The International Journal of Biostatistics

Volume 6, Issue 1

2010

Article 24

Simple Nonparametric Confidence Regions for the Evaluation of Continuous-Scale Diagnostic Tests

Gianfranco Adimari, *University of Padua*

Monica Chiogna, *University of Padua*

Recommended Citation:

Adimari, Gianfranco and Chiogna, Monica (2010) "Simple Nonparametric Confidence Regions for the Evaluation of Continuous-Scale Diagnostic Tests," *The International Journal of Biostatistics*: Vol. 6: Iss. 1, Article 24.

DOI: 10.2202/1557-4679.1256

Simple Nonparametric Confidence Regions for the Evaluation of Continuous-Scale Diagnostic Tests

Gianfranco Adimari and Monica Chiogna

Abstract

The evaluation of the ability of a diagnostic test to separate diseased subjects from non-diseased subjects is a crucial issue in modern medicine. The accuracy of a continuous-scale test at a chosen cut-off level can be measured by its sensitivity and specificity, i.e. by the probabilities that the test correctly identifies the diseased and non-diseased subjects, respectively.

In practice, sensitivity and specificity of the test are unknown. Moreover, which cut-off level to use is also generally unknown in that no preliminary indications driving its choice could be available.

In this paper, we address the problem of making joint inference on pairs of quantities defining accuracy of a diagnostic test, in particular, when one of the two quantities is the cut-off level. We propose a technique based on an empirical likelihood statistic that allows, within a unified framework, to build bivariate confidence regions for the pair (sensitivity, cut-off level) at a fixed value of specificity as well as for the pair (specificity, cut-off level) at a fixed value of sensitivity or the pair (sensitivity, specificity) at a fixed cut-off value.

A simulation study is carried out to assess the finite-sample accuracy of the method. Moreover, we apply the method to two real examples.

KEYWORDS: empirical likelihood, likelihood ratios, ROC curve, sensitivity, specificity

Author Notes: This work was supported by University of Padova under Grant 075919, 2007.

1 Introduction and motivation

Diagnostic tests, commonly used to attempt to distinguish diseased from non-diseased patients, are important tools in modern medicine. A continuous test produces numeric values on a continuous scale. Without loss of generality, we assume that higher test values indicate a higher likelihood of disease. In order to classify a result of the test as positive or negative, a cut-off level must be chosen. The result is called positive if the value exceeds the cut-off level, and negative otherwise. A positive result indicates the presence of disease.

The evaluation of the ability of a test to separate diseased from non-diseased subjects is a crucial issue, as the precise diagnosis of disease can lead to a pertinent medical treatment to patients and reduce medical mishap. The accuracy of a diagnostic test can be evaluated by its sensitivity and specificity, which are defined as the probabilities that the test correctly identifies the diseased and non-diseased subjects, respectively. Let X be the result of a continuous diagnostic test for a non-diseased subject and Y the result of the test for a diseased subject. For a given cut-off level τ , the sensitivity and the specificity of the test are

$$\theta(\tau) = \Pr\{Y > \tau\} = 1 - F_Y(\tau),$$

$$\eta(\tau) = \Pr\{X \leq \tau\} = F_X(\tau),$$

respectively, where $F_X(\cdot)$ denotes the distribution function of X and $F_Y(\cdot)$ the distribution function of Y . Therefore, the sensitivity depends only on the diseased population. On the contrary, the specificity depends only on the non-diseased population. Choosing a high cut-off 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 cut-off level gives the opposite results. The trade-off between sensitivity and specificity of a diagnostic test can be captured, in a single graph, by the Receiver Operating Characteristic (ROC) curve. The ROC curve is a plot of sensitivity against 1-specificity as the cut-off value varies over its entire possible range. It is often used to define single indices which summarize the overall accuracy of the test, of which the most popular is the area under the ROC curve (AUC).

Clearly, when a test is under evaluation, its sensitivity and specificity at a fixed cut-off value τ are unknown, being typically unknown the distribution of test results, for both diseased and non-diseased patients. Moreover, the cut-off level to use is generally unknown: its choice may depend on specific objectives and on the available data at hand. Inference on such quantities – and summaries derived from them – is, hence, an important topic. Various aims could be stated. In some circumstances, for example, a minimally acceptable value for the specificity of the test

(i.e., a maximal value for the probability of false positive errors) could be required. Thus, the objective would be to make an inference on the range of sensitivity of the test at this given level of specificity. In other situations, one may want to control the probability of false negative errors; in this case, the objective would be to find the range of specificity of the test for a given level of sensitivity. In both cases, it could be of interest also to collect information regarding the choice of a cut-off level. On the contrary, if a threshold selection rule was known, an objective could be to find the range of both the sensitivity and specificity of the test corresponding to a given cut-off level. Finally, one could be simply interested in making inference on a single summary index of accuracy, such as the AUC.

Many statistical methods for the evaluation of diagnostic tests are reviewed in two monographies by Zhou *et al.* (2002) and Pepe (2003). We refer the interested reader to these books and references therein. Moreover, various recent articles propose new inferential procedures or refinements of existing techniques. See Claeskens *et al.* (2003), Pepe and Cai (2004), Zhou and Qin (2005), Qin and Zhou (2006), Adimari (2006), Adimari and Chiogna (2006), Kim and Qin (2008), Qin *et al.* (2007), among others. Some of the mentioned works develop inferential techniques based on the empirical likelihood function (see Owen, 2001, as a general reference).

Despite the richness of the literature, the problem of making joint inference on pairs of quantities defining the accuracy of a test, in particular when one of the two quantities is the cut-off level, appears, to our knowledge, not to have been sufficiently explored. In this paper, we try to fill in this gap by proposing an empirical likelihood-based approach for the construction of nonparametric bivariate confidence regions. Empirical likelihood methods are especially appealing for the purpose of building confidence regions, for various reasons. For instance, they involve no pre-specified assumptions about the shape of the confidence regions, which is determined automatically by the data. Moreover, unlike the symmetric confidence regions derived by a normal approximation method, empirical likelihood confidence regions do not need to estimate a variance term. Finally, confidence regions based on likelihood ratios are range-respecting, too.

The method that we discuss allows us to address, within a unified framework, inferential problems such as those previously described. In particular, it works well in making inference about the pair (sensitivity, cut-off level), at a fixed value of specificity, as well as about the pair (specificity, cut-off level), at a fixed value of sensitivity, or about the pair (sensitivity, specificity), at a fixed cut-off value. The new confidence regions are computationally simple and easy to implement in practice.

The paper is organized as follows. Section 2 describes the proposed method, giving theoretical justification. Section 3 presents some results of a simulation study

carried out to assess the finite-sample performance of the method. In Section 4, we illustrate the method with two real examples. Section 5 contains some final remarks.

2 Bivariate nonparametric confidence regions

Let $x_1, \dots, x_m$ be a random sample from X , i.e., the test results from m patients known to be non-diseased, and $y_1, \dots, y_n$ a random sample from Y , i.e., the test results from n patients known to be diseased. 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 $y_1, \dots, y_n$.

Consider the empirical likelihood function based on the two independent samples $x_1, \dots, x_m$ and $y_1, \dots, y_n$

$$L(\mathbf{p}, \mathbf{q}) = \prod_{i=1}^m p_i \prod_{j=1}^n q_j.$$

Here, $\mathbf{p} = (p_1, \dots, p_m)$ and $\mathbf{q} = (q_1, \dots, q_n)$ are probability vectors, representing multinomial distributions on $x_1, \dots, x_m$ and $y_1, \dots, y_n$, respectively. Let F_X^{-1} denote the inverse function of F_X and $I(\cdot)$ the indicator function. Since the relation $1 - \theta = F_Y(F_X^{-1}(\eta))$ holds, the distributions of variables X and Y must be such that the η -quantile of X equals the $(1 - \theta)$ -quantile of Y . Then, one could maximize $L(\mathbf{p}, \mathbf{q})$ subject to the constraints

$$\sum_{i=1}^m p_i I(x_i \leq \tau) = \eta, \quad \sum_{j=1}^n q_j I(y_j \leq \tau) = 1 - \theta. \quad (1)$$

The constrained maximum is given by the product

$$\sup_{\mathbf{p}: \sum_{i=1}^m p_i I(x_i \leq \tau) = \eta} \prod_{i=1}^m p_i \times \left(\sup_{\mathbf{q}: \sum_{j=1}^n q_j I(y_j \leq \tau) = 1 - \theta} \prod_{j=1}^n q_j \right), \quad (2)$$

where the term in the left hand side is simply the empirical likelihood for the η -quantile of X , and the term in the right hand side is the empirical likelihood for the $(1 - \theta)$ -quantile of Y . It is well known that each term in (2) reduces to

$$m^{-m} \left(\frac{\eta}{\hat{F}_X(\tau)} \right)^{m\hat{F}_X(\tau)} \left(\frac{1 - \eta}{1 - \hat{F}_X(\tau)} \right)^{m(1 - \hat{F}_X(\tau))} \quad \text{if } \eta \in (0, 1) \text{ and } \tau \in [x_{(1)}, x_{(m)}],$$

and

$$n^{-n} \left(\frac{1 - \theta}{\hat{F}_Y(\tau)} \right)^{n\hat{F}_Y(\tau)} \left(\frac{\theta}{1 - \hat{F}_Y(\tau)} \right)^{n(1 - \hat{F}_Y(\tau))} \quad \text{if } \theta \in (0, 1) \text{ and } \tau \in [y_{(1)}, y_{(n)}],$$

respectively, where $x_{(i)}$, $i = 1, \dots, m$, and $y_{(j)}$, $j = 1, \dots, n$, denote the order statistics from the samples. See Owen (2001, Section 3.6). Moreover, the unconstrained maximum of $L(\mathbf{p}, \mathbf{q})$ is $m^{-m}n^{-n}$. It follows that the empirical likelihood statistic (i.e., minus twice the maximum empirical log-likelihood ratio) corresponding to the maximization of $L(\mathbf{p}, \mathbf{q})$ subject to (1) reduces to

$$\begin{aligned} \ell(\theta, \eta, \tau) = & 2m \left\{ \hat{F}_X(\tau) \log \frac{\hat{F}_X(\tau)}{\eta} + [1 - \hat{F}_X(\tau)] \log \frac{1 - \hat{F}_X(\tau)}{1 - \eta} \right\} \\ & + 2n \left\{ \hat{F}_Y(\tau) \log \frac{\hat{F}_Y(\tau)}{1 - \theta} + [1 - \hat{F}_Y(\tau)] \log \frac{1 - \hat{F}_Y(\tau)}{\theta} \right\}, \quad (3) \end{aligned}$$

for $\theta \in (0, 1)$, $\eta \in (0, 1)$, $\tau \in \mathcal{T}$, where $\mathcal{T} = [x_{(1)}, x_{(m)}] \cap [y_{(1)}, y_{(n)}]$, i.e., $\mathcal{T} = [\max\{x_{(1)}, y_{(1)}\}, \min\{x_{(m)}, y_{(n)}\}]$. When $\tau \notin \mathcal{T}$, is adequate to set $\ell(\theta, \eta, \tau) = +\infty$. See also Adimari (1998).

The following result holds.

Theorem 1 *Let F_X and F_Y be continuous and strictly increasing in a neighbourhood of some value τ_0 . Let $\theta_0 = 1 - F_Y(\tau_0)$ and $\eta_0 = F_X(\tau_0)$ be the sensitivity and the specificity levels corresponding to the threshold τ_0 . Then, when $\min\{m, n\} \rightarrow +\infty$,*

$$\ell(\theta_0, \eta_0, \tau_0) \xrightarrow{d} \chi_2^2,$$

where χ_2^2 indicates the chi-squared distribution with 2 degrees of freedom.

Proof. By standard results on the asymptotic behaviour of sample extrema, it follows that $\Pr\{\tau_0 \in \mathcal{T}\} \rightarrow 1$ as $\min\{m, n\} \rightarrow +\infty$. Therefore, $\ell(\theta_0, \eta_0, \tau_0)$ is finite with probability tending to 1, as $\min\{m, n\} \rightarrow +\infty$. Since $\hat{F}_X(\tau_0) = \eta_0 + O_p(m^{-1/2})$ and $1 - \hat{F}_Y(\tau_0) = \theta_0 + O_p(n^{-1/2})$, by using the expansion $\log(z) = z - 1 - (z - 1)^2/2 + o((z - 1)^2)$ in (3) with $\theta = \theta_0$, $\eta = \eta_0$ and $\tau = \tau_0$, we obtain, after some algebra,

$$\ell(\theta_0, \eta_0, \tau_0) = m \frac{\{\hat{F}_X(\tau_0) - \eta_0\}^2}{\eta_0(1 - \eta_0)} + n \frac{\{\hat{F}_Y(\tau_0) - 1 + \theta_0\}^2}{\theta_0(1 - \theta_0)} + o_p(1).$$

The result follows immediately from the asymptotic normality and independence of $m^{1/2}\{\hat{F}_X(\tau_0) - \eta_0\}$ and $n^{1/2}\{\hat{F}_Y(\tau_0) - 1 + \theta_0\}$.

Theorem 1 establishes that the quantity $\ell(\theta, \eta, \tau)$ is asymptotically pivotal, under very weak conditions. Such a result can be used to construct a nonparametric

confidence region for the parameter pairs (θ_0, τ_0) , (η_0, τ_0) or (θ_0, η_0) , for a fixed value of the third remaining parameter. In particular, for example, the set

$$\mathcal{R}_\alpha = \{(\theta, \tau) : \ell(\theta, \eta_0, \tau) \leq c_\alpha\},$$

where $\alpha \in (0, 1)$ and c_α is such that $\Pr\{\chi_2^2 > c_\alpha\} = 1 - \alpha$, represents a confidence region with nominal coverage probability $1 - \alpha$ for the pair (sensitivity, cut-off level), at a fixed value η_0 of specificity. Thus, at a confidence level $1 - \alpha$, $\mathcal{R}_\alpha$ gives the pairs of values for sensitivity and cut-off level, which are compatibles with the selected level η_0 of specificity. Approximate confidence regions, with asymptotically correct coverage probability $1 - \alpha$, for the pair (specificity, cut-off level), at a fixed sensitivity θ_0 , can be obtained in an analogous way. The same applies to the pair (sensitivity, specificity) at a fixed cut-off level τ_0 .

Regions obtained by $\ell(\theta, \eta, \tau)$ retain all good features of empirical likelihood confidence regions. In addition, they are very simple to compute, since the function $\ell(\theta, \eta, \tau)$ essentially involves the empirical distribution functions $\hat{F}_X$ and $\hat{F}_Y$ only. For a fixed nominal level $1 - \alpha$, the χ_2^2 calibration gives a threshold c_α . Then, a value for the pair of parameters of interest will belong to the region $\mathcal{R}_\alpha$ if the corresponding ℓ -value is not greater than c_α . In practice, confidence regions are obtained by using any software routine for graphing the level curves of $\ell(\theta, \eta, \tau)$. For example, in the R environment, visualization of level curves can be produced with the `contour` command. It should be noted that, because of the non-smoothness of $\hat{F}_X$ and $\hat{F}_Y$, the confidence regions have no smooth contours when the cut-off level τ is one of the two parameters of interest.

We conclude this Section with three remarks.

REMARK 1. The unconstrained maximum of $L(\mathbf{p}, \mathbf{q})$ is attained when $\mathbf{p} = (1/m, \dots, 1/m)$ and $\mathbf{q} = (1/n, \dots, 1/n)$. As a consequence, when the cut-off is fixed at the level τ_0 , the maximum empirical likelihood estimates of θ and η are simply the empirical counterparts of $1 - F_Y(\tau_0)$ and $F_X(\tau_0)$. Hence, $\hat{\theta} = 1 - \hat{F}_Y(\tau_0)$ and $\hat{\eta} = \hat{F}_X(\tau_0)$. Similarly, for example, the point estimates of the cut-off τ and the sensitivity θ at a fixed value η_0 of specificity are the usual nonparametric estimates, $\hat{\tau} = \hat{F}_X^{-1}(\eta_0)$ and $\hat{\theta} = 1 - \hat{F}_Y(\hat{\tau})$, with $\hat{F}_X^{-1}(u)$ denoting the sample u -quantile of X .

REMARK 2. Other measures which can be used to describe the performance of a diagnostic test are likelihood ratios. The positive likelihood ratio of the 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 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.

Therefore, likelihood ratios are functions of the sensitivity and the specificity of the test. More specifically, for the positive likelihood ratio λ^+ , we have

$$\lambda^+(\tau) = \frac{\theta(\tau)}{1 - \eta(\tau)}.$$

For the negative likelihood ratio λ^- , we have

$$\lambda^-(\tau) = \frac{1 - \theta(\tau)}{\eta(\tau)}.$$

Then, we can express the sensitivity θ and the specificity η as functions of λ^+ and λ^-

$$\theta = \frac{\lambda^+(1 - \lambda^-)}{\lambda^+ - \lambda^-}, \quad \eta = \frac{\lambda^+ - 1}{\lambda^+ - \lambda^-}.$$

Hence, the quantity (3) can be formulated in terms of λ^+ , λ^- and τ and can be used to obtain nonparametric confidence regions, for example for the pair (positive likelihood ratio, negative likelihood ratio) at a fixed cut-off level τ_0 .

REMARK 3. In some circumstances, the test values indicating a higher likelihood of disease are the lowest ones. In these cases, for a given cut-off level τ , the sensitivity and the specificity of the test are

$$\theta(\tau) = \Pr\{Y \leq \tau\} = F_Y(\tau),$$

$$\eta(\tau) = \Pr\{X > \tau\} = 1 - F_X(\tau),$$

respectively. Then, it is easy to show that the function $\ell(\theta, \eta, \tau)$ takes the form

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

It is worth noting that one could change the signs of the variables X and Y (and, therefore, of the observed data and of the cut-off level τ), and then use (3).

3 Some simulation results

In this section, we report the results of a simulation study carried out to evaluate the goodness of the χ^2_2 approximation to the tail of the distribution of the quantity

$\ell(\theta, \eta, \tau)$, so that the finite-sample accuracy of the confidence regions (for every parameter pairs) obtained by (3) can be assessed.

In the simulation work, we have considered three different parametric models for generating the data, i.e. the test responses of non-diseased and diseased patients: (i) a Gaussian model ($N(\mu, \sigma)$), (ii) a Beta model ($Be(\gamma, \beta)$), and (iii) an Exponential model ($Exp(\gamma)$). To perform the simulation experiments, random samples for the X and the Y values have been generated, respectively, from a $N(0, 1)$ and a $N(\mu, 1/2)$ in scenario (i), from a $Be(1, 3)$ and a $Be(\gamma, 1/3)$ in scenario (ii) and from a $Exp(1)$ and a $Exp(\gamma)$ in scenario (iii), for various combinations of the parameters μ and γ . Each simulation experiment is based on 10000 replications.

For three levels of nominal coverage $1 - \alpha$, Tables 1, 2 and 3 give the estimated coverage probabilities of the confidence regions obtained by using the proposed empirical likelihood method in scenario (i), (ii) and (iii), respectively. The reported values for μ in scenario (i) and γ in scenarios (ii) and (iii), together with the cut-off value τ_0 , depend on the choice of reference values (θ_0, η_0) for the true pair (sensitivity, specificity). To exemplify, consider scenario (ii), where $X \sim Be(1, 3)$ and $Y \sim Be(\gamma, 1/3)$, and choose for the pair (θ_0, η_0) the value $(0.95, 0.90)$. In this case, $\tau_0 = F_X^{-1}(0.90) = 0.536$ and γ is fixed so as to satisfy $0.95 = 1 - F_Y(0.536)$.

The tables report the simulation results for different settings of the sample sizes m and n . For each reference pair (θ_0, η_0) , the setting showing the smallest sample sizes has been chosen so as to present actual coverage probabilities reasonably close to the nominal ones. Clearly, due to the nonparametric nature of the proposed technique, sample sizes matching this criterion strongly depend on the reference values (θ_0, η_0) , needing to be larger for reference values approaching 1. In fact, for instance, in the previous exemplification of scenario (ii), with $(\theta_0, \eta_0) = (0.95, 0.90)$ and sample sizes $m = n = 20$, the method generates confidence regions with very poor actual coverages (0.536, 0.540, 0.558, for the nominal levels 0.90, 0.95, 0.99, respectively). The poor performance is due, in particular, to the high frequency of replications in which the statistic (3) is not finite. Of course, the occurrence of this phenomenon tends to be neglectable with increasing sample sizes, as tables' results show.

Overall, the simulation experiments achieve satisfying results. As expected, the accuracy of the method increases with increasing sample sizes. Moreover, the smallest sample sizes reported in the tables may be indicative of the effective sizes needed to guarantee sufficiently accurate coverages.

Table 1: Actual coverage probabilities of the $1 - \alpha$ confidence regions obtained by the empirical likelihood statistic $\ell(\theta, \eta, \tau)$. In each simulation experiment, the m test responses for non-diseased subjects are generated from a $N(0, 1)$ and the n test responses for diseased subjects are generated from a $N(\mu, 1/2)$.

τ_0	μ	η_0	θ_0	m	n	$1 - \alpha$		
						0.99	0.95	0.90
0.842	1.263	0.8	0.80	20	20	0.971	0.939	0.899
				50	20	0.981	0.945	0.894
				20	50	0.980	0.943	0.892
				50	50	0.990	0.948	0.897
				100	100	0.989	0.953	0.904
			0.90	30	30	0.947	0.913	0.873
				50	30	0.951	0.927	0.883
				30	50	0.986	0.944	0.880
				50	50	0.988	0.945	0.892
				100	100	0.990	0.951	0.896
			0.95	65	65	0.959	0.930	0.894
				100	65	0.957	0.930	0.885
				65	100	0.987	0.951	0.890
				100	100	0.986	0.951	0.893
				150	150	0.988	0.949	0.899
1.281	1.702	0.9	0.80	30	30	0.953	0.915	0.877
				50	30	0.986	0.941	0.878
				30	50	0.957	0.933	0.888
				50	50	0.988	0.947	0.890
				100	100	0.988	0.951	0.897
			0.90	40	40	0.966	0.941	0.892
				65	40	0.981	0.948	0.888
				40	65	0.976	0.940	0.883
				65	65	0.989	0.949	0.873
				100	100	0.988	0.946	0.894
			0.95	75	75	0.970	0.927	0.889
				100	75	0.973	0.940	0.900
				75	100	0.986	0.943	0.889
				100	100	0.986	0.947	0.890
				150	150	0.989	0.943	0.895

Table 2: Actual coverage probabilities of the $1 - \alpha$ confidence regions obtained by the empirical likelihood statistic $\ell(\theta, \eta, \tau)$. In each simulation experiment, the m test responses for non-diseased subjects are generated from a $Be(1, 3)$ and the n test responses for diseased subjects are generated from a $Be(\gamma, 1/3)$.

τ_0	γ	η_0	θ_0	m	n	$1 - \alpha$		
						0.99	0.95	0.90
0.416	0.857	0.8	0.80	20	20	0.971	0.942	0.905
				50	20	0.980	0.946	0.894
				20	50	0.980	0.945	0.893
				50	50	0.991	0.951	0.896
				100	100	0.988	0.950	0.893
			0.90	30	30	0.948	0.917	0.885
				50	30	0.951	0.926	0.879
				30	50	0.987	0.950	0.880
				50	50	0.989	0.946	0.881
				100	100	0.988	0.947	0.890
			0.95	65	65	0.960	0.926	0.889
				100	65	0.956	0.929	0.887
				65	100	0.985	0.950	0.891
				100	100	0.987	0.952	0.884
				150	150	0.990	0.950	0.901
0.536	1.117	0.9	0.80	30	30	0.951	0.916	0.875
				50	30	0.985	0.944	0.882
				30	50	0.954	0.928	0.882
				50	50	0.987	0.946	0.890
				100	100	0.990	0.949	0.896
			0.90	40	40	0.968	0.946	0.893
				65	40	0.976	0.944	0.886
				40	65	0.978	0.944	0.893
				65	65	0.989	0.952	0.878
				100	100	0.991	0.954	0.900
			0.95	75	75	0.971	0.934	0.896
				100	75	0.973	0.940	0.898
				75	100	0.985	0.944	0.891
				100	100	0.987	0.949	0.892
				150	150	0.986	0.940	0.895

Table 3: Actual coverage probabilities of the $1 - \alpha$ confidence regions obtained by the empirical likelihood statistic $\ell(\theta, \eta, \tau)$. In each simulation experiment, the m test responses for non-diseased subjects are generated from a $Exp(1)$ and the n test responses for diseased subjects are generated from a $Exp(\gamma)$.

τ_0	γ	η_0	θ_0	m	n	$1 - \alpha$		
						0.99	0.95	0.90
1.609	0.138	0.8	0.80	20	20	0.973	0.946	0.904
				50	20	0.981	0.945	0.896
				20	50	0.981	0.946	0.900
				50	50	0.991	0.948	0.895
				100	100	0.990	0.948	0.898
				30	30	0.952	0.918	0.883
				50	30	0.952	0.925	0.881
				30	50	0.986	0.945	0.881
				50	50	0.987	0.946	0.889
				100	100	0.989	0.948	0.894
	0.065	0.90	0.90					
				30	30	0.952	0.918	0.883
				50	30	0.952	0.925	0.881
				30	50	0.986	0.945	0.881
				50	50	0.987	0.946	0.889
				100	100	0.989	0.948	0.894
	0.032	0.95	0.95	65	65	0.959	0.926	0.885
				100	65	0.956	0.933	0.892
				65	100	0.985	0.951	0.888
				100	100	0.983	0.946	0.886
				150	150	0.987	0.947	0.894
2.302	0.097	0.9	0.80	30	30	0.954	0.920	0.879
				50	30	0.989	0.946	0.882
				30	50	0.950	0.923	0.879
				50	50	0.988	0.943	0.888
				100	100	0.991	0.947	0.894
				40	40	0.968	0.943	0.896
				65	40	0.978	0.946	0.890
				40	65	0.977	0.940	0.892
				65	65	0.987	0.951	0.881
				100	100	0.989	0.950	0.898
	0.046	0.90	0.90					
				40	40	0.968	0.943	0.896
				65	40	0.978	0.946	0.890
				40	65	0.977	0.940	0.892
				65	65	0.987	0.951	0.881
				100	100	0.989	0.950	0.898
	0.022	0.95	0.95	75	75	0.968	0.930	0.893
				100	75	0.973	0.941	0.901
				75	100	0.984	0.938	0.885
				100	100	0.986	0.947	0.886
				150	150	0.986	0.940	0.897

4 Two illustrations

4.1 Predicting recovery from severe head injuries

Creatinine kinase brain isoenzyme (CK-BB) is the most plentiful isoenzyme found in the brain. Clinical use of CK-BB, in relation to head injuries, suggests that an elevation of serum CK-BB can predict poor recovery or death after a severe head injury. Our first dataset (Hans *et al.*, 1985) presents levels of CK-BB in 60 patients measured on the first day after suffering a severe head trauma: 19 patients eventually showed an acceptable recovery, and 41 a poor recovery or no recovery.

For these data, Zhou *et al.* (2002) and Zhou and Qin (2005) provide nonparametric confidence intervals for the sensitivity of CK-BB in predicting poor recovery, at a fixed 0.90 level of its specificity. In particular, Zhou and Qin (2005), adopting a resampling procedure, suggest to use (0.429, 0.836) as 0.95 confidence interval for the sensitivity of the test based on the serum CK-BB. Moreover, Zhou *et al.* (2002) apply the method in Schafer (1994) to find a decision threshold value that gives, with a confidence of 0.95, a specificity of at least 0.90 and a sensitivity of at least 0.312.

Proper joint inference on the pair (sensitivity, cut-off level) at a fixed specificity value requires the construction of bivariate confidence regions. To tackle joint inference on sensitivity and cut-off level of CK-BB in predicting poor recovery, at a fixed 0.90 specificity value, we apply the technique described in Section 2. Figure 1 shows the contour curves of $\ell(\theta, 0.90, \tau)$, giving confidence regions for the pair (sensitivity, cut-off level) at different coverage rates. The solid lines correspond to the standard coverages 0.90, 0.95 and 0.99. The plot shows that, at the nominal 0.95 confidence level, there is a wide variety of pairs of values for (θ, τ) compatible with the target 0.90 specificity, some of which are markedly far away from the point estimate $(\hat{\theta}, \hat{\tau}) = (0.561, 281)$. This indicates a wide variability in the data, partly due to the small sample size, in particular for subjects showing a good recovery. Evidently, compared to the marginal confidence interval (0.429, 0.836) provided in Zhou and Qin (2005), the region given in Figure 1 provides a lot of additional information. For example, it shows that, at the nominal 0.95 confidence level, the highest values for the sensitivity, compatible with the target specificity, can be attained for cut-off values ranging roughly between 150 and 200.

Figure 2 adds to the 0.95 confidence region for the pair (sensitivity, cut-off level) at the fixed 0.90 level of specificity, the plot of the 0.95 confidence region for the pair (specificity, cut-off level) at the fixed 0.80 level of sensitivity. The overlapping area gives an indication about cut-off levels which are compatible with both a specificity of 0.90 and a sensitivity of 0.80. The inspection of such an area can drive the choice of the cut-off value. In particular, values of τ ranging from 150

Figure 1: Severe head injuries data. Contour curves of $\ell(\theta, 0.90, \tau)$, giving confidence regions for the pair (sensitivity, cut-off level) at the fixed 0.90 level of specificity. The solid lines correspond to the standard coverages 0.90, 0.95, 0.99.

to 200 appear to be all compatible with both a specificity of 0.90 and a sensitivity of 0.80. The choice of the cut-off value, then, may depend on the primary role of the test in the clinical practice. If the point of the test was to rule out poor recovery, sensitivity would be the parameter deserving crucial consideration. In this case, a small value of τ would be preferable, as it would give rise to higher levels of sensitivity. In the general practice, inspection of figures like the one just shown in this example can allow refinements of the commonly used cut-off levels.

The left panel in Figure 3 shows the contour curves of $\ell(\theta, \eta, 200)$, highlighting the joint 0.95 confidence region for the pair (sensitivity, specificity) when the cut-off level is set equal to 200. The right panel of the figure shows the anal-

Figure 2: Severe head injuries data. In red, the 0.95 confidence region for the pair (sensitivity, cut-off level) at the fixed 0.90 level of specificity. In black, the 0.95 confidence region for the pair (specificity, cut-off level) at the fixed 0.80 level of sensitivity.

ogous 0.95 confidence region when the cut-off value is fixed at 281, i.e., the non-parametric point estimate value. For the two choices of the cut-off level, the plots show, with a confidence of 0.95, the various performance regimes, in terms of sensitivity and specificity, of the CK-BB serum levels for prediction of poor recovery. It can be seen, in particular, that, when the cut-off is set to 281, the specificity and the sensitivity never fall below 0.733 and 0.371, respectively, whereas, when the cut-off is set to 200, the specificity and sensitivity never fall below 0.516 and 0.494, respectively.

Figure 3: Severe head injuries data. Contour curves of $\ell(\theta, \eta, \tau)$, giving the confidence regions for the pair (sensitivity, specificity) when the cut-off level is set equal to 200 (left panel) and to 281 (right panel). The solid lines correspond to the standard coverages 0.90, 0.95, 0.99.

4.2 Detecting carriers of Duchenne muscular dystrophy

Duchenne muscular dystrophy (DMD) is one of the most prevalent types of muscular dystrophy and is characterized by a rapid progression of muscle degeneration that occurs early in life. It is a genetically transmitted disease, passed from a mother to her children. Unfortunately, no cure has yet been discovered, so the screening of females who could be potential carriers is of great importance. Our second dataset (Andrews and Herzberg, 1985) concerns serum pyruvate kinase (PK) activity measured on 67 female carriers and on 127 healthy female controls. It is known from the literature that carriers tend to show increased serum levels of PK activity. This suggests using PK levels as a diagnostic test in detecting carriers of DMD.

In order to evaluate the performance of the test based on PK levels, we applied the method proposed in Section 2. In particular, Figure 4 shows the nominal 0.95 confidence regions for both the pair (sensitivity, cut-off level) and the pair (specificity, cut-off level), at the fixed 0.80 level of specificity and sensitivity, respectively.

The plot shows that a cut-off level compatible with the chosen reference value 0.80 for the sensitivity and specificity is around 15.5, which is slightly higher than the value (15.0) which is frequently found in the medical literature. For such

Figure 4: Duchenne muscular dystrophy data. In red, the 0.95 confidence region for the pair (sensitivity, cut-off level) at the fixed 0.80 level of specificity. In black, the 0.95 confidence region for the pair (specificity, cut-off level) at the fixed 0.80 level of sensitivity.

a cut-off level, Figure 5 shows, in the left panel, the 0.90, 0.95 and 0.99 confidence regions for the pair (sensitivity, specificity) of the test. The simulation results given in the paper lead us to think that the actual coverages of the regions are very close to the nominal ones under these working conditions.

The right panel zooms to the interesting area of the plot and compares the 0.95 confidence region based on $\ell(\theta, \eta, 15.5)$ with the nominal 0.95 rectangular confidence region obtained by the asymptotic normality of the nonparametric estimators $(\hat{\eta}, \hat{\theta})$, with logistic transformation. Rectangular regions are frequently used for their simplicity, although they impose artificial symmetry constraints to

Figure 5: Duchenne muscular dystrophy data. Contour curves of $\ell(\theta, \eta, 15.5)$, giving the confidence regions for the pair (sensitivity, specificity) when the cut-off level is set equal to 15.5. The solid lines correspond to the standard coverages 0.90, 0.95, 0.99. The right panel zooms and compares the 0.95 confidence region based on $\ell(\theta, \eta, 15.5)$ with the nominal 0.95 rectangular confidence region obtained by the asymptotic normality of the nonparametric estimators $(\hat{\eta}, \hat{\theta})$, with logistic transformation.

the shape of the confidence region. The picture shows that our elliptical region is smaller than the rectangular one and, evidently, the two confidence regions could lead to different decisions on hypotheses on the pair (sensitivity, specificity). For example, if the null hypothesis $H_0 : (\theta, \eta) = (0.8, 0.85)$ was tested, this hypothesis would be rejected - at the 5% significance level - according to our confidence region, whereas it would be accepted - at the same significance level - on the basis of the rectangular confidence region.

5 Conclusions

In this article, we have tackled the construction of nonparametric bivariate confidence regions for pairs of quantities defining the accuracy of a diagnostic test. In particular, we have proposed a technique based on an empirical likelihood statistic that allows, within a unified framework, to make inference about the pair (sensitivity, cut-off level), at a fixed value of specificity, as well as about the pair (speci-

ficity, cut-off level), at a fixed value of sensitivity, or about the pair (sensitivity, specificity), at a fixed cut-off value.

The proposed technique works under very weak assumptions and it is easy to implement, being essentially based on the computation of the empirical distribution functions from the data. The simulation study shows that the derived confidence regions are accurate for moderate to high sample sizes. Of course, for small to moderate sample sizes, the accuracy is greatly influenced by the true values for the pair (sensitivity, specificity).

Due to the non-smoothness of $\hat{F}_X$ and $\hat{F}_Y$, the confidence regions have rough contours when the the cut-off level τ is one of the two parameters of interest. However, one could easily obtain smooth regions by using kernel-based estimators of F_X and F_Y .

An R code implementing the proposed technique is available from the authors.

References

- Adimari, G. (1998): An empirical likelihood statistic for quantiles. *J. Statist. Comput. Simul.* **60**, 85–95.
- Adimari, G. (2006): Nonparametric confidence intervals for the area under the ROC curve (in Italian). *Statistica* **65**, 39–49.
- Adimari, G., M. Chiogna (2006): Partially parametric interval estimation of $\Pr\{Y > X\}$. *Comp. Statist. Data Analysis* **51**, pp. 1875–1891.
- Andrews, D.F., A.M. Herzberg (1985): *Data*. Springer-Verlag, New York.
- Claeskens, G., B.Y. Jing, L. Peng, W. Zhou (2003): Empirical likelihood confidence regions for comparison distributions and ROC curves. *The Canadian Journal of Statistics* **31**, pp. 173–190.
- Hans, O., A. Albert, J. Born, J. Chapelle (1985): Derivation of a bioclinical index in severe head trauma. *Intensive Care Med.* **11**, pp. 186–191.
- Kim, J., G.S. Qin (2008): Logit-transformation based confidence intervals for the sensitivity of a continuous-scale diagnostic test. *International Conference on BioMedical Engineering and Informatics*, pp. 768–772.
- Owen, A.B. (2001): *Empirical likelihood*. Chapman and Hall, London.
- Pepe, M.S. (2003): *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press.
- Pepe, M.S., T. Cai, (2004): The analysis of placement values for evaluating discriminatory measures. *Biometrics* **60**, pp. 528–535.
- Qin, G.S., X.H. Zhou (2006): Empirical likelihood inference for the area under the ROC curves. *Biometrics* **62**, pp. 613–622.

- Qin, G.S., A.E. Davis, B.Y. Jing (2009): Empirical likelihood-based confidence intervals for the sensitivity of a continuous-scale diagnostic test at a fixed level of specificity. *Statistical Methods in Medical Research*, forthcoming article, doi:10.1177/0962280209105512.
- Schafer, H. (1994): Efficient confidence bounds for ROC curve. *Stat. Med.* **13**, pp. 1551–1561.
- Zhou, X.H., N.A. Obuchowski, D.M. McClish (2002): *Statistical methods in diagnostic medicine*. Wiley & Son.
- Zhou, X.H., G.S. Qin (2005): Improved confidence intervals for the sensitivity of a continuous-scale diagnostic test at a fixed level of specificity. *Statistics in Medicine* **24**, pp. 465–477.