does not provide satisfactory results. In Section 2, we use the delta method based on a parametric approach for the construction of these confidence regions and compare with a bootstrap approach and a non-parametric approach that uses spline methodology. A large simulation study is conducted in Section 3. Methods are applied to a dataset from a study on a brucellosis diagnostic marker in Section 4 and to a standard reference dataset of a pancreatic marker study that has been used extensively in the two-class ROC literature (Wieand et al., 1989). We conclude with a discussion.

2. Construction of Confidence Regions in the ROC Space

Suppose that a diagnostic marker is administered to n_X non-diseased subjects and to n_Y diseased subjects. The ROC curve and related statistics of interest can be calculated non-parametrically based on empirical estimates or parametrically, for example, based on normality assumptions. In the following subsections, the classical approach for the construction of confidence regions in the ROC space is reviewed assuming a known cut-off point. When the cut-off point is chosen via the Youden index and estimated from the data, a parametric approach based on delta method is proposed. Based on the distributional properties of the data, other alternatives are also considered.

2.1. *The Classical Approach for Given Cut-Off Point*

In the general case, data from diseased and healthy subjects are independent so that the confidence interval for specificity is independent of that for sensitivity. Therefore, the empirical estimators of sensitivity and specificity are proportions based on binomial distributions. For illustrative purposes and for the depiction of the results based on the standard definition of the ROC curve, 1-specificity will be used instead of specificity. Then for a given cut-off point c, $\hat{sens}(c) = \frac{\sum_{i=1}^{n_Y} I(Y_i>c)}{n_Y}$ and $1-\hat{spec}(c) = \frac{\sum_{i=1}^{n_X} I(X_i>c)}{n_X}$, where $I(\cdot)$ is the indicator function. If $(1-spec_l, 1-spec_u)$ and $(sens_l, sens_u)$ are 97.5% univariate confidence intervals for $1-spec(c)$ and $sens(c)$, respectively, then the rectangle $(1-spec_l, 1-spec_u)\times(sens_l, sens_u)$ is a 95% rectangular confidence region for $(1-spec(c), sens(c))$, where $sens_l = \hat{sens}(c) - 2.24\cdot\sqrt{\frac{\hat{sens}(c)\cdot(1-\hat{sens}(c))}{n_Y}}$, $sens_u = \hat{sens}(c) + 2.24\cdot\sqrt{\frac{\hat{sens}(c)\cdot(1-\hat{sens}(c))}{n_Y}}$ and similarly for $1-spec_l$ and $1-spec_u$ (Pepe, 2003, Section 2.2).

The described approach yields rectangular confidence regions that are easily communicated to and understood by clinicians. If the normal approximation to the binomial proportions is not appropriate for a specific application, the corrected Wald method that employs a logit or a probit transformation can be employed. This approach is described in Zou et al. (2012), Section 3.6. Elliptical confidence regions for $(1-spec(c), sens(c))$ can also be constructed in this way. Namely, a joint 95% elliptical confidence region is defined by

$$\begin{bmatrix} 1-spec-(1-\hat{spec}(c)) \\ sens-\hat{sens}(c) \end{bmatrix} \cdot\mathbf{V}^{-1}\cdot \begin{bmatrix} 1-spec-(1-\hat{spec}(c)) \\ sens-\hat{sens}(c) \end{bmatrix} = q^2,$$

where q^2 is the 95% percentile of a χ^2_2 distribution and $\mathbf{V}$ is a diagonal 2×2 matrix with $(\frac{\hat{spec}(c)\cdot(1-\hat{spec}(c))}{n_X}, \frac{\hat{sens}(c)\cdot(1-\hat{sens}(c))}{n_Y})$ in the main diagonal (Kosinski, Chen, and Lyles, 2010).

2.2. *An Approach Based on the Delta Method When the Cut-Off Point is Estimated from the Data*

Under the assumption that measurements for non-diseased and diseased subjects follow normal distributions $N(\mu_X, \sigma_X^2)$ and $N(\mu_Y, \sigma_Y^2)$, respectively, it follows that, $sens(c) = \Phi(\frac{\mu_Y - c}{\sigma_Y}) = \Phi(\delta_e)$ and $spec(c) = \Phi(\frac{c-\mu_X}{\sigma_X}) = \Phi(\delta_p)$ (Pepe, 2003, p. 82). Then, $1-spec(c) = \Phi(\frac{\mu_X - c}{\sigma_X})$ and

$$ROC(t) = \Phi\left(\frac{\mu_Y - \mu_X + \sigma_X\Phi^{-1}(t)}{\sigma_Y}\right),\ t\in[0,1].$$

Following Fluss et al. (2005), the maximum of the Youden index is defined as,

$$J = max_c\left\{\Phi\left(\frac{c-\mu_X}{\sigma_X}\right) + \Phi\left(\frac{\mu_Y - c}{\sigma_Y}\right) - 1\right\}$$

and the corresponding optimal cut-off point, c^*, has the following closed-form expression,

$$c^* = \frac{\sigma_X^2\mu_Y - \sigma_Y^2\mu_X - \sigma_X\sigma_Y\sqrt{(\mu_X-\mu_Y)^2 + (\sigma_X^2-\sigma_Y^2)\log\left(\frac{\sigma_X^2}{\sigma_Y^2}\right)}}{\sigma_X^2-\sigma_Y^2},$$

if $\sigma_X \neq \sigma_Y$.

Otherwise, $c^* = \frac{\mu_X+\mu_Y}{2}$. J, c^* and other parameters of interest are estimated by substituting for the unknown μ_X, μ_Y, σ_X, σ_Y in the above formulae their sample means and standard deviations. The resulting estimators are denoted with $\hat{\mu}_X$, $\hat{\mu}_Y$, $\hat{\sigma}_X$, $\hat{\sigma}_Y$. Sensitivity and specificity are proportions and thus they are bounded between zero and one. As a result, the normal approximation for the construction of confidence intervals described in the classical approach can be inadequate for small samples and may also result in intervals which exceed the bounds. To obtain a $(1-\alpha)$% confidence interval for $sens(c^*)$ we apply standard normal asymptotic theory on $\hat{\delta}_e = \frac{\hat{\mu}_Y - c^*}{\hat{\sigma}_y}$,

which is not bounded, and use $\Phi(\hat{\delta}_e \pm z_{1-\alpha/2}\sqrt{Var(\hat{\delta}_e)})$ and similarly for $1-spec(c^*)$. Since $\hat{\mu}_X$, $\hat{\mu}_Y$, $\hat{\sigma}_X$, $\hat{\sigma}_Y$ are all independent, using the delta method (van der Vaart, 2000) we obtain,

$$Var(\hat{\delta}_e) \approx \left(\frac{\partial\hat{\delta}_e}{\partial\mu_X}\right)^2 Var(\hat{\mu}_X) + \left(\frac{\partial\hat{\delta}_e}{\partial\sigma_X}\right)^2 Var(\hat{\sigma}_X) + \left(\frac{\partial\hat{\delta}_e}{\partial\mu_Y}\right)^2 Var(\hat{\mu}_Y) + \left(\frac{\partial\hat{\delta}_e}{\partial\sigma_Y}\right)^2 Var(\hat{\sigma}_Y).$$

Similarly for $Var(\hat{\delta}_p)$. Numerical estimates can be obtained based on $Var(\hat{\mu}_X) = \frac{\sigma_X^2}{n_X}$, $Var(\hat{\mu}_Y) = \frac{\sigma_Y^2}{n_Y}$, $Var(\hat{\sigma}_X) = \frac{\sigma_X^2}{2(n_X-1)}$, $Var(\hat{\sigma}_Y) = \frac{\sigma_Y^2}{2(n_Y-1)}$ as given in Schisterman and Perkins (2007). The Bonferroni adjustment implies that if $(1-spec_l(c^*), 1-spec_u(c^*))$ and $(sens_l(c^*), sens_u(c^*))$ are 97.5% univariate confidence intervals for $1-spec(c^*)$ and $sens(c^*)$, respectively, then the rectangle $(1-spec_l(c^*), 1-spec_u(c^*)) \times (sens_l(c^*), sens_u(c^*))$ is a 95% rectangular confidence region for $(1-spec(c^*), sens(c^*))$, where $sens_l(c^*) = \Phi(\hat{\delta}_e - 2.24\cdot\sqrt{Var(\hat{\delta}_e)})$, $sens_u(c^*) = \Phi(\hat{\delta}_e + 2.24\cdot\sqrt{Var(\hat{\delta}_e)})$ and similarly for $1-spec_l(c^*)$ and $1-spec_u(c^*)$, respectively. Estimation of $Var(\hat{c}^*)$ and $Var(\hat{J})$ using the delta method as described here has been developed in Schisterman and Perkins (2007) and was used to obtain approximate confidence intervals for c^* and J. In order to take into account the correlation between $\hat{\delta}_e$ and $\hat{\delta}_p$ one can construct elliptical confidence regions based on bivariate normality properties. The covariance between $\hat{\delta}_e$ an $\hat{\delta}_p$ can be readily obtained as

$$Cov(\hat{\delta}_e, \hat{\delta}_p) \approx (\frac{\partial\hat{\delta}_e}{\partial\mu_X})(\frac{\partial\hat{\delta}_p}{\partial\mu_X})Var(\hat{\mu}_X) + (\frac{\partial\hat{\delta}_e}{\partial\sigma_X})(\frac{\partial\hat{\delta}_p}{\partial\sigma_X})Var(\hat{\sigma}_X) + (\frac{\partial\hat{\delta}_e}{\partial\mu_Y})(\frac{\partial\hat{\delta}_p}{\partial\mu_Y})Var(\hat{\mu}_Y) + (\frac{\partial\hat{\delta}_e}{\partial\sigma_Y})(\frac{\partial\hat{\delta}_p}{\partial\sigma_Y})Var(\hat{\sigma}_Y).$$

Let

$$\hat{\Sigma} = \begin{pmatrix} Var(\hat{\delta}_e) & Cov(\hat{\delta}_e, \hat{\delta}_p) \\ Cov(\hat{\delta}_e, \hat{\delta}_p) & Var(\hat{\delta}_p) \end{pmatrix}.$$

The ellipse defined by $(\mathbf{x}-\mathbf{a})^t\hat{\Sigma}^{-1}(\mathbf{x}-\mathbf{a}) = q^2$, where $\mathbf{a} = (\hat{\delta}_e, \hat{\delta}_p)$ and q^2 is the 95% percentile of a χ_2^2, is an approximate 95% confidence region for (δ_e, δ_p). Formulas for the partial derivatives that appear in the expressions of $Var(\hat{\delta}_p)$, $Var(\hat{\delta}_e)$ and $Cov(\hat{\delta}_e, \hat{\delta}_p)$ are given in Web Appendix A. The elliptical confidence regions are expected to perform better than the rectangular ones, especially if the correlation between $\hat{\delta}_e$ and $\hat{\delta}_p$ is relatively high. Once we transform back to the ROC space (i.e., to the $(1-spec, sens)$ space), we obtain a more irregular shape which will be called an egg-shaped region in the sequel. It represents a 95% confidence region for $(1-spec(c^*), sens(c^*))$.

2.3. *Bootstrap Alternative*

The proposed delta method-based approach results in cumbersome closed form expressions in the formulas for $Var(\hat{\delta}_p)$, $Var(\hat{\delta}_e)$ and $Cov(\hat{\delta}_e, \hat{\delta}_p)$. Alternatively, a standard bootstrap approach can be used. The bootstrap approach may also have better small sample properties than the delta-based approach, since the delta-based approximations of $Var(\hat{\delta}_p)$, $Var(\hat{\delta}_e)$ and $Cov(\hat{\delta}_e, \hat{\delta}_p)$ are expected to perform well when the variances of $\hat{\mu}_X$, $\hat{\mu}_Y$, $\hat{\sigma}_X$, $\hat{\sigma}_Y$ are in fact small. Specifically, X and Y are sampled with replacement from the respective empirical distributions and δ_e and δ_p are estimated for each bootstrap sample assuming normality. Based on the bootstrap pairs of the estimated δ_e and δ_p, we evaluate an estimate of the corresponding covariance matrix, namely $\hat{\Sigma}_{(boots)}$. We thus use the bootstrap not for the estimation of the sampling distributions of the deltas but rather only for the estimation of the covariance matrix needed for the construction of the confidence regions. The proposed bootstrap-based 95% confidence regions for $(1-spec(c^*), sens(c^*))$ are defined by transforming the ellipse $(\mathbf{x}-\mathbf{a})^t\hat{\Sigma}_{(boots)}^{-1}(\mathbf{x}-\mathbf{a}) = q^2$ back to the ROC space, where $\mathbf{a} = (\hat{\delta}_e, \hat{\delta}_p)$ and q^2 is the 95% percentile of a χ_2^2. That is, for the derivation of the corresponding egg-shaped and rectangular bootstrap-based confidence regions we simply replace $\hat{\Sigma}$ with $\hat{\Sigma}_{(boots)}$ and follow the procedure described in Section 2.2.

2.4. *Box–Cox Approach for Non-Normal Data*

The normality assumption is quite restrictive and can lead to false results when it is significantly violated. However, the use of the binormal model, which assumes that a monotone transformation makes both distributions marginally normal, is a popular method of extending the parametric approach (see, e.g., Zou et al., 2012, Section 3.5) and has been shown to be robust to departures from normality (Hanley, 1996). The Box–Cox transformation as a way of modelling the monotone transformation in the binormal model has been shown to perform very well in the ROC context (e.g., Fluss et al., 2005; Zou et al., 2012, Chapter 3). The Box–Cox approach for ROC curves has been introduced in Zou and Hall (2000). Define

$$X^{(\lambda)} = \begin{cases} \frac{X^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log(X), & \lambda = 0. \end{cases}$$

$Y^{(\lambda)}$ can be similarly defined. Assume that both $X^{(\lambda)}$ and $Y^{(\lambda)}$ are independently normally distributed with parameters $\mu_{X^{(\lambda)}}$,$\sigma_{X^{(\lambda)}}$ and $\mu_{Y^{(\lambda)}}$, $\sigma_{Y^{(\lambda)}}$, respectively. The common λ can be estimated by maximizing the following profile log-likelihood:

$$l(\lambda) = -\frac{n_X}{2}\log\left[\frac{\Sigma_{i=1}^{n_X}\left(X_i^{(\lambda)} - \frac{\Sigma_{i=1}^{n_X}X_i^{(\lambda)}}{n_X}\right)^2}{n_X}\right] - \frac{n_Y}{2}\log\left[\frac{\Sigma_{j=1}^{n_Y}\left(Y_j^{(\lambda)} - \frac{\Sigma_{j=1}^{n_Y}Y_j^{(\lambda)}}{n_Y}\right)^2}{n_Y}\right] + (\lambda-1)\left(\Sigma_{i=1}^{n_X}\log X_i + \Sigma_{j=1}^{n_Y}\log Y_j\right) + k \qquad (1)$$

where k is constant. The full log-likelihood can be written as in Hernandez and Johnson (1980) and reduces to the profile likelihood given in (1), by simply plugging in the maximum likelihood estimates of the means and variances. For a given application, the parameter λ is estimated by the available data. Note that λ is a parameter in the likelihood function and thus affects the information matrix which in this case is not diagonal. Hence, the variability of the estimate of λ must be taken into account during the construction of confidence regions when the Box–Cox transformation is employed. The corresponding information matrix is of the form (details are shown in Web Appendix B):

$$\mathbf{I} = -\begin{pmatrix} \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \mu^2_{X^{(\lambda)}}} & 0 & 0 & 0 & \frac{\partial l(\boldsymbol{\theta})}{\partial \mu_{X^{(\lambda)}} \partial \lambda} \\ 0 & \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \sigma^2_{X^{(\lambda)}}} & 0 & 0 & \frac{\partial l(\boldsymbol{\theta})}{\partial \sigma_{X^{(\lambda)}} \partial \lambda} \\ 0 & 0 & \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \mu^2_{Y^{(\lambda)}}} & 0 & \frac{\partial l(\boldsymbol{\theta})}{\partial \mu_{Y^{(\lambda)}} \partial \lambda} \\ 0 & 0 & 0 & \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \sigma^2_{Y^{(\lambda)}}} & \frac{\partial l(\boldsymbol{\theta})}{\partial \sigma_{Y^{(\lambda)}} \partial \lambda} \\ \frac{\partial l(\boldsymbol{\theta})}{\partial \mu_{X^{(\lambda)}} \partial \lambda} & \frac{\partial l(\boldsymbol{\theta})}{\partial \sigma_{X^{(\lambda)}} \partial \lambda} & \frac{\partial l(\boldsymbol{\theta})}{\partial \mu_{Y^{(\lambda)}} \partial \lambda} & \frac{\partial l(\boldsymbol{\theta})}{\partial \sigma_{Y^{(\lambda)}} \partial \lambda} & \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \lambda^2} \end{pmatrix},$$

where $\boldsymbol{\theta} = (\mu_{X^{(\lambda)}}, \sigma_{X^{(\lambda)}}, \mu_{Y^{(\lambda)}}, \sigma_{Y^{(\lambda)}}, \lambda)$ and $l(\boldsymbol{\theta})$ is the full log-likelihood as presented in Web Appendix B. The information matrix is evaluated at $\hat{\theta}$, the maximum likelihood estimate of θ. An estimate of the corresponding covariance matrix can thus be derived by inversion of the above observed information matrix. Observe that the obtained covariance matrix will not be diagonal since the estimates of $\mu_{X^{(\lambda)}}$, $\sigma_{X^{(\lambda)}}$, $\mu_{Y^{(\lambda)}}$, and $\sigma_{Y^{(\lambda)}}$ are not independent. We then simply use the upper left 4×4 part of this covariance matrix denoting it by $\hat{\Sigma}^{(\lambda)}$. The covariance of the estimated δ_e and δ_p based on the transformed samples, that is $\hat{\delta}_e^{(\lambda)}$ and $\hat{\delta}_p^{(\lambda)}$, respectively, can be obtained by:

$$Cov(\hat{\delta}_e^{(\lambda)}, \hat{\delta}_p^{(\lambda)}) \approx \left(\frac{\partial \hat{\delta}_e^{(\lambda)}}{\partial \mu_{X^{(\lambda)}}}, \frac{\partial \hat{\delta}_e^{(\lambda)}}{\partial \sigma_{X^{(\lambda)}}}, \frac{\partial \hat{\delta}_e^{(\lambda)}}{\partial \mu_{Y^{(\lambda)}}}, \frac{\partial \hat{\delta}_e^{(\lambda)}}{\partial \sigma_{Y^{(\lambda)}}} \right) \hat{\Sigma}^{(\lambda)} \left(\frac{\partial \hat{\delta}_p^{(\lambda)}}{\partial \mu_{X^{(\lambda)}}}, \frac{\partial \hat{\delta}_p^{(\lambda)}}{\partial \sigma_{X^{(\lambda)}}}, \frac{\partial \hat{\delta}_p^{(\lambda)}}{\partial \mu_{Y^{(\lambda)}}}, \frac{\partial \hat{\delta}_p^{(\lambda)}}{\partial \sigma_{Y^{(\lambda)}}} \right)'. \quad (2)$$

and the corresponding variance of $\hat{\delta}_e$ is given by

$$Var(\hat{\delta}_e^{(\lambda)}) \approx \left(\frac{\partial \hat{\delta}_e^{(\lambda)}}{\partial \mu_X^{(\lambda)}}, \frac{\partial \hat{\delta}_e^{(\lambda)}}{\partial \sigma_X^{(\lambda)}}, \frac{\partial \hat{\delta}_e^{(\lambda)}}{\partial \mu_Y^{(\lambda)}}, \frac{\partial \hat{\delta}_e^{(\lambda)}}{\partial \sigma_Y^{(\lambda)}} \right) \hat{\Sigma}^{(\lambda)} \left(\frac{\partial \hat{\delta}_e^{(\lambda)}}{\partial \mu_X^{(\lambda)}}, \frac{\partial \hat{\delta}_e^{(\lambda)}}{\partial \sigma_X^{(\lambda)}}, \frac{\partial \hat{\delta}_e^{(\lambda)}}{\partial \mu_Y^{(\lambda)}}, \frac{\partial \hat{\delta}_e^{(\lambda)}}{\partial \sigma_Y^{(\lambda)}} \right)'.$$

The expression for $Var(\hat{\delta}_p)$ is similarly obtained. Based on the estimates of $Cov(\hat{\delta}_e^{(\lambda)}, \hat{\delta}_p^{(\lambda)})$, $Var(\hat{\delta}_e^{(\lambda)})$, $Var(\hat{\delta}_p^{(\lambda)})$ we can derive the rectangular and egg-shaped confidence regions as described in Section 2.2. For the bootstrap alternative, since the variability of $\hat{\lambda}$ must be taken into account, we sample X and Y with replacement and perform the Box–Cox transformation by maximizing the profile likelihood given in (1) for each bootstrap sample. Some authors (Molodianovitch, Faraggi, and Reiser, 2006; Schisterman, Reiser, and Faraggi, 2006; Molanes-Lopez and Leton, 2011) did not take the variability due to estimating the Box–Cox transformation into account when studying certain aspects of ROC curve analysis and found by simulation that their results were satisfactory. We found this approach to be very unsatisfactory for our problem and thus examined the effect of estimating λ.

2.5. *A Spline-Based Non-Parametric Approach*

Sometimes in practice, common parametric assumptions may not be justified by the available data even after applying the Box–Cox transformation. In such situations, empirical approaches are preferable. A straightforward empirical approach would be to implement the bootstrap procedure to the empirical ROC curve and work with the logit transformation of the estimated sensitivity and specificity that correspond to the optimal cut-off point (in order to improve the use of the normal approximation). Such an approach is summarized in the following simple steps:

- Step 1: Sample with replacement from X and Y and construct the empirical ROC curve. Obtain the optimal estimated cutoff, $\hat{c}^*$, as well as the corresponding pair of $\widehat{sens}(\hat{c}^*)$ and $\widehat{spec}(\hat{c}^*)$.
- Step 2: Use the logit transformation to obtain $logit_{\hat{se}^*} = log(\widehat{sens}(\hat{c}^*)/(1 - \widehat{sens}(\hat{c}^*))$ and $logit_{\hat{sp}^*} = log(\widehat{spec}(\hat{c}^*)/(1 - \widehat{spec}(\hat{c}^*))$.
- Step 3: Repeat steps 1 and 2 m times. Then, based on the m bootstrapped values of $logit_{\hat{se}^*}$ and $logit_{\hat{sp}^*}$ derive the bootstrap estimates of $Var(logit_{\hat{se}^*})$, $Var(logit_{\hat{sp}^*})$, and $Cov(logit_{\hat{se}^*}, logit_{\hat{sp}^*})$.
- Step 4: Construct the rectangular and elliptical confidence regions for $(logit_{se^*}, logit_{sp^*})$ in the logit space, while assuming $(logit_{\hat{se}^*}, logit_{\hat{sp}^*})$ is approximately distributed as a bivariate normal.
- Step 5: Construct the rectangular and egg-shaped confidence regions by transforming back to the *ROC* space.

However, for small to moderate sample sizes, bootstrapped $logit_{\hat{se}^*}$ and $logit_{\hat{sp}^*}$ take on values from a quite small set of possible values. This happens because the optimal cut-off point, especially for large values of J, tends to be the exact same point for a large number of bootstrap samples. As a result, the obtained bootstrap estimates of $Var(logit_{\hat{se}^*})$, $Var(logit_{\hat{sp}^*})$, and $Cov(logit_{\hat{se}^*}, logit_{\hat{sp}^*})$ are not valid, since the normality of $logit_{\hat{se}^*}$ and $logit_{\hat{sp}^*}$ completely collapses. Such a fully non-parametric approach based on empirical estimates does not result in valid estimates for $\hat{\Sigma}$ (for small to moderate sample sizes). We thus selected to work with a smooth estimate of the ROC curve. We chose a spline-based approach, namely the logspline approach, initially introduced by Kooperberg and Stone (1992). The logspline approach mainly focuses on density estimation and has been shown to outperform competing approaches for heavy-tailed and mixture distributions (Takada, 2008). We thus resort to this approach when the normality assumption is not valid, even after applying the Box–Cox transformation. We briefly outline the methodology:

Let the integer $K \geq 3$, and the knot sequence $\tau_1, \ldots, \tau_K$ with $-\infty \leq L < \tau_1 < \ldots < \tau_K \leq U \leq \infty$, where L and U are numbers that define the support of the marker measurements. Knot selection is discussed in Kooperberg and Stone (2004). The logspline density model is

$$f(x;\boldsymbol{\theta}) = exp(\theta_1 B_1(x) + \cdots + \theta_p B_p(x) - \mathcal{C}(\boldsymbol{\theta})), \;\; L < x < U, \tag{3}$$

where

$$\mathcal{C}(\boldsymbol{\theta}) = log\left(\int_L^U exp(\theta_1 B_1(x) + \cdots + \theta_p B_p(x))\mathrm{d}x\right)$$

is the normalizing constant, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ are the unknown spline parameters to be estimated and $B_1(x)$, $B_2(x)$,..., $B_p(x)$ are the spline basis functions. These can appropriately be defined in order to form a natural cubic spline for which the condition

$$\int_L^U exp(\theta_1 B_1(x) + \cdots + \theta_p B_p(x))\mathrm{d}x < \infty$$

is satisfied. The corresponding logspline survival function is then given by

$$S(x;\boldsymbol{\theta}) = 1 - \int_L^x f(z,\boldsymbol{\theta})\mathrm{d}z, \quad L < x < U.$$

The fitting procedure involves the maximization of the corresponding likelihood. By fitting the logspline model separately for the measurements of diseased and healthy subjects we obtain the corresponding logspline estimates of the underlying survival functions, namely $\hat{S}_X(\cdot)$ and $\hat{S}_Y(\cdot)$. The smooth logspline-based ROC curve is the plot defined by:

$$\hat{ROC}(\cdot) = \{(\hat{S}_X(c), \hat{S}_Y(c)), -\infty < c < \infty\}.$$

The corresponding Youden index-based optimal cut-off point is then obtained numerically. The proposed spline-based approach for the derivation of the corresponding egg-shaped and rectangular confidence regions for sensitivity and 1-specificity at the optimal Youden index-based cut-off point involves the five-step procedure described above where, in the first step, the logspline estimated ROC curve and its associated optimal cut-off point is obtained. The smooth logspline approach will result in m unique bootstrap values of $logit_{\hat{se}^*}$ and $logit_{\hat{sp}^*}$. As a result, the implementation of the proposed approach is straightforward. For the implementation of the logspline approach we used the `polspline` package of `R` with knot addition and deletion (see also Hansen et al., 1997). The codes for implementing our approaches are provided as Supplementary Material in the online version of the article.

3. Simulation Study

In this section we present a large simulation study in order to evaluate the proposed methods. The scenarios considered are shown in Table 1. Initially, we explored a setting which involves data generated by normal distributions. We evaluated our approaches, that is, the construction of the rectangle and egg shaped confidence regions, in terms of the average areas of these regions and estimated coverage levels. For each scenario we computed both 90% and 95% nominal confidence regions for the $(1 - spec(c^*), sens(c^*))$ point along with the corresponding observed coverage probabilities (percentage out of 1000 confidence regions for each scenario that contains the true point). Since the conclusions for both confidence levels were quite similar we only report on the 95% level.

We initially compared our approaches with the classical procedure, considering both rectangular and elliptical confidence regions, where the optimal cut-off point is estimated based on the binormal assumption and then is considered to be fixed and known at its estimated value, thus not taking into account the correlation between sensitivity and specificity at the Youden index-based optimal cut-off point. We explored scenarios for which the optimal Youden index equals to J =0.2, 0.4, 0.6, 0.8 and for sample sizes of $(n_X, n_Y) = (15, 35), (50, 100), (30, 30), (50, 50), (100, 100), (200, 200)$. Here, for brevity, we only present results for sample sizes $(n_X, n_Y) = (30, 30)$, $(50, 50)$, $(100, 100)$. Results for the remaining cases

Table 1
Parameter values of the distributions used in all settings

	μ_x	σ_x^2	σ_y^2	μ_y			
				$J = 0.2$	$J = 0.4$	$J = 0.6$	$J = 0.8$
Normal equal variances	6.5	0.09	0.09	6.652	6.814	7.005	7.269
Normal unequal variances	6.5	0.09	0.25	6.617	6.873	7.143	7.505
Lognormal	2.5	0.09	0.25	2.617	2.873	3.143	3.505
Power normal ($X^{-1/3}$, $Y^{-1/3}$)	3.5	0.09	0.25	3.383	3.127	2.857	2.495
Mixture: $X \sim N(10, 1^2)$, $Y \sim 0.5N(\mu_y, 1^2) + 0.5N(\mu_y + 4, \sqrt{5}^2)$:					9.980	11.120	12.160
	$Scale_x$	$Shape_x$	$Shape_y$		$Scale_y$		
Gamma	0.5	2	2	0.344	0.230	0.142	0.072

Table 2

Simulation results of 1000 replications for the proposed approaches. The delta method approach assuming normality (Delta) and the corresponding bootstrap approach (Boots) are compared to the classical approach. The classical approach will result in an ellipse while the delta method in a more irregular shape, referred to as 'egg'. X and Y are generated from two normal distributions such that $J = 0.2, 0.4, 0.6, 0.8$. For the bootstrap, we considered 1000 bootstrap samples for each simulation replication.

			Normal, equal variances				Normal, unequal variances			
			Coverage		Area		Coverage		Area	
n_X, n_Y	J	Methods	Rect.	Egg/Ell.	Rect.	Egg/Ell.	Rect.	Egg/Ell.	Rect.	Egg/Ell.
		Delta	0.8900	0.9010	0.3068	0.1775	0.9460	0.9490	0.1129	0.0802
	0.2	Boots	0.8850	0.8940	0.2478	0.1503	0.9370	0.9410	0.0957	0.0736
		Classical	0.7000	0.6230	0.1434	0.1343	0.8850	0.8740	0.1043	0.0981
		Delta	0.9360	0.9400	0.1165	0.0995	0.9340	0.9410	0.0718	0.0647
30, 30	0.4	Boots	0.9280	0.9130	0.1113	0.0919	0.9270	0.9210	0.0709	0.0618
		Classical	0.9060	0.8810	0.1324	0.1240	0.9180	0.9050	0.1109	0.1043
		Delta	0.9500	0.9470	0.0532	0.0477	0.9470	0.9440	0.0415	0.0372
	0.6	Boots	0.9440	0.9190	0.0524	0.0429	0.9430	0.9180	0.0414	0.0340
		Classical	0.9490	0.9310	0.0999	0.0937	0.8960	0.8820	0.0817	0.0864
		Delta	0.9610	0.9470	0.0225	0.0161	0.9610	0.9480	0.0191	0.0138
	0.8	Boots	0.9560	0.9170	0.0233	0.0139	0.9570	0.9140	0.0199	0.0119
		Classical	0.9980	0.9540	0.0483	0.0462	0.9440	0.9430	0.0447	0.0430
		Delta	0.9040	0.9130	0.2249	0.1247	0.9550	0.9580	0.0613	0.0455
	0.2	Boots	0.8920	0.9040	0.2001	0.1127	0.9470	0.9520	0.0543	0.0424
		Classical	0.6310	0.6530	0.0900	0.0843	0.8760	0.8960	0.0644	0.0603
		Delta	0.9310	0.9350	0.0706	0.0619	0.9440	0.9430	0.0426	0.0389
50, 50	0.4	Boots	0.9300	0.9150	0.0684	0.0589	0.9380	0.9320	0.0424	0.0379
		Classical	0.9040	0.9019	0.0814	0.0763	0.9220	0.9310	0.0698	0.0654
		Delta	0.9430	0.9440	0.0321	0.0287	0.9460	0.9440	0.0249	0.0223
	0.6	Boots	0.9420	0.9220	0.0316	0.0267	0.9440	0.9260	0.0247	0.0210
		Classical	0.9210	0.9400	0.0621	0.0582	0.9220	0.9370	0.0560	0.0524
		Delta	0.9590	0.9480	0.0134	0.0093	0.9640	0.9460	0.0113	0.0079
	0.8	Boots	0.9490	0.9230	0.0135	0.0084	0.9510	0.9210	0.0115	0.0072
		Classical	0.9720	0.9300	0.0326	0.0307	0.8950	0.9290	0.0297	0.0282
		Delta	0.9120	0.9210	0.1259	0.0706	0.9650	0.9600	0.0292	0.0221
	0.2	Boots	0.9090	0.9170	0.1229	0.0677	0.9330	0.9530	0.0267	0.0209
		Classical	0.6830	0.6210	0.0465	0.0435	0.9040	0.9080	0.0325	0.0304
		Delta	0.9330	0.9440	0.0354	0.0317	0.9450	0.9440	0.0214	0.0197
100, 100	0.4	Boots	0.9220	0.9290	0.0345	0.0307	0.9400	0.9400	0.0212	0.0193
		Classical	0.9640	0.9540	0.0415	0.0389	0.9420	0.9360	0.0356	0.0333
		Delta	0.9490	0.9450	0.0161	0.0143	0.9490	0.9400	0.0125	0.0112
	0.6	Boots	0.9490	0.9290	0.0158	0.0137	0.9480	0.9300	0.0124	0.0108
		Classical	0.9640	0.9540	0.0316	0.0296	0.9610	0.9620	0.0290	0.0271
		Delta	0.9620	0.9400	0.0066	0.0045	0.9600	0.9390	0.0056	0.0038
	0.8	Boots	0.9600	0.9290	0.0066	0.0042	0.9540	0.9320	0.0056	0.0036
		Classical	0.9600	0.9350	0.0177	0.0166	0.9610	0.9490	0.0164	0.0154

are given in Web Appendix C. The results for these simulations are shown in Table 2 (and Table 1 of Web Appendix C). The classical approach yields poor coverage for $J = 0.2$ and usually has coverage farther from the nominal 95% than the delta method and its corresponding bootstrap approach. This is to be expected since the classical approach does not account for the correlation of $(\hat{\delta}_e, \hat{\delta}_p)$. We indicate that the classical regions can be very large, often more than twice as large as the parametric based areas. The classical approach is expected to perform better only in cases where the cut-off point is actually known and fixed. As a result, we did not consider the classical approach any further.

Results in Table 2 show that the proposed methodologies result in the anticipated coverage for moderate and larger sample sizes. For smaller sample sizes the coverage is somewhat lower than the nominal level.

Next, we considered implementing our approaches using the Box–Cox transformation in various settings. The scenarios considered involve cases where the data were generated from normal distributions with equal and unequal variances, lognormal distributions, power normal distributions (that are also members of the Box–Cox family), as well as distributions outside the Box–Cox family such as Gammas and a mixture model. Theoretical values of the parameters for each

scenario are presented in Table 1. We note that, for the mixture model, we only considered $J = 0.4, 0.6, 0.8$ because for $J = 0.2$ there is extensive overlap for the underlying densities of diseased and healthy subjects. We explored the proposed methodologies which involve the Box–Cox transformation and replaced the classical approach with the logspline technique. Results are presented in Tables 3 and 4 (and in Tables 2–4 in Web Appendix C, respectively). In the binormal setting, when the Box–Cox transformation is not applied, the areas of the confidence regions are slightly smaller compared to the corresponding regions obtained after the implementation of the Box–Cox transformation (Table 2 vs. Table 3; the same seed was used). This is expected due to the extra parameter that has to be estimated by the available data. In addition, when the transformation is applied, the delta based rectangles and egg-shaped regions perform slightly better than their bootstrap-based alternatives. This, however, is less evident in the case where no transformation is applied. In this case, the bootstrap has similar performance to the fully parametric approaches in terms of coverage. In Table 3, results for the case of normal distributions with equal variances are less satisfactory than for the case with unequal variances. This is perhaps

Table 3

Simulation results of 1000 replications when the Box–Cox transformation is applied to the data. The delta method approach after applying the Box–Cox transformation (Delta BC), Bootstrap approach after applying the Box–Cox transformation (Boots BC), and the logspline approach (Logspline) are considered. X and Y are generated from two normal distributions (with equal or unequal variances) or two lognormals such that $J = 0.2, 0.4, 0.6, 0.8$. For the bootstrap approach we obtained 1000 bootstrap samples, while for the logspline 400 bootstrap samples were obtained for each simulation replication.

			Normal, equal variances				Normal, unequal variances				Log-normals			
			Coverage		Area		Coverage		Area		Coverage		Area	
n_X, n_Y	J	Methods	Rect.	Egg	Rect.	Egg	Rect.	Egg	Rect.	Egg	Rect.	Egg	Rect.	Egg
		Delta BC	0.8830	0.8990	0.3074	0.1783	0.9500	0.9550	0.1194	0.0845	0.9420	0.9520	0.1170	0.0823
	0.2	Boots BC	0.8740	0.8920	0.2507	0.1581	0.9520	0.9490	0.1067	0.0828	0.9520	0.9430	0.1061	0.0827
		Logspline	0.9870	0.9390	0.5565	0.2756	0.9530	0.9520	0.3419	0.2042	0.9410	0.9500	0.3473	0.2257
		Delta BC	0.9280	0.9260	0.1323	0.1114	0.9350	0.9360	0.0857	0.0771	0.9280	0.9340	0.0810	0.0724
	0.4	Boots BC	0.9210	0.9050	0.1221	0.1016	0.9290	0.9210	0.0855	0.0744	0.9300	0.9230	0.0853	0.0741
30, 30		Logspline	0.9910	0.9490	0.3234	0.1854	0.9790	0.9590	0.2435	0.1529	0.9630	0.9700	0.2530	0.1618
		Delta BC	0.9250	0.9220	0.0701	0.0643	0.9340	0.9300	0.0592	0.0530	0.9340	0.9290	0.0518	0.0476
	0.6	Boots BC	0.9260	0.9020	0.0638	0.0539	0.9280	0.9110	0.0545	0.0458	0.9280	0.9150	0.0542	0.0455
		Logspline	0.9740	0.9530	0.1653	0.1027	0.9840	0.9640	0.1434	0.1434	0.9780	0.9780	0.1587	0.0998
		Delta BC	0.9330	0.9090	0.0331	0.0284	0.9440	0.9280	0.0298	0.0251	0.9380	0.9310	0.0239	0.0211
	0.8	Boots BC	0.9280	0.8990	0.0284	0.0213	0.9340	0.8980	0.0257	0.0191	0.9300	0.9020	0.0253	0.0187
		Logspline	0.9760	0.9760	0.0594	0.0381	0.9800	0.9730	0.0594	0.0382	0.9780	0.9700	0.0606	0.0378
		Delta BC	0.9010	0.9120	0.2285	0.1265	0.9550	0.9600	0.0638	0.0474	0.9520	0.9580	0.0630	0.0466
	0.2	Boots BC	0.8880	0.8990	0.2019	0.1167	0.9510	0.9600	0.0583	0.0461	0.9530	0.9610	0.0581	0.0461
		Logspline	0.9830	0.9320	0.4773	0.2151	0.9560	0.9460	0.2695	0.1516	0.9500	0.9590	0.2560	0.1641
		Delta BC	0.9420	0.9290	0.0823	0.0700	0.9440	0.9470	0.0497	0.0454	0.9360	0.9440	0.0479	0.0437
	0.4	Boots BC	0.9240	0.9100	0.0768	0.0654	0.9370	0.9290	0.0501	0.0447	0.9390	0.9280	0.0500	0.0446
50, 50		Logspline	0.9830	0.9340	0.2493	0.1356	0.9640	0.9480	0.1862	0.1116	0.9610	0.9710	0.1946	0.1198
		Delta BC	0.9440	0.9350	0.0418	0.0387	0.9460	0.9470	0.0341	0.0310	0.9360	0.9420	0.0312	0.0287
	0.6	Boots BC	0.9290	0.9120	0.0392	0.0343	0.9390	0.9220	0.0323	0.0281	0.9380	0.9250	0.0322	0.0279
		Logspline	0.9830	0.9500	0.1251	0.0751	0.9810	0.9630	0.1011	0.0654	0.9720	0.9630	0.1189	0.0731
		Delta BC	0.9410	0.9310	0.0180	0.0160	0.9490	0.9390	0.0162	0.0140	0.9440	0.9400	0.0143	0.0125
	0.8	Boots BC	0.9310	0.9130	0.0168	0.0133	0.9370	0.9180	0.0149	0.0117	0.9380	0.9200	0.0148	0.0115
		Logspline	0.9770	0.9710	0.0412	0.0272	0.9770	0.9740	0.0385	0.0262	0.0473	0.9630	0.0408	0.0269
		Delta BC	0.9110	0.9230	0.1296	0.0720	0.9650	0.9560	0.0301	0.0230	0.9650	0.9550	0.0298	0.0227
	0.2	Boots BC	0.9090	0.9090	0.1235	0.0692	0.9340	0.9570	0.0279	0.0221	0.9360	0.9560	0.0278	0.0221
		Logspline	0.9870	0.9320	0.3750	0.1452	0.9650	0.9640	0.1997	0.0976	0.9680	0.9790	0.1675	0.0978
		Delta BC	0.9300	0.9330	0.0422	0.0363	0.9460	0.9450	0.0248	0.0229	0.9430	0.9420	0.0240	0.0222
	0.4	Boots BC	0.9200	0.9120	0.0394	0.0343	0.9340	0.9280	0.0246	0.0224	0.9320	0.9270	0.0246	0.0224
100, 100		Logspline	0.9780	0.9380	0.1670	0.0846	0.9630	0.9580	0.1229	0.0703	0.9760	0.9850	0.9850	0.0741
		Delta BC	0.9400	0.9400	0.0210	0.0196	0.9500	0.9430	0.0168	0.0154	0.9440	0.9380	0.0158	0.0146
	0.6	Boots BC	0.9210	0.9120	0.0199	0.0180	0.9300	0.9170	0.0161	0.0144	0.9370	0.9200	0.0161	0.0143
		Logspline	0.9780	0.9480	0.0714	0.0447	0.9690	0.9520	0.0627	0.0406	0.9730	0.9700	0.0815	0.0491
		Delta BC	0.9500	0.9440	0.0087	0.0078	0.9410	0.9420	0.0077	0.0068	0.9370	0.9300	0.0071	0.0061
	0.8	Boots BC	0.9310	0.9090	0.0083	0.0069	0.9330	0.9120	0.0073	0.0060	0.9350	0.9180	0.0072	0.0059
		Logspline	0.9790	0.9590	0.0256	0.0256	0.9750	0.9650	0.0228	0.0160	0.9770	0.9620	0.0274	0.0184

Table 4

Simulation results of 1000 replications when the Box–Cox transformation is applied. The delta method approach after applying the Box–Cox transformation (Delta BC), Bootstrap approach after applying the Box–Cox transformation (Boots BC), and the logspline approach (Logspline) are considered. X and Y are generated from two gammas, or two power normals, or from a mixture model such that $J = 0.4, 0.6, 0.8$. For the bootstap we consider 1000 bootstrap samples, while for the logspline 400 bootstrap samples were obtained for each simulation replication.

			Power normals				Gammas			
			Coverage		Area		Coverage		Area	
n_X, n_Y	J	Methods	Rect.	Egg	Rect.	Egg	Rect.	Egg	Rect.	Egg
		Delta BC	0.9390	0.9500	0.1176	0.0829	0.8960	0.9100	0.2964	0.1715
	0.2	Boots BC	0.9330	0.9400	0.1046	0.0822	0.8790	0.8940	0.2431	0.1530
		Logspline	0.9490	0.9550	0.3545	0.2036	0.9760	0.9600	0.4613	0.2694
		Delta BC	0.9350	0.9320	0.0826	0.0727	0.9200	0.9190	0.1147	0.0981
	0.4	Boots BC	0.9180	0.8980	0.0843	0.0727	0.9150	0.9080	0.1109	0.0940
30, 30		Logspline	0.9700	0.9520	0.2556	0.1571	0.9780	0.9650	0.2954	0.1824
		Delta BC	0.9400	0.9330	0.0514	0.0476	0.9270	0.9260	0.0575	0.0532
	0.6	Boots BC	0.9250	0.8880	0.0525	0.0438	0.9220	0.9030	0.0574	0.0487
		Logspline	0.9670	0.9570	0.1523	0.0965	0.9780	0.9610	0.1724	0.1040
		Delta BC	0.9380	0.9320	0.0229	0.0208	0.9330	0.9090	0.0241	0.0216
	0.8	Boots BC	0.9150	0.8930	0.0237	0.0174	0.9400	0.8960	0.0249	0.0186
		Logspline	0.9630	0.9650	0.0579	0.0373	0.9760	0.9660	0.0601	0.0367
		Delta BC	0.9540	0.9570	0.0640	0.0473	0.9250	0.9260	0.2106	0.1192
	0.2	Boots BC	0.9460	0.9540	0.0584	0.0465	0.9180	0.9160	0.1867	0.1097
		Logspline	0.9690	0.9680	0.2763	0.1475	0.9680	0.9650	0.3794	0.2021
		Delta BC	0.9420	0.9380	0.0489	0.0444	0.9420	0.9380	0.0715	0.0621
	0.4	Boots BC	0.9290	0.9210	0.0499	0.0455	0.9370	0.9100	0.0681	0.0589
50, 50		Logspline	0.9590	0.9570	0.1964	0.1158	0.9730	0.9640	0.2430	0.1387
		Delta BC	0.9390	0.9420	0.0315	0.0293	0.9490	0.9460	0.0355	0.0328
	0.6	Boots BC	0.9280	0.9130	0.0317	0.0274	0.9340	0.9100	0.0343	0.0300
		Logspline	0.9740	0.9640	0.1122	0.0703	0.9710	0.9660	0.1374	0.0791
		Delta BC	0.9450	0.9410	0.0140	0.0125	0.9430	0.9460	0.0145	0.0129
	0.8	Boots BC	0.9400	0.9010	0.0141	0.0109	0.9310	0.9050	0.0143	0.0112
		Logspline	0.9720	0.9680	0.0391	0.0265	0.9610	0.9470	0.0432	0.0267
		Delta BC	0.9530	0.9510	0.0298	0.0227	0.9370	0.9370	0.1158	0.0677
	0.2	Boots BC	0.9170	0.9480	0.0273	0.0219	0.9270	0.9160	0.1096	0.0638
		Logspline	0.9780	0.9710	0.1933	0.0950	0.9680	0.9790	0.2926	0.1353
		Delta BC	0.9450	0.9480	0.0245	0.0225	0.9440	0.9530	0.0364	0.0322
	0.4	Boots BC	0.9410	0.9340	0.0245	0.0223	0.9320	0.9270	0.0340	0.0302
100, 100		Logspline	0.9680	0.9720	0.1285	0.0731	0.9790	0.9800	0.1673	0.0896
		Delta BC	0.9480	0.9480	0.0160	0.0148	0.9550	0.9520	0.0180	0.0167
	0.6	Boots BC	0.9320	0.9250	0.0158	0.0140	0.9400	0.9150	0.0170	0.0153
		Logspline	0.9610	0.9550	0.0673	0.0434	0.9790	0.9780	0.0938	0.0535
		Delta BC	0.9490	0.9470	0.0070	0.0061	0.9550	0.9500	0.0072	0.0063
	0.8	Boots BC	0.9390	0.9160	0.0069	0.0056	0.9430	0.9120	0.0069	0.0056
		Logspline	0.9840	0.9700	0.0226	0.0162	0.9770	0.9730	0.0295	0.0190

due to the fact that we are simulating from equal variances but are not using this information when actually computing confidence intervals.

We observe that the proposed approaches have nice coverage properties in most cases. The logspline approach turns out to be slightly conservative yielding coverage above 0.95 in many cases. As expected, the logspline approach yields essentially larger areas compared to the delta and bootstrap-based approaches that assume normality after applying the Box–Cox transformation. This is the price to pay for not making any strict parametric assumptions. We observe that the egg-shaped regions are in most cases smaller in terms of the area of the confidence region as compared to the corresponding rectangular regions. This is more evident in cases where the absolute value of the correlation of $\hat{\delta}_e$ and $\hat{\delta}_p$ is high. This finding is illustrated in Figure 1 in Web Appendix C, where the ratio *rectangular area/egg − shaped area* is plotted versus the theoretical correlation for the normal distributions with unequal variances scenario. Furthermore, we observe that for the logspline approach also, the egg-shaped regions are smaller than the rectangular regions. For the mixture scenario, where there are significant departures from normality, we note that the Box–Cox transformation drastically failed to result in an appropriate coverage. We, thus, only present results of

the logspline approach (Table 4 in Web Appendix C). The logspline approach turns out to be robust in these scenarios and has nice coverage properties. The price to pay is that the logspline results in larger areas than other methods.

A reviewer has raised the issue that confidence intervals and coverages are less relevant if the estimates of sensitivity and specificity are biased. We thus added a simulation study in Web Appendix D (Tables 5 through 7 therein) showing that bias and mean squared error (MSE) for estimates of sensitivity and specificity at the optimal J are quite small under different estimation scenarios. Specifically, bias is lower when J increases, probably due to the better separation of the distributions. Larger J are more relevant in practice since these correspond to useful diagnostic markers. The empirical and Box–Cox approaches result in better estimates overall, however, the logspline approach is competitive when sampling from gamma distributions. Although the gamma distribution is not in the Box–Cox transformation family the Box–Cox procedure still exhibits small biases in this scenario.

4. Applications

4.1. *Brucellosis Data*

Brucellosis is an intra-cellular bacterial disease transmitted to humans primarily by contact with infected animals or by ingestion of unpasteurized dairy products. Defects of the cellular immunity and T-lymphocytes have been described in brucellosis patients. Differences between healthy and diseased subjects on the proliferation and the stimulation of T-lymphocyte subsets from phytohemagglutinin (PHA) cultured peripheral blood, have also been studied (Zhan, 1995). The dataset of 35 brucellosis patients and 15 controls has been analyzed before (e.g., in Nakas et al., 2003; Skendros et al., 2007). Stimulation of CD3-positive lymphocytes is a significant marker for the detection of brucellosis.

The normality assumption cannot be rejected for either the healthy or diseased groups (it p-values of the Shapiro Wilk test are 0.894 and 0.258, respectively). The QQ-plot for the data is presented in Figure 2 in Web Appendix E. All proposed methodologies are illustrated: analysis under the normality assumption, analysis using the Box–Cox transformation as well as the logspline approach. The corresponding graphs are presented in Figure 1.

Under the normality assumption, the estimated optimal cut-off point along with the 95% confidence interval is 81.0082 (78.3866, 83.6298). This confidence interval is obtained using the delta method for the derivation of the variance of the cut-off following Schisterman and Perkins (2007). The associated estimates for the marginal sensitivity and specificity are 0.6595 (0.4948, 0.7981) and 0.8339 (0.7025, 0.9204). The corresponding 95% bootstrap based CIs are (0.4803, 0.8083) and (0.7250, 0.9101), respectively. The areas of the delta-based rectangle and egg-shaped regions are 0.0857 and 0.0802, respectively. For the corresponding bootstrap methods we obtain areas equal to 0.0770 and 0.0637, respectively. Figure 1 illustrates that the egg-shaped and rectangular regions are similar for the bootstrap and the delta based approaches. We note here that the classical rectangle and ellipse (Figure 2) yield confidence regions with areas equal to 0.1161, 0.1125, respectively, which are larger than the delta based areas given above and extend beyond the ROC space.

Results are quite similar after applying the Box–Cox transformation. The delta-based CIs for the sensitivity and specificity are (0.4861, 0.8103) and (0.7193, 0.9693), respectively, while the bootstrap yields (0.5028, 0.7988) and (0.7901, 0.9499). The delta-based rectangular and egg-shaped areas are 0.1058 and 0.0842, respectively, while the corresponding bootstrap areas are equal to 0.0617 and 0.0473, respectively. The classical approach yields larger regions (Figure 2). This is also seen in our simulation studies when comparing the results

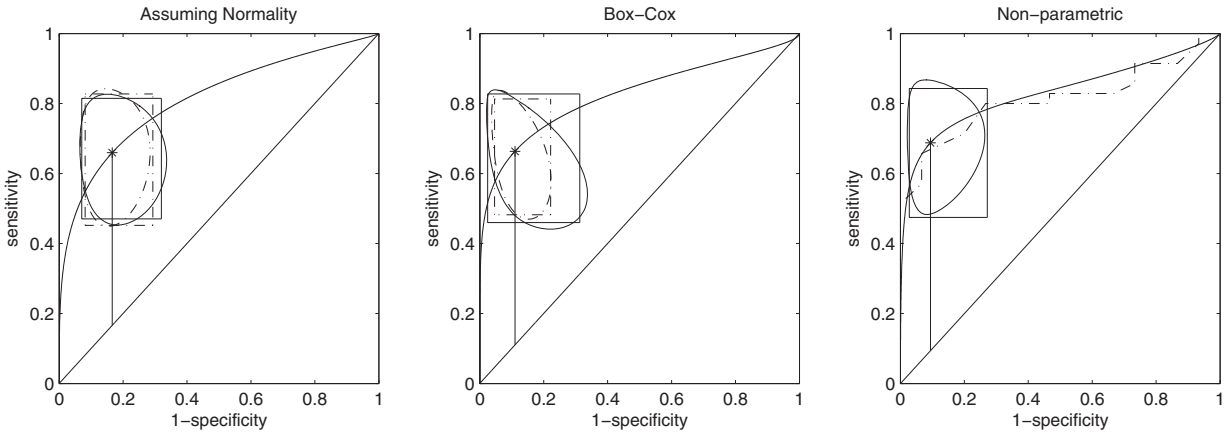

Figure 1. 95 confidence regions for the brucellosis data set. Left: ROC curve obtained by assuming normality for the diseased and the controls (AUC = 0.8169). The solid rectangle and egg-shaped regions refer to the delta-based approach while the dashed confidence regions refer to the bootstrap-based approach. Middle: The corresponding ROC curve and confidence regions after the Box–Cox transformation (AUC = 0.8301). The solid regions refer to the delta-based approach and the dashed to the bootstrap-based approach. Right: The logspline-based ROC curve (AUC = 0.8262) and corresponding confidence regions. The dashed step function refers to the empirical ROC (AUC = 0.8114).

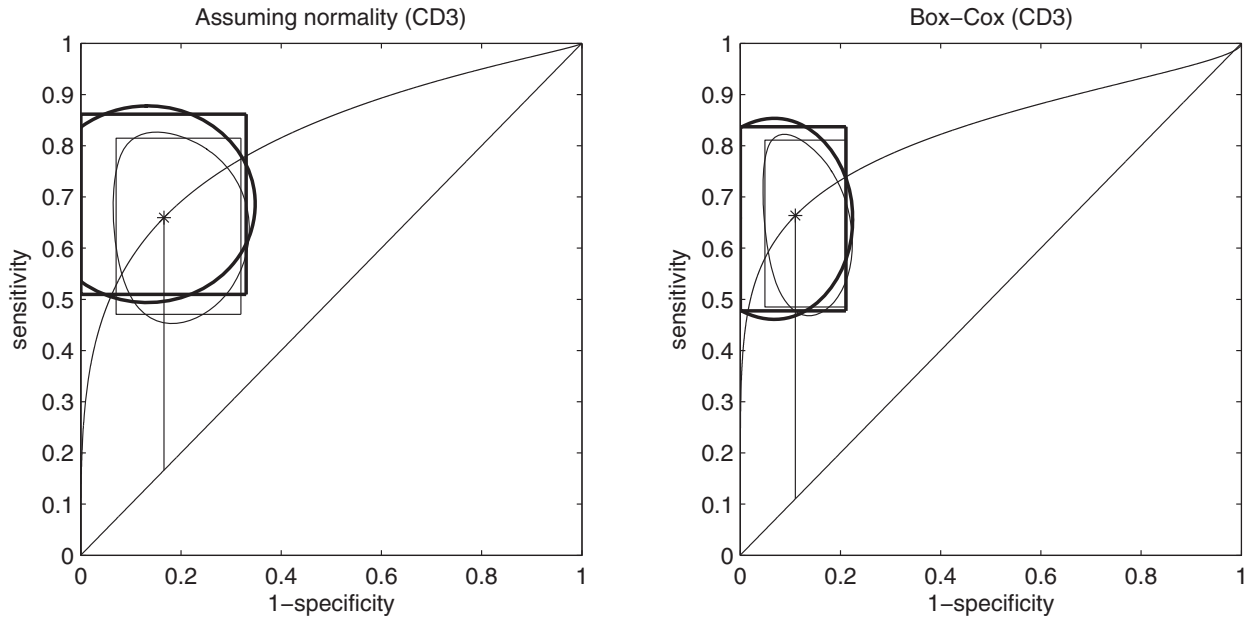

Figure 2. Classical (thick lines) and delta based (thin lines) 95% confidence regions. The delta based confidence regions are re-plotted here (cf. Figure 1) for comparison purposes with the corresponding classical ones. Left: confidence regions for the CD3 data assuming normality. The classical rectangle and ellipse yield areas of 0.1161, 0.1125, respectively. Classical regions are extending beyond the ROC space. Right: confidence regions for the CD3 data after using the Box–Cox transformation. The classical rectangle and ellipse yield areas of 0.0759, 0.0740, respectively.

of Tables 2 and 3. It is a trivial fact that, when the normality assumption holds, there is no need for a Box–Cox transformation. The logspline egg-shaped confidence region yields an area equal to 0.0717 while the corresponding rectangular area equals to 0.0898.

4.2. *Pancreatic Cancer Data*

We also analyze a well known data set of patients with pancreatic cancer versus a control group. Wieand et al. (1989) present a case control study of 90 patients with pancreatic cancer and 51 controls. This data is graphically presented in Molodianovitch et al. (2006). Here, for illustrative purposes, we apply our methods to the measurements of marker CA-125 (a cancer antigen). Results are shown in Web Appendix F.

5. Discussion

The standard procedures available in textbooks and the literature for obtaining confidence intervals for sensitivity and specificity assume that the sensitivity and specificity estimators are independent and are computed at a known or given cut-off point. They do not account for the variation and correlation introduced by estimating the optimal cut-off point via the maximum of the Youden index. Resulting $(1-\alpha)\%$ confidence intervals can be highly inaccurate in practice as illustrated in our initial simulations (Table 2 here and Table 1 in Web Appendix C).

We fill this literature gap by proposing two different approaches. The first approach focuses on the frequently made binormal model which assumes that a common monotonic transformation to normality exists for both the healthy and diseased populations. More specifically we estimate this monotonic transformation using the Box–Cox family of transformations and then use the delta method in order to obtain confidence regions for the sensitivity-specificity pair at the optimal cut-off point. The second approach makes no parametric assumptions but uses a spline based technique (logspline) to estimate the densities for both populations which is followed by bootstrapping for statistical inference (as in Kooperberg and Stone, 2004). A kernel approach is a possible alternative to the logspline but in initial simulations we found it to be less effective and do not discuss it further in this paper. The performance of our proposed approaches is illustrated through an extensive simulation study. Our procedures have nice coverage properties and are applicable for small sample sizes even when no strict parametric assumptions are used. Bayesian smoothing approaches were not considered being beyond the scope of this research.

Our proposed methodologies are also applicable when J^*, the generalized Youden index is used for selecting the optimal cut-off point. Although we focus on the commonly used Youden index-based cut-off point, this issue arises with any method of choosing a cut-off point which is based on the observed data. The application of our methods to other possible metrics such as those discussed in Zou et al. (2013) is an issue of future research.

In contrast to other authors who used the Box–Cox transformation approach for ROC curve analyses we found it necessary to take into account the variability due to estimating lambda. This may be due to the fact that previous authors (e.g., Molanes-Lopez and Leton, 2011) considered

one dimensional problems such as confidence intervals for the Youden Index while our construction of confidence regions in the ROC space is a two dimensional problem. The Box–Cox approach works well even for some distributions such as the gamma which is not in the Box–Cox family and does not satisfy the assumptions of the binormal model. This property of the Box–Cox approach in the ROC context has already been pointed out in the literature (e.g., Molodianovitch et al. 2006; Schisterman et al., 2006). It should be noted that the gamma distribution has a shape similar to the log-normal which is in the Box–Cox family. For such distributions the logspline approach results in too large confidence regions. However for extreme departures from binormality such as mixture distributions the Box–Cox approach breaks down while the logspline performs better.

We have illustrated our findings with a dataset for which the normality assumption is reasonable (Brucellosis data) and through a reference dataset in diagnostic accuracy studies (Pancreatic cancer) where the normality assumption fails. The superiority of the proposed methodologies as compared to the classical approach is evident (Figure 2). Our study is limited to continuous diagnostic markers. Markers with ordinal outcomes or with probability mass at zero (Schisterman et al., 2006) have not been considered and require further research.

Although rectangular regions are easily communicated, egg-shaped regions provide smaller areas with coverage close to the nominal level and are thus preferable. Practitioners are encouraged to provide confidence intervals along with point estimates both for the optimal cut-off point and for the corresponding sensitivity-specificity pairs, though this may require the involvement of a professional statistician.

6. Supplementary Materials

Web Appendices, Tables and Figures referenced in Sections 2.2, 2.4, 3, 4.1, 4.2, 5 are available with this paper at the Biometrics website on Wiley Online Library. Matlab or R code for the implementation of the methodologies and reproduction of the applications are also available with this paper at the *Biometrics* website on Wiley Online Library.

Acknowledgements

The authors thank three anonymous referees, the associate editor and the co-editor for very useful comments that improved the presentation of the paper.

References

Chen, K. C., Yeh, C. J., Kuo, J. F., Hsieh, C. L., Yang, S. F., and Wang, and C. H. (2011). Footprint analysis of flatfoot in preschool-aged children. *European Journal of Paediatrics* **170**, 611–617.

Fluss, R., Faraggi, D., and Reiser, B. (2005). Estimation of the Youden index and its associated cutoff point. *Biometrical Journal* **47**, 458–472.

Greiner, M., Pfeiffer, D., and Smith, R. M. (2000). Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. *Preventive Veterinary Medicine* **45**, 23–41.

Hanley, J.A. (1996). The use of the 'binormal' model for parametric ROC analysis of quantitative diagnostic tests. *Statistics in Medicine* **15**, 1575–1585.

Hansen, M., Kooperberg, C., Truong, Y. K., and Stone, C. J. (1997). The use of polynomial splines and their tensor products in extended linear modeling (with discussion). *The Annals of Statistics* **25**, 1371–1470.

Hernandez, F. and Johnson, R. A. (1980). The large sample behavior of transformations to normality. *Journal of the American Statistical Association* **75**, 855–861.

Kooperberg, C. and Stone, C. J. (1992). A study of logspline density estimation. *Computational Statistics & Data Analysis* **12**, 327–348.

Kooperberg, C. and Stone, C. J. (2004). Comparison of parametric and bootstrap approaches to obtaining confidence intervals for logspline density estimation. *Journal of Computational and Graphical Statistics* **13**, 106–122.

Kosinski, A. S., Chen, Y., and Lyles, R. H. (2010). Sample size calculations for evaluating a diagnostic test when the gold standard is missing at random. *Statistics in Medicine* **29**, 1572–1579.

McClish, D. K. (2012). Evaluation of the accuracy of medical tests in a region around the optimal point. *Academic Radiology* **19**, 1484–1490.

Molanes-López, E. M. and Letón, E. (2011). Inference of the Youden index and associated threshold using empirical likelihood for quantiles. *Statistics in Medicine* **30**, 2467–2480.

Molodianovitch, K., Faraggi, D., and Reiser B. (2006). Comparing the areas under two correlated ROC curves: Parametric and non-parametric approaches. *Biometrical Journal* **48**, 45–757.

Nakas, C., Yiannoutsos, C. T., Bosch, R. J., and Moyssiadis, C. (2003). Assessment of diagnostic markers by goodness-of-fit tests. *Statistics in Medicine* **22**, 2503–2513.

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

Rhemrev, S. J., Beeres, F. J. P., van Leerdam, R. H., Hogervorst, M., and Ring, D. (2010). Clinical prediction rule for suspected scaphoid fractures—A prospective cohort study. *Injury* **41**, 1026–1030.

Schisterman, E. F. and Perkins, N. (2007). Confidence intervals for the Youden index and corresponding optimal cut-point. *Communications in Statistics – Simulation and Computation* **36**, 549–563.

Schisterman, E. F., Reiser, B., and Faraggi, D. (2006). ROC analysis for markers with mass at zero. *Statistics in Medicine* **25**, 623–638.

Skaltsa, K., Jover, L., and Carrasco, J. L. (2010). Estimation of the diagnostic threshold accounting for decision costs and sampling uncertainty. *Biometrical Journal* **52**, 676–697.

Skendros, P., Boura, P., Chrisagis, D., and Raptopoulou-Gigi, M. (2007). Diminished percentage of CD4+ T-lymphocytes expressing interleukine-2 receptor alpha in chronic brucellosis. *Journal of Infection* **54**, 192–197.

Takada, T. (2008). Asymptotic and qualitative performance of nonparametric density estimators: a comparative study. *Econometrics Journal* **11**, 573–592.

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

Wieand, S., Gail, M. H., James, B. R., and James, K. L. (1989). A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika* **76**, 585–592.

Zhan, Y. (1995). Differential activation of brucella-reactive CD4+T cells by brucella infection or immunization with antigenic extracts. *Infection Immunology* **63**, 969–975.

Zou, K. H. and Hall, W. J. (2000). Two transformation models for estimating an ROC curve from continuous data. *Journal of Applied Statistics* **27**, 621–631.

Zou, K. H., Liu, A., Bandos, A. I., Ohno-Machado, L., and Rockette, H. E. (2012). *Statistical Evaluation of Diagnostic Performance: Topics in ROC Analysis*. Boca Raton: CRC Press.

Zou, K. H., Yu, C. R., Liu, K., Carlsson, M. O., and Cabrera, J. (2013). Optimal thresholds by maximizing various metrics via ROC-type analysis. *Academic Radiology* **20**, 807–815.

Received December 2012. *Revised May* 2013.
Accepted August 2013.