STATISTICS IN MEDICINE
Statist. Med. 2002; **21**:3093–3106 (DOI: 10.1002/sim.1228)

Estimation of the area under the ROC curve

David Faraggi* and Benjamin Reiser

Department of Statistics, University of Haifa, Haifa 31905, Israel

SUMMARY

The area under the receiver operating characteristic curve is frequently used as a measure for the effectiveness of diagnostic markers. In this paper we discuss and compare estimation procedures for this area. These are based on (i) the Mann–Whitney statistic; (ii) kernel smoothing; (iii) normal assumptions; (iv) empirical transformations to normality. These are compared in terms of bias and root mean square error in a large variety of situations by means of an extensive simulation study. Overall we find that transforming to normality usually is to be preferred except for bimodal cases where kernel methods can be effective. Copyright © 2002 John Wiley & Sons, Ltd.

KEY WORDS: diagnostic markers; kernel density estimation; Mann–Whitney statistic; power transformation

1. INTRODUCTION

This paper compares several methods of estimating the area under the receiver operating characteristic (ROC) curve for continuous diagnostic markers.

Plotting the ROC curve is a popular way of displaying the discriminatory accuracy of a diagnostic test for detecting whether or not a patient has a disease or condition. ROC methodology is derived from signal detection theory [1] where it is used to determine if an electronic receiver is able to satisfactory distinguish between signal and noise. It has been used in medical imaging and radiology [2], psychiatry [3], non-destructive testing [4] and manufacturing inspection systems [5].

Recently there has been an increased use of ROC curves for assessing the effectiveness of continuous diagnostic markers in distinguishing between diseased and healthy individuals [6–9]. A person is assessed as diseased (positive) or healthy (negative) depending on whether the corresponding marker value is greater than or less than or equal to a given threshold value. Associated with any threshold value is the probability of a true positive (sensitivity) and the probability of a true negative (specificity). The theoretical ROC curve is a plot of $q = \text{sensitivity}$ versus $p = 1\text{-specificity}$ for all possible threshold values. ROC curves can be estimated under parametric or non-parametric assumptions [9].

* Correspondence to: David Faraggi, Department of Statistics, University of Haifa, Haifa 31905, Israel

Received April 2000
Accepted November 2001

The most commonly used global index of diagnostic accuracy is the area under the ROC curve (AUC). Let X and Y denote the diagnostic marker measurements for diseased and healthy subjects, respectively. Bamber [10] showed that $\mathrm{AUC} = \mathrm{Prob}(X > Y)$. This can be interpreted as the probability that in a randomly selected pair of healthy and diseased individuals the diagnostic marker value is higher for the diseased subject. Values of AUC close to 1.0 indicate that the marker has high diagnostic accuracy. This index arises in many problems not connected to diagnostic markers [11, 12]. For example, interpreting X as the strength of a mechanical system to which stress Y is applied, gives $\mathrm{Prob}(X > Y)$ as the resulting reliability. Wolfe and Hogg [13] recommend the use of this index as a general measure for the differences between two distributions.

In this paper we review two parametric and two non-parametric approaches for estimating the AUC. The non-parametric approaches are (i) use of the Mann–Whitney statistic (MW); (ii) fit a smooth ROC curve using kernel smoothing and then estimate the AUC by integration (K). The parametric approaches are (i) assume that the marker values for both healthy and diseased populations follow the normal distribution and then estimate the AUC using standard parametric methods (N); (ii) apply a Box–Cox type [14] power transformation to the data and after obtaining the appropriate transformation use normal theory (NT). For the kernel method we use the standard normal kernel and consider two different approaches to setting the bandwidth. In addition we examine the use of data transformation before applying the kernel. The above methods have all been suggested in the literature and are described in Section 3.

Different estimation methods will of course provide different estimated AUC values. Goddard and Hinberg [7] examine data for several markers on prostate cancer. They estimated the AUC using methods MW and N as well as log transforming the data and then applying N (a special case of NT). They found that the MW and N methods will in some cases differ considerably while applying N after a log transformation provided results closer to MW. A further example is discussed in Section 2. Since the different estimation methods can provide a span of estimated AUC values on the same data set, their properties needed to be examined in order to provide a recommendation as to which approach is to be preferred.

Estimating the AUC using the MW approach follows naturally from estimating the ROC curve as a step-function based on empirical cumulative distribution functions. The other approaches are obtained by calculating the ROC curve as a smooth function, the AUC being estimated as the area under this smooth curve. It has been argued [6, 15] that it is advantageous to use a smooth ROC curve estimate. When this is used it is natural to use the corresponding AUC estimator. Although the MW estimator is known to be unbiased, it might be hoped that some of the alternative approaches could be effectively unbiased and may provide some gain in efficiency as measured by root mean square error. Consequently the approaches MW, K, N and NT are compared in terms of bias and root mean square error (RMSE) by means of an extensive simulation study which is discussed in Section 4. Another approach (RC) due to Metz *et al.* [16] assumes that some unknown monotone transformation of the marker values for both the healthy and diseased populations results in normally distributed variables. They provide a complex algorithm which first discretizes the continuous data into I categories (with I usually quite large) and then numerically optimizes a likelihood function over $I + 1$ parameters. A computer program (ROCKIT) which carries out this procedure can be obtained at http://www-radiology.uchicago.edu/krl/toppage11.htm#software. This method is compared to the other methods in Section 4.4. The example is re-examined in Section 5 while Section 6 provides a concluding discussion.

2. EXAMPLE: DUCHENNE MUSCULAR DYSTROPHY DATA

Duchenne muscular dystrophy (DMD) is a progressive recessive muscle disorder passed from a mother to her children. With the lack of an effective treatment the screening of females as potential carriers is of great importance. Percy *et al.* [17] discuss data gathered on four different markers as part of a program to develop an effective method of screening. Complete data are available on 127 serum samples from healthy females controls (Y) and 67 samples from carriers (X) [18]. We consider for illustrative purposes only the measurements of blood serum creatine kinase (CK). Figure 1(a) provides a histogram of this data. The non-normality of the

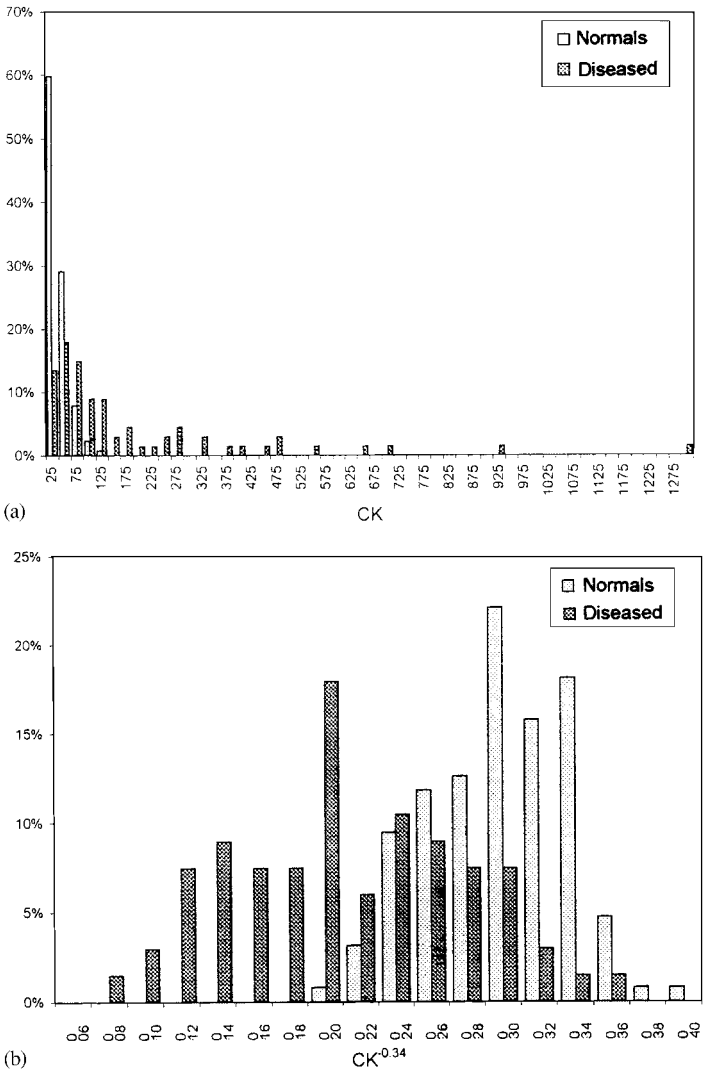

Figure 1. Histograms of CK: (a) before transformation; (b) after transformation.

data is quite apparent. The CK values taken to the power of -0.34 are shown in Figure 1(b). This power was obtained by applying the Box–Cox method of estimating transformations (see Section 3.2). Clearly the transformed data appear more symmetrically distributed. Applying the AUC estimation methods MW, N, NT, K1, K1T, K2, K2T and RC to the CK data results in the estimated area being 0.870, 0.740, 0.873, 0.787, 0.852, 0.843, 0.857 and 0.875, respectively. K1 and K2 denote the kernel method using two different bandwidth calculations. When these kernel methods are applied to the transformed CK data they are referred to as K1T and K2T, respectively.

There are large differences between these methods. In order to better understand the differences between these methods and to help decide between them we carried out a simulation study which is reported in Section 4.

3. ESTIMATION OF THE AUC

Suppose that diagnostic test results $x_1,\ldots,x_m$ and $y_1,\ldots,y_n$ are available from the diseased and healthy population having cumulative distribution functions F and G, respectively. Then at threshold c, $q=1-F(c)$ and $p=1-G(c)$. The theoretical ROC curve is a plot of $(1-G(c),1-F(c))$ for all possible values of c or, equivalently, a plot of (p,q) where p ranges from 0 to 1 and

$$q=1-F(G^{-1}(1-p)) \tag{1}$$

The AUC is the area under this curve. Different estimates of the AUC arise from different approaches to estimating the ROC curve. Below we describe different estimators of the AUC.

3.1. Non-parametric approaches

The simplest non-parametric estimation method for the ROC curve involves replacing F and G in (1) by their empirical cumulative distribution functions $\hat{F}_m(t)$ and $\hat{G}_n(t)$, respectively. The empirical cumulative distribution function is defined for any given value t, to be the observed percentage of sample values less than or equal to t. The resulting estimated ROC curve is an increasing step function on the unit square that can be quite jagged. The area under this curve is equal to the Mann–Whitney U statistic [10, 19] and provides an unbiased non-parametric estimator for the AUC. We will denote this estimator by MW.

Several authors [6, 15, 20, 21] discuss refining the non-parametric approach to provide a smoothed ROC curve using the kernel method. Following Zou *et al.* [20] we choose the Gaussian kernel and estimate the probability density function (PDF) $f(t)=F'(t)$ by

$$\hat{f}(t)=\frac{1}{m}\sum_{i=1}^{m}\frac{1}{h_x}\phi\left(\frac{t-x_i}{h_x}\right) \tag{2}$$

where ϕ is the PDF of the standard normal distribution.

The choice of h_x, the bandwidth which controls the amount of smoothing, is discussed below. The PDF $g(t)=G'(t)$ is estimated similarly. From (2) it follows by integration that the estimator of F is

$$\hat{F}(t)=\frac{1}{m}\sum_{i=1}^{m}\Phi\left(\frac{t-x_i}{h_x}\right) \tag{3}$$

where Φ is the standard normal cumulative distribution function, $\hat{G}$ is obtained similarly. Using $\hat{F}$ and $\hat{G}$ in (1) results in a smoothed estimator of the ROC curve. Lloyd [15] shows that the resulting kernel estimate of the AUC can be written as

$$K = \frac{1}{nm} \sum_{i=1}^{m} \sum_{j=1}^{n} \Phi\left(\frac{x_i - y_i}{\sqrt{(h_x^2 + h_y^2)}} \right) \tag{4}$$

Zou *et al.* [20] use for the bandwidth

$$h_x = 0.9 \min(s_x, iqr_x/1.34) m^{-1.5} \tag{5}$$

where s_x and iqr_x are the standard deviation and the quartile inter range, respectively, of the m test results on the diseased sample. h_y is obtained similarly. This choice of bandwidth has been recently found to be effective in estimating the overlapping coefficient [22], which is an alternative measure of the difference between two distributions. Silverman [23] also recommends it as doing 'very well for a wide range of densities'. We denote the estimate of AUC found by using (5) in (4) as K1.

Lloyd and Yong [21] describe a more complex selection of bandwidth procedure in smoothing ROC curves. Their method is an extension of the 'two-stage plug-in' procedure of Wand and Jones [24]. We denote the AUC estimate obtained by their algorithm as K2.

Sometimes, before the kernel is applied, data is transformed to be more symmetrical [15]. Consequently we examine the use of the Box–Cox power transformation before applying the kernel method and denote the resulting AUC estimates as K1T and K2T.

3.2. *Parametric approaches*

A simple parametric approach is to assume the X and Y are independent normal variates with $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$. Consequently

$$q = \Phi\left(\frac{\mu_X - c}{\sigma_X} \right), \quad p = \Phi\left(\frac{\mu_Y - c}{\sigma_Y} \right)$$

and the ROC is estimated by plotting q versus p for all possible values of c with the unknown parameters being replaced by their usual estimates. Furthermore

$$\mathrm{AUC} = \Phi\left(\frac{\mu_X - \mu_Y}{\sqrt{(\sigma_X^2 + \sigma_Y^2)}} \right) \tag{6}$$

and can be estimated by substituting sample means and standard deviations into (6). We denote this estimator by N. Some alternative estimators are considered by Reiser and Guttman [25] under the normality assumptions and are found to be inferior.

When data analysis indicates that the normal assumption is untenable, an *ad hoc* transformation such as the log is often suggested [7, 8]. Recently several authors [20, 26, 27] have recommended using the data to fit a power transformation of the Box–Cox type and then applying the N estimator to the transformed data. In this situation X and Y, the diagnostic marker measurements on the diseased and healthy subjects, respectively, are not assumed to be normally distributed but rather it is assumed that transformed versions of them are normally

distributed. More specifically define

$$X^{(\lambda)} = \begin{cases} \frac{X^{\lambda}-1}{\lambda} & \lambda \neq 0 \\ \log(X) & \lambda = 0 \end{cases} \tag{7}$$

where it is assumed that $X^{(\lambda)} \sim \mathrm{N}(\mu_1, \sigma_1^2)$. $Y^{(\lambda)}$ is defined similarly and it is assumed that $Y^{(\lambda)} \sim \mathrm{N}(\mu_2, \sigma_2^2)$. Note that

$$\mathrm{AUC} = P(X > Y) = P(X^{(\lambda)} > Y^{(\lambda)}) \tag{8}$$

Based on the observations on the diseased and healthy subjects, an appropriate likelihood function can be constructed [20] and maximized giving $\hat{\lambda}$, the maximum likelihood of estimate λ. It follows from (8) that the AUC can be estimated by first transforming the data using $\hat{\lambda}$ and then applying the N estimation method. We denote this estimator by NT.

4. SIMULATION STUDY FOR COMPARING THE AUC ESTIMATORS

We compare the estimators of AUC discussed above, namely MW, N, NT, K1, K1T, K2, K2T, in terms of their RMSE and bias, for a wide variety of cases, by means of an extensive simulation study. The RC method is discussed in Section 4.4.

Our simulations cover a wide variety of different distributional shapes, a sampling of which is presented in Figure 2. Many different distributional combinations were considered in this study, each for several choices of AUC and with sample size $n = m = 20, 50, 100$. For the purpose of computing RMSE and bias, 1000 simulations of each scenario were computed. The results are summarized in Tables I–IX. Each cell of the tables presents first the bias and then the RMSE.

The distributions in Figure 2 are standardized to give an $\mathrm{AUC} = 0.90$. For other cases the distributions for the healthy and diseased populations would move closer together. The variety of cases illustrated in Figure 2 covers the symmetric, asymmetric and bimodal situations often seen in real data [7] and is similar to the shapes considered in the simulation study of Hajian-Tilaki *et al.* [28]. A program was written in Gauss in order to compute the various estimators and carry out the simulations. If, while computing the NT procedure any of the simulated data was found to be negative, then both diseased and healthy sample values were shifted equally so that the minimum value was slightly larger than zero before the transformation parameter λ was estimated.

4.1. Simulations with normal distributions

We first consider samples with normal populations of equal variance (Table I) and then of unequal variance (Table II). For these tables μ_X, the expected value of the diseased population, is chosen to correspond to the AUC values.

In addition to the tabulated sample sizes (20, 100) we also considered the intermediate sample size of 50 which produced intermediate results and was not tabulated for brevity. Computations were also carried out for other choices of mean and variance but as they produce very similar results are not reported. Although the distribution for the healthy population is the same in Tables I and II, simulations were carried out independently so that the two tables are

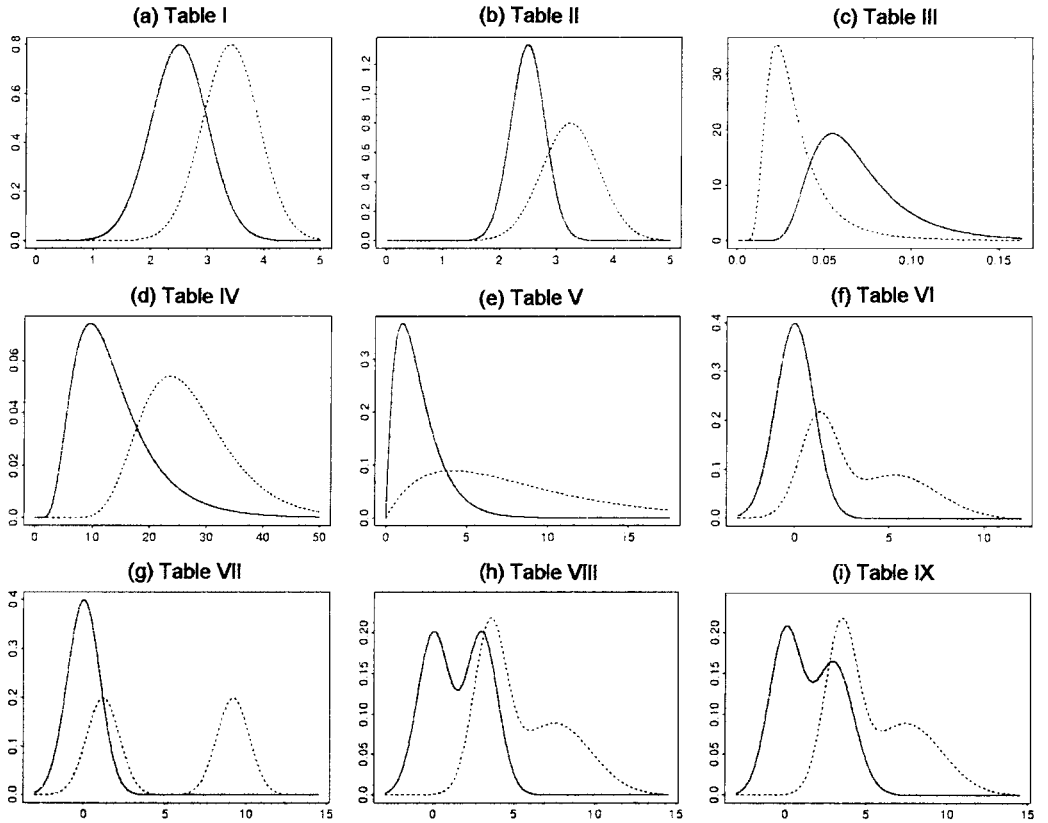

Figure 2. Distributions used in the simulation study with AUC = 0.9: scenarios as in Tables I–IX.

Table I. Simulated bias and RMSE for AUC estimators: $Y \sim \mathrm{N}(2.5, 0.25)$, $X \sim \mathrm{N}(\mu_X, 0.25)$.

Methods	$n = m = 20$			$n = m = 100$		
	AUC = 0.7	AUC = 0.8	AUC = 0.9	AUC = 0.7	AUC = 0.8	AUC = 0.9
MW	−0.001 0.081	−0.000 0.068	0.001 0.048	−0.001 0.036	−0.001 0.031	−0.001 0.021
N	−0.002 0.078	−0.003 0.066	−0.002 0.047	−0.000 0.035	−0.001 0.030	−0.001 0.021
NT	−0.000 0.078	−0.000 0.066	−0.000 0.046	−0.000 0.035	−0.001 0.030	−0.001 0.020
K1	−0.017 0.076	−0.021 0.068	−0.019 0.053	−0.010 0.036	−0.014 0.033	−0.014 0.026
K1T	−0.016 0.076	−0.021 0.068	−0.019 0.053	−0.010 0.036	−0.014 0.033	−0.014 0.026
K2	−0.015 0.076	−0.018 0.068	−0.016 0.052	−0.006 0.036	−0.008 0.031	−0.007 0.023
K2T	−0.014 0.076	−0.018 0.068	−0.017 0.052	−0.006 0.036	−0.008 0.031	−0.007 0.023

independent. Tables I and II show, as expected, that the N estimator usually has the smallest RMSE when the data are normal but that its advantage is often small although it can run as high as 6 per cent when compared with the non-parametric MW procedure. An exception is the case of AUC = 0.7 and $n = m = 20$ for which the kernel methods have the smallest RMSE. The MW estimator is known to be unbiased and the small bias estimates appearing

Table II. Simulated bias and RMSE for AUC estimators: $Y \sim \mathrm{N}(2.5, 0.09)$, $X \sim \mathrm{N}(\mu_X, 0.25)$.

Methods	$n=m=20$			$n=m=100$		
	AUC = 0.7	AUC = 0.8	AUC = 0.9	AUC = 0.7	AUC = 0.8	AUC = 0.9
MW	0.005 0.081	−0.001 0.073	−0.002 0.052	−0.000 0.039	0.000 0.031	−0.000 0.023
N	0.002 0.078	−0.002 0.070	−0.005 0.049	−0.000 0.037	−0.000 0.030	−0.000 0.022
NT	0.005 0.080	0.001 0.070	−0.002 0.048	−0.000 0.037	0.000 0.030	−0.000 0.022
K1	−0.010 0.075	−0.021 0.072	−0.023 0.057	−0.010 0.038	−0.012 0.033	−0.013 0.027
K1T	−0.010 0.076	−0.020 0.072	−0.023 0.057	−0.010 0.038	−0.012 0.033	−0.012 0.027
K2	−0.008 0.076	−0.018 0.072	−0.020 0.056	−0.005 0.038	−0.006 0.031	−0.006 0.024
K2T	−0.008 0.076	−0.018 0.072	−0.020 0.056	−0.005 0.038	−0.006 0.032	−0.006 0.024

Table III. Simulated bias and RMSE for AUC estimators: $Y^{1/3} \sim \mathrm{N}(2.5, 0.09)$, $X^{1/3} \sim \mathrm{N}(\mu_X, 0.25)$.

Methods	$n=m=20$						$n=m=100$					
	AUC = 0.7		AUC = 0.8		AUC = 0.9		AUC = 0.7		AUC = 0.8		AUC = 0.9	
MW	−0.000	0.083	0.000	0.073	0.001	0.049	0.000	0.038	−0.001	0.031	−0.000	0.021
N	−0.064	0.109	−0.056	0.109	−0.045	0.082	−0.090	0.103	−0.075	0.091	−0.053	0.064
NT	−0.000	0.081	0.002	0.070	0.000	0.047	0.000	0.038	−0.000	0.030	−0.000	0.020
K1	−0.025	0.083	−0.026	0.077	−0.025	0.057	−0.016	0.041	−0.019	0.037	−0.016	0.027
K1T	−0.015	0.078	−0.018	0.072	−0.019	0.053	−0.009	0.038	−0.014	0.034	−0.013	0.025
K2	−0.017	0.081	−0.017	0.074	−0.016	0.052	−0.005	0.039	−0.007	0.032	−0.005	0.022
K2T	−0.013	0.078	−0.016	0.072	−0.016	0.052	−0.004	0.038	−0.007	0.032	−0.006	0.023

in the tables are a result of simulation noise. This comment also applies to all the simulations discussed below. Since the observed biases for the N estimates are of the same order as those of the MW procedure, the N method can be considered to be effectively unbiased for normal data. The NT procedure gives results essentially equivalent to N. In fact transforming the data first does not improve the kernel procedures either. The kernel procedures are somewhat downward biased. The K1 and K2 procedures perform similarly although K2 appears to be consistently slightly less biased. The more complex K2 procedure is only clearly better than K1 for AUC = 0.9 and $n=m=100$. Except for the case of AUC = 0.7 and $n=m=20$ the kernel methods do not improve on the non-parametric MW estimator and for large AUC do worse. Overall the results for equal and unequal variances are very similar.

4.2. Simulations with skewed distributions

We obtained skewed data by first generating normal variates which where then taken to the power of (-3). This power was chosen to correspond to the power transformation $(-0.34 \approx -1/3)$ found to be best for the CK data. The relevant probability density functions are graphed in Figure 2(c). Bias and RMSE are provided in Table III. Not surprisingly, the N procedure, which assumes normality, performs worst here both in terms of RMSE and bias. This bias does not decrease with a larger sample size and is largest for AUC = 0.7. NT provides the best results being effectively unbiased and having consistently the lowest RMSE except for AUC = 0.7 and $n=m=20$ where the K1T and K2T have a lower RMSE. It should

Table IV. Simulated bias and RMSE for AUC estimators: $\log(Y) \sim \mathrm{N}(2.5, 0.25)$, $\log(X) \sim \mathrm{N}(\mu_X, 0.09)$.

Methods	$n = m = 20$						$n = m = 100$					
	AUC = 0.7		AUC = 0.8		AUC = 0.9		AUC = 0.7		AUC = 0.8		AUC = 0.9	
MW	−0.001	0.084	−0.003	0.072	−0.002	0.050	−0.001	0.037	−0.000	0.032	−0.000	0.022
N	−0.038	0.095	−0.035	0.091	−0.020	0.062	−0.049	0.065	−0.034	0.053	−0.019	0.033
NT	0.001	0.082	−0.001	0.069	−0.002	0.046	−0.000	0.036	0.000	0.031	−0.000	0.021
K1	−0.021	0.082	−0.027	0.076	−0.023	0.057	−0.015	0.039	−0.015	0.036	−0.013	0.026
K1T	−0.015	0.079	−0.023	0.072	−0.023	0.055	−0.011	0.037	−0.013	0.034	−0.013	0.026
K2	−0.016	0.082	−0.020	0.074	−0.017	0.054	−0.007	0.037	−0.006	0.033	−0.005	0.022
K2T	−0.013	0.080	−0.021	0.072	−0.020	0.054	−0.006	0.037	−0.006	0.033	−0.007	0.023

Table V. Simulated bias and RMSE for AUC estimators: Gamma AUC = 0.9, $p = 2$, $p1 = 2$.

Methods	$n = m = 20$				$n = m = 100$			
	$\lambda_Y = 0.5$		$\lambda_Y = 1$		$\lambda_Y = 0.5$		$\lambda_Y = 1$	
MW	0.001	0.048	0.001	0.050	0.000	0.022	−0.001	0.021
N	−0.041	0.061	−0.041	0.061	−0.047	0.052	−0.048	0.053
NT	0.001	0.045	0.001	0.046	0.001	0.021	−0.000	0.020
K1	−0.040	0.061	−0.040	0.063	−0.029	0.037	−0.031	0.038
K1T	−0.019	0.052	−0.018	0.054	−0.012	0.025	−0.014	0.026
K2	−0.028	0.054	−0.028	0.055	−0.010	0.024	−0.012	0.024
K2T	−0.017	0.051	−0.016	0.053	−0.005	0.023	−0.008	0.023

be noted that the non-parametric MW method produces similar but somewhat higher RMSE values with a difference of about 4 per cent for a sample size of 20. The kernel procedures are again biased, but in contrast with the results in Section 4.1, here transforming the data before applying the kernel smoothing generally improves both the RMSE and bias. This improvement is more substantial for K1. Once the data are transformed there is little difference between the two different bandwidth selection methods. For AUC = 0.7 and $n = m = 20$, K1T and K2T somewhat improve on both NT and MW in terms of RMSE. However, this advantage fades with increasing AUC and for AUC = 0.9 they are substantially worse than NT and somewhat worse than MW.

We further examined skewed data in Table IV that summarizes results for the log-normal distribution. The data were obtained by first simulating normal variates and then exponentiating these values. Again, Tables III and IV are constructed independently. Examining Table IV leads us to the same conclusions as those given in Table III. We examined other transformations in the power family and found that if the transformation produced distributions that were very skewed the results paralleled those of Tables III and IV while for transformations which gave approximately symmetric bell shaped distributions the results were similar to those of Tables I and II.

Since the distributions used were obtained by applying power transformations to normal data and these fall into the Box–Cox transformation family, it is perhaps not surprising that NT does better than the kernel methods. In Table V we consider a simulation which is not in this family: the gamma distribution. Denoting the gamma density function as $\mathrm{e}^{-Z/\lambda} Z^{p-1}/(\lambda^p \Gamma(p))$

Table VI. Simulated bias and RMSE for AUC estimators: $Y \sim \mathrm{N}(0,1)$, $X \sim 0.5\mathrm{N}(\mu,1) + 0.5\mathrm{N}(\mu+4,5)$.

Methods	$n=m=20$						$n=m=100$					
	AUC = 0.7		AUC = 0.8		AUC = 0.9		AUC = 0.7		AUC = 0.8		AUC = 0.9	
MW	0.004	0.081	−0.001	0.069	0.000	0.049	0.000	0.038	−0.000	0.031	0.000	0.021
N	0.041	0.076	0.006	0.052	−0.024	0.048	0.043	0.052	0.009	0.026	−0.023	0.029
NT	0.027	0.081	0.008	0.059	−0.012	0.043	0.032	0.047	0.012	0.029	−0.012	0.021
K1	−0.003	0.069	−0.026	0.064	−0.038	0.060	−0.002	0.034	−0.017	0.033	−0.026	0.033
K1T	−0.004	0.073	−0.022	0.065	−0.030	0.056	−0.003	0.035	−0.014	0.032	−0.020	0.029
K2	−0.001	0.072	−0.019	0.063	−0.028	0.054	−0.000	0.036	−0.006	0.031	−0.008	0.022
K2T	−0.003	0.074	−0.018	0.064	−0.024	0.052	−0.000	0.037	−0.006	0.031	−0.007	0.022

Table VII. Simulated bias and RMSE for AUC estimators: $Y \sim \mathrm{N}(0,1)$, $X \sim 0.5\mathrm{N}(\mu,1) + 0.5\mathrm{N}(\mu+8,5)$.

Methods	$n=m=20$						$n=m=100$					
	AUC = 0.7		AUC = 0.8		AUC = 0.9		AUC = 0.7		AUC = 0.8		AUC = 0.9	
MW	0.001	0.082	−0.003	0.073	−0.000	0.049	0.000	0.039	0.001	0.031	−0.000	0.020
N	0.099	0.116	0.040	0.069	−0.017	0.048	0.103	0.107	0.048	0.054	−0.011	0.022
NT	0.077	0.111	0.037	0.073	−0.006	0.042	0.089	0.096	0.047	0.054	−0.004	0.017
K1	0.021	0.070	−0.022	0.064	−0.058	0.074	0.015	0.036	−0.013	0.030	−0.041	0.045
K1T	0.015	0.075	−0.018	0.064	−0.044	0.062	0.011	0.037	−0.010	0.028	−0.032	0.037
K2	0.016	0.073	−0.017	0.065	−0.040	0.061	0.003	0.037	−0.001	0.030	−0.008	0.022
K2T	0.012	0.076	−0.013	0.066	−0.030	0.054	0.002	0.038	−0.000	0.030	−0.006	0.021

with subscripts X and Y added to the parameters to distinguish the two populations, we generated data for $p_X = p_Y = 2$, $\lambda_Y = 0.5$ and 1.0 and λ_X being chosen to give AUC = 0.9. The resulting distributional shape is presented in Figure 2(e).

Our conclusions are similar to those found in Tables III and IV for distributions in the power family. N performs the worst both in terms of RMSE and bias while NT is the best. For the skewed distributions N is biased downwards. Again NT is effectively unbiased and improves on the MW procedure in terms of RMSE, in cases of small sample size by as much as 8 per cent. K1 does as bad as N for small sample sizes. K2 is better than K1 both in terms of RMSE and bias. Transforming the data before applying the kernel improves K1 substantially and K2 only marginally but these still do worse than MW and NT. For the normal and skewed distributions discussed above the kernel procedures result in estimates that tend to be too low.

4.3. Simulations with mixtures of normal distributions

In order to consider additional scenarios which are not generated by the power family transformations on normal variates, we examined mixtures of normal distributions. These can produce bimodal forms. In Tables VI and VII we consider the X (diseased) to come from a mixture while the Y (healthy) data follow a normal distribution. In Tables VIII and IX both populations are considered as mixtures. The particular distributions used for the simulations are given in the tables with the parameter μ varying according to the specified AUC val-

Table VIII. Simulated bias and RMSE for AUC estimators: $Y \sim 0.5\mathrm{N}(0,1) + 0.5\mathrm{N}(3,1)$, $X \sim 0.5\mathrm{N}(\mu,1) + 0.5\mathrm{N}(\mu+4,5)$.

Methods	$n=m=20$						$n=m=100$					
	AUC = 0.7		AUC = 0.8		AUC = 0.9		AUC = 0.7		AUC = 0.8		AUC = 0.9	
MW	0.000	0.079	0.002	0.068	−0.000	0.047	−0.000	0.036	0.001	0.029	−0.000	0.020
N	0.021	0.071	0.015	0.057	−0.008	0.039	0.024	0.039	0.016	0.029	−0.006	0.018
NT	0.023	0.073	0.020	0.058	−0.005	0.037	0.024	0.040	0.021	0.032	−0.001	0.016
K1	−0.012	0.071	−0.015	0.062	−0.027	0.051	−0.008	0.035	−0.008	0.028	−0.016	0.025
K1T	−0.009	0.071	−0.012	0.061	−0.027	0.051	−0.006	0.034	−0.006	0.027	−0.015	0.025
K2	−0.009	0.072	−0.010	0.062	−0.020	0.048	−0.003	0.035	−0.001	0.028	−0.006	0.020
K2T	−0.007	0.072	−0.009	0.062	−0.022	0.048	−0.002	0.035	−0.001	0.028	−0.006	0.020

Table IX. Simulated bias and RMSE for AUC estimators: $Y \sim 0.5\mathrm{N}(0,1) + 0.5\mathrm{N}(3,1.5)$, $X \sim 0.5\mathrm{N}(\mu,1) + 0.5\mathrm{N}(\mu+4,5)$.

Methods	$n=m=20$						$n=m=100$					
	AUC = 0.7		AUC = 0.8		AUC = 0.9		AUC = 0.7		AUC = 0.8		AUC = 0.9	
MW	0.003	0.081	0.001	0.064	−0.000	0.047	−0.000	0.036	−0.000	0.029	−0.000	0.020
N	0.018	0.073	0.010	0.054	−0.007	0.041	0.017	0.036	0.011	0.027	−0.005	0.017
NT	0.020	0.075	0.016	0.055	−0.003	0.039	0.018	0.037	0.016	0.029	−0.000	0.016
K1	−0.013	0.073	−0.017	0.060	−0.025	0.052	−0.010	0.035	−0.012	0.030	−0.016	0.025
K1T	−0.009	0.073	−0.014	0.059	−0.025	0.051	−0.008	0.034	−0.009	0.029	−0.015	0.024
K2	−0.009	0.074	−0.012	0.060	−0.019	0.049	−0.004	0.035	−0.004	0.028	−0.006	0.020
K2T	−0.007	0.074	−0.011	0.059	−0.021	0.049	−0.004	0.035	−0.003	0.028	−0.007	0.020

ues. These distributions are graphed in Figures 2(f)–(i). As can be seen from the graphs the results in Table VII are based on a stronger bimodality in the diseased population than those in Table VI. In Table IX the healthy population exhibits a weaker bimodality than in Table VIII. From Tables VI and VII we can see that similar to the cases of skewed distributions considered previously N is biased, with the bias being unaffected by increasing sample size but being largest for AUC = 0.7. In terms of RMSE, the N procedure is quite variable, sometimes producing the lowest RMSE value and sometimes the highest. With strong bimodality (Table VII), N is particularly bad for small AUC. NT does not provide any consistent improvement except for AUC = 0.9. In addition, transforming the data before applying the kernel approach usually improves the RMSE and bias results somewhat except for AUC = 0.7 and $n=m=20$. K2 is superior to or the same as K1 again except for AUC = 0.7. For small sample sizes the kernel methods can improve substantially on MW in terms of RMSE (up to 8 per cent) except for AUC = 0.9, where the kernel does much worse.

In Tables VIII and IX the N procedure does much better. Although it is clearly biased it often has the lowest RMSE. The NT results are quite similar to that of N. Whenever AUC = 0.9 the NT method was effectively unbiased and had the lowest RMSE. Transformation does not lead to improvement in the kernel methods. K1 and K2 are similar except for AUC = 0.9 where K2 is clearly superior. For AUC = 0.7 or 0.8 the kernel approach has RMSE as good as or better than the MW method.

Table X. Simulated bias and RMSE for RC estimates.

Scenario	Bias	RMSE
Table I, AUC = 0.9, $n = m = 100$	0.000	0.020
Table III, AUC = 0.9, $n = m = 100$	0.002	0.024
Table V, AUC = 0.9, $\lambda_Y = 1$, $n = m = 100$	−0.001	0.023
Table VII, AUC = 0.7, $n = m = 100$	0.038	0.050

4.4. The RC method

A referee recommended that the RC [16] method also be examined. Hajian-Tilaki *et al.* [28] carried out a simulation study which compared the RC and MW procedures in terms of bias and variance for various distributions. They concluded that the two approaches gave very similar results. Consequently our comparisons of the MW and other methods should carry over to RC.

In order to examine this somewhat further we carried out a small simulation of RC (using 100 simulations) for a few of the scenarios described above. These are given in Table X. In the first three scenarios RC performs very similarly to MW and NT while in the fourth it is quite different. This scenario deals with a very strong bimodal distribution (Figure 2(g)) for the diseased population. Although Hajian-Tilaki *et al.* [28] do consider bimodal mixtures, they do not consider a case as strongly bimodal as this. Here we see a clear bias in RC. For the fourth scenario RC is certainly preferable to NT both in terms of RMSE and bias but is strongly outperformed by MW and the kernel methods (see Table VII).

5. THE EXAMPLE REVISITED

Figure 1(a) shows that the data is quite skewed. The normal assumption is untenable. It is not surprising that the parametric estimation method assuming normality gives an estimate quite different from the others. The fact that it is much lower than the unbiased MW estimate corresponds to the negative biases noted in our simulations on skewed distributions. Our simulations indicate that NT should be appropriate in this situation. The transformed data (Figure 1(b)) is 'closer' to normality. NT, MW and RC are quite similar in correspondence to our simulations. Furthermore, our simulations (see Tables III, IV and V) lead us to expect that for skewed distributions all the kernel estimates will tend to be too low with K1 being the worst while K1T and K2T will be similar to each other and closer to being unbiased. This seems to correspond well with what we observe for the CK data.

6. DISCUSSION

The RC and NT approaches are both based on quite similar assumptions. The RC method assumes that some monotonic transformation of the data will result in normality while the NT method is less general in that it assumes that the transformation belongs to a particular family, that is, the family of power transformations. Thus it is not surprising that the two often produce similar results. They can, however, differ substantially for strongly bimodal distributions where

RC is better but neither are appropriate. It should be noted that computationally NT is much simpler than RC.

From the above results it is clear that the NT method is the preferred method unless the marker distributions for the healthy and/or diseased populations are suspected of being mixtures. It should be noted that although MW is usually not the best in terms of RMSE it is often close to the best. In the unimodal situations we considered NT was effectively unbiased and in terms of RMSE generally did somewhat better than the non-parametric MW approach, sometimes as much as 8 per cent better. The NT approach has the additional useful property of providing a continuous ROC curve. For unimodal distributions of various shapes we did not find the kernel approach generally useful although for small AUC accompanied by small sample sizes they did lead to a slight improvement in RMSE at the expense of greater bias. In dealing with mixtures we found that if the two populations are well separated ($\mathrm{AUC} = 0.9$) NT still performs the best. Apparently in such a situation the actual details of the distributional forms are not of great importance. For less separated populations ($\mathrm{AUC} = 0.7,\ 0.8$) based on mixtures there was no clear winning method. In this type of situation the kernel approach proved superior to the MW method and when compared to N or NT the kernel approach varied from much better (for example, Table VII, $n = m = 20$, $\mathrm{AUC} = 0.7$) to slightly worse (for example, Table VIII, $n = m = 20$, $\mathrm{AUC} = 0.8$). For mixture distributions with moderate AUC there was very little difference between the various kernel approaches and the K1 could be recommended due to its simplicity.

A referee suggested that standard errors of the simulated indices be computed in order to help interpret the tables. This would require repeating each of our simulation sets, which are of size 1000, many times and hence would be extremely burdensome in terms of time. We carried out a limited number of repetitions in a few cases and found that the standard error for the RMSE is about 2.5 per cent of its calculated value. This supports the above interpretations of the simulation results.

ACKNOWLEDGEMENTS

We would like to express our thanks to the editor and referees whose comments led to substantial improvements in this paper. In addition we want to thank Julia Vider for carrying out the RC computations.

REFERENCES

1. Green DM, Swets JA. *Signal Detection Theory and Psychophysics*. Wiley: New York, 1966.
2. Metz CE. Some practical issues of experimental design and data analysis in radiological ROC studies. *Investigation Radiology* 1989; **24**:234–245.
3. Hsiao JK, Bartko JJ, Potter WZ. Diagnosing diagnoses receiver operating characteristic methods and psychiatry. *Archives of General Psychiatry* 1989; **46**:664–667.
4. Nockemann C, Heidt H, Thomsen N. Reliability in NDT: ROC study of radiographic weld inspections. *Nondestructive Testing and Evaluation International* 1991; **24**:235–245.
5. Somoza E, Mossman D, McFeeters L. The info-ROC technique: a method for comparing and optimizing inspection systems. In *Review of Progress in Quantitative Nondestructive Evaluation*, Thompson DO, Chimenti DE (eds). Plenum Press: New York, 1990.
6. Zou KH, Hall WJ, Shapiro DE. Smooth non-parametric receiver-operating characteristic (ROC) curves for continuous diagnostic tests. *Statistics in Medicine* 1997;**16**;2143–2156.
7. Goddard MJ, Hinberg I. Receiver operator characteristic (ROC) curves and non-normal data: an empirical study. *Statistics in Medicine* 1990; **9**:325–337.
8. Reiser B, Faraggi D. Confidence intervals for the generalized ROC criterion. *Biometrics* 1997; **53**:644–652.

9. Shapiro DE. The interpretation of diagnostic tests. *Statistical Methods in Medical Research* 1999; **8**:113–134.
10. Bamber DC. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology* 1975; **12**:387–415.
11. Reiser B, Guttman I. Statistical inference for $Pr(Y<X)$: the normal case. *Technometrics* 1986; **28**:253–257.
12. McCool JI. Inference on $P(Y<X)$ in the Weibull case. *Communications in Statistics – Simulation* 1991; **20**:129–148.
13. Wolfe DA, Hogg RV. On constructing statistics and reporting data. *American Statistician* 1971; **25**:27–30.
14. Box GEP, Cox DR. An analysis of transformations. *Journal of the Royal Statistical Society*, *Series B* 1964; **26**:211–243.
15. Lloyd CJ. Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. *Journal of the American Statistical Association* 1998; **93**:1356–1364.
16. Metz CE, Herman BA, Shen J. Maximum likelihood estimator of receiver operator characteristic (ROC) curves from continuously-distributed data. *Statistics in Medicine* 1998; **17**:1033–1053.
17. Percy ME, Andrews DF, Thompson MW. Duchene muscular dystrophy carrier detection using logistic discrimination: serum creatine kinase, hemopexin, pyravite kinase and lactate dehydrogeneze in combination. *American Journal of Medical Genetics* 1982; **13**:27–38.
18. Andrews DF, Herzberg AM. *Data*. Springer-Verlag: New York, 1985.
19. Hanley JA, McNeil BJ. The meaning and use of the area under the receiver operating characteristic (ROC) curve. *Radiology* 1982; **143**:29–36.
20. Zou KH, Tempany CM, Fielding JR, Silverman SG. Original smooth receiver operating characteristic curves estimation from continuous data: statistical methods for analyzing the predictive value of spiral CT of ureteral stones. *Academic Radiology* 1998; **5**:680–687.
21. Lloyd CJ, Yong Z. Kernel estimators of the ROC curves are better than empirical. *Statistics and Probability Letters* 1999; **44**:221–228.
22. Stine RA, Heyse JF. Nonparametric estimates of overlap. *Statistics in Medicine* 2001; **20**:215–236.
23. Silverman BW. *Density Estimator for Statistics and Data Analysis*. Chapman and Hall: London, 1986.
24. Wand MP, Jones MC. *Kernel Smoothing*. Chapman and Hall: London, 1995.
25. Reiser B, Guttman I. A comparison of three point estimates for $Pr(Y<X)$ in the normal case. *Computational Statistics and Data Analysis* 1987; **5**:59–66.
26. Kramar A, Faraggi D, Ychou M, Reiser B, Grenier J. Criteres ROC generalises pour l'evaluation de plusieurs marqueurs tumoraux. *Revue d'Epidemiologie ei de Sante Publique* 1999; **47**:217–226.
27. Schisterman E. Lipid peroxidation and antioxidant biomarkers and biomarker disease. PhD thesis, State University of New York, Buffalo, 1999.
28. Hajian-Tilaki KO, Hanley JA, Joseph L, Collet J. A comparison of parametric and nonparametric approaches to ROC analysis of quantitative diagnostic tests. *Medical Decision Making* 1997; **17**:94–102.

