2009

Extending the Skill Test for Disease Diagnosis

Shu-Chuan Lin

Paul H. Kvam
University of Richmond, pkvam@richmond.edu

Jye-Chyi Lu

Extending the Skill Test for Disease Diagnosis

Shu-Chuan Lin* **Paul H. Kvam** **Jye-Chyi Lu**

H. Milton Stewart School of Industrial & Systems Engineering

Georgia Institute of Technology

Abstract

For diagnostic tests, we present an extension to the skill plot introduced by Briggs and Zaretski [1]. The method is motivated by diagnostic measures for osteopetrosis in a study summarized by Hans et al. [2]. Diagnostic test accuracy is typically defined using the area (or partial area) under the receiver operator characteristic (ROC) curve. If partial area is used, the resulting statistic can be highly subjective because the focus region of the ROC curve corresponds to a set of low false-positive rates that are chosen by the experimenter. This paper introduces a more objective diagnostic test for which the focus region depends on a skill score, which in turn depends on the loss functions associated with misdiagnosis. More specifically, the skill-based diagnostic test serves as a more objective version of the nonparametric test introduced by Dodd and Pepe [3].

1 Introduction

The classification and prediction of dichotomous events has been a cornerstone in biostatistics research and is becoming increasingly important in many other areas of science, including meteorology, economics and computer science. Sing et al. [4] show how pattern classification, scoring and ranking predictors are vital in a wide range of biological problems. Examples include predicting phenotypic properties of HIV-1 from genotypic information, microarray analysis for the prediction of tissue condition based on gene expression, predicting bio-availability or toxicity of drug

*Contract/grant sponsor: National Science Foundation, CMMI 0700131

compounds and gauging treatment effect in clinical trials (Brumback et al. [5]). In many cases, robustness and efficiency of the markers of a diagnostic test are critical and the cost of misclassification is a primary factor for classification and prediction.

Diagnostic tests that use markers to determine whether a patient is diseased or healthy are standard tools in medical screening. Early detection is considered essential to effective treatment. Finding new markers that are less invasive, less expensive, and more accurate than existing measures are important in disease prevention [3]. For the diagnosis of many modern diseases, the difference in marker measurements used to screen healthy patients from diseased patients can be subtle, and statistical researchers work to develop the most effective tool to discern this difference. Misclassification costs are often asymmetric; that is, the cost of misclassifying a healthy patient into the diseased group (a *false positive* result) is often less than the cost of misclassifying a diseased patient into the healthy group. One tool that has been especially useful in recent decades is the receiver operating characteristic (ROC) curve.

1.1 ROC Curve

To describe the ROC, let Y equal one if the patient actually has the disease, and Y is zero otherwise. Let $p = P(Y = 1)$ represents the proportion of diseased patients in the total population. Suppose each patient has diagnostic marker response X. If the patient is a member of the healthy population then X has cumulative distribution function (CDF) F, and if the patient is from the diseased group then X has CDF G. Without loss of generality, we will assume the patient is diagnosed as diseased if $X \geq c$ for some fixed threshold c. In many such cases, it might be assumed that $F(x) \geq G(x)$ for all values of x.

The ROC curve is a plot of the true-positive rate (TPR = sensitivity) versus the false-positive rate (FPR = 1 − specificity), for a classification rule based on a continuously increasing sequence of threshold values. The graph of TPR ($P(X \geq c|Y = 1)$) vs. FPR ($P(X \geq c|Y = 0)$) defines the ROC curve:

$$\mathrm{R}(t) = 1 - G\left(F^{-1}(1-t)\right) \tag{1}$$

where $0 < t < 1$. The curve shows the inherent trade-off between FPR and TPR. A test can be judged according to how its corresponding ROC curve arches over the 45^o line - the more concave the better.

A typical diagnostic test classifies patients according to single marker, and the misclassification rates depend on the threshold value that distinguishes the two screening outcomes. However, it is not always certain how to determine the optimal cutoff point. Chapter 4 of Pepe [6] suggests factors such as health care resources, invasive examination etc., can influence the choice of threshold. In general, the choice of the cutoff point depends on:

(1.) The fixed cost k_{10} for classifying a diseased person as a healthy one.

(2.) The fixed cost k_{01} for classifying a healthy person as a diseased patient.

(3.) The overall misclassification probability, $p(1 - \mathrm{TPR}) + (1 - p)\mathrm{FPR}$. By minimizing the overall misclassification probability, the slope of the ROC curve at the optimal cutoff point is $(1 - p)/p$.

(4.) The expected cost of misclassification, $k_{10}p(1 - \mathrm{TPR}) + k_{01}(1 - p)\mathrm{FPR}$. By minimizing the expected cost of misclassification, the slope of the ROC curve at the optimal cutoff point is $(k_{01}(1 - p))\,/\,(k_{10}p)$.

Theoretical results for ROC curves are well established. Pepe [7] developed a semiparametric estimator for ROC curves within the generalized linear model framework for binary regression. Hsieh and Turnbull [8] consider nonparametric estimators based on empirical distribution functions and derive asymptotic properties. Lloyd and Yong [9] showed smooth kernel-based estimators outperform this strictly empirical estimator. Claeskens et al. [10] and Hall et al. [11] study nonparametric methods, e.g. empirical likelihood method or bootstrap, for constructing confidence intervals and confidence bands for estimators of ROC curves.

1.2 Applying the ROC

There are numerous ways of summarizing the ROC curve into an objective test statistic. The area under the ROC curve (AUC), defined as $\int_0^1 \mathrm{R}(t)dt$ was one of the first commonly used measures of test quality. It can be easily shown that if $Z_0 \sim F$ and $Z_1 \sim G$ are independent, then

$$\int_0^1 \mathrm{R}(t)dt = P(Z_0 \leq Z_1). \tag{2}$$

For continuous data, AUC is equivalent to the probability that a random observation coming from the diseased population (Z_1) is larger than that from the non-diseased population (Z_0). A diagnostic test that produces AUC $\leq 1/2$ is considered non informative.

The AUC is the most commonly used method of summarizing a diagnostic test's overall accuracy. However, the AUC summarizes test performance over regions of the ROC space that may be of no practical interest. The partial area under the curve (PAUC) restricts the AUC-integration to an area of interest, based on FPRs that are considered clinically relevant:

$$\mathrm{A}(t_0, t_1) = \mathrm{PAUC} = \int_{t_0}^{t_1} \mathrm{ROC}(t)dt, \tag{3}$$

where the interval (t_0, t_1) denotes the false-positive rates of interest. Corresponding to the case in which $F = G$, if a diagnostics test has an $\mathrm{A}(t_0, t_1)$ which equals to $(t_1^2 - t_0^2)/2$, it has no ability for classifying patients correctly.

Dodd and Pepe [3] describe the significance of the PAUC through the odds,

$$\Lambda(t_0, t_1) = \frac{\mathrm{A}(t_0, t_1)}{(t_1 - t_0) - \mathrm{A}(t_0, t_1)}. \tag{4}$$

This is the odds of the probability of a correct classification to the probability of an mistaken classification, given the test result is from the healthy population in the region (t_0, t_1). Note if the test has an odds of $(t_0 + t_1)/(2 - (t_0 + t_1))$, then the test conveys no information. If the test is perfect, the odds will increase to infinity.

1.3 Subjectivity in Diagnostic Tests

Dodd and Pepe [3] emphasized that although the partial AUC is a more clinically relevant summary measure of test accuracy, the choice of the appropriate restricted region may be controversial. Reasonable choices depend on information about the cost associated with true- and false-positive diagnoses. For example, if a diagnostic test is not particularly efficient at screening a disease that has affected most of the at-risk set, the naive guess that every patient has the disease might actually be cost effective.

The PAUC can, in fact, be deliberately misused. FPRs could be unscrupulously chosen in such a way as to maximize the significance of the PAUC statistic, for example. Choices for acceptable FPRs are implicitly a function of the relative loss associated with the type I error (healthy patient diagnosed as diseased) and type II error (diseased patient diagnosed as healthy).

In this paper, we introduce a more objective method of selecting relevant FPRs, based on utility/loss criteria associated with the type I and type II errors. Although this advantage is evident, there is added uncertainty to the diagnostic statistics due to estimation based on the known loss functions. We do not expound on choices for loss functions here. Instead, we refer the reader to Schervish [12], where loss functions are addressed along with their affect on diagnostic tests.

In Section 2, we investigate the "skill statistic" and consider tests at FPRs that produce skillful tests. That is, the PAUC will use FPR values that correspond with skillful diagnostic tests. Properties of the skill statistic are investigated in Section 3. In Section 4, we introduce the motivating example to this research, featuring data for $5,662$ women being diagnosed for osteoporosis. In Section 5, we discuss the practical use of skill scores in diagnostic tests by pointing out its strength and weaknesses with respect to traditional tests.

2 The Skill Score

Mozer and Briggs [13] developed a *skill score* as a method to evaluate probabilistic forecasts of binary events as "skilled" or "not skilled" by integrating the loss from misclassification. Properties for the skill score were further developed in Briggs and Ruppert [14]. They define a diagnostic test as *skillful* if it is more effective in screening disease than the optimal naive guess. As mentioned earlier, the optimal naive guess will classify all patients as healthy or all patients as diseased based solely on the cost functions. It makes sense to use only threshold values that correspond to skillful tests, and the skill score is useful because it considers both the cost of the forecast and the loss of making incorrect forecasts. The *skill plot* [1] summarizes the diagnostic skill over a range of threshold values and offers a novel alternative to the ROC curve for describing disease diagnosis.

The skill score is based on simple loss function. If we define

$$\theta = \frac{k_{01}}{k_{01} + k_{10}},$$

then θ is the (relative) loss when $Y = 0$ and $X \geq c$, and $1-\theta$ is the loss when $Y = 1$ and $X < c$. Without loss of generality, we will assume the misclassification cost of k_{01} is less than k_{10}, so that $\theta \leq 1/2$.

Recall $p = p_{1+} = P(Y = 1)$ is the proportion of diseased patients in the total population. Let $p_{+1} = P(X \geq c) = p\bar{G}(c) + (1-p)\bar{F}(c) =$ proportion of people classified as diseased, where $\bar{G}(c) \equiv 1 - G(c) =$ P(correctly classify patient, given they have the disease), and $\bar{F}(c) \equiv 1 - F(c) =$ P(classify as diseased person given patient is actually healthy). Probabilities of individual outcomes are summarized in Table 1.

	$Y = 1$	$Y = 0$	
$X \geq c$	p_{11}	p_{01}	p_{+1}
$X < c$	p_{10}	p_{00}	p_{+0}
	$p = p_{1+}$	$1 - p = p_{0+}$	

Table 1: Contingency table.

Without information from X, the optimal naive classification rule is based solely on comparing p and θ. There are two possible actions: classify all subjects as healthy people if $p < \theta$ or classify all subjects as diseased patients if $p \geq \theta$. The expected loss for this rule is $E_N = p(1-\theta)I(p < \theta) + (1-p)\theta I(p \geq \theta)$. With the information from X, the expert classification rule is based on a critical cutoff point c. Subjects with $X \geq c$ are classified as diseased, and the others as healthy. A skill score can be constructed based upon the relative difference in expected loss between the optimal naive and the expert classification:

$$K_{p,\theta}(c) = \frac{E_N - E_E(c)}{E_N - E_P}, \tag{5}$$

where $E_E(c)$ is the expected loss from the expert guess based on the a cutoff point c, and E_P is the expected loss from a perfect classification (we will assume $E_P = 0$). If E_E is based on a diagnostic classification with threshold c, then $E_E(c) = p_{01}\theta + p_{10}(1-\theta)$, and the skill score simplifies to

$$K_{p,\theta}(c) = \frac{p_{11} - p_{+1}\theta}{p(1-\theta)} I(p < \theta) + \frac{p_{+0}\theta - p_{10}}{(1-p)\theta} I(p \geq \theta). \tag{6}$$

The skill plot simply plots $K_{p,\theta}(c)$ versus threshold value of c.

Briggs and Zaretski [1] applied the skill score to achieve an optimal threshold value by finding the value of c that maximizes $\hat{K}(c)$. In a similar vein, Baker [15] considered a simple linear utility function of FPR and TPR in order to create an optimal test. While the maximum skill score provides an optimal decision tool regarding an individual, in this research we are more interested in the quality of the overall inference, and for this purpose we focus on the AUC and PAUC statistics.

2.1 Skillful Diagnostic Tests

Because the skill score provides an effective loss-based metric for diagnostic test performance, it seems intuitive that the PAUC should be based only on skillful tests. Instead of integrating the ROC over an arbitrarily chosen range of threshold values, we use only the set of values c for which $K_{p,\theta}(c) \geq 0$. This seems at first to

be an obvious objective, but we will see not all rational diagnostic tests are skillful, especially if asymmetric costs of misclassification are involved. Note that, if $p < \theta$, a positive skill score occurs if TPR $\geq \theta p_{+1}/p_{1+}$, or FPR $\leq (1-\theta)p_{+1}/p_{0+}$. In terms of Y, $K_{p,\theta}(c) \geq 0$ if

$$P(Y = 1|X \geq c) = \frac{\bar{G}(c)}{p\bar{G}(c) + (1-p)\bar{F}(c)} = \frac{p_{11}}{p_{+1}} \geq \theta. \tag{7}$$

For the case $p < \theta$, the PAUC in (3) becomes

$$\int_{K_{p,\theta}(t)\geq 0} \mathrm{R}(t)dt.$$

In the less common scenario when $p \geq \theta$, a positive skill score occurs if TPR $\geq 1 - \theta p_{+0}/p_{1+}$, or equivalently, if $P(Y = 1|X < c) = p_{10}/p_{+0} < \theta$. Figure 1 shows a skill score based on $F(t) = \Phi(t)$, $G(t) = \Phi((t-1.5)/1.2)$, where Φ represents the standard Normal CDF. For this figure, the distribution G was chosen to match the simulated distributions in Dodd and Pepe [3]. The relative loss associated by the costs of misclassification determines the range of FPR values (t_0, t_1), which in turn determine the PAUC statistic.

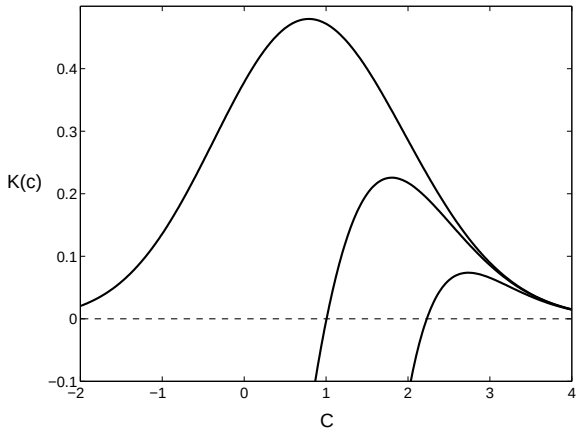

Figure 1: Skill scores based on $F(t) = \Phi(t)$, $G(t) = \Phi((t-1.5)/1.2)$, and $\theta = 0.5$. From left to right, $K(c)$ corresponds to $p = 0.5, 0.2, 0.05$.

3 Diagnostic Statistics

Suppose we observe $\{(X_1, Y_1), \cdots, (X_n, Y_n)\}$ as a training sample consisting of n paired observations where Y_i equals one if the i^{th} person has the disease and equals zero otherwise. The empirical distribution function (EDF) for G, denoted by $G_{n_1}(t)$, is based on $n_1 = \sum_{i=1}^{n} Y_i < n$, and the EDF for F, denoted by $F_{n_0}(t)$, is based on a sample of $n_0 = \sum_{i=1}^{n}(1 - Y_i)$. Note that $\hat{p}_{+1} = n^{-1}\sum_{i=1}^{n} I(X_i \geq c)$. In this paper, we will assume that the sample sizes are such that $n_1/n \to p > 0$ as $n \to \infty$.

The plug-in estimator for the ROC, $\hat{R}(t) = 1 - G_{n_1}(F_{n_0}^{-1}(1-t))$ simplifies to the proportion of the sample designated as diseased that have marker scores larger than $n_0 p$ out of n_0 sample observations that were classified as healthy (i.e., the p^{th} percentile of F_{n_0}). The nonparametric plug-in estimator $R(t; F_{n_0}, G_{n_1})$ creates a jagged ROC curve (see Figure 3) due to the discrete jumps of F_{n_0} and G_{n_1} at the observations. From (2), the AUC's empirical analog is the Mann-Whitney U-statistic [8, 16]. We construct the empirical analog to the skill score, K, in a similar manner. In terms of F and G, we can write (6) as

$$K_{p,\theta}(c) = \left(\bar{G}(c) - \frac{(1-p)\theta}{p(1-\theta)}\bar{F}(c)\right) I(p < \theta) + \left(F(c) - \frac{p(1-\theta)}{(1-p)\theta}G(c)\right) I(p \geq \theta). \quad (8)$$

Because we are assuming the healthy population has lower marker scores, if $F(c) \geq G(c)$ and $\theta = p$, then $K_{p,\theta}(c) = F(c) - G(c) \geq 0$. Let H_n be the EDF of the full sample (ignoring group membership) so that

$$\bar{H}_n(c) = \hat{p}_{+1} = \frac{1}{n}\sum_{i=1}^{n} I(X_i \geq c) = \frac{1}{n}\left(n_1 \bar{G}_{n_1}(c) + n_0 \bar{F}_{n_0}(c)\right).$$

Then the plug-in estimator to (8) is simply

$$\begin{aligned} \hat{K}_{\hat{p},\theta}(c) &= \left(\bar{G}_{n_1}(c) - \frac{(1-\hat{p})\theta}{\hat{p}(1-\theta)}\bar{F}_{n_0}(c)\right) I(\hat{p} < \theta) + \left(F_{n_0}(c) - \frac{\hat{p}(1-\theta)}{(1-\hat{p})\theta}G_{n_1}(c)\right) I(\hat{p} \geq \theta), \\ &= \frac{\bar{G}_{n_1}(c)\hat{p} - \bar{H}_n(c)\theta}{\hat{p}(1-\theta)} I(\hat{p} < \theta) + \frac{H_n(c)\theta - G_{n_1}(c)\hat{p}}{(1-\hat{p})\theta} I(\hat{p} \geq \theta), \end{aligned} \quad (9)$$

where $\hat{p} = n_1/n$.

Unlike the smooth curves in Figure 1, the estimated skill score is generally jagged. Although it is typically concave around its peak, there can also be small anomalies that defy concavity, especially with small samples. Because the skill score is the key for extending the skill test for disease diagnosis, we will investigate some of its basic properties before applying it in the evaluation of the ROC.

3.1 Properties of the Skill Statistic

Confidence intervals for the skill score can be constructed using normal approximations, and examples in the next section show these intervals to be effective with sufficiently large samples. In the following theorems, we describe the asymptotic properties of $\hat{K}_{\hat{p},\theta}(c)$ along with estimators for $P(Y = 1|X \geq c)$ and $P(Y = 0|X < c)$. The proof is relegated to the appendix.

Theorem 1. Assume F and G are continuous distributions and twice differentiable, have finite mean and variance, and for some $\epsilon > 0$, $\epsilon \leq \theta < 1/2$. Then $\sqrt{n}(\hat{K}_{\hat{p},\theta}(c) - E[\hat{K}_{p,\theta}(c)]) \to N(0, \sigma^2)$, where

$$E[\hat{K}_{\hat{p},\theta}(c)] \approx \frac{\bar{G}(c)p - \bar{H}(c)\theta}{p(1-\theta)} I(p < \theta) + \frac{H(c)\theta - G(c)p}{(1-p)\theta} I(p \geq \theta),$$

and

$$\begin{aligned} \sigma^2 &\approx \frac{1}{p(1-\theta)^2}\left((1-2\theta)\bar{G}(c)G(c) + \frac{\theta^2}{p}\bar{H}(c)H(c) + \frac{1-p}{p^2}\bar{H}^2(c)\theta^2\right) I(p < \theta) \\ &+ \frac{1}{(1-p)^2}\left(\frac{p-2p\theta}{\theta^2}\bar{G}(c)G(c) + \bar{H}(c)H(c) + \frac{p}{(1-p)\theta^2}(H(c)\theta - G(c))^2\right) I(p \geq \theta). \end{aligned}$$

The conditional probability from (7) provides an alternative way to characterize a skillful diagnostic test. The probability of observing a diseased patient conditional

on positive diagnosis results $\eta_1(c) = P(Y = 1|X \geq c)$ is expressed in terms of F and G as

$$\eta_1(c) = \frac{\bar{G}(c)p}{\bar{G}(c)p + \bar{F}(c)(1-p)}. \tag{10}$$

Theorem 2. Under the regularity conditions of Theorem 1, for the plug-in estimator of $\eta_1(c)$ in (10)

$$\hat{\eta}_1(c) = \frac{\bar{G}_{n_1}(c)\hat{p}}{\bar{G}_{n_1}(c)\hat{p} + \bar{F}_{n_0}(c)(1-\hat{p})}, \tag{11}$$

we have $\sqrt{n}\left(\hat{\eta}_1(c) - E(\hat{\eta}_1(c))\right) \to N(0, \sigma^2_{\eta_1})$, where $E(\hat{\eta}_1(c)) \approx \eta_1(c)$ and

$$\sigma^2_{\hat{\eta}_1(c)} \approx \frac{\bar{G}(c)\bar{F}(c)p(1-p)}{(\bar{G}(c)p + \bar{F}(c)(1-p))^4}\left(G(c)\bar{F}(c)(1-p) + \bar{G}(c)F(c)p + \bar{G}(c)\bar{F}(c)\right).$$

When $p < \theta$, recall that $\hat{K} \geq 0$ occurs if $\hat{\eta}_1(c) \geq \theta$. Similarly, when $p \geq \theta$, the probability of observing a healthy subject conditional on observing a negative diagnosis is expressed in terms of F and G as

$$\eta_0(c) = P(Y = 0|X < c) = \frac{F(c)(1-p)}{G(c)p + F(c)(1-p)}. \tag{12}$$

Theorem 3. Under the regularity conditions of Theorem 1, for the plug-in estimator of $\eta_0(c)$ in (12)

$$\hat{\eta}_0(c) = \frac{F_{n_0}(c)(1-p)}{G_{n_1}(c)p + F_{n_0}(c)(1-p)}, \tag{13}$$

we have $\sqrt{n}(\hat{\eta}_0(c) - E(\hat{\eta}_0(c))) \to N(0, \sigma^2_{\eta_0})$, where $E(\hat{\eta}_0(c)) \approx \eta_0(c)$ and

$$\sigma^2_{\hat{\eta}_0(c)} \approx \frac{G(c)F(c)p(1-p)}{(G(c)p + F(c)(1-p))^4}\left(\bar{G}(c)F(c)(1-p) + G(c)\bar{F}(c)p + G(c)F(c)\right).$$

In this case, we assess the skill of a diagnostic test based on whether $\hat{\eta}_0(c) > 1-\theta$.

The skill score from (5) has a more general form called the *Total Expected Misclassification Cost* (TEMC). For example, T_E, the TEMC from the expert forecast, is $np_{01}k_{01} + np_{10}k_{10}$. Under $p < \theta$, the naive guess is that all subjects are healthy, and TEMC based on the naive guess, $T_N = npk_{10}$. The skill score is alternatively expressed as

$$\begin{aligned} \hat{K}_{T,\hat{p},\theta}(c) &= \frac{np_{11}k_{10} - np_{01}k_{01}}{npk_{10}} = \frac{p_{11}k_{10} - p_{01}k_{01}}{pk_{10}} \\ &= \bar{G}_{n_1}(c) - \frac{(1-\hat{p})k_{01}}{pk_{10}}\bar{F}_{n_0}(c). \end{aligned} \tag{14}$$

Note that (14) is the skill score defined by Expected Misclassification Cost Per Person (EMCPP), which does not depend on the total sample size. These skill score outcomes are summarized in Table 2. The estimates of $\sigma^2_{T_0}$ and $\sigma^2_{T_1}$ in Table 2 are

$$\hat{\sigma}^2_{T_0} \approx \frac{1}{np}\left(\bar{G}(c)G(c) + \frac{1-p}{p}\frac{k_{01}^2}{k_{10}^2}\bar{F}(c)F(c) + \frac{1-p}{p^2}\frac{k_{01}^2}{k_{10}^2}\bar{F}^2(c)\right),$$

and

$$\hat{\sigma}^2_{T_1} \approx \frac{1}{n(1-p)}\left(\frac{1-p}{p}\frac{k_{01}^2}{k_{10}^2}\bar{G}(c)G(c) + \bar{F}(c)F(c) + p(1-2p)^2\frac{k_{01}^2}{k_{10}^2}\bar{G}^2(c)\right).$$

3.2 Estimating PAUC

We know from (1) that the PAUC is estimated based on a subjectively chosen set of FPR. Dodd and Pepe [3] admit that controversy is unavoidable with such a subjective choice. With this in mind, the skill score $K(c)$ offers a more objective and coherent way of selecting the subset of FPR according to these fixed loss functions. To eliminate the inherent subjectivity in the PAUC, we consider the set of FPR such that $K(c) \geq 0$, which corresponds to $FPR \leq (1-\theta)p_{+1}/p_{0+}$. The corresponding empirical estimator of PAUC,

Expert Forecast		
Expected loss	$p_{10}(1-\theta)+p_{01}\theta$	
TEMC	$np_{01}k_{01}+np_{10}k_{10}$	
EMCPP	$p_{01}k_{01}+p_{10}k_{10}$	
Situation	$p<\theta$	$p\geq\theta$
Naive guess	all healthy	all diseased
Expected loss	$p(1-\theta)$	$(1-p)\theta$
TEMC	npk_{10}	$n(1-p)k_{01}$
EMCPP	pk_{10}	$(1-p)k_{01}$
Skill score by TEMC	$\frac{p_{11}k_{10}-p_{01}k_{01}}{pk_{10}}$	$\frac{p_{00}k_{01}-p_{10}k_{10}}{(1-p)k_{01}}$
Estimator	$\bar{G}_m(c)-\bar{F}_{n_0}(c)\frac{1-p}{p}\frac{k_{01}}{k_{10}}$	$F_{n_0}(c)-G_{n_1}(c)\frac{p}{1-p}\frac{k_{10}}{k_{01}}$
Variance	$\hat{\sigma}^2_{T_0}$	$\hat{\sigma}^2_{T_1}$
Skillful	$\frac{p_{11}}{p_{01}}\geq\frac{k_{01}}{k_{10}}$	$\frac{p_{00}}{p_{10}}\geq\frac{k_{10}}{k_{01}}$

Table 2: The skill score defined by TEMC and EMCPP.

$$\hat{A}_K=\int_{\hat{K}_{p,\theta}(t)\geq 0} R(t)dt, \tag{15}$$

is based on the estimated skill score in (14).

To show how the estimator is constructed, we introduce a practical application in the following section. The results show that the skill score can vary greatly depending on the loss function, so costs for misclassification should not be chosen arbitrarily by the practitioner.

With this infusion of extra empirical information on the skill score comes added uncertainty. In turn, this new estimator would be an inferior choice to the regular PAUC estimator based on expert opinion, as long as the expert opinion is accurate. While this is not the norm, Pepe [6] includes actual case studies in which past data can aid in deciding valid FPR values. To make a fair examination of $\hat{A}_K$, in Section 5 we construct examples using different costs and populations.

4 Osteoporosis Study

We illustrate the PAUC estimator with an EPIDOS prospective study of 5,662 elderly French women for diagnosis of osteoporosis [2]. Hip bone mineral density

(BMD) is the considered the "gold standard" in detecting osteoporosis by using dual photon X-ray absorptiometry (DPXA) technology. Expert analysis can model osteoporosis as a function of multiple factors via statistical learning techniques (e.g., discriminant analysis, tree classifiers, neural networks). However, diagnostic tests must sometimes rely on single markers to present a simple and effective diagnostic tool for practitioners. Given this approach, we focus on BMD as a primary marker.

The EPIDOS study group was recruited for a 2-year follow-up study between January 1992 and January 1994. There were 115 fractures recorded during two years (a 2.07 % rate), and Figure 2 shows density plots for BMD scores grouped by whether fractures occurred or not. Density estimates are based on the R function `density` to generate kernel density estimates with a Gaussian smoothing kernel and the default (rule-of-thumb) bandwidth selection.

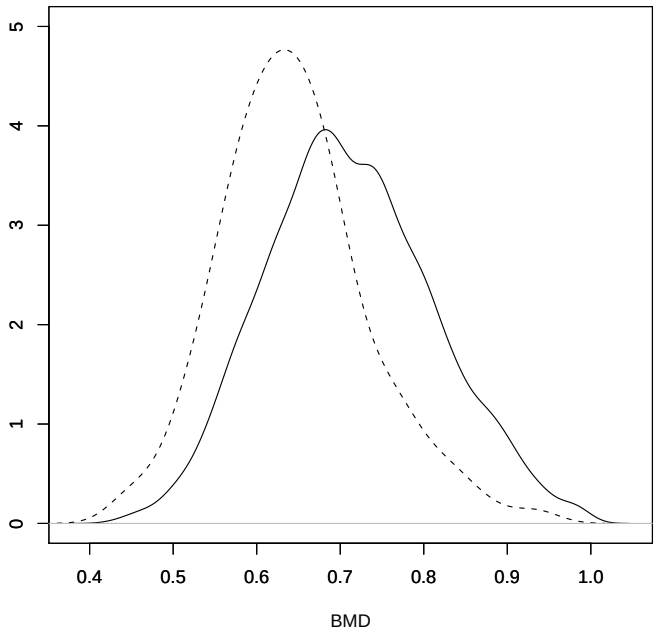

Figure 2: Density plots for BMD scores. The solid line represents hip fracture group and the dashed line is non hip fracture group.

Figure 3 shows the ROC curve for the fracture data. The AUC statistic is 0.6949, and for testing the hypothesis of equal distributions (i.e., the ROC is a 45°

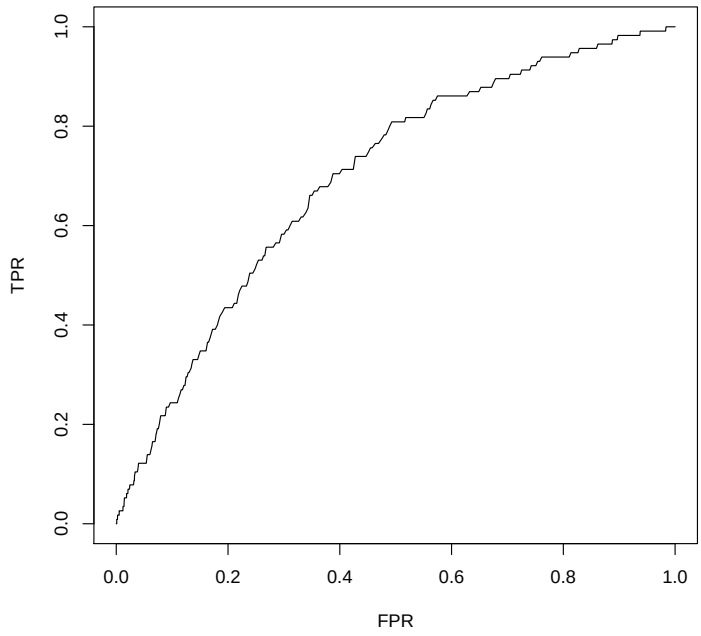

Figure 3: ROC curve for BMD for hip fractures.

line), the p-value corresponding to the Mann-Whitney test is $5.38e^{-13}$. Using the variance estimation method of De Long et al. [17], the variance estimate for the AUC is approximately

$$\frac{Var(P(Z_0 \geq Z_1))}{n_1} + \frac{Var(P(Z_1 \geq Z_0))}{n_0}.$$

An approximate 95% confidence interval for the AUC is calculated to be (0.6099, 0.7799).

Figure 4 shows the Skill plots for the BMD hip fracture marker under three different relative loss values: θ= 0.01, 0.02 and 0.1. Depending on the relative loss, the test can be nearly everywhere skillful ($\theta = 0.02$), skillful for half the values ($\theta = 0.01$) and almost nowhere skillful ($\theta = 0.10$).

Figure 5 shows the Skill Plot for θ= 0.01 along with an approximate 95% confidence interval based on various estimates from Section 3.1. The lower half of the figure displays density plots of BMD for hip fractures under θ= 0.01. The solid line

represents hip-fracture group and the dashed line is for the non hip-fracture group.

From the plot, we see that the test is skillful for to values of $c \geq 0.68$. The skillful region ($\hat{K}(c) \geq 0$) corresponds to the set FPRs $\in (0.388, 1)$. By integrating the ROC only over the FPR values that produce a skillful test, we calculate the PAUC to be $\hat{A}(0.388, 1) = 0.5235$. To estimate the variance of PAUC based on DeLong et al. [17],

$$\begin{aligned} Var(\hat{A}(t_0, t_1)) &\approx \frac{Var(P(Z_0 \geq Z_1, Z_0 \in (u_0, u_1))}{n_1} + \frac{Var(P(Z_1 \geq Z_0, Z_0 \in (u_0, u_1)))}{n_0} \\ &\approx \hat{A}(t_0, t_1)(1 - \hat{A}(t_0, t_1))/n_1 + \hat{A}(t_0, t_1)(1 - \hat{A}(t_0, t_1))/n_0, \end{aligned}$$

where $t_0 = P(X \geq u_1 | Y = 0)$, and $t_1 = P(X \geq u_0 | Y = 0)$.

The corresponding 95% confidence interval for the PAUC is $(0.4313, 0.6157)$. If the test conveys no information, the PAUC would be 0.4248, which is not included in the 95% confidence interval (the p-value for this test hypothesis is 0.01797). The PAUC odds Λ defined in (4) equals 5.9018, compared to the odds for a test that conveys no information which is 2.2668. $\hat{K}(c)$ is maximized at $c = 0.73$, corresponding to FPR = 0.575 and TPR = 0.861. A 95% confidence interval for maximized skill score is $(-0.0006, 0.2800)$.

5 Discussion

By using a proper loss function to decide where a diagnostic test is skillful, we can also construct a more objective test based on the PAUC. The traditional method involves subjectively choosing sensible FPR values that correspond to a region of the ROC, which in turn is used to compute the PAUC. Information from the loss function, along with the necessary information from the sample (required to estimate the prevalence of the disease in the population) combine to make up the skill statistic. The region of interest corresponds to where the skill statistic is positive, and the diagnostic test is considered skillful. In regions where the test is not skillful, the cost-saving naive classification rule is preferred (i.e., either treat or not treat the

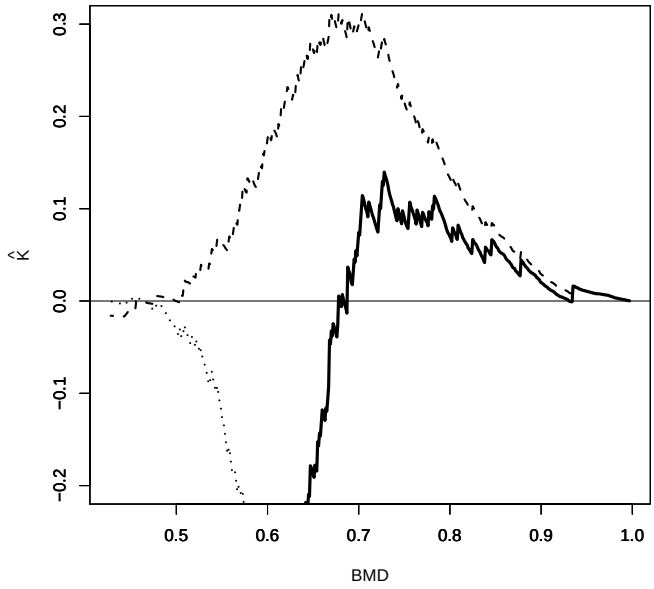

Figure 4: The Skill Plot of BMD for hip fractures under θ= 0.01, 0.02 and 0.1. The wide solid line represents θ= 0.01, the dash line is θ= 0.02 and the dot line is θ= 0.1.

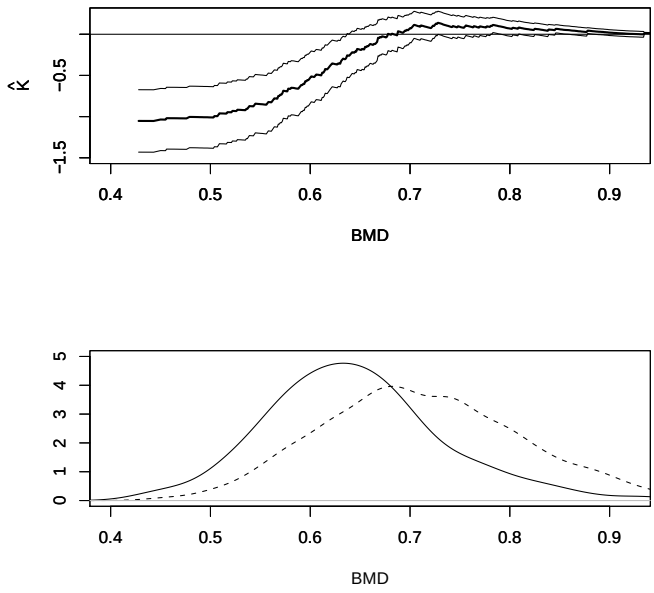

Figure 5: Top:The Skill Plot and 95 % CI for θ= 0.01. Bottom: Density plots of BMD for hip fractures for hip fractures group (solid line) and non hip fracture group (dashed line).

entire population).

The skill score hinges on the relationship between the disease prevalence (p) and the assigned loss (θ). It is easy to show that if $p \gg \theta$, the skill score will always be less than zero, meaning the test is nowhere skillful, and it will be cost efficient to treat every patient, no matter what results from the diagnostic test.

The four typical sets of FPR that are considered in Dodd and Pepe [3] are (t_0, t_1) = (0.0,0.1), (0.0,0.2), (0.1,0.2), (0.1,0.3). By choosing to minimize the false-positive rate, we are implying a loss function where θ is close to zero, so the disease prevalence might be larger ($p \geq \theta$). Since the ROC curve plots TPR versus FPR, by restricting FPR to values close to zero also will restrict the TPR to its lowest values. If the θ is large enough, compared to p, increasing the TPR becomes more crucial than lowering the FPR, and we can end up using intervals such as (0.5,1) or (0.8,1).

Figure 6 shows the range of values (t_0, t_1) that are considered skillful for the case θ = 0.50 and different factors of prevalence ($0 \leq p \leq 1$). Although most realistic cases would lead to a smaller value of θ, this figure illustrates the relationship between loss and prevalence in terms of determining the skillful region. The bars indicate the FPR values used for the PAUC, so in the case $p \leq \theta$, large FPR values are preferred (and with the trade-off, rates for the true-positive rate increase). The bar at p=θ=0.5 is darkened to show that when the prevalence matches the relative loss, the PAUC will consider all values between (0,1).

The criteria for constructing the PAUC was based on requiring skillful tests, i.e., $\hat{K} \geq 0$. However, alternative criteria might be preferred, such as requiring more skill ($\hat{K} \geq \epsilon$) or just eliminating the more unskilled tests ($\hat{K} > -\epsilon$). Figure 7 displays the values of the ROC curve considered for the PAUC that meet a slightly higher skill requirement ($\hat{K} \geq 0.05$) resulting in narrower intervals compared to those in Figure 6.

The properties discussed in Section 3 suggest that estimating the skill score will provide satisfactory results with large samples. However, with smaller samples, the variability from $\hat{K}$ can have a great affect on the values (t_0, t_1) chosen for the PAUC computation.

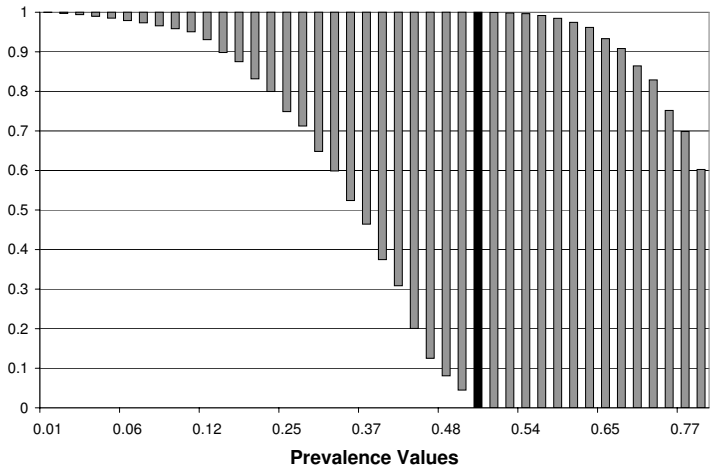

Figure 6: Skillful values used to compute PAUC (vertical bars) based on θ=0.5 and varying prevalence values.

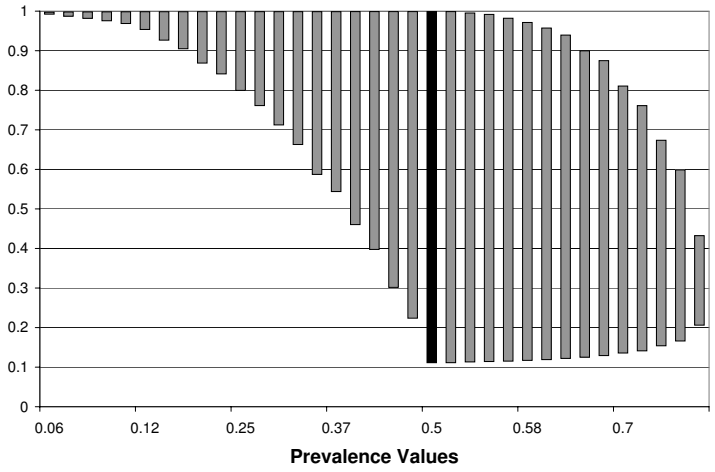

Figure 7: PAUC values of (t_0, t_1) for which the skill score $\hat{K} \geq 0.05$, based on θ=0.5 and varying prevalence values.

In Table 3, we summarize simulation results on the comparison of the mean squared error (MSE) for the PAUC estimate versus the MSE of the Dodd and Pepe [3] estimator. We consider prevalence values of $p = 1/6$ and $p = 1/2$, where expected sample sizes for the healthy population are 50, 100 and 500. Each result in Table 3 is based on 1000 simulations, which are based on $F(t) = \Phi(t)$, $G(t) = \Phi((t-1.5)/1.2)$, chosen to match the simulated distributions in Dodd and Pepe [3] (see Figure 1 in Section 2.1). Because they use different interval values under the ROC curve, the ratio of the MSEs do not necessarily converge to one with large sample size, but Table 3 indicates that the regular Dodd and Pepe estimator has better MSE with smaller samples.

Briggs and Zaertski [1] pointed out advantages of the skill score in regard to finding optimal cutoff points for a diagnostic test, but this paper shows the skill score can also be used in assessing the test. Although the PAUC statistic can be criticized for leading the practitioner to choose false-positive rates arbitrarily, the skill score can be used to guide this choice by using objectively constructed loss functions. The skill plot, in fact, makes the interpretation of the relative loss function easier because it also includes the effect of disease prevalence in the plot.

	$p = 1/6$			$p = 1/2$		
	MSE($\hat{A}_K$)/MSE($\hat{A}_{DP}$)			MSE($\hat{A}_K$)/MSE($\hat{A}_{DP}$)		
θ	n=60	n=120	n=600	n=60	n=120	n=600
0.01	1.68	1.68	1.20	0.13	0.05	0.01
0.05	1.52	1.52	1.12	0.21	0.12	0.10
0.1	1.06	1.06	1.00	0.52	0.51	0.64
0.167	0.92	0.96	1.00	0.13	0.05	0.01
0.2	0.96	0.96	1.00	0.99	1.15	1.09
0.3	1.12	1.12	1.01	0.86	0.96	0.99
0.5	2.27	2.36	1.08	0.71	0.84	0.96
0.6	2.94	2.99	1.15	0.81	0.89	0.98

Table 3: MSE for skill-based PAUC estimator ($\hat{A}_K$) over the MSE of the Dodd and Pepe [3] estimator ($\hat{A}_{DP}$), based on various values of θ and sample size n.

Appendix: Asymptotic Properties of the Skill score

Under standard regularity conditions for probability densities ($p293$, [18]), by the Central Limit Theorem,

$$\sqrt{n}[F_n(x) - F(x)] \to N[0, F(x)(1 - F(x))],$$

Therefore, $E(\bar{G}_{n_1}(c)) = \bar{G}(c)$, and $Var(\bar{G}_{n_1}(c)) = \bar{G}(c)(1-\bar{G}(c))/(n_1) = \bar{G}(c)G(c)/(n_1)$. The same applies to $E(\bar{H}_n(c)) = \bar{H}(c)$, and

$$Var(\bar{H}_n(c)) = \frac{1}{n}\bar{H}(c)(1 - \bar{H}(c)) = \frac{1}{n}\bar{H}(c)H(c).$$

$$Cov(\bar{G}_{n_1}(c), \bar{H}_n(c)) = Cov\{\bar{G}_{n_1}(c), [[(n_1)\bar{G}_{n_1}(c) + (n_0)\bar{F}_{n_0}(c)]/n]\}$$

$$= \frac{n_1}{n}Var(\bar{G}_{n_1}(c)) = \frac{1}{n}\bar{G}(c)G(c).$$

Let $\hat{p} = n_1/n \sim$ Binomial$(p), E(\hat{p}) = p, Var(\hat{p}) = p(1-p)/n$, $Cov(\hat{p}, \bar{H}_n(c)) = 0$, and $Cov(\hat{p}, \bar{G}_{n_1}(c)) = 0$. Let $W_n^{*T} = [\bar{H}_n, \bar{G}_{n_1}, \hat{p}]$, therefore, $\sqrt{n}(W_n^{*T} - E[W_n^{*T}]) \to N[0, \Sigma]$, where $E[W_n^{*T}] = [\bar{G}(c), \bar{H}(c), p]$, and,

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{pmatrix},$$

with $\sigma_{11} = \bar{G}(c)G(c)/p, \sigma_{22} = \bar{H}(c)H(c), \sigma_{12} = \sigma_{21} = \bar{G}(c)G(c), \sigma_{33} = p(1-p)$, $\sigma_{13} = \sigma_{31} = 0$, and $\sigma_{23} = \sigma_{32} = 0$.

Using a Taylor expansion, $E[f(X,Y,Z)] \approx f(X_0, Y_0, Z_0)$, we have conditional on $X = X_0, Y = Y_0, Z = Z_0$,

$$Var[f(X,Y,Z)] \approx \begin{pmatrix} \frac{\partial f}{\partial X} & \frac{\partial f}{\partial Y} & \frac{\partial f}{\partial Z} \end{pmatrix} \Sigma \begin{pmatrix} \frac{\partial f}{\partial X} \\ \frac{\partial f}{\partial Y} \\ \frac{\partial f}{\partial Z} \end{pmatrix},$$

Thus,

$$\sqrt{n}(\hat{K}_{p,\theta}(c) - E[\hat{K}_{p,\theta}(c)]) \to N(0, \sigma^2),$$

where,

$$E[\hat{K}_{p,\theta}(c)] \approx \frac{\bar{G}(c)p - \bar{H}(c)\theta}{p(1-\theta)} I(p < \theta) + \frac{H(c)\theta - G(c)p}{(1-p)\theta} I(p \geq \theta),$$

and

$$\begin{aligned}\sigma^2 &\approx \frac{1}{p(1-\theta)^2}\left((1-2\theta)\bar{G}(c)G(c) + \frac{\theta^2}{p}\bar{H}(c)H(c) + \frac{1-p}{p^2}\bar{H}^2(c)\theta^2\right) I(p < \theta) \\ &+ \frac{1}{(1-p)^2}\left(\frac{p-2p\theta}{\theta^2}\bar{G}(c)G(c) + \bar{H}(c)H(c) + \frac{p}{(1-p)\theta^2}(H(c)\theta - G(c))^2\right) I(p \geq \theta).\end{aligned}$$

References

[1] Briggs WM, Zaretzki R. The Skill Plot: a Graphical Tecnique for Evaluating the Predictive Usefulness of Continuous Diagnostic Tests (with discussion). *Biometrics* 2008; **64**: 250-263.

[2] Hans D, Dargent-Molina P, Schott AM, Sebert JL, Cormier C, Kotzki PO. Delmas P. D. Pouilles J. M. Breart G. and Meunier P. J. Ultrasonographic Heel Measurements to Predict Hip Fracture in elderly Women: the EPIDOS Prospective Study *The Lancet* 1996; **143**: 29-36.

[3] Dodd LE, and Pepe MS, Partial AUC Estimation and Regression *Biometrics* 2003; **59**: 614-623.

[4] Sing T, Sander O, Beerenwinkel N, Lengauer T, ROCR: Visualizing Classifier Performance in R. *Bioinformatics* 2005; **21**: 3940-3941.

[5] Brumback LC, Pepe MS, Alonzo TA, Using the ROC Curve for Gauging Treatment Effect in Clinical Trials. *Statistics in Medicine* 2006; **25**: 575-590.

[6] Pepe MS, *The Satistical Evaluation of Medical Tests for classification and Prediction.* Oxford University Press: New York, 2003.

[7] Pepe MS. An Interpretation for ROC Curve and Inference Using GLM Procedures. *Biometrics* 2000; **56**: 352-359.

[8] Hsieh F, Turnbull BW. Nonparametric and Semiparametric Estimation of the Receiver Operating Characteristic Curve. *The Annals of Statistics* 1996; **24**: 25-40.

[9] Lloyd CJ, Yong Z. Kernel estimators of the ROC curve are better than empirical. *Statistics & Probability Letters* 1999; **44**: 221-228.

[10] Claeskens G, Jing BY, Peng L, Zhou W. Empirical Likelihood Confidence Regions for Comparison Distributions and ROC Curves. *The Canadian Journal of Statistics* 2003; **31**: 173-190.

[11] Hall P, Hyndman RJ, and Fan Y. Nonparametric Confidence Intervals for Receiver Operating Characteristic Curves. *Biometrika* 2004; **91**: 743-750.

[12] Schervish M. A General Method for Comparing Probability Assessors. *The Annals of Statistics* 1989; **17**: 1856-1879.

[13] Mozer JB, Briggs WM. Skill in Real-time Solar Wind Shock Predictions. *Journal of Geophysical Research* 2003; **108**: 1-9.

[14] Briggs WM, and Ruppert D. Assessing the Skill of Yes/No Predictions. *Biometrics* 2005; **61**: 799-807.

[15] Baker SG. Identifying Combinations of Cancer Markers for Further Study as Triggers of Early Intervention. *Biometrics* 2000; **56**: 1082-1087.

[16] Hanley JA, McNeil BJ. The Meaning and Use of the Area under a Receiver Operating Characteristic(ROC) Curve. *Radiology*1982; **143**: 29-36.

[17] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics* 1988; **44**: 837-845.

[18] Roussas GG. *A Course in Mathematical Statistics* Academic Press. 1997.