Estimation of the Youden Index and it's associated cutoff point

Ronen Fluss[*]**, David Faraggi,** and **Benjamin Reiser**

Department of Statistics, University of Haifa 31905, Israel

Summary

The Youden Index is a frequently used summary measure of the ROC (Receiver Operating Characteristic) curve. It both, measures the effectiveness of a diagnostic marker and enables the selection of an optimal threshold value (cutoff point) for the marker. In this paper we compare several estimation procedures for the Youden Index and its associated cutoff point. These are based on (1) normal assumptions; (2) transformations to normality; (3) the empirical distribution function; (4) kernel smoothing. These are compared in terms of bias and root mean square error in a large variety of scenarios by means of an extensive simulation study. We find that the empirical method which is the most commonly used has the overall worst performance. In the estimation of the Youden Index the kernel is generally the best unless the data can be well transformed to achieve normality whereas in estimation of the optimal threshold value results are more variable.

Key words: Diagnostic markers, Kernel smoothing, Power transformation, Sensitivity Specificity

1 Introduction

The ROC (Receiver Operating Characteristic) curve is a popular graphical method of displaying the discriminatory accuracy of a marker (diagnostic test) for distinguishing between two populations. It is used in many scientific areas such as: radiology (Metz, 1989), psychiatry (Hsiao et al., 1989), epidemiology (Aoki et al., 1997) and manufacturing inspecting systems (Somoza et al., 1990). Recently there has been an increased use of the ROC curve for biomedical problems examining the effectiveness of continuous diagnostic markers in distinguishing between diseased and healthy individuals (Strike, 1995; Shapiro, 1999; Greiner et al., 2000). A person is assessed as diseased (positive) if the tested marker value is greater than a given threshold value, otherwise the subject is diagnosed as healthy (negative). The accuracy of any given threshold value can be measured by the probability of a true positive (sensitivity) and the probability of a true negative (specificity).

The ROC curve is a plot of the sensitivity (Se(c)) versus 1-specificity (1-Sp(c)) over all possible threshold values (c) of the marker. To evaluate the discriminatory ability of a marker it is common to summarize the information of the ROC curve into a single global value or index. Several such indices are found in the literature and have been used in various applications (Shapiro, 1999; Greiner et al., 2000).

Although the area under the ROC curve (AUC) is the most commonly used global index of diagnostic accuracy the Youden Index (Youden, 1950) is also frequently used in practice (see for example Aoki et al., 1997;Grmec & Gasparovic, 2001). This index can be defined as $J = \max_c \{Se(c) + Sp(c) - 1\}$ and ranges between 0 and 1. Complete separation of the distributions of the marker values for the diseased and healthy populations results in J=1 whereas complete overlap gives J=0. The Youden Index has an attractive feature not present in the AUC. J provides a criterion for choosing the "optimal" threshold value (c*), the threshold value for which Se(c)+Sp(c)-1 is maximized (Greiner et al., 2000). There are other criteria for obtaining an optimal threshold such as: efficiency, misclassification-cost, odds ratio and the kappa index (Greiner et al., 2000) which are not considered further here. The YI is the easiest to apply and does not require further information such as prevalence rates and decision error costs.

[*] Corresponding author: e-mail: rfluss@stat.haifa.ac.il, Phone: +00 972 48 345 539, Fax: +00 972 48 253 849

There are several approaches for estimating the ROC curve and consequently it's associated Youden Index. Faraggi and Reiser (2002) examined two parametric and two non-parametric approaches for estimating the AUC. We review these procedures and study their application to estimating J and c*. These procedures differ due to various ways of estimating the cumulative distribution functions (cdf's) of the marker values, based on sample data taken from both the healthy and diseased groups. Using these estimated cdfs we can estimate Se and Sp for any c and thereby J. The first non-parametric method, referred to as the empirical method (EMP), estimates the cdf of the marker with the empirical cdf of the sample. This is a very popular method due to its simplicity. The second non-parametric method uses a kernel smoothing function on the sample cdf. The Kernel method (K) has the advantage of providing a smooth estimate of the cdf. The first parametric approach (N), assumes that the marker values for both healthy and diseased populations follow the normal distribution and estimates J using standard parametric methods. The second parametric approach (TN), assumes that a monotone transformation exists so that the transformed marker values follow the normal distribution. After estimating the transformation and applying it to the data normal theory is used. As a by-product of estimating J by each of these approaches a corresponding estimate of c* is obtained.

Since different estimation methods will provide different estimated J and c* values for the same data a comparison of their performances needs to be carried out. Faraggi and Reiser (2002) used an extensive simulation study to examine the effectiveness of the different approaches of estimating the AUC. They studied the bias and the root mean square error (RMSE) of different estimation procedures. They found the TN method to be the preferred approach unless the marker values follow a bimodal distribution. In this case the K procedure seemed to be best. In our study we want to extend their work by comparing the effectiveness of the different approaches (N,TN,K,EMP) in estimating the Youden Index (J) and its associated optimal threshold (c*). In Section 2 we provide the estimation formulae for J and c* based on the four different methods. Faraggi and Reiser (2002) illustrate AUC estimation with an example of a blood serum marker for muscular dystrophy. We use the same data to illustrate the estimation of J and c* in Section 3. These four estimation methods are compared in Section 4 through an extensive simulation study. A concluding discussion in Section 5 completes this paper.

2 Estimation Methods

Suppose that results on a diagnostic test (marker) $x_1, x_2, \ldots, x_m$ and $y_1, y_2, \ldots, y_n$ are available from two random samples on the diseased and healthy populations having cumulative distribution functions G_D and F_H respectively. For any given threshold c, $Se(c)=1-G_D(c)$ and $Sp(c)=F_H(c)$. Therefore the Youden Index is

$$J=\max_c\{Se(c)+Sp(c)-1\}=\max_c\{F_H(c)-G_D(c)\}. \quad (1)$$

The value of c that achieves this maximum will be considered the optimal threshold c*. The estimation of J is carried out by estimating G_D and F_H ($\hat{G}_D, \hat{F}_H$) and substituting these estimates in equation (1) i.e. $\hat{J} = \max_c\{\hat{F}_H(c) - \hat{G}_D(c)\}$. As mentioned above the different approaches of estimating G_D and F_H, will result in different estimates of J and *c**. We discuss several estimation procedures below.

2.1 Parametric methods

Let *X* and *Y* denote the diagnostic marker measurements for the diseased and healthy subjects respectively. A simple parametric approach for estimating J is to assume that both *X* and *Y* have independent normal distributions with different variances $X \sim N(\mu_D, \sigma_D^2)$ and $Y \sim N(\mu_H, \sigma_H^2)$. Without loss of generality assume that $\mu_D > \mu_H$ (otherwise take the negative of the marker values). Consequently

$$Se(c) = 1 - \Phi\left(\frac{c-\mu_D}{\sigma_D}\right),\ Sp(c) = \Phi\left(\frac{c-\mu_H}{\sigma_H}\right),\ S(c) = \Phi\left(\frac{c-\mu_H}{\sigma_H}\right) - \Phi\left(\frac{c-\mu_D}{\sigma_D}\right) \quad \text{and}$$

$$J \equiv \max_c \{S(c)\}. \tag{2}$$

where Φ is the standard normal cumulative distribution function. In order to carry out the maximization in (2) we compute the first derivative of S(c), set it to zero and solve the resulting quadratic equation. The root

$$c^* = \frac{\left(\mu_D \sigma_H^2 - \mu_H \sigma_D^2\right) - \sigma_H \sigma_D \sqrt{(\mu_H - \mu_D)^2 + \left(\sigma_H^2 - \sigma_D^2\right)\log\left(\sigma_H^2 / \sigma_D^2\right)}}{\left(\sigma_H^2 - \sigma_D^2\right)} \tag{3}$$

can be shown to provide the maximum. When assuming equal variances, $\sigma_H^2 = \sigma_D^2 = \sigma^2$, $c^* = (\mu_D + \mu_D)/2$. Substituting c* in formula (2) we obtain the solution

$$J = \Phi\left(\frac{c^* - \mu_H}{\sigma_H}\right) - \Phi\left(\frac{c^* - \mu_D}{\sigma_D}\right) \tag{4}$$

J and c* are estimated by substituting for the unknown parameters μ_D, μ_H, σ_D, σ_H in formulae (3) and (4) their corresponding samples means and standard deviation. Estimating c* and J in this manner will be referred to as the Normal method (N). This normality assumption for the marker values will be questionable in many cases (e.g. the DMD data).

A less restrictive approach assumes that there exists some monotonic transformation t(.) such that t(*X*) and t(*Y*) are normally distributed. Note that the ROC curve is invariant under such a transformation. After applying the transformation to the sample data, we can use the N estimates (c^*_t, J_t) for the transformed data. Estimation for the original sample data follows with: $\hat{J} = \hat{J}_t$, and $\hat{c}^* = t^{-1}(\hat{c}_t^*)$.

For using this method one must define t(.). Recently several authors (Zou and Hall, 2000; Zou and Hall, 2002; Faraggi and Reiser, 2002; O'Malley and Zou, 2002) have recommended using the data to fit a power transformation of the Box-Cox type:

$$t(y) = y^{(\lambda)} = \begin{cases} (y^\lambda - 1)/\lambda & \lambda \neq 0 \\ \log(y) & \lambda = 0 \end{cases}$$

The above authors have found this approach to be useful for a wide variety of situations. Based on the assumption that $y^{(\lambda)}$ and $x^{(\lambda)}$ follow the normal distribution, λ can be estimated using the Maximum Likelihood Estimation procedure. Using the resulting $\hat{\lambda}$ to transform the data and applying the Normal method gives the Transformed Normal (TN) procedure.

2.2 Non-parametric methods

The simplest non-parametric approach uses the empirical cdfs as estimates of the cdfs of *X* and *Y*. These are known to be consistent estimates (Knight, 2000). The empirical cdfs can be written as:

$$\hat{G}_D(c) = \frac{1}{m}\sum_{i=1}^{m} I(x_i \le c), \quad \hat{F}_H(c) = \frac{1}{n}\sum_{i=1}^{n} I(y_i \le c) \quad \text{where} \quad I(u \le c) = \begin{cases} 1 & u \le c \\ 0 & u > c \end{cases}$$

resulting in

$$\hat{J} = \max_c \left\{\hat{F}_H(c) - \hat{G}_D(c)\right\} \quad c \in \{x_1, ..., x_m, y_1, ..., y_n\} \tag{5}$$

For estimating c* we examine two possible approaches: (i) The observation where the maximum was found.(ii) Suppose both samples are merged and sorted in an ascending order denoted by: $d_1, \ldots, d_{m+n}$ and that the maximum was reached at d_j. Since the value of $\hat{J}$ is constant in the interval $[d_j, d_{j+1})$ then it is reasonable to take $(d_j + d_{j+1})/2$ as an estimate of c*. In our simulations (Section 4) we found only small

differences between these two procedures with a slight preference to the second approach. Consequently we only present results for the second approach which we denote by EMP.

The Kernel method (K) is another non-parametric estimating method that uses a smoothing kernel function on the empirical cdfs. Following Zou et al. (1998) we use the Gaussian kernel function $\hat{\mathrm{F}}_{\mathrm{H}}(t) = \frac{1}{n}\sum_{i=1}^{n}\Phi\left(\frac{t-y_i}{h_y}\right)$, $\hat{G}_{\mathrm{D}}(t) = \frac{1}{m}\sum_{i=1}^{m}\Phi\left(\frac{t-x_i}{h_x}\right)$. Further, to control the amount of smoothing, we use the bandwidths $h_y = 0.9\cdot\min\left\{s_y, iqr_y/1.34\right\}\cdot n^{-0.2}$ and $h_x = 0.9\cdot\min\left\{s_x, iqr_x/1.34\right\}\cdot m^{-0.2}$ where $s_y = \left(\frac{1}{n-1}\sum_{1}^{n}(y_i-\bar{y})^2\right)^{0.5}$ is the standard deviation, and $\mathrm{iqr_y}$ is the inter quartile range of the healthy sample. We define $\mathrm{s_x}$ and $\mathrm{iqr_x}$ similarly for the diseased sample. This choice of the bandwidth has been recommended by Silverman (1986) as doing "very well for a wide range of densities". Faraggi and Reiser (2002) compared several different bandwidth procedures in the context of AUC estimation and found that more complex procedures did not lead to any improvement. The Youden Index estimated by the Kernel method is

$$\hat{J} = \max_c\left\{\hat{F}_{\mathrm{H}}(c) - \hat{G}_{\mathrm{D}}(c)\right\} \tag{6}$$

Iterative numerical methods are used to find the maximum in (6) as well as the maximizing threshold (c*).

3. Example: Duchene Muscular Dystrophy

Duchene muscular dystrophy (DMD) is a progressive recessive disorder passed from a mother to her children. With the lack of an effective treatment for the disease the child dies at an early age, and therefore screening of potential female carriers is of great interest. Percy et al. (1982) discuss data gathered on four different markers as part of a program to develop an effective screening procedure. Complete data (Andrews and Herzberg, 1985) are available on these markers for n=127 blood serum samples from a healthy female control group and m=67 samples from carriers. Faraggi and Reiser (2002) considered only the Creatine Kinase (CK) marker and showed that these marker values are non-normally distributed for both the control and carrier groups while the CK values taken to the power of -0.34 are much more normal like. The power coefficient was obtained from the Box-Cox method of estimating transformations (see Section 2.1)

We applied the four different methods for estimating the Youden Index (J) and the optimal cutoff point (c*) to the CK data. The resulting estimates and their bootstrap standard errors (in parentheses) are presented in Table 1.

Table 1 Estimation of J and c* for CK.

	N	TN	EMP	K
J	0.665 (0.03)	0.613 (0.05)	0.612 (0.06)	0.591 (0.04)
c*	82.07 (5.35)	58.12 (3.94)	56.50 (10.36)	73.36 (8.42)

Figure 1(a) presents the ROC curve for the CK marker calculated according to the different methods and indicating where J is obtained. The TN and EMP procedures give quite similar ROC curves. Figure 1(b) gives Se(c)+Sp(c)-1 as a function of c for these methods again indicating the point on each curve where the Youden Index is obtained. Due to the different shapes involved it is clear that even similar J's can have quit different c*'s. The Normal method produces an ROC curve quite different than the others. Note that in Figure 1(b) the curve for the K procedure has a large plateau with a wide range of c values having quite imilar Se(c)+Sp(c)-1 values. This is also reflected in the large standard error of c* for the K method. In order to better understand these differences and to compare them an extensive simulation study was carried out and is reported in Section 4.

4. Simulation studies

We wish to evaluate and compare the finite sample properties of the four different estimators of the Youden Index (J) and its corresponding cutoff point (c*) discussed above namely N, TN, EMP, K. A simulation study was performed and these methods were compared in terms of root mean square error (RMSE) and bias. The normal assumption will frequently not be appropriate for marker data. We consider the N method in order to examine its robustness and to provide a baseline comparison for the TN procedure.

The simulations cover a wide variety of different distributional shapes i.e. symmetric, skewed and bimodal situations often seen in real data (see for example Goddard and Hinberg, 1990). These distributional shapes are similar to those used in Faraggi and Reiser (2002) who examined several AUC estimation methods. Figure 2 presents some of these distributions, standardized to give a Youden Index of J=0.8. A detailed description of the distributions and the appropriate formulae for J and c* used in the simulations is presented in the Appendix and their parameters are presented in Table 2.

Table 2: Parameter values of models used in simulation study[2]

Distribution					μ_H		σ^2_H	σ^2_D	μ_D (corresponding to J)			
									0.2	0.4	0.6	0.8
Normal equal variances					6.5		0.25	0.25	6.753	7.024	7.342	7.782
Normal non equal variances					6.5		0.09	0.25	6.617	6.873	7.143	7.505
Normal$^{-1/3}$					3.5		0.09	0.25	3.383	3.127	2.857	2.495
Lognormal					2.5		0.09	0.25	2.617	2.873	3.143	3.505
Gamma (ν, r instead of μ, σ^2)					0.5		2	2	0.344	0.23	0.142	0.072
Mix Models	P_D	P_H	μ_{H1}	σ^2_{H1}	μ_{H2}	σ^2_{H2}	σ^2_{D1}	σ^2_{D2}	μ_{D1} (adjusted to J) $\mu_{D2}=\mu_{D1}+4$			
									0.2	0.4	0.6	0.8
mix1	0.5	1	10	1	-	-	1	5	7.64	9.98	11.12	12.16
mix2	10	1	-	-	1	5	1	0.8	10.22	10.85	11.53	12.44
mix3	10	1	13	1	1	5	0.5	0.5	10.28	12.39	13.58	14.74
mix4	10	1	13	1.5	1	5	0.5	0.5	10.55	12.24	13.58	14.91
mix5	0.8	0.5	10	1	13	1	1	5	11.4	12.64	13.9	15.00
mix6	0.8	0.5	10	1	13	1.5	1	5	11.37	12.51	13.86	15.16

We used several choices of J (0.2, 0.4, 0.6, 0.8) spanning a practical range of Youden Index values along with sample sizes n=m=20,50,100. Following a referee's suggestion we also considered the following unequal sample sizes: (m,n) = (20,60), (50,150) (100,300). For brevity we do not report on these unbalanced cases since the comparative results for both J and c* are essentially equivalent to the reported below for the equal sample sizes. For the purpose of estimating the RMSE and bias, we used a thousand simulations of each scenario. The simulations and the computation of the estimators were programmed using the R statistical software package (Dalgaard, 2002). For each simulated data sets J and c* were estimated by all the four methods described above and were used to compute the estimated RMSE and bias of each method.

4.1 Simulation results – The Youden Index

4.1.1 Simulations with normal distributions

[2] See definition of mixed models in Appendix.

We first consider samples from normal populations. We examined both equal and non-equal variances, but since the overall results for both cases where quite similar we present only the latter (Table 3). For this Table μ_D, the expected value of the diseased population is chosen to correspond to the J value. The Table indicates that all four methods show a decrease in bias as sample size increases. N, TN and K have similar very small biases although K's bias is usually negative while the others have a positive bias. The EMP method is, in terms of bias, the worst method with substantially higher values of bias. N and TN have similar values of RMSE for all levels of separation, regardless of sample sizes. For J=0.2 K has lower RMSE values than N and TN whereas for J=0.8 K has slightly higher values than N and TN. The EMP method always has the highest RMSE which can be as much as 70% more than the smallest RMSE.

Table 3 Bias and RMSE for J estimators: Y~N(6.5, 0.09), X~N(μ_D, 0.25)

	20				50				100			
J	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
						Bias						
N	-0.037	0.010	0.001	0.002	-0.014	0.003	0.005	-0.001	-0.003	-0.001	0.001	-0.001
TN	-0.047	0.007	0.003	0.005	-0.020	0.002	0.005	0.001	-0.003	-0.001	0.001	0.000
EMP	0.101	0.092	0.073	0.053	0.059	0.055	0.048	0.032	0.040	0.035	0.030	0.023
K	0.016	-0.005	-0.024	-0.029	0.004	-0.011	-0.020	-0.028	0.001	-0.012	-0.021	-0.023
						RMSE						
N	0.127	0.116	0.102	0.072	0.075	0.074	0.064	0.046	0.054	0.051	0.045	0.033
TN	0.130	0.118	0.103	0.071	0.076	0.075	0.064	0.046	0.054	0.051	0.045	0.033
EMP	0.178	0.160	0.136	0.097	0.110	0.099	0.086	0.064	0.079	0.071	0.059	0.046
K	0.109	0.113	0.106	0.086	0.073	0.075	0.070	0.057	0.055	0.054	0.053	0.044

4.1.2 Simulations with skewed distributions

The similarity between the N and TN procedures is not surprising for normal data. We would expect a greater difference for skewed distributions. We produced skewed data by first generating normal variates which were then taken to the power of -3. The corresponding probability density functions are presented in Figure 2(c). The power -3 was chosen to correspond to the power transformation (-.34 ≈ -1/3) found to be best for the CK data. The simulated data is examined in Table 4.

Table 4 Bias and RMSE for J estimators: $Y^{-1/3}$~N(3.5,0.09), $X^{-1/3}$~N(μ_D, 0.25)

	20				50				100			
J	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
						Bias						
N	0.097	0.107	0.057	-0.044	0.118	0.110	0.045	-0.069	0.122	0.108	0.038	-0.082
TN	-0.034	0.014	0.020	0.011	-0.015	0.007	0.007	0.001	-0.003	0.002	0.002	0.002
EMP	0.100	0.090	0.080	0.056	0.063	0.054	0.047	0.033	0.041	0.036	0.030	0.024
K	0.037	0.023	0.006	-0.033	0.022	0.013	-0.005	-0.037	0.014	0.007	-0.006	-0.030
						RMSE						
N	0.192	0.151	0.088	0.076	0.150	0.126	0.062	0.085	0.138	0.115	0.050	0.094
TN	0.189	0.128	0.102	0.074	0.118	0.072	0.064	0.047	0.066	0.050	0.045	0.033
EMP	0.154	0.150	0.133	0.096	0.096	0.093	0.085	0.063	0.066	0.065	0.059	0.045
K	0.114	0.115	0.093	0.076	0.073	0.072	0.061	0.059	0.052	0.050	0.045	0.044

For this highly skewed scenario, the N method, which assumes normality, not surprisingly usually performs worst in terms of bias. Further this bias does not decrease as the sample size increases. The TN and K methods have very similar low biases which are practically 0 when sample sizes are large. For

small sample sizes the EMP method's bias is slightly higher than that of N but for n≥50 it is lower although still not as good as K and TN.

In parallel with the bias results, N has higher RMSE values than TN and is usually the worst method. However, for J=0.6 N is similar to both TN and K leaving EMP as the worst method. The performances of TN and K vary accordingly to J. When J =0.2 K has the best performance. When J=0.4 and 0.6 they perform equally whereas for J=0.8 TN has RMSE values lower than K. EMP is usually better than N and worse than both K and TN except for J=0.2 where it is usually better than TN. In this skewed situation all methods perform, in terms of bias and RMSE, similarly to the normal scenario discussed in Section 4.1.1 except for the N procedure which is much worse here.

We next examine skewed data obtained from a log-normal distribution (Figure 2(d)). Table 5 provides conclusions similar to those given for Table 4 with the N procedure being an exception. For J≤0.6 in the log-normal scenario, N performs slightly better in both bias and RMSE. However, N is still worse than EMP for large sample sizes. When J=0.8 N is significantly better becoming the best method.

Table 5 Bias and RMSE for J estimators: log(Y)~N(2.5,0.09), log(X)~N(μ_D, 0.25)

	20				50				100			
J	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
						Bias						
N	0.075	0.088	0.052	-0.005	0.088	0.088	0.047	-0.012	0.089	0.089	0.046	-0.016
TN	-0.033	0.013	0.015	0.010	-0.017	0.006	0.004	0.004	-0.009	0.004	0.002	0.001
EMP	0.102	0.093	0.076	0.056	0.058	0.056	0.044	0.034	0.037	0.037	0.028	0.022
K	0.034	0.019	-0.003	-0.027	0.016	0.009	-0.011	-0.028	0.008	0.006	-0.011	-0.026
						RMSE						
N	0.180	0.139	0.096	0.059	0.126	0.110	0.069	0.039	0.109	0.100	0.058	0.180
TN	0.187	0.121	0.106	0.075	0.118	0.074	0.065	0.046	0.075	0.052	0.047	0.187
EMP	0.154	0.150	0.132	0.096	0.095	0.096	0.082	0.063	0.066	0.068	0.059	0.154
K	0.113	0.112	0.098	0.077	0.073	0.073	0.063	0.054	0.052	0.052	0.047	0.113

It is not surprising that TN does reasonably well in Tables 4-5 since the distributions used fall in the Box-Cox transformation family. We examined a simulation scenario which is not in this family, namely the gamma distribution. The shape of the distribution is indicated in Figure 2(e). The results were very similar to the log-normal case and are not presented for brevity.

4.1.3 Simulations with mixtures of normal distributions

In order to consider additional scenarios not obtained from the Box-Cox family we examined mixtures of two normal distributions which result in bimodality. We examined different mixtures with different degrees of bimodality. The parameters of the distributions considered are described in the Appendix and the parameter values used are given in Table 2.

The results for Bias and RMSE of mix1,3,4 are quite similar as are those of mix2,5,6 (See Table 2 for definitions). Note that the first set has a stronger bimodality in the diseased population than the second set (Figure 2). For brevity we only present the simulation results for mix1 (Table 6) and mix2 (Table 7).

In Table 6 both N and TN have a very similar pattern in the bias results. When the level of separation is low (J=0.2) they are very biased and as the separation increases the bias decreases significantly. For J=0.4 and 0.6 TN is better than N. K has an opposite pattern. For J≤0.6 it is practically unbiased with the best performance of all methods but for J=0.8 it has more bias (about 0.05) than N or TN. For N, TN and K the bias is not affected by sample size. The bias of EMP decreases as the sample size increases and no pattern is found in connection to the separation level (J values). N and TN have similar RMSEs, with

Table 6 Bias and RMSE for J estimators: mix1

	20		50		100	

J	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
						Bias						
N	-0.331	0.104	0.067	-0.021	-0.376	0.109	0.061	-0.020	-0.417	0.103	0.062	-0.019
TN	-0.447	0.050	0.047	-0.017	-0.490	0.052	0.036	-0.018	-0.509	0.046	0.036	-0.019
EMP	0.057	0.076	0.088	0.057	0.037	0.047	0.052	0.035	0.022	0.025	0.033	0.025
K	0.004	0.028	-0.002	-0.057	0.000	0.024	-0.012	-0.054	-0.004	0.014	-0.012	-0.046
						RMSE						
N	0.436	0.141	0.100	0.062	0.438	0.125	0.077	0.043	0.449	0.112	0.071	0.033
TN	0.518	0.126	0.100	0.066	0.515	0.093	0.068	0.046	0.517	0.070	0.054	0.036
EMP	0.119	0.134	0.134	0.094	0.073	0.087	0.086	0.062	0.052	0.057	0.060	0.046
K	0.087	0.108	0.086	0.091	0.055	0.073	0.059	0.071	0.042	0.050	0.044	0.057

very high values when J=0.2 (about 0.5), which decrease as J increases reaching about 0.05 when J=0.8. As with bias, K has the best performance when J≤0.6 but has higher RMSE values when J=0.8. EMP's RMSE pattern is similar to its bias pattern.

Table 7 Bias and RMSE for J estimators: mix2

	20				50				100			
J	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
						Bias						
N	0.135	0.074	0.002	-0.055	0.148	0.071	-0.006	-0.063	0.149	0.070	-0.007	-0.067
TN	0.067	0.037	-0.001	-0.033	0.066	0.024	-0.016	-0.042	0.064	0.021	-0.018	-0.047
EMP	0.135	0.113	0.089	0.062	0.083	0.067	0.051	0.037	0.055	0.043	0.034	0.023
K	0.056	0.009	-0.025	-0.041	0.031	-0.007	-0.030	-0.038	0.018	-0.011	-0.026	-0.033
						RMSE						
N	0.184	0.123	0.076	0.084	0.165	0.093	0.050	0.074	0.158	0.081	0.034	0.073
TN	0.142	0.112	0.087	0.077	0.097	0.069	0.058	0.060	0.081	0.049	0.043	0.056
EMP	0.176	0.167	0.137	0.099	0.111	0.104	0.087	0.063	0.077	0.070	0.059	0.044
K	0.118	0.113	0.100	0.091	0.075	0.073	0.073	0.063	0.054	0.052	0.052	0.049

For mix2 TN is usually less biased than N. For n>20 N usually is the most biased method except for J=0.6 where both N and TN are the optimal methods and virtually unbiased. K has little bias and except for J=0.6 it is usually the least biased method. EMP bias values are very similar to those found in mix1. For small sample sizes (n= 20) it is the worst method but improves for n>20 being usually less biased than N and for J=0.8 it is the least biased.

For mix2 the relative performances of the methods in terms of RMSE are similar to the bias results. TN usually has lower values than N with the exception of J=0.6 where N has the lowest RMSE of all methods. There is no consistency in the relative performances of the methods. Generally for J≤0.4 K exhibits good performance. For J=0.8 and sample size larger than 20 TN, K and EMP all are very similar.

Compared to mix1, N and TN show a large improvement in RMSE when J=0.2 (about 0.35 difference), but as J increases the improvement decreases becoming slightly worse than in mix1 when J=0.8. EMP has higher values when J≤0.4 but is almost identical for larger J. TN does well overall for mix2 having the lowest RMSE in five cases and being reasonably close to the lowest in the other seven. K also performs well and is frequently similar to TN.

4.2 Simulation results – The cutoff point

Although J and c* are strongly related a good method for estimating the one is not necessarily good for the other since most properties of estimators, such as bias, are not preserved under non-linear monotonic transformations. In this Section we examine our findings on the performances of the different estimation methods in estimating the cutoff point c*. In the following Tables the row, labelled c*, lists the values of the correct cutoff points as calculated using the true parameter values, corresponding to the appropriate values of J.

4.2.1 Simulations with normal distributions

The results for the normal case are presented in Table 8.

Table 8 Bias and RMSE for c* estimators: Y~N(6.5, 0.09), X~N(μ_D, 0.25)

	20				50				100			
J	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
c*	6.829	6.806	6.850	6.950	6.829	6.806	6.850	6.950	6.829	6.806	6.850	6.950
						Bias						
N	-0.150	-0.009	-0.003	-0.002	-0.055	-0.005	-0.002	-0.002	-0.017	0.000	-0.001	0.001
TN	-0.167	-0.021	-0.014	-0.008	-0.067	-0.011	-0.008	-0.006	-0.016	-0.003	-0.005	-0.001
EMP	-0.054	-0.024	-0.016	-0.009	-0.025	-0.024	-0.012	-0.007	-0.011	-0.009	-0.005	-0.008
K	0.027	0.016	0.015	0.013	0.025	0.009	0.008	0.009	0.024	0.013	0.010	0.007
						RMSE						
N	0.286	0.081	0.064	0.066	0.184	0.050	0.039	0.040	0.105	0.035	0.029	0.029
TN	0.300	0.091	0.071	0.069	0.199	0.056	0.047	0.043	0.099	0.038	0.034	0.031
EMP	0.201	0.155	0.130	0.129	0.144	0.118	0.100	0.095	0.111	0.092	0.080	0.076
K	0.201	0.127	0.100	0.092	0.119	0.084	0.067	0.059	0.088	0.062	0.050	0.043

All methods improve as sample sizes increase and for n>20 the bias of all methods is very small. When J=0.2, N and TN are the most biased methods especially when sample sizes are small (n=20) while K and EMP are the least biased. However for J≥0.4, N has the lowest bias and TN is just slightly more biased than N. K and EMP have similar biases. N has the lowest RMSE while TN has similar values. EMP has the highest RMSEs. For J=0.2 generally both N and TN have the highest RMSEs whereas K has the lowest.

4.2.2 Simulations with skewed distributions

In Table 9 we examine skewed data generated from normal data transformed by the power transformation -3.All methods except for N show improvement in bias with increasing sample size. N always has the largest bias except for n=20 and J=0.2 where it has the smallest bias whereas TN has an exceptionally high bias. For J≥0.4 both TN and EMP are the best performing methods with very similar biases. However for J=0.2 TN has a larger bias than EMP. For J≥0.4, TN has the lowest RMSE whereas N has the highest. K and EMP perform similarly in terms of RMSE. When J=0.2 N is generally the best performing method in terms of RMSE while all other methods are similar.

The results for both the log-normal and gamma distribution were very similar and therefore are not presented.

Table 9[*] Bias and RMSE for c* estimators: $Y^{-1/3}$~N(3.5,0.09), $X^{-1/3}$~N(μ_D, 0.25)

[*] RMSE, bias and c* in this table should be multiplied by 10^{-2}

	20				50				100			
J	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
c*	3.136	3.069	3.201	3.525	3.136	3.069	3.201	3.525	3.136	3.069	3.201	3.525
						Bias						
N	0.035	0.350	0.435	0.363	0.163	0.405	0.476	0.420	0.194	0.426	0.510	0.450
TN	-0.310	-0.025	0.003	0.010	-0.138	-0.004	-0.005	0.000	-0.036	-0.009	0.001	0.005
EMP	-0.086	-0.045	-0.003	0.022	-0.061	-0.017	-0.019	-0.009	-0.028	-0.008	0.006	-0.014
K	0.175	0.135	0.204	0.236	0.085	0.102	0.154	0.177	0.055	0.074	0.130	0.159
						RMSE						
N	0.436	0.459	0.537	0.517	0.279	0.447	0.518	0.498	0.238	0.454	0.537	0.490
TN	0.610	0.254	0.219	0.244	0.417	0.146	0.140	0.156	0.234	0.110	0.102	0.114
EMP	0.543	0.419	0.399	0.448	0.433	0.316	0.308	0.318	0.337	0.258	0.246	0.262
K	0.650	0.370	0.384	0.450	0.423	0.246	0.261	0.303	0.286	0.192	0.211	0.249

4.2.3 Simulations with mixtures of normal distributions

In Mix1, N and TN have the largest bias which does not improve with increasing sample size. For J≤0.4 N has lower bias values than TN while for J≥0.6 it is vice versa. Overall, EMP is the least biased procedure. When J=0.4 K is the best performing in terms of bias but otherwise it is second best. In terms of RMSE the relative performances are very varied. When J=0.2 both N and TN have very high RMSE values. N has smaller RMSE than TN when J≤0.4 and vice-versa for J≥0.6. For J=0.2 K has the lowest RMSE regardless of sample size. For J≥0.6 TN is usually the best method. When J=0.4, N generally has the lowest RMSE and K is second best.

Table 10 Bias and RMSE for c* estimators: mix1

	20				50				100			
J	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
c*	11.73	11.66	11.1	11.37	11.73	11.66	11.1	11.37	11.73	11.66	11.1	11.37
						Bias						
N	-2.241	-0.280	0.386	0.296	-2.568	-0.258	0.409	0.300	-2.844	-0.255	0.412	0.308
TN	-2.568	-0.472	0.140	0.126	-2.915	-0.446	0.171	0.139	-3.051	-0.434	0.187	0.145
EMP	-0.045	-0.343	-0.011	0.019	-0.046	-0.203	-0.020	-0.008	-0.029	-0.138	-0.005	-0.008
K	0.161	-0.094	0.310	0.192	0.126	-0.077	0.274	0.140	0.108	-0.093	0.236	0.111
						RMSE						
N	2.578	0.398	0.464	0.403	2.816	0.310	0.441	0.341	2.988	0.282	0.428	0.329
TN	2.785	0.582	0.348	0.327	3.001	0.493	0.271	0.232	3.080	0.458	0.239	0.197
EMP	0.593	0.819	0.489	0.379	0.435	0.597	0.369	0.293	0.334	0.471	0.298	0.222
K	0.484	0.511	0.506	0.386	0.327	0.338	0.381	0.251	0.245	0.267	0.304	0.185

For the mix2 scenario N is usually the most biased method with bias increasing with sample size. TN performs better than N and its bias does decrease slightly with an increase in sample size. EMP generally has the lowest bias. No other consistent pattern is apparent in bias. TN usually has the lowest RMSE. K is usually better than EMP which is the most biased when J=0.2,0.8.

Table 11 Bias and RMSE for c* estimators: mix2

	20				50				100			
J	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
c*	10.91	10.66	10.9	11.31	10.91	10.66	10.9	11.31	10.91	10.66	10.9	11.31
						Bias						
N	0.186	0.484	0.365	0.133	0.28	0.543	0.383	0.155	0.296	0.558	0.403	0.164
TN	-0.207	0.049	-0.021	-0.096	-0.116	0.056	-0.048	-0.106	-0.102	0.055	-0.045	-0.111
EMP	-0.260	-0.007	0.005	-0.007	-0.159	-0.024	-0.022	-0.010	-0.121	0.012	-0.007	-0.004
K	0.163	0.174	0.098	0.023	0.185	0.113	0.056	0.003	0.180	0.104	0.042	0.006
						RMSE						
N	0.464	0.581	0.453	0.269	0.346	0.575	0.421	0.214	0.329	0.574	0.420	0.196
TN	0.543	0.340	0.254	0.230	0.327	0.241	0.187	0.177	0.244	0.187	0.138	0.150
EMP	0.831	0.489	0.368	0.331	0.711	0.384	0.296	0.265	0.621	0.333	0.244	0.207
K	0.964	0.541	0.340	0.256	0.736	0.339	0.217	0.165	0.618	0.275	0.152	0.121

5. Discussion and final remarks

From the results of our study it is clear that none of the four methods examined is superior, in terms of bias and RMSE, for all the distributional scenarios considered. Furthermore, a good performance in estimating J did not necessarily indicate a good performance in estimating the corresponding c*. This is not surprising since most properties of estimators, such as bias, are not preserved under non-linear monotonic transformations. Generally, in all scenarios, the performances of the methods for estimating both J and c* are highly dependant on the correct value of the Youden Index, in addition to the effect of the distributional shape.

In estimating the Youden Index we found that the EMP procedure, which is the most commonly used method (see for example: Greiner et al., 2000; Grmec, 2001), has the worst performance and is not recommended unless sample size is very large. For data which is unimodal and has a normal distribution N, TN and K perform well. When data does not have a normal distribution N has higher bias and RMSE values than TN and K. The K procedure was found to be the optimal method for small to moderate levels of separation ($J \leq 0.6$) especially for small sample size. When the level of separation is high TN is the best performing procedure. For moderate separation and large sample sizes ($n \geq 50$) TN is similar to K. In bimodal situations K is generally the best method for $J \leq 0.6$ while TN is the optimal method for $J=0.8$. Apparently for well separated populations the details of the distributional forms are not of great importance and TN will provide a reasonable solution. This is in spite of the fact that the transformation to normality is not effective for bimodal distributions.

In estimating the cutoff point both EMP and N are generally worse than TN and K. When the distribution of the marker values is unimodal TN was generally found to have the best performance. For bimodal distributions the results are ambivalent. Generally for a high level of separation ($J=0.8$) K is usually best or nearly best and for moderate level of separation ($J=0.4, 0.6$) TN is usually the optimal procedure. For small level of separation no preference was found.

Re-examining the Duchene data of Section 3, note that as the normal assumption is untenable it is not surprising that N provides estimates of J and c* quite different from the others. The estimates of the four methods, all indicate that the correct value of J is around 0.6. Similar to our findings on the J estimates in the skewed data case in Section 4 (Table 4) we find that for $J=0.6$ and $n \geq 50$ N is considerably higher than TN while K is slightly lower. EMP is practically identical to TN whereas in our simulation study EMP has a larger bias than TN. This can be a result of the difference between a single sample estimate and a mean of a thousand samples estimates. Our simulations suggest that TN should be used in this situation. Our simulations for estimating c* for skewed distributions (Table 9) indicate that N has the largest bias followed by K with a considerably smaller bias and then TN and EMP with negative values.

Note that generally in these cases TN has the smallest (in absolute value) bias. This relative ordering is also found for the CK data (Table 1). Since TN tended to have the smallest RMSE in our simulations we conclude that the TN estimate of c* be used for the CK data.

Since the K method was found to have good properties in many distributional scenarios it may be useful to examine further the optimal smoothing function and bandwidth to use for estimating J and c*. Faraggi and Reiser (2002) discussed bandwidth selection only in the context of AUC estimation. Recently Hall and Hyndman (2003) have discussed improving bandwidth selection when estimating ROC curves. It would be interesting to examine if the procedures they studied improve the K method for estimating J and c*.

In this work we have dealt with the point estimation of J and c*. A few studies have considered confidence intervals for the Youden Index and the corresponding cutoff point. Barkan (2001) discusses confidence intervals using the EMP method and Faraggi (2003) considers confidence intervals using the N method. Both dealt with bootstrap based intervals. Our work indicates that as point estimates K and TN are to be preferred, depending on the scenario. This suggests that it would be useful to examine confidence intervals based on these methods.

Appendix

Distributions used in study

a) The normal distributions calculations are described in detail in Section 2.1.

b) Skewed distributions:

i) The power transformation to normality: $Y^a \sim N(\mu_H, \sigma^2_H)$ and $X^a \sim N(\mu_D, \sigma^2_D)$

It can readily be seen that the optimal threshold after transformation (c_a^*) can be obtained from (3) while the corresponding J is given by (4) using c_a^* instead of c*. The optimal threshold in the original scale is $c^* = (c_a^*)^{1/a}$

ii) The lognormal follow the same pattern only with log(.) instead of $(.)^a$.

iii)The Gamma distribution: $Y \sim G(\nu_H, r)$ and $X \sim G(\nu_D, r)$ where $f_G(u;\nu,r) = \Gamma^{-1}(r)\nu^r x^{r-1} e^{-\frac{x}{\nu}}$. It can readily be seen that

$$c^* = r\frac{\log(\nu_D) - \log(\nu_H)}{\nu_D - \nu_H}$$

$$J = F_H\left(c^*; r, \nu_H\right) - G_D\left(c^*; r, \nu_D\right)$$

where F and G are Gamma cdfs. c* was obtained by finding the maximum defined in formula (1) of J in Section 3.

c) The mixture of normal distributions: The mixture probability density function can be written as:

$$f(x;\mu_1,\sigma_1,\mu_2,\sigma_2,p) = \frac{p}{\sqrt{2\pi}\sigma_1}\exp\left\{\frac{-(x-\mu_1)^2}{2\sigma_1^2}\right\} + \frac{1-p}{\sqrt{2\pi}\sigma_2}\exp\left\{\frac{-(x-\mu_2)^2}{2\sigma_2^2}\right\}$$

Using subscripts H and D on the parameters to distinguish the two populations we obtain:

$$J = \max_c\left\{p_H \cdot \Phi\left(\frac{c-\mu_{H1}}{\sigma_{H1}}\right) + (1-p_H)\Phi\left(\frac{c-\mu_{H2}}{\sigma_{H2}}\right) - p_D \cdot \Phi\left(\frac{c-\mu_{D1}}{\sigma_{D1}}\right) + (1-p_D)\Phi\left(\frac{c-\mu_{D2}}{\sigma_{D2}}\right)\right\}.$$

There is no closed form for calculating J and its corresponding *c**. These must be found by numerical methods. We used a combination of Newton Raphson and a simple linear search to obtain these results.

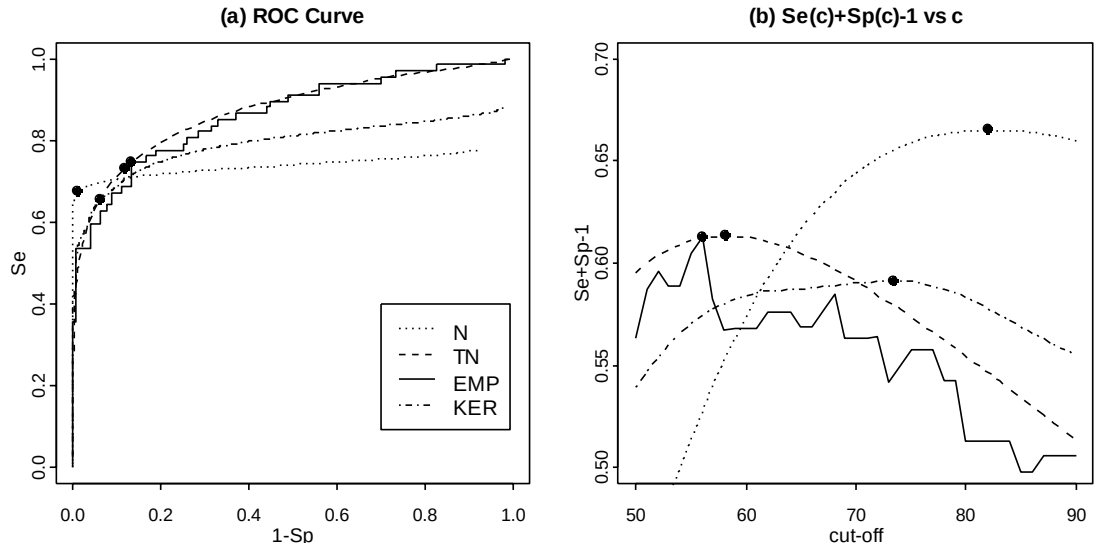

Fig. 1 Analysis of CK marker data indicating the location where Youden Index is obtained

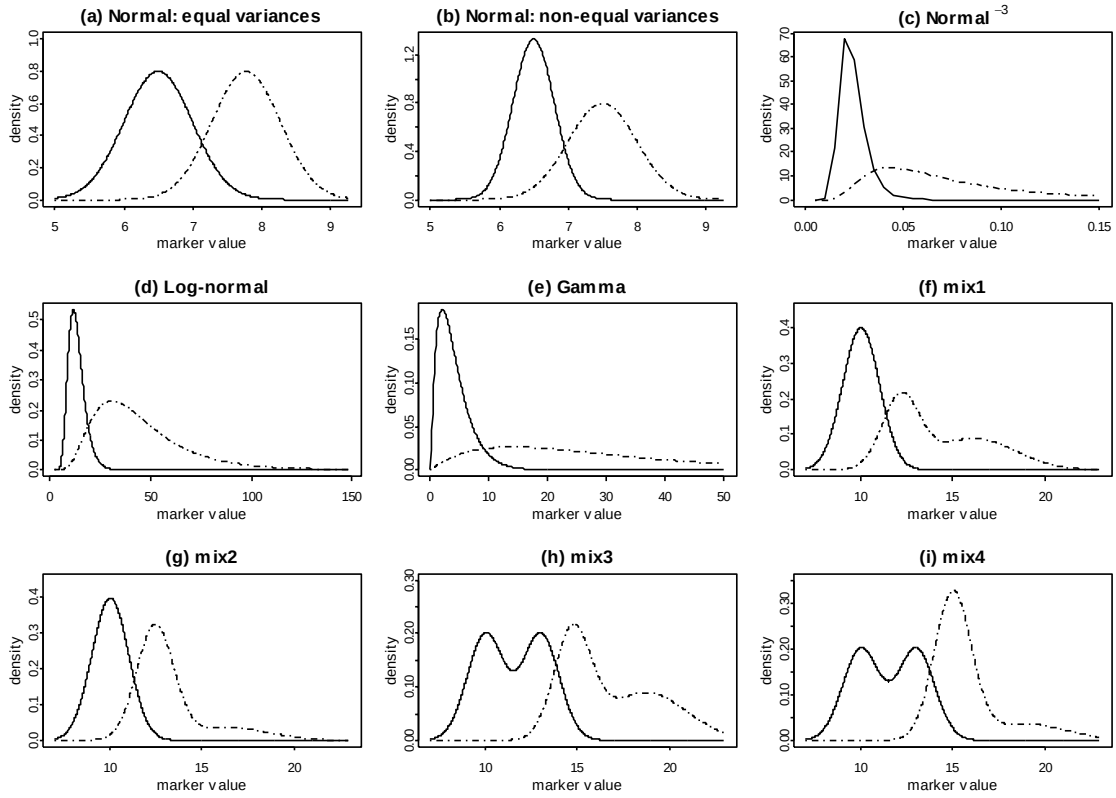

Fig. 2 Distributions[1, 2] used in the simulation study with J=0.8

[1] Solid line: healthy ; dashed line: diseased.
[2] See definition of mixed models in Table 2.

Acknowledgements: The authors thank the referee for his helpful comments.

References

Andrews, D.F. and Herzberg, A.M. (1985). Data. Springer-Verlag, New York.

Aoki, K., Misumi, J., Kimura, T., Zhao, W. and Xie, T. (1997). Evaluation of cutoff levels for screening of gastric cancer using serum pepsinogens and distributions of levels of serum pepsinogens I, II and of PG I/PG II ratios in a gastric cancer case-control study. *Journal of Epidemiology* **7** 143-151.

Barkan, N. (2001). Statistical inference on "r·specificity+sensitivity" (In Hebrew), PhD Thesis, Haifa University.

Dalgaard, P. (2002). Introductory Statistics with R. Springer, New York.

Faraggi, D. and Reiser, B. (2002). Estimation of the area under the ROC curve. *Statistics in Medicine* **21**, 3093-3106.

Faraggi, D. (2003). Adjusting receiver operating characteristic curves and related indices for covariates. *The Statistician* **52**, Part 2, 179-192.

Goddard, M.J. and Hinberg, I. (1990) Receiver Operator Characteristic (ROC) curves and non-normal data: an empirical study. *Statistics in Medicine* **9**, 325-337.

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

Grmec, S. and Gasparovic, V. (2001). Comparison of APACHE II, MEES and Glasgow Coma Scale in patients with nontraumatic coma for prediction of mortality. *Critical Care* **5**, 19-23.

Hall, P.G. and Hyndman, R.J. (2003). Improved methods for bandwidth selection when estimating ROC curves. *Statistics & Probability Letters* **64**, 181-189.

Hsiao, J.K., Barko, J.J. and Potter, W.Z. (1989). Diagnosing diagnoses: receiver operating characteristic methods and psychiatry. *Archives of General Psychiatry* **46**, 664-667.

Knight, K. (2000). Mathematical Statistics. Chapman &Hall/CRC.

Metz, C.E. (1989). Some practical issues of experimental design and data analysis in radiological ROC studies. *Investigation Radiology* **24**, 234-245.

O'Malley, A. and Zou, K.H. (2002). Baysian hierarchical transformation models for receiver operating characteristic curve analysis. *ASA Proceeding of the Joint Statistical Meetings*, 2541-2549.

Percy, M.E., Andrews D.F. and Thompson M.W. (1982). Duchene muscular dystrophy carrier detection using logistic discrimination: serum creatine kinase, hemopexin, pyravite kinase and lactate dehydrogeneze in combination. *American Journal of Medical Genetics* **13**, 27-38.

Shapiro, D.E. (1999). The interpretation of diagnostic tests. *Statistical Methods in Medical Research* **8**, 113:134.

Silverman, B.W. (1986). Density Estimator for Statistics and Data Analysis. Chapman and Hall, London.

Somoza, E., Mossman, D. and McFeeters, L. (1990). The info-ROC technique: a method for comparing and optimizing inspection systems. In *Review of Progress in Quantitative Nondestructive Evaluation*, Thomson DO, Chimenti DE (eds). Plenum Press, New York.

Strike, P.W. (1995). Measurement in laboratory medicine: a primer control and interpretation. Butterworth-Heinemann, Oxford.

Youden, W.J. (1950). Index for rating diagnostic tests. *Cancer* **3**, 32-35.

Zou, K.H., Tempany, C.M., Fielding, J.R. and Silverman, S.G. (1998). Original smooth receiver operating characteristic curves estimation from continuous data: statistical methods for analysing the predictive value of spiral CT of urethral stones. *Academic Radiology* **5**, 680-687.

Zou, K.H. and Hall W.J. (2000). Two transformation models for estimating an ROC curve derived from continuous data. *Journal of Applied Statistics* **27 (5)**, 621-631.

Zou, K.H. and Hall W.J. (2002). Semiparametric and parametric transformation models for comparing diagnostic markers with paired design. *Journal of Applied Statistics* **29 (6)**, 803-816.